Wulf, Christoph [Hrsg.]

Evaluation. Beschreibung und Bewertung von Unterricht, Curricula und Schulversuchen

München : R. Piper & Co. Verlag 1972, 419 S. - (Erziehung in Wissenschaft und Praxis; 18)

Quellenangabe/ Reference:
Wulf, Christoph [Hrsg.]: Evaluation. Beschreibung und Bewertung von Unterricht, Curricula und Schulversuchen. München : R. Piper & Co. Verlag 1972, 419 S. - (Erziehung in Wissenschaft und Praxis; 18) - URN: urn:nbn:de:0111-opus-15135 - DOI: 10.25656/01:1513

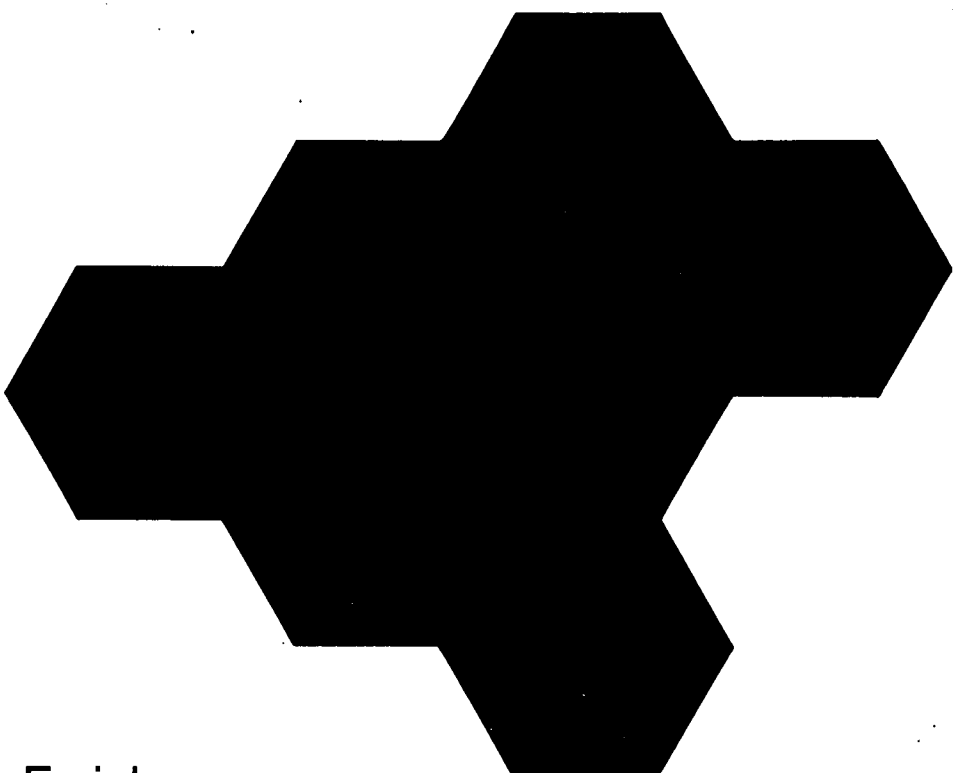
https://nbn-resolving.org/urn:nbn:de:0111-opus-15135
https://doi.org/10.25656/01:1513

Evaluation

Evaluation

Texte
herausgegeben
von
Christoph Wulf

Beschreibung
und Bewertung
von Unterricht, Curricula
und Schulversuchen



18
Piper

Erziehung
in Wissenschaft
und Praxis

Piper

Evaluation

Beschreibung und Bewertung von Unterricht,
Curricula und Schulversuchen

Texte
herausgegeben von Christoph Wulf

R. Piper & Co. Verlag
München

ISBN 3-492-01985-4

Gesamtherstellung Clausen & Bosse, Leck/Schleswig
Umschlagentwurf Gerhard M. Hotop
Printed in Germany

Inhalt

Vorwort

Der hier vorgelegte Band zur Evaluation bietet eine Auswahl der wichtigsten Beiträge der amerikanischen und englischen Evaluationsliteratur der letzten Jahre (1967–1972). In diesem Zeitraum entstand in den USA auf dem Hintergrund der gewaltigen Anstrengungen zur Reform des Bildungswesens in den sechziger Jahren ein neues und komplexes Verständnis von Evaluation, das sich von den vorangegangenen Ansätzen durch eine größere Vielfalt der Zielsetzungen, Verfahren und Methoden unterscheidet. Im Rahmen dieser Entwicklung wurde Evaluation allmählich zu einem zentralen Bereich angewandter pädagogischer Forschung. Die führenden erziehungswissenschaftlichen Zeitschriften brachten Sonderhefte heraus, die sich mit Evaluation befaßten. Evaluation wurde zu einem integralen Bestandteil der gegenwärtigen Innovationen im Bereich der Curriculumentwicklung, in der Lehrerbildung und im Feld der Schulversuche. Das entscheidende Anliegen bestand darin, herauszufinden, inwieweit die unter großem Kostenaufwand durchgeführten Projekte zu einer tatsächlichen Verbesserung der Curricula und darüber hinaus der Schulwirklichkeit führen. Diese Frage wurde um so drängender, als die Bildungsreformer sich von seiten der Öffentlichkeit einem verstärkten Druck zur Rechtfertigung der Reformen ausgesetzt sahen. Die Rechtfertigung sollte sich einmal auf die Wirksamkeit der Reformen im Hinblick auf die intendierten Ziele, zum anderen wegen der knapper werdenden Ressourcen auch auf die Relation zwischen finanziellem und personellem Aufwand und der tatsächlich erreichten Verbesserung der Schulwirklichkeit beziehen. Diese Forderungen nach genauer Beschreibung der Innovationen und ihrer Bewertung fanden auch eine starke Resonanz, weil sich immer stärkere Zweifel an der lückenlosen Vorausplanbarkeit von Innovationen mit Hilfe vorgängig entwickelter Theorien einstellten – nicht zuletzt auf Grund der Erfahrungen mit den umfangreichen Reformen (vgl. z. B. Schwab 1970, 1971; Eisner 1971). Somit wurde Evaluation ein zentrales Element der Entwicklungsprozesse für Reformprojekte selbst.

Evaluation erfolgte in einer fortwährenden Überprüfung der Realisierung von Intentionen bzw. normativen Kriterien und als Entwicklung von Alternativen auf Grund unbefriedigender Ergebnisse, wobei die Modifikationen sich häufig auf die Mittel der Realisierung, oft genug aber auch auf die Intentionen selbst bezogen.

Im ersten Teil dieses Auswahlbandes (I) soll ein Überblick über den gegenwärtigen Diskussionsstand im Bereich der Evaluation gegeben werden (vgl. auch Wulf 1972c). Dabei wird sich zeigen, daß die wesentlichen Beiträge zur Konzeptualisierung von Evaluation im Rahmen der »Evaluation von Curricula« entstanden sind. Zum Teil erfolgt dabei eine erhebliche Ausweitung des Begriffs Curriculum, der zum Synonym für Bildungsprogramm wird und der damit als ein Oberbegriff für alle kodifizierten Bildungsintentionen dient. In diesem Teil wird zugleich eine Einführung in die Gesamtproblematik der Evaluation gegeben, die das Verständnis der weiteren Beiträge des Sammelbandes erleichtern soll.

Im zweiten Teil (II) sollen in erster Linie die Grundfragen erörtert werden, die Einfluß auf die Ziele, Formen und Verfahren der Evaluation haben und die z. T. ganz neue Aspekte für den Bereich liefern, der in der BRD häufig Begleitforschung genannt wird. Im Unterschied zur Beurteilung individueller Schülerleistungen zielt Evaluation auf die Beschreibung und Beurteilung von Bildungsprogrammen, deren Verbesserung eine entscheidende Voraussetzung für relevante Schülerleistungen ist. Die starke Konzentration auf die Fragen der Evaluationsforschung hat in den USA dazu geführt, daß Evaluation bereits in der Planungsphase größerer Curriculumprojekte und Schulversuche als »Evaluation der Voraussetzungen« bzw. als »Kontextevaluation« eine wichtige Rolle spielt. Dabei ist es Aufgabe der Evaluatoren, die Bedürfnisse der Adressaten zu erheben, damit eine bessere Planung von Innovationen stattfinden kann. Eine weitere wichtige Funktion kommt der Evaluation während der Entwicklung von Reformprojekten als »formative Evaluation« zu; auch ihr wurde in letzter Zeit verstärkte Aufmerksamkeit zugewandt. Sie soll die Mängel der Reformprojekte (Curricula) so rechtzeitig aufzudecken versuchen, daß sie noch vor Abschluß des Projekts beseitigt werden können. Schließlich gilt es, neue Funktionen und neue Verfahren für die »Evaluation der Endprodukte der Curricula bzw. Schulversuche« zu entwerfen und zu entwickeln. Allein im Zusammenhang mit diesen drei Funktionen der Evaluation

- in der Planungsphase,
- in der Realisierungsphase,
- in der Überprüfungsphase der Ergebnisse

tauchen zahlreiche Fragen auf, die oft schwer lösbar sind. So ist es z. B.

schwierig zu klären, welches die speziellen Aufgaben der »Evaluation der Voraussetzungen« sind oder welche Untersuchungsverfahren man für die »formative Evaluation« wählen soll. Dabei spielt eine zentrale Rolle die Frage, ob man sich für die Vergleichsgruppenuntersuchungen oder Detailuntersuchungen einzelner Gruppen entscheiden soll. In diesem Zusammenhang wird immer wieder das Problem erörtert, ob nicht die Funktion und die Verfahren gegenwärtiger pädagogischer Forschung im Rahmen der Durchführung von Bildungsreformen neu definiert werden müssen, damit man den komplexen Bedingungen pädagogischer Innovationen gerecht werden kann. Mit der Frage nach der Funktion der Evaluation erhebt sich zugleich das zentrale Problem, inwieweit Evaluation eine *Beschreibung* von pädagogisch relevanten Sachverhalten ist oder in welchem Maß sie zur *Bewertung* der Reformen und ihrer Ergebnisse beiträgt oder inwieweit Evaluation als *Entscheidungsvorbereitung* und Entscheidungshilfe für Entscheidungsträger dienen kann.

Es wird sich zeigen, daß das »Konzept der Evaluation« in hohem Maße ideologieanfällig ist. Auf der einen Seite läuft Evaluation Gefahr, als *Auftragsforschung* so verwendet zu werden, daß bildungspolitische oder pädagogische Entscheidungen im nachhinein durch wissenschaftlich gewonnene Ergebnisse gerechtfertigt werden können. Auf der anderen Seite kann man durch »Evaluation« radikal jedes Ergebnis unter Ideologieverdacht stellen und somit fragwürdig machen.

Man wird die Rolle der Evaluation in jedem Fall zwischen diesen Extremen klar definieren müssen. Dementsprechend werden auch die für die Evaluation verwendeten Verfahren empirisch bzw. hermeneutisch-ideologiekritisch sein müssen; hermeneutische bzw. ideologiekritische Verfahren sind im Rahmen der amerikanischen Evaluationsliteratur bisher wenig ausgearbeitet worden.

Die Frage nach den Verfahren und Methoden der Evaluation wirft weiter das Problem auf, welche Form und welches Ausmaß der Evaluation im Hinblick auf die beteiligten Personen und den Anteil der Kosten im Verhältnis zum Gesamtprojekt zukommen soll. In einigen Fällen kann es – wie z. B. in den USA und in England – sinnvoll sein, im Rahmen eines Großprojekts ein Evaluationsteam auszugliedern; es kann aber auch Evaluation ohne organisatorische Ausgliederung im Rahmen der Entwicklung des Projekts durchgeführt werden (vgl. z. B. Meyer 1972); in einigen Fällen wird Evaluation möglicherweise sogar von allen am Versuch Beteiligten zur Verbesserung ihrer eigenen Arbeit durchgeführt; dann ist ein Ziel dabei, die beteiligten Lehrer und Wissenschaftler zur *Selbstevaluation* zu bringen. So wichtig diese Form der Evaluation ist, so begrenzt ist hier die Möglichkeit, Erkenntnisse zu generalisieren.

In der Frage der *Generalisierbarkeit* liegt ein weiteres Problem der Evaluationsforschung, das wissenschaftstheoretisch schwierige Fragen aufwirft, sobald der klassische Versuchsplan mit Versuchs- und Kontaktgruppen aufgegeben wird. Dies sind nur einige der zahlreichen Probleme, die im Rahmen der Grundfragen der Evaluation behandelt werden und die einen wichtigen Beitrag zu einem komplexen Verständnis für die Funktionen von Begleitforschung bilden.

Im dritten Teil (III) wird mit der Evaluation von *Unterrichtsprozessen* bzw. *-interaktionen* ein weiteres Feld pädagogischer Evaluation behandelt. Es ist bisher unter evaluativer Fragestellung nur z. T. erforscht worden; dennoch gibt es zahlreiche Unterrichtsbeobachtungssysteme und -untersuchungen, die für eine Evaluation von Unterrichtsprozessen relevante Forschungsergebnisse erbracht haben. Dieser Bereich ist besonders wichtig, da zahlreiche Forschungen ergeben haben, daß viele Innovationsversuche bereits infolge traditioneller »Interaktionspattern« an enge Grenzen der Wirksamkeit stoßen, die, wenn sie durch eine Evaluationsuntersuchung festgestellt werden, systematisch abgebaut oder verändert werden können. Die Evaluation von Unterrichtsprozessen ist in letzter Zeit in stärkerem Maße Teil der Evaluationsforschung geworden, die zuvor auf formative oder summative Evaluation im Sinne von Leistungsüberprüfung eines Bildungsprogramms begrenzt war. Das wurde durch die Tradition der verhaltensorientierten Ansätze verursacht, nach denen nur das meßbare Ergebnis von Unterrichtsprozessen, nicht aber ihr Prozeßcharakter selbst Gegenstand der Evaluation war. Unter dem Einfluß verstärkter Kritik im Rahmen der Evaluationsliteratur an rigiden behavioristischen Modellen, bei denen die Überprüfung der Lernerfolge (verstanden als das Erreichen der vorher formulierten Lernziele) oft mit Evaluation verwechselt wird, hat die Prozeßdimension im Unterricht und damit auch in der Evaluation mehr Beachtung gefunden. Dennoch sind gerade in diesem Bereich der Evaluation noch viele Fragen offen, die weiterer Klärung bedürfen.

Schließlich galt es, einige ausgewählte Beispiele für Evaluationsuntersuchungen in den Band mitaufzunehmen. Im Teil IV sollen konkrete Fälle für die Bewältigung einiger der angesprochenen Probleme der Evaluation in einem bestimmten pädagogischen Kontext geboten werden, bei denen auch manche Unzulänglichkeiten dieser Untersuchungen sichtbar werden, die zur Kritik und Verbesserung herausfordern können. Zudem werden bei einer solchen Konkretisierung Probleme sichtbar, die in die theoretischen Reflexionen noch nicht eingehen oder eingegangen sind, die jedoch für die Planung von Evaluationsuntersuchungen von außerordentlicher Wichtigkeit sind. Das dürfte bei den unterschiedlichen methodi-

schen Ansätzen der einzelnen Evaluationsuntersuchungen besonders deutlich werden.

Von deutscher Seite aus gesehen, verkürzen die angelsächsischen Evaluationsmodelle und -untersuchungen die politische Dimension der Evaluation und kommen daher in Gefahr, die technologische Dimension zu sehr in den Vordergrund zu rücken, inhaltlich aber vor einer eindeutigen Wert- und Kriteriensetzung zurückzuscheuen. Daher wurde für Teil V ein Beitrag aus der BRD ausgewählt, der in knapper Form neben vielen anderen zum Teil bereits angesprochenen Fragen die *gesellschaftspolitische Dimension* der Evaluation an Hand eines Beispiels aus der Vorschulerziehung durch das Engagement für die Emanzipation von Unterschichtkindern deutlich macht.

Für alle, die sich eingehender mit Evaluation, vor allem im curricularen Bereich, befassen wollen, dürfte die speziell für die deutschen Verhältnisse überarbeitete klassifizierte Bibliographie von Wert sein (VI).

Um das Lese-Verständnis zu erleichtern, wurde ein Glossar (VII) erarbeitet, das dem Leser bei schwierigen Begriffen eine erste Orientierung geben soll.

Schließlich sei den zahlreichen amerikanischen Kollegen gedankt, die mich im Verlauf von zwei längeren Forschungsaufenthalten in den USA bei der Gestaltung dieses Bandes beraten haben. Mein Dank gilt auch den zahlreichen Mitarbeitern des Deutschen Instituts für Internationale Pädagogische Forschung, die bei der Übersetzung der Beiträge mitgearbeitet haben und die mich bei Überwindung der manchmal erheblichen sprachlichen Schwierigkeiten unterstützt haben. Ferner möchte ich G. Skoupil und H. Kohl für das wiederholte Schreiben der Manuskripte danken.

Christoph Wulf

Anmerkung:

Im allgemeinen sind schwierige amerikanische Begriffe, wenn sie einen bestimmten Eigenwert haben, bei der ersten Verwendung in Klammern hinzugefügt worden.

Die kursiv gedruckten Seitenzahlen in den bibliographischen Angaben beziehen sich auf diesen Auswahlband.

I Einführung in die Problematik der Evaluation, dargestellt am Beispiel der Curriculumevaluation

CHRISTOPH WULF

Curriculumevaluation

1. Die Notwendigkeit der Curriculumevaluation [1]

In der Bundesrepublik findet heute Curriculumentwicklung an vielen Orten, zugleich aber auf verschiedenen Ebenen statt:

- Im Rahmen der *Lehrplanrevisionen und Richtlinienarbeit,* die von den Ländern betrieben wird [2],
- in vielen unmittelbar auf die Schule bezogenen *Initiativgruppen,* die an traditionellen Schulen und Gesamtschulen gebildet werden [3],
- als sorgfältige Entwicklung einiger *Teilcurricula,* wie sie an mehreren Universitäten und Forschungsinstituten geplant oder bereits erfolgt ist [4].

Diese Bemühungen sind unterschiedlich zu bewerten, was ihren jeweiligen theoretischen Ansatz, den erreichten Grad der theoretischen Reflexion und die Qualität der verschiedenen Ergebnisse betrifft. Bereits in diesem für die BRD noch frühen Stadium der Curriculumentwicklung sollte ein weiterer Schritt vollzogen werden. Man sollte die Aufgaben und Ziele der Curriculumevaluation reflektieren und für ihre Realisierung ein methodisches Instrumentarium entwickeln. Die systematische Erweiterung der Curriculumarbeit in dieser Hinsicht ist notwendig: ein neues Curriculum, über dessen Wert nichts zu erfahren ist, erscheint wertlos. Es bedarf aber nicht nur der kritischen Bewertung dessen, was als Produkt der Curriculumentwicklung entsteht, sondern auch der kritischen Beurteilung des ganzen Prozesses, in dem es entstand. So gesehen, ist Evaluation integraler Bestandteil aller curricularen Arbeit überhaupt. Daher muß Evaluation von Anfang an in der ganzen Breite ihrer Aufgaben, ihrer Bedeutung, ihrer Verfahren begriffen werden. Es besteht aber durchaus die Gefahr, daß der Begriff der Evaluation zu eng und oberflächlich verstanden wird, z. B. wenn Evaluation nur auf individuelle Leistungsmessung begrenzt wird (vgl. z. B. Bloom/Hastings/Madaus 1971), wodurch ihre wichtige Funktion für den Prozeß der Curriculumentwicklung übersehen wird. Dieser Gefahr sind tatsächlich große Teile der amerikanischen Curriculumentwicklung seit dem Anfang der sechziger Jahre

erlegen. Sollen ähnliche Mißerfolge in der Bundesrepublik vermieden werden, dürfte zunächst ein Überblick über den augenblicklichen Stand der Evaluationsforschung in den USA angebracht sein; dadurch läßt sich ein guter Ausgangspunkt für die Entwicklung von Modellen, Methoden und einer Technologie der Curriculumevaluation gewinnen.

In den letzten Jahren ist Evaluation zu einem zentralen Thema der amerikanischen Curriculumforschung geworden [5]. Die Kontroversen über die Funktion der Evaluation haben wesentliche Beiträge zur Curriculumentwicklung geleistet (vgl. Review of Educational Research 1970). Gekennzeichnet ist die Entwicklung der letzten Jahre durch eine zunehmende Vielfalt der Aufgaben, die der Evaluation zufallen. Dabei ist es zu grundlegenden Veränderungen in der Konzeptualisierung und Methodologie gekommen, ohne daß es jedoch gelungen wäre, die dabei aufgeworfenen Probleme zu lösen. Summarisch läßt sich sagen: *Curriculumevaluation zielt auf die Sammlung, Verarbeitung und Interpretation von Daten mit dem Ziel, Entscheidungen über ein Curriculum zu fällen. Das impliziert: (1) objektive Beschreibungen von Zielen, Umwelt, Personal, Methoden und Inhalt und Ergebnissen; und (2) persönliche Urteile über die Qualität und Angemessenheit dieser Ziele, der Umwelt usw.* (vgl. Stake 1967 b).

Scriven hat auf die Unterscheidung zwischen Ziel und Rolle der Evaluation aufmerksam gemacht (vgl. Scriven S. 60 ff). Ziel ist die Sammlung von Informationen zum Zweck rational begründeter Entscheidungen über »etwas«. Die Rolle der Evaluation hängt davon ab, was dieses »etwas« ist und von *wem* und *welche* Maßstäbe dabei angelegt werden. Aufgabe der Evaluation ist es z. B., zur Konstruktion eines Curriculum, zur Antizipation seines Erfolges oder zu seiner Verbesserung beizutragen. In der Regel erfolgt Evaluation für verschiedene Adressaten, z. B. Politiker, Fachleute, Lehrer, Eltern, Schüler und Curriculumentwickler. Mit Rücksicht auf ihr unterschiedliches Erkenntnisinteresse ist die *Form* und die *Sprache*, in der die Ergebnisse eines Evaluationsprozesses dargestellt werden, von Fall zu Fall zu variieren.

Um zu einer angemessenen Konzeptualisierung von Evaluation beizutragen, muß man

- ein Vorstellungsschema liefern, das die Gebiete und Probleme klassifiziert, die evaluiert werden sollen,
- Strategien und Methoden der Evaluation entwickeln und
- ein System von Generalisationen für den Gebrauch der verschiedenen Verfahren und Techniken entwerfen (Alkin 1969, 2) [6].

Das ist in den USA im Laufe der letzten Jahre in verstärktem Maße geschehen. So entwarf Scriven bereits 1967 eine Methodologie der Eval-

uation und Stake 1967 ein Evaluationsmodell. 1969 folgten Alkins Evaluationstheorie, Maguires Evaluationsmethodologie, Provus' Diskrepanzmodell und Stufflebeams CIPP-Evaluationsmodell (Context Input Process Product).

In mehreren dieser Evaluationsmodelle wird im Rahmen der Theorienbildung als Hauptaufgabe der Evaluation die *Bereitstellung von Daten für Entscheidungen* bezeichnet. Sie finden in Situationen statt, in denen zwischen mehreren Alternativen auf Grund unterschiedlicher Bewertung gewählt wird. Im Prozeß der Entwicklung und Einführung eines Curriculum in die Schule lassen sich mehrere Phasen kennzeichnen, in denen Entscheidungen von Entscheidungsträgern auf verschiedenen Ebenen zu fällen sind. Das bedeutet: es werden je nach Phase mehrere Funktionen der Entscheidung angesprochen und dementsprechend unterschiedliche Informationen benötigt; es müssen daher mehrere Evaluationsfelder unterschieden werden (Alkin 1969 a, 2):

1. Erhebung der *Werte* und *Bedürfnisse* der Schüler und der Gesellschaft[7], um die Bildungs- und Lernziele auszuwählen, die in einem Curriculum in einer bestimmten historisch-gesellschaftlichen Situation realisiert werden sollen.

2. *Programmplanung* mit dem Ziel der Auswahl von Curricula. Hier gilt es, denen, die Entscheidungsbefugnis haben, möglichst objektive Informationen über verschiedene Curricula zu liefern, damit sie entscheiden können, welche speziellen Curricula ihren besonderen Bedürfnissen entsprechen.

3. *Programmimplementation*: Die Einführung curricularer Programme in die Schule und die Evaluation des Ausmaßes, in dem die Intentionen der eingeführten Curricula verwirklicht werden, und die Feststellung, inwieweit sie den Bedürfnissen entsprechen, die als Voraussetzung für die Entscheidung über die Einführung formuliert worden sind.

4. Evaluation der betreffenden Curricula mit dem Ziel, *sie zu verbessern, während sie noch im Prozeß der Entwicklung sind.*

5. Evaluation der Ergebnisse der betreffenden Curricula mit dem Ziel der *Programmbestätigung* (programm certification). Hier gilt es, eine abschließende Evaluation der Curricula vorzunehmen, um den generellen Wert und ihre Verwendbarkeit bestimmen zu können.

Auf die Evaluation im fünften Evaluationsfeld folgt wieder die Evaluation im ersten Feld, die auf z. T. veränderte Bedürfnisse und Interessen stoßen dürfte, wodurch ein neuer Ausgangspunkt für die Curriculumrevision gebildet wird.

Im ganzen bedeutet das, daß in jeder Phase der curricularen Arbeit evaluiert werden muß. Daraus folgt wieder, daß die Rückmeldung der

Ergebnisse Einfluß haben muß auf die Gestaltung der jeweils nächsten Phase oder, um dieses gedankliche Modell zu benutzen, daß Evaluation in einem Regelkreis mit allen anderen Verfahren im Curriculumgefüge zusammengekoppelt ist. Ihre Funktion ist es, die Optimierung des Curriculum zu steuern.

Am Beispiel eines Teils der Evaluation, der Datenerhebung, exemplifiziert, heißt das:

1. Daten über Bedürfnisse haben Einfluß auf die Bestimmung der Ziele.
2. Daten über die Gebiete, auf denen die Ziele liegen, haben Einfluß auf die Planung des Programms.
3. Daten über die schulische Einführung des Programms haben Einfluß auf die Einschätzung der Bedürfnisse, Ziele, Gebiete, Programmplanungen.
4. Die veränderte Einschätzung hat Einfluß auf die Umgestaltung des bisherigen Programms.
5. Daten über die Gesamtergebnisse des Programms haben Einfluß auf die vorläufig abschließende Einschätzung aller vorausliegenden Phasen und auf das Gesamturteil über das Programm und so fort.

Um die angestrebten Informationen optimal zu erhalten, bedarf es auch der Erarbeitung einer *Technologie der Evaluation* (Stake 1967 b), die die Konzeptualisierung und Methodologie der Evaluation ergänzen muß. Diese müßte Elemente der psychometrischen Testtechnologie (vgl. Buros 1965) [8], der sozialwissenschaftlichen Erhebungstechnologie, der Kommunikationstechnologie und der Unterrichtstechnologie einschließen. Das Fehlen einer solchen Technologie macht sich um so stärker bemerkbar, als die vorhandenen Tests sich selten für die Zwecke der Evaluation eignen. Die meisten von ihnen sind standardisierte Leistungstests, die zwar zur Feststellung von individuellen Leistungsunterschieden bei Schülern geeignet sind, nicht aber zur Evaluation eines Curriculum.

Will man sich einen Überblick über das gesamte Problemfeld der Evaluation verschaffen, so lassen sich drei Problemkreise der Evaluationsforschung, die im folgenden zu behandeln sind, abgrenzen:

1. die verschiedenen Formen und Rollen der Evaluation,
2. die Versuche, ein Evaluationsmodell zu erstellen, das einen Rahmen für weitere Bemühungen um eine Konzeptualisierung und Methodologie abgibt,
3. die Wert- und Entscheidungsprobleme der Evaluation und Methoden zu ihrer Lösung.

II. Rollen und Formen der Evaluation

Die erste und einfachste Unterscheidung zur Begriffs- und Funktionsbestimmung der Evaluation ist bereits getroffen. Es ist Scrivens Distinktion zwischen Zielen und Rollen. Dabei liegt für unsere Darstellung das Hauptgewicht zunächst auf den möglichen verschiedenen Rollen; denn als Ziel kann einheitlich die Absicht gesehen werden, Antworten auf bestimmte Fragen zu erhalten, die bestimmte Entscheidungen ermöglichen. Befragt werden muß alles, was in dem Gesamtprozeß der Curriculumentwicklung relevant ist, z. B.

- die Personen, die mitarbeiten,
- die Ziele, die für ihre Arbeit gesetzt werden oder gelten,
- die Verfahren, mit deren Hilfe man die Ziele anstrebt,
- die Programme, von denen man sich dabei leiten läßt,
- die Instrumente, die dabei benutzt werden,
- die Urteile und Wertschätzungen, die die Arbeit steuern und die sich naturgemäß auf Personen, Ziele, Themen, Verfahren usw. beziehen können.

In Frage gestellt werden kann schließlich (auf einer weiteren Reflexionsstufe) die Evaluation selbst (und was aus ihr als Verfahren folgt oder gefolgt ist). Einige Klärungen innerhalb dieses weiten Problem-, Sach- und Begriffsfeldes sind bereits erfolgt. Von Belang sind m. E. die Distinktionen zwischen

- formativer und summativer Evaluation,
- Mikro- und Makroevaluation,
- innerer und äußerer Evaluation,
- vergleichender und nicht vergleichender Evaluation,
- intrinsischer und Ergebnisevaluation.

Sie werden im folgenden kurz dargestellt (vgl. Flechsig 1970).

1. Formative und summative Evaluation

Scrivens Unterscheidung zwischen formativer und summativer Evaluation zur Kennzeichnung zweier Rollen von Evaluation hat Verbreitung gefunden. Formative Evaluation erfolgt, wenn das Curriculum sich noch im Prozeß der Entstehung (Formung) befindet. Summative Evaluation wird nach der Beendigung der Entwicklung eines Curriculum benötigt, wenn eine (vorläufig) abschließende Beschreibung und Wertung gewünscht wird. Entsprechend diesen beiden Rollen der Evaluation sind das Erkenntnisinteresse und die Methoden der Evaluation verschieden. Cronbach hatte 1963 in seinem bekannten Aufsatz »Evaluation for Cour-

se Improvement« die große Bedeutung der formativen Evaluation für die Curriculumentwicklung deutlich gemacht, ohne jedoch terminologisch zwischen formativer und summativer Evaluation zu unterscheiden. Die Aufgabe formativer Evaluation ist es, Schwächen der in der Entwicklung befindlichen Curricula aufzudecken und zu beseitigen (Scriven 1967). Sie ist also als *der* Teil der Curriculumentwicklung zu begreifen, der Auskünfte über die Qualität des in der Entwicklung begriffenen Curriculum bietet. Auf Grund formativer Evaluation kann das Curriculum oder können einige seiner Teile revidiert werden, bis es seine Endfassung findet, die dann summativer Evaluation unterliegt. Grobman hat zu Recht darauf aufmerksam gemacht, daß »es keine scharf abgrenzbare Unterscheidung zwischen diesen beiden Phasen gibt; die formative Evaluation hört nicht auf, bevor die summative Evaluation anfangen kann« (1968, 14). Es ist die Aufgabe formativer Evaluation, auf jede nur erdenkliche Weise Informationen über die Wirkungen des Curriculum zu erhalten; dabei sollten nicht nur die strengen Maßstäbe eines Forschungsplans (research design) angelegt werden, vielmehr gilt es, einen Evaluationsplan zu entwerfen. Um möglichst vielfältige Informationen über das Curriculum zu erhalten, schlägt Cronbach statt der Verwendung von Tests, die auf individuelle Unterschiede zielen, die Einrichtung einer Gesamtheit von Items vor. Aus einer beispielsweise 600 Items umfassenden Gesamtheit sollten besser je 50 Items 12 Gruppen von 50 Schülern gegeben werden als nur 50 Items 600 Schülern.

Die Entwicklung einer Gesamtheit von Items, die es erlauben, qualifizierte Aussagen innerhalb der formativen Evaluation zu erhalten, sollte mit der Formulierung von Lernzielen Hand in Hand gehen. Die Operationalisierung der Lernziele in Items legt ja obendrein eine permanente Revision der Lernziele und Items nahe. Es gilt die Übereinstimmung zwischen den formulierten Lernzielen, den impliziten und denen, die durch Items getestet werden, zu erreichen. Dabei ist das Validitätsproblem schwierig zu lösen. Es ist notwendig, Lernziele und Curriculuminhalte, Lernziele und Inhalte der Prüfung, Curriculuminhalt und Prüfungsinhalt so zusammenzubringen, daß sie wechselseitige Identifikation zulassen (Scriven 1967). Die Herstellung einer derartigen Entsprechung ist eine wichtige Aufgabe des Evaluators, der er sich unterziehen muß, bevor er das Curriculum im Rahmen der formativen Evaluation hinsichtlich seiner Wirkungen auf Schüler mit dem Ziel der Verbesserung evaluieren kann.

Lindvall und Cox formulieren vier Fragen, die die Funktion von Kategorien für formative Evaluation haben; jede Kategorie enthält mehrere Unterkategorien, die ebenfalls in Form von Fragen entworfen wer-

den. Mit der Verwendung dieses *Kategoriensystems* kann formative Evaluation beginnen. Es wurde von den Autoren vor allem im Hinblick auf die formative Evaluation der *Individually Prescribed Instruction* entwikkelt, beansprucht aber zugleich allgemeine Verwendungsfähigkeit:

1. Welche Ziele soll das Programm erreichen?
   a) Sind die formulierten Ziele wirklich Lernziele?
   b) Sind die formulierten Lernziele die wirklichen Ziele des Programms?
   c) Sind die Ziele wertvoll?
   d) Sind die Ziele erreichbar?
2. Wie ist der Plan, diese Lernziele zu erreichen?
   a) Verspricht der Plan, zur Erreichung der Lernziele beizutragen?
   b) Ist der Plan genügend detailliert entwickelt?
   c) Können Plan und Verfahren ohne Schwierigkeiten von den Leuten verstanden werden, die sie verwirklichen sollen?
   d) Ist es wahrscheinlich, daß der Plan verwirklicht werden kann?
3. Stellt das funktionierende Programm eine richtige Einführung des Plans dar?
   a) Welches sind die spezifischen Punkte, die in einer Analyse der Funktion zu beobachten sind?
   b) Werden die Aktivitäten wirklich entsprechend dem Plan ausgetragen?
   c) Wie kann erreicht werden, daß Plan und Verwirklichung korrespondieren?
   d) Führt die Untersuchung der wirklichen Funktion zu irgendwelchen Veränderungsvorschlägen?
4. Erreicht das Programm, wenn es entwickelt und verwirklicht ist, sein erwünschtes Ziel?
   a) Sehen die Pläne die Evaluation aller Programmziele vor?
   b) Sind die Evaluationsverfahren reliabel?
   c) Ist der gesamte Evaluationsprozeß umfassend genug, um das benötigte Gesamtbild der Programmergebnisse zu liefern?
   d) Welches sind die Folgerungen aus den Ergebnissen für die Modifikation des Programms? (Lindvall/Cox, 1970, 5–11) [9].

Summative Evaluation, auch Ergebnisevaluation oder Produktevaluation genannt, ist die abschließende Evaluation eines Curriculum. Scriven hat ihre prinzipielle Gleichwertigkeit mit formativer Evaluation betont. Ihre Rolle ist es, nach der Evaluation des Curriculum Daten zur Verfügung zu stellen, die seinen Wert erkennen lassen und die »den Schulsystemen helfen, Entscheidungen über die Adaptation und den Gebrauch von Materialien zu treffen« (Grobman 1968, 14). Evaluation kann dabei einmal auf das Curriculummaterial selbst bezogen werden und seine Validität im Hinblick auf seine Gestaltung und auf seine Lernziele prüfen.

Sie kann aber auch das Verhalten der Schüler untersuchen, die mit diesem Curriculummaterial arbeiten, um festzustellen, welche Fähigkeiten und welche Veränderungen im Verhalten sie erreicht haben.

2. *Mikro- und Makroevaluation*

Innerhalb einer abschließenden Evaluation sind zwei Schwerpunkte möglich. Den einen bildet die *Mikroevaluation*, also die Evaluation von Einzelteilen unter einer bestimmten Fragestellung, den anderen die *Makroevaluation*, die auf eine Beantwortung von allgemeinen Fragen, z. B. über die Benutzbarkeit des Materials, zielt. Summative Evaluation wird im allgemeinen auf die Kombination von Mikroevaluation und Makroevaluation zielen, um zu einer fundierten abschließenden Beschreibung und Wertung des Curriculum zu kommen.

3. *Innere und äußere Evaluation*

Formative Evaluation erfolgt als Teil der Entwicklung eines Curriculum. Sie wird am besten von den an der Entwicklung Beteiligten ausgeführt (innere Evaluation). Da sie in der Entwicklungsarbeit stehen, haben sie Kenntnis vom Gegenstand der Evaluation. In einem stärkeren Maße mit dem Projekt identifiziert als Evaluatoren, die nicht mit der Projektherstellung verbunden sind, werden sie nicht so leicht durch formalen Dogmatismus den kreativen Schwung der übrigen Curriculumhersteller bremsen. Auch können sie eher bei der rechtzeitigen deutlichen Formulierung der Lernziele helfen. Der dynamische Charakter der Curriculumentwicklung fordert eine schnelle Anpassungsbereitschaft an die sich wandelnden Fragen und Probleme der Curriculumhersteller, die nur ein mit dem Projekt mitarbeitender Evaluator leisten kann. Bei formativer Evaluation ist der Zeitpunkt ihres Beginns und die Art ihrer Durchführung zu bedenken. Atkin hat auf diesbezügliche Probleme hingewiesen (1963, 129–132). So kann z. B. das Testen eines schwierigen Vorstellungszusammenhangs zu einem zu frühen Zeitpunkt negative Folgen im Hinblick auf ein weiteres vertieftes Verständnis dieses Zusammenhangs haben.

Summative Evaluation kann von Evaluatoren geleistet werden, die nicht in Verbindung mit Curriculumentwicklung oder -realisierung stehen (äußere Evaluation). Das hat den Vorteil, daß die Evaluationskriterien oft objektiver gewählt werden, mithin auch größere Objektivität bei umfassenden Wertungen über das Curriculum gegeben ist. Auch ist es für nicht an der Projektentwicklung beteiligte Evaluatoren leichter, Vergleichsuntersuchungen verschiedener Curricula vorzunehmen. Ein solches Eval-

uationsteam kann sich aus verschiedenen Spezialisten zusammensetzen, so daß ein großes Spektrum von methodischen Ansätzen gewährleistet wird.

4. Vergleichende und nicht vergleichende Evaluation

Betrachtet man die Ergebnisse der Evaluation neuer Curricula in den USA, die im Vergleich zu Kontrollgruppen erfolgt ist, die mit altem Unterrichtsmaterial arbeiten, fällt im allgemeinen das gute Abschneiden der neuen Curricula und die Unzulänglichkeit der alten Materialien ins Auge; zuweilen allerdings ergeben sich keine signifikanten Unterschiede. Das hat u. a. dazu geführt, daß die Zulänglichkeit der Kriterien des Vergleichs angezweifelt wurde, da sie vor allem im Hinblick auf die Lernziele der neuen Curricula erstellt wurden und für vergleichende Evaluation nicht angemessen waren. Infolgedessen hat man oft dem Urteil von Fachleuten über Inhalt und Ziele und Unterrichtsprozesse mehr Wert zugemessen als empirisch festgestellten geringen Unterschieden von Schülern in Experimentier- und Kontrollgruppen. Cronbach schlägt die genaue Untersuchung der »Leistungen einer genau umschriebenen Gruppe am Ende eines Curriculum im Hinblick auf die wesentlichen Ziele und Nebenwirkungen« (1963, 43, *48*) vor. Durch eine genaue Evaluation des Verhaltens *einer* Gruppe glaubt Cronbach zuverlässigere Daten zu erhalten als durch den methodisch schwierigeren Vergleich verschiedenartig unterrichteter Gruppen. Scriven wendet sich gegen diese Annahme, indem er darauf hinweist, daß die Evaluation eines Curriculum nicht nur die Beschreibung impliziert, sondern auch seine Wertung im Hinblick auf andere einschließt. Ob vergleichende oder nicht vergleichende Evaluation gewählt wird, richtet sich nach der Ausgangssituation und dem Erkenntnisinteresse. Evaluation mit Hilfe des Vergleichs empfiehlt sich dann, wenn die allgemeinen Zielsetzungen gleich sind, d. h. wenn z. B. zwischen zwei Curricula oder zwei Teilen eines Curriculum gewählt werden soll, die auf verschiedenen Wegen das gleiche Ziel erreichen wollen. Eine nicht vergleichende Evaluation ist dann angebracht, wenn die Unterschiede zwischen den Curricula auf Grund verschiedener Zielsetzungen so groß sind, daß die Summe der Gemeinsamkeiten zu klein ist.

5. Intrinsische Evaluation und Ergebnisevaluation

Unter intrinsischer Evaluation wird die Evaluation von beispielsweise Inhalten, Lernzielen, Zensierungsverfahren verstanden, ohne daß ihre Auswirkungen auf die Schüler untersucht werden. Ihre Kriterien sind im

allgemeinen nicht operationalisiert und beziehen sich auf den zu evaluierenden Gegenstand selbst. Die Schwierigkeiten bei der intrinsischen Evaluation liegen darin, daß sie die Kriterien und Ziele formulieren muß, durch die Inhalt, Lehrerverhalten usw. evaluiert werden, zumal die von den Curriculumentwicklern formulierten Lernziele und Kriterien oft nicht gleich den impliziten Lernzielen und Kriterien sind [10].

Unter der reinen Ergebnisevaluation wird dagegen die Untersuchung der Effekte des Curriculum auf die Schüler verstanden. Sie werden z. B. durch Vor- und Nachtest in der Experimentiergruppe festgestellt. Dabei wird oft davon ausgegangen, daß »die Berücksichtigung von Ziel- und Inhaltbewertung oder einer anderen sekundären Bewertung für die Curriculumevaluation nicht nur irrelevant, sondern auch äußerst unzuverlässig« ist (Scriven 1967, 59, 79). Für diese Art von Evaluation ist es weniger wichtig, was die Lehrer und Schüler als ihr Tun *bezeichnen*, wichtig ist, was sie *wirklich tun*. Die Schwierigkeiten bei der reinen Ergebnisevaluation liegen darin, *mit einem sinnvollen Grad an Allgemeinheit* zu beschreiben, was der Schüler gelernt hat.

Aufgabe der Evaluation ist es jedoch nicht nur, die Ergebnisse eines Curriculum festzustellen; Evaluation muß auch die Frage beantworten, wie diese im Verhältnis zu anderen Curricula zu werten sind. Dazu bedarf es der Spezifizierung des Lernerfolges, seiner Begründung und der Formulierung von Zielen und Kriterien. In der Ergebnisevaluation muß entschieden werden, »welches Verhalten zu werten ist als adäquates Verständnis für Schüler auf einem bestimmten Niveau für ein bestimmtes Fach, und es bedarf der Anwendung dieser Entscheidung auf Daten über das spätere Verhalten der Schüler, um eine Gesamtevaluation zu ermöglichen« (Scriven 1967, 61). Es muß also das Verhalten der Schüler analysiert werden, um zu bestimmen, wo Unzulänglichkeiten im Hinblick auf bestimmte Kriterien oder Lernziele liegen. Im allgemeinen verbindet man intrinsische und Ergebnisevaluation, um zu möglichst umfassenden Ergebnissen zu kommen.

III. Modelle der Curriculumevaluation

1. Das Stakesche Evaluationsmodell

Einen Schritt weiter in die Problematik der Konzeptualisierung von Evaluation führt der Entwurf von Modellen für die Curriculumevaluation. Sie liegen auf einer höheren begrifflichen Stufe als die Distinktionen zur Beschreibung unterschiedlicher Evaluationsformen und Evaluationsfunktionen. Geht man wieder davon aus, daß die Verfahren der Evaluation

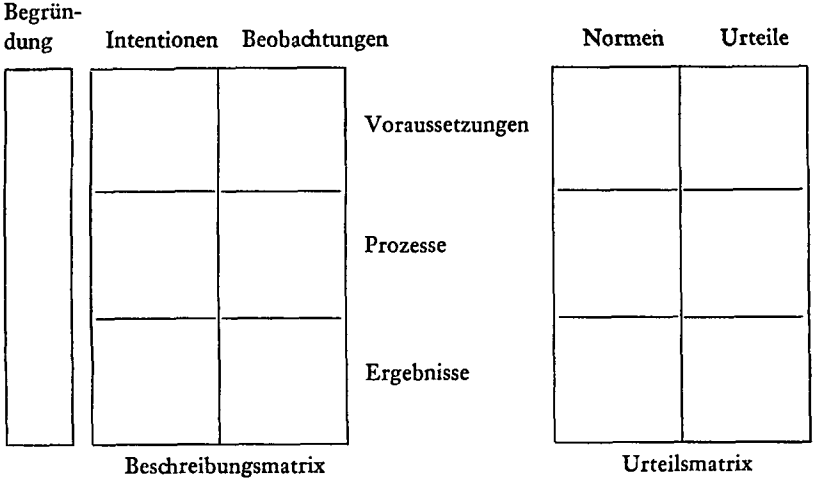
beabsichtigen, relevante Informationen für die Verbesserung von Curricula zu erhalten, so sind vor allem drei Felder wichtig:
- das Vorfeld des Unterrichts,
- das Aktionsfeld des Unterrichts,
- das Feld seiner Wirkungen.

Etwas detaillierter beschrieben, heißt das: Das erste Evaluationsfeld enthält alle Bedingungen, die vor dem Unterricht bestehen und Einfluß auf seine Ergebnisse haben. Zu diesen gehört die Erhebung der sozialen Situation der Schüler, ihrer Bedürfnisse und Interessen und ihrer Umwelt mit ihren Wert- und Zielvorstellungen.

Zum zweiten Evaluationsfeld gehören alle Bedingungen innerhalb der Schule, die für die Erreichung der Ergebnisse wichtig sind. Es schließt den Prozeß der Einbringung des Curriculum in die Schule ein, die Einführung des Lehrers in die Arbeit mit den curricularen Materialien und vor allem die Unterrichtsprozesse, die Interaktionen der Lehrer mit den Schülern, der Schüler untereinander und der Schüler mit dem Unterrichtsmaterial. Das dritte Evaluationsfeld beinhaltet die augenblicklichen, mittelfristigen und langfristigen Ergebnisse des gesamten Unterrichts (vgl. Sjogren 1970).

Stake (1967 a) hat die drei Evaluationsfelder *Voraussetzungen, Prozesse, Ergebnisse* genannt und eine verstärkte Evaluation der Voraussetzungen und der Prozesse gefordert. Zu diesem Zweck entwickelt er eine Eval-

Evaluationsmatrix nach Stake



uationsmatrix, indem er über die drei Evaluationsfelder zwei Beschreibungskategorien, »Intentionen« und »Beobachtungen«, und zwei Urteilskategorien, »Normen« und »Urteile«, legt. Dadurch entsteht eine zwölffeldige Evaluationsmatrix. Ergänzende Informationen für ihre Anwendung bietet die Kenntnis der rationalen Begründung des Curriculum, dessen Zielsetzungen mit ihrer Hilfe besser zu erfassen sind.

Unter »Intentionen« werden z. B. die geplante Erarbeitung von Sachgebieten oder das erwartete Schülerverhalten verstanden. Sie sollten möglichst präzise verbalisiert werden. Die Diskrepanz zwischen Lernzielen und Lernergebnissen ist nicht ohne weiteres Zeugnis für einen negativen Verlauf des Unterrichts. Es können unvorhergesehene Ergebnisse erzielt worden sein, die durchaus positiv zu werten sind. Ob Lernziele *immer* vor dem Unterricht festgelegt und ob sie *immer* in Verhaltenstermini bestimmt werden müssen (vgl. Wulf 1972 b), ist in der amerikanischen Curriculumforschung umstritten [11]. »Beobachtungen« bilden die zweite Beschreibungskategorie der Matrix. Hier geht es um die Beobachtung dessen, was in den drei Evaluationsfeldern beschreibenswert erscheint. Es setzt die Auswahl von Kriterien voraus, unter denen zu beobachten ist. Diese ist eine Entscheidung des Evaluators, bei der es im Curriculum formulierte Kriterien und von außen herangetragene zu berücksichtigen gilt. Es bieten sich zwei Wege an, die Daten der beiden Beschreibungskategorien zu verarbeiten: *»Man muß die Kontingenzen zwischen den Voraussetzungen, Prozessen und Ergebnissen und die Kongruenz zwischen den Intentionen und Beobachtungen finden«* Stake 1967 a, *104*). Die Kontingenzen zwischen den einzelnen Variablen der drei Evaluationsfelder bedürfen der Beachtung: »Insofern, als Evaluation die Suche nach Beziehungen ist, die die Verbesserung der Erziehung ermöglichen, ist es Aufgabe des Evaluators, Ergebnisse zu identifizieren, die mit bestimmten Voraussetzungen und Unterrichtsprozessen kontingent sind« (a. a. O.).

Zwischen den drei Evaluationsfeldern besteht in der ersten Beschreibungskategorie, »Intentionen«, eine logische Kontingenz, im Bereich der zweiten, »Beobachtungen«, eine empirische Kontingenz. Vollständige Kongruenz der Daten eines Curriculum besteht, wenn alle beabsichtigten Voraussetzungen, Prozesse und Ergebnisse eintreten. Das ist selten der Fall. Im allgemeinen wird der Evaluator Kongruenz lediglich bei einigen Teilzielen feststellen können. Kongruenz besagt nur, daß das, was beabsichtigt war, eintrat, nicht jedoch, daß die Ergebnisse *reliabel* oder *valide* sind.

Die Urteilsmatrix besteht aus zwei Kategorien, den *»Normen«* und den *»Urteilen«*. Mit welchen Methoden und Strategien die Normen zu erstellen sind, ist ein zentrales Problem der Evaluation. Normen sind im allgemeinen gebunden an die historisch-gesellschaftliche Situation. Mit

den Lernzielen können sie nicht gleichgesetzt werden, wie es in einigen Evaluationsprojekten geschehen ist, da dann Evaluation nur Überprüfung der Erfüllung von Lernzielen wäre. Vielmehr müssen auch die Lernziele an Normen gemessen werden. Zwei Arten von Normen sind denkbar, *absolute* und *relative*. Letztere sind beispielsweise in einem alternativ vorliegenden Curriculum enthalten. Wenigstens an einem der beiden Normengefüge muß das Curriculum durch Urteilen gemessen werden, um seinen Wert zu bestimmen.

2. Das Provussche Diskrepanzmodell

Provus macht darauf aufmerksam, daß die entscheidenden evaluativen Informationen durch die Feststellung von Diskrepanz gewonnen werden. Danach besteht der Prozeß der Evaluation darin,

- Normen festzulegen,
- die Diskrepanz zwischen dem Curriculum und den Normen festzustellen und
- die Diskrepanzinformation dazu zu benutzen, die Schwächen des Curriculum zu identifizieren.

Die Evaluation eines Curriculum soll vier Entwicklungsstadien durchlaufen und dabei drei größere Inhaltskategorien einschließen. Das erste Entwicklungsstadium ist die *Definition* des Curriculum, das zweite seine *Einführung* in die Schule (installation), das dritte die curricularen *Prozesse*, das vierte seine Fertigstellung als *Produkt*. Nach diesen vier kann als fünftes Entwicklungsstadium die *Kosten-Effektivitäts-Berechnung* erfolgen. In den vier bzw. fünf Entwicklungsstadien »bleiben die inhaltlichen Spezifikationen dieselben: *Eingabe*, *Prozeß* und *Ergebnis*, relativ zu Zeit und Geld« (Provus 1969, 244).

Unter *Definition* eines Curriculum wird seine genaue Beschreibung im Hinblick auf seine Lernziele, die Schüler, Lehrer, Medien und den Prozeß verstanden, der zur Erreichung der Lernziele führt. Sie dient als Norm zur Evaluation des Curriculum. Werden im (zweiten) Entwicklungsstadium, der *Einführung* des Curriculum, Informationen gewonnen, die eine Diskrepanz zwischen den Erwartungen hinsichtlich der Einführung des als Norm definierten Curriculum und dem tatsächlichen Lehrer- und Schülerverhalten anzeigen, muß entweder das Verhalten so korrigiert werden, daß es der Norm entspricht, oder es muß die Norm verändert werden. In dem (dritten) Entwicklungsstadium, dem *Prozeß*, bildet der Teil der Curriculumdefinition die Norm, der die Intentionen hinsichtlich des Unterrichtsprozesses enthält. Wird eine Diskrepanzinformation zwischen Intentionen und tatsächlichem Prozeß erhalten, muß sie zur Neudefini-

tion der Intentionen oder zu einer besseren Durchführung des Unterrichtsprozesses führen. Im (vierten) Entwicklungsstadium, in dem es um das *Produkt* des Unterrichts geht, gilt es festzustellen, ob das Curriculum seine in der Definition als Norm gesetzten End- und Gesamtlernziele erreicht hat oder ob Diskrepanz festgestellt werden muß.

Mit diesem Evaluationsmodell wird die Evaluationskategorie der *Diskrepanz* gewonnen. Darüber hinaus bietet Provus brauchbare Hinweise auf die methodischen Probleme seiner Anwendung und die spezifischen Aufgaben von Evaluationsgruppen, die hier nicht mehr berücksichtigt werden können.

Da es ein Modell für die Evaluation laufender curricularer Programme ist, fehlt die Reflexion auf die Aufgaben der Evaluation zur »Erhebung der Werte und Bedürfnisse«. Es fehlt auch die Reflexion auf die Notwendigkeit einer auf den Ergebnissen der Produktevaluation fußenden Revision des Curriculum. Außerdem werden die in den einzelnen Entwicklungsphasen notwendigen Entscheidungen nicht als Probleme der Wertung innerhalb der Evaluation gesehen.

3. Das Stufflebeamsche Evaluationsmodell

Gerade die Transparenz von Entscheidungsprozessen bildet das Schwergewicht der Bemühungen Stufflebeams (1969; vgl. auch: Phi Delta Kappa National Study Committee on Evaluation 1971). In diesem Rahmen werden sieben Entscheidungsvariablen unterschieden, die in den verschiedenen Entscheidungssituationen heranzuziehen sind:

- die Ebene der Entscheidungen,
- der Schwerpunkt der Entscheidungen,
- die Inhalte der Entscheidungen,
- die Funktion der Entscheidungen,
- die Gegenstände der Entscheidungen,
- der Zeitpunkt der Entscheidungen,
- der Grad an kritischer Reflektiertheit der Entscheidungen.

Diese Variablen gilt es auf den drei von Stufflebeam unterschiedenen interdependenten Ebenen der Evaluation (Gemeinde, Bundesstaat, Nation) zu bedenken. In diesem Modell werden vier Situationen und die entsprechenden Entscheidungsfunktionen (Planung, Programmentwicklung, Implementation, Modifikation) genannt, für die vier Strategien entwickelt werden. Die erste zielt auf die Evaluation des *Kontexts*. Hierbei gilt es, – ähnlich der ersten der fünf oben unterschiedenen Phasen der Curriculumentwicklung – die Bedürfnisse im Vorfeld zu erheben. Die zweite Strategie, die auf die *Eingabe* (input) gerichtet ist und deren

Aufgabe die Analyse alternativer Verfahrensentwürfe ist, will Daten für Entscheidungen über die Spezifizierung von Lernzielen, Verfahren, Materialien usw. sammeln. Die dritte beabsichtigt die Erhebung von Daten, die über den *Prozeß* gewonnen werden und die zur Verbesserung des Curriculum helfen sollen. Die vierte Evaluationsstrategie zielt darauf ab, die *Gesamtergebnisse* zu messen.

Darüber hinaus hat Stufflebeam (1969, 70, *143*) versucht, die Struktur eines Evaluationsentwurfs zu entwickeln, der sechs Kategorien und jeweils einige Unterkategorien enthält. Er dürfte wegen seines formalen Charakters als Leitlinie für Evaluationsprojekte Beachtung verdienen, obwohl er nicht die unter den einzelnen Kategorien subsumierten Probleme der Wertung und immanenten Entscheidungen reflektiert.

A. Evaluationsschwerpunkt
1. Identifikation der wichtigsten Entscheidungsebenen (z. B. örtliche, einzelstaatliche und/oder bundesstaatliche)
2. Planung und Beschreibung aller Entscheidungssituationen auf jeder Entscheidungsebene in bezug auf ihren Schwerpunkt, die kritische Reflektiertheit, den Zeitpunkt und die Komposition der Alternativen
3. Bestimmung der Kriterien für jede Entscheidungssituation durch Spezifikation der Variablen für die Messungen und der Normen für die Beurteilung von Alternativen
4. Definition der Grundsätze und Richtlinien, innerhalb derer die Evaluation erfolgen soll

B. Informationssammlung
1. Spezifikation des Ursprungs der zu sammelnden Informationen
2. Bestimmung der Instrumente und Methoden für die Sammlung der erforderlichen Informationen
3. Spezifikation des anzuwendenden Stichprobenverfahrens
4. Spezifikation der Bedingungen und des Zeitplans für die Informationssammlung

C. Informationsorganisation
1. Erstellung eines Plans für die Informationen, die gesammelt werden sollen
2. Bestimmung der Mittel zur Kodierung, Organisation, Speicherung und zum Wiederabruf der Informationen

D. Informationsanalyse
1. Auswahl der analytischen Verfahren, die angewendet werden sollen

2. Bestimmung der Mittel zur Durchführung der Analyse
3. Spezifikation des Ausmaßes und der Form der Evaluationsberichte
4. Zeitplan des Informationsberichts

E. Informationsbericht
1. Definition der Adressatengruppe
2. Bestimmen der Mittel der Informationsvermittlung
3. Festlegen des Formats des Evaluationsberichts
4. Planung der Elemente für die Darstellung der Information

F. Administration der Evaluation
1. Zusammenfassung des Evaluationsplans
2. Bestimmung der für die Evaluation erforderlichen Mitarbeiterstellen und Finanzen
3. Spezifikation der Mittel, um die Evaluation gemäß ihren Grundsätzen und Richtlinien durchzuführen
4. Evaluation der Möglichkeiten des Evaluationsplans, valide, reliable, zuverlässige, aktuelle und überzeugende Informationen zu liefern
5. Spezifikation und zeitliche Planung der Mittel, um den Evaluationsplan regelmäßig auf den neuesten Stand zu bringen
6. Breitstellung eines Etats für das ganze Evaluationsprogramm.

IV. Das Wert- und Entscheidungsproblem in der Evaluation

Das dritte Problemfeld der Evaluationsforschung führt in das auch wissenschaftstheoretisch strittige Feld der Wertung (vs Wertfreiheit bzw. Wertneutralität). Damit kommt auch die in der BRD geläufige Opposition von deskriptiven und präskriptiven Verfahren ins Spiel.

In der Curriculumevaluation befaßt man sich erst seit kurzem mit der Problematik der Wertimplikationen beim Urteilen und bei Entscheidungsprozessen. So kann Stake (1970, 186) urteilen: »Die meisten Arbeiten über die Methodologie der Evaluation erwähnen kein Verfahren, Urteilsdaten zu sammeln und zu analysieren«.

Westbury kennzeichnet das Verhältnis von Evaluation und Beschreibung: »Evaluation kann (und muß wahrscheinlich) Beschreibung einschließen, aber Beschreibung schließt nicht notwendigerweise Evaluation ein« (1970, 241). Nach Larkins/Shaver liegt die Aufgabe der Evaluation neben Datensammlung und Analyse in der Berücksichtigung von Werten, Wertinhalten und -kriterien zur Beurteilung eines Curriculum.

»Ein adäquater Evaluationsentwurf fordert sowohl Verfahren, um ad-

äquate Schätzungsdaten (assessment data) – Schätzungsentwurf – zu erhalten, als auch Verfahren, um den Wert der Schätzungsdaten abzuwägen. Diese letzteren Verfahren sollen von nun an als *Wertanalyse* bezeichnet werden. Ein Modell für den Wertanalyseteil eines Evaluationsentwurfs schließt substantielle und Verfahrensüberlegungen ein. Die grundlegenden substantiellen Überlegungen sind der Satz von Werten, der benutzt wird, das Curriculum zu beurteilen. Die Verfahrensüberlegungen schließen ein, wie relevante Werte zu entwickeln, wie Wertkonflikte zu lösen, wie Entscheidungskriterien festzusetzen sind, um zu bestimmen, ob ein Curriculum einem Wert angemessen genügt« (Larkins/Shaver 1969, 6).

Jensen (1950) entwickelt ein Verfahren, die Rationalität der Auswahl von Lernzielen unter Zugrundelegung einer bestimmten Wertposition zu belegen. Auf jeden Fall sollte der Evaluator Kenntnis von verschiedenen Schemata zur Kategorisierung von Werten haben [12], weil dadurch die Spezifizierung von Lernzielen vereinfacht wird.

1. Werte, Prioritäten, Normen

Unterschiede in den strittigen öffentlichen Fragen beruhen auch auf unterschiedlichen Auffassungen von ethischen Werten. Neben den Wertvorstellungen, die aus direkten Interessen entstehen, bilden die ethischen Werte eine wichtige Basis für die Aufstellung von Bildungszielen und Lernzielen. Auch das ist bisher in der Curriculumforschung kaum reflektiert worden, so daß relativ wenig von den Wertvorstellungen der an der Erziehung Beteiligten bekannt ist. Der verbreitete Tylersche Weg der Lernzielbestimmung berücksichtigt kaum die den Lernzielen vorgegebenen Werte und ihre kritische Sichtung (vgl. Smith/Tyler 1942 u. Tyler 1969 b). Die Hinzuziehung von Sozialwissenschaftlern für die Erforschung der Wertvorstellungen wird wiederholt gefordert. In diesem Zusammenhang beginnt die Pädagogik auch die lange vernachlässigten Beziehungen zwischen Erziehungswissenschaft und Ethik wieder aufzunehmen [13]. Unter dem Gesichtspunkt der Sammlung von Informationen für die Evaluation lassen sich verschiedene Arten von Wertdaten unterscheiden. Außer *Lernzielen, summativen (Wert-)Urteilen* sind noch *Prioritäten* und *Normen* als Wertdaten zu nennen. Die Bestimmung der Prioritäten hat im Vergleich zu der genauen Lernzielbestimmung wenig Beachtung gefunden. Es liegt bisher kein Modell vor, um begründet festlegen zu können, was in einer Folge von Lernzielen oder Ereignissen den Vorrang hat. Hier bedarf es weiterer Theorienbildung und Forschung.

Maguire (1968) hat versucht, empirisch die Beziehungen zwischen Wer-

ten, Lernzielen und Prioritäten zu untersuchen. Er erhielt Werteinschätzungen von einem heterogenen Satz von Lernzielen durch Lehrer und sodann von jeweils denselben Lehrern verschiedene Urteile darüber, welchen Lernzielen sie die Priorität geben wollten. Dabei stellten sich vor allem vier verschiedene Wert-Dimensionen der Lernziele in den Augen der Lehrer heraus: fachwissenschaftlicher Wert, motivationale Qualität, Einfachheit der Implementation und semantische Eigenschaften. Eine andere Kategorie von Beurteilungsdaten bilden »Standards«, die von Stake »als ein erwünschtes Qualitätsniveau für etwas, wie es von einer Autorität angegeben wird« (1970, 185), begriffen werden. Ihnen entspricht im deutschen Sprachgebrauch am ehesten der Begriff Norm.

Im Anschluß an Stake lassen sich drei Situationen identifizieren, in denen der Evaluator Daten über Werte, Prioritäten, Normen und Lernziele sammeln kann:

- die Befragung von Personen nach einem standardisierten Protokoll,
- die Beobachtung des Personenverhaltens,
- die Analyse von Curricula durch Experten.

Je nach Zielsetzung der Evaluation kann sich eine Kombination der Situationen empfehlen.

Vier empirische Methoden bieten sich an, um die Wertvorstellungen von Befragten zu ermitteln:

- Survey,
- Skalierung,
- Q-Technik,
- Semantisches Differential (Polaritätsprofil).

Survey-Befragungen empfehlen sich dort, wo gute Daten durch direkte Befragung erhalten werden können [14]. Wenn jedoch die einzelnen Fragen mehrdeutig sind oder das Ziel der Frage unklar ist, empfehlen sich stärker redundante Verfahren wie Skalierung [15]. Eine besondere Art ihrer Verarbeitung ist die 1953 von Stephenson entwickelte Q-Technik, die sich zur Ermittlung von Werten eignet [16]. Das semantische Differential ist ebenfalls ein spezielles Skalierungsverfahren. Mit ihm kann festgestellt werden, welche Bewertung Menschen bestimmten Begriffen und Vorstellungen, wie z. B. »Gewaltenteilung«, zuordnen. Es wurde 1957 von Osgood, Suci und Tannebaum entwickelt. Von Geis wurde dieses Verfahren 1968 zur Evaluation eines Projekts zur Verbesserung eines Curriculum verwendet. Taylor und Maguire benutzten es, um Biologielernziele zu untersuchen (1967).

Um Personenverhalten im Bereich der Schule zu beschreiben und zu bewerten, sind in den letzten Jahren viele Beobachtungssysteme entworfen und erprobt worden. Sie haben verschiedene Ansatzpunkte und beto-

nen unterschiedliche Aspekte. Mehrere versuchen, aus der Beobachtung des Verhaltens die ihm immanenten Werte, Normen und Ziele zu erfassen; andere begnügen sich mit der Beschreibung der Interaktionen zwischen Lehrer und Schülern oder untersuchen die Logik des Lehrens. Eine wertvolle Sammlung von Instrumenten zur Beobachtung, Beschreibung und Bewertung von Verhalten (vor allem im Unterrichtsprozeß) enthalten die Bände »Mirrors for Behavior« (Simon/Boyer 1967, 1970). Sie geben eine detaillierte Übersicht über 79 Beobachtungsverfahren, so daß auf eine Darstellung einzelner Systeme hier verzichtet wird.

Für die dritte Situation empfehlen sich die Techniken der Inhaltsanalyse, mit deren Hilfe sowohl Aufschluß über den Inhalt selbst als auch über seine Beziehung zum Produzenten und zum Konsumenten geliefert werden soll [17].

2. *Der programmatische und der öffentlich-politische Bereich*

Ein Kategorisierungsschema für Wertvorstellungen, Lern- und Bildungsziele, für Prioritäten und vor allem Ergebnisse von Curricula im Hinblick auf ihren politischen Charakter liefert Berlak (1970). Er unterscheidet zwischen *programmatischen* (programatic) und *öffentlich-politischen* (public-policy) Ergebnissen (outcomes) von Curricula. Diese Bereichsunterscheidung geschieht im Bewußtsein des eminent politischen Charakters der Erziehung. Das Politische wird erstens der Maßstab, mit dessen Hilfe die Ergebnisse des Unterrichts in solche von öffentlich-politischem Interesse und solche programmatischen Charakters unterteilt werden, zweitens der Maßstab, Wertvorstellungen von großer öffentlich-politischer Bedeutung von solchen geringerer politischer Relevanz zu unterscheiden, drittens der Maßstab für das Recht des Evaluators, Urteile abzugeben. Daß auch der programmatische Bereich nicht ohne politische Implikationen ist, wird bei der Unterscheidung der beiden Bereiche durch Berlak nicht übersehen. Die programmatischen Ergebnisse von Curricula werden durch »Lernziele« angestrebt, die öffentlich-politischen durch »Bildungsziele«.

Der Evaluator muß sich bewußt werden, ob er »Lernziele« im programmatischen Bereich oder »Bildungsziele« im öffentlich-politischen Bereich evaluiert. Danach richtet es sich, ob er beschreibt, Urteilskriterien empfiehlt oder urteilt. Dabei stellt sich das Problem, wie die Grenze zwischen den beiden Bereichen bestimmt wird. Berlak nennt vier Kriterien zur Bestimmung des öffentlich-politischen Bereichs, die er in Form von Fragen formuliert:

1. Ändert des Programm direkt oder indirekt das Verhältnis der Macht zwischen Bürger und Staat?

2. Betrifft das Programm sofort oder später den Status einer Person oder die Macht, die sie in dem sozialen System ausüben kann?
3. Hat das Programm einen Effekt, der danach strebt, politische oder soziale Spannungen zu vermehren oder abzubauen?
4. Bewirkt das Programm einen Wechsel im Selbstverständnis oder Selbstgefühl des Individuums?

Öffentlich-politische und programmatische Ergebnisse können *geplant* oder *ungeplant* und, wenn ungeplant, noch halbwegs *erwartet* oder völlig *unerwartet* auftreten. Je nach der Zugehörigkeit zum öffentlich-politischen oder programmatischen Bereich bedarf es unterschiedlicher Evaluationsstrategien, die von verschieden ausgebildeten Personen angewendet werden. Eine sorgfältige Unterscheidung der beiden Bereiche bietet sich aus politischen und finanziellen Gründen an. Soll die Erreichung oder Verfehlung von vorher formulierten Lernzielen evaluiert werden, handelt es sich im allgemeinen um den programmatischen Bereich. Werden dagegen die Wertvorstellungen und Meinungen der Eltern über einschneidende schulorganisatorische Veränderungen evaluiert, befindet man sich im öffentlich-politischen Bereich.

Für den Evaluator schlägt Berlak folgendes Verhalten vor: Im programmatischen Evaluationsbereich ist der Evaluator aufgerufen, Werturteile abzugeben. Im Bereich der öffentlich-politischen Fragen ist es beim jetzigen Wissenschaftsstand angemessen, das Schwergewicht auf die Beschreibung zu legen, da es für diesen besonderen Bereich noch keine Evaluationsexperten gibt. Bei der Evaluation kann es in beiden Bereichen zu einander widersprechenden Ergebnissen kommen, mithin auch zu Konflikten, die ausgetragen werden müssen. So könnte z. B. die Evaluation von Curricula im programmatischen Bereich in den Gesamtschulen zeigen, daß die kognitiven Ergebnisse geringer sind als die in herkömmlichen Schulen. Dennoch kann die Evaluation im öffentlich-politischen Bereich zeigen, daß diese Schulart erfolgreich ist, z. B. bei der Verwirklichung des Bildungsziels der optimalen Sozialisation des einzelnen.

3. Aspekte der Urteils- und Entscheidungsproblematik der Evaluation

Umstritten ist in der pädagogischen Diskussion, ob Wertkonflikte rational lösbar sind, ob Werturteile bewiesen und widerlegt werden können, »ob ethische Urteile in objektiver Weise gerechtfertigt werden können« (Berlak 1970, 269). Sind Wertkonflikte nicht rational lösbar, so können diejenigen, die Entscheidungsmacht haben, Ergebnisse der Evaluation auf Grund ihrer Wertvorstellungen ablehnen. Geiger (1961) und Scriven (1966) treten entschieden für die Lösbarkeit ethischer Differenzen mit-

tels rationaler Strategien ein, wie sie in den Sozialwissenschaften verwendet werden. Berlak (1970) nimmt im Anschluß an Stevenson (1944) eine Zwischenposition ein. Er tritt für die Lösbarkeit einiger Aspekte von ethisch strittigen Fragen ein, andere hält er nicht für lösbar. Schwierig dürfte die Bildung von Kriterien zur Zuordnung der Aspekte zu einer der Wertkategorien sein. Die Frage nach der Rechtfertigung von Werturteilen in objektiver Weise berührt die komplizierte Problematik von Entscheidungen im curricularen Bereich. Nach einer Unterscheidung von Cronbach und Suppes (1969) zielt eine Gruppe von Evaluationsuntersuchungen auf ein besseres Verständnis des untersuchten Gegenstandes, eine andere auf die wissenschaftliche Vorbereitung von Entscheidungsprozessen.

Es besteht Übereinstimmung darüber, daß das letztere eine zentrale Aufgabe der Evaluation ist. Um so erstaunlicher ist es, daß erst seit kurzem dieses Problem ausdrücklich thematisiert wird und daß bisher kaum Untersuchungen über Entscheidungsabläufe in diesem Bereich vorliegen. Es werden deskriptive Studien der Entscheidungsabläufe im pädagogischen Bereich gebraucht, die dazu beitragen können, den Entscheidungsprozeß in curricularen Fragen durchsichtig zu machen. Entscheidungsträger müssen befragt und beobachtet werden. Einige von ihnen sind fähig, den Prozeß, der zu den Entscheidungen führt, zu verbalisieren, bei anderen wird man nur versuchen, die Prinzipien und Kriterien der Urteile von ihrem Verhalten abzulesen. Beide Arten der Untersuchung können zur Entwicklung von Modellen und Kriterien der Entscheidungsabläufe führen. Darüber hinaus dürfte auch die Untersuchung von Entscheidungsabläufen unter experimentellen Bedingungen wertvolle Hilfe zur Entwicklung derartiger Modelle und Kriterien bieten. Ihrer Kenntnis bedarf der Evaluator, damit er die Ergebnisse der Evaluation so formulieren kann, daß sie den individuellen Bedingungen der Entscheidungssituation Rechnung tragen und die Informationen enthalten, die für die betreffenden Entscheidungen wichtig sind.

Gooler (o. J.) hat ein Verfahren entwickelt, das es einerseits dem Entscheidungsträger erleichtert, diejenigen Informationen zu finden, die er für die Entscheidung eines komplexen Problems benötigt, das andererseits dem Evaluator eine Möglichkeit bietet, die Datensammlung so anzulegen, daß sie die für den Entscheidungsprozeß relevanten Daten enthält. Es umfaßt in fünf Schritten die Identifikation

- der Gruppe von Hauptinteressenten,
- von Schlüsselfragen, die diese Gruppe hat,
- derjenigen Daten, die benötigt werden, um die Schlüsselfragen zu beantworten,

- von Quellen für diese Daten,
- von Methoden, diese Daten zu erhalten.

Je präziser die Vorstellungen des Entscheidungsträgers in einzelnen Punkten sind, desto größer ist die Hilfe für die Evaluatoren, die für die Entscheidung wichtigen Daten zu sammeln. Dabei empfiehlt sich eine Kooperation von Entscheidungsträgern und Evaluatoren zu einem möglichst frühen Zeitpunkt. Die hier entwickelte Methode bietet wenig mehr als eine Formalisierung des Datensammlungsprozesses, dessen Ergebnisse eine begründete Entscheidung ermöglichen. Weitere Untersuchungen zum Beitrag der Evaluation zu Entscheidungsprozessen werden dringend benötigt (vgl. Wulf 1972c). Dabei muß u. a. die von Stufflebeam und Gooler nicht problematisierte Kompetenz und Legitimität der Entscheidungsträger und die Effektivität ihrer Entscheidungen im Hinblick auf ihre Folgen thematisiert werden. Auch gilt es Prinzipien und Regeln zu entwickeln, die den Entscheidungsprozeß leiten, bzw. Kriterien, auf Grund derer er beurteilt werden kann. Darüber hinaus gilt es zu untersuchen, inwieweit in der Evaluation gewonnene Daten überhaupt Einfluß auf Entscheidungen haben.

V. Probleme der Curriculumevaluation

Curriculumevaluation steht in den USA und in der BRD vor gleichen Aufgaben und Problemen. Im Augenblick erscheint es notwendig, Begriffe und Modelle, Methoden und Instrumente der Evaluation zu entwikkeln, um so die immer neu und präzise zu formulierenden Zielsetzungen der Evaluation einzelner Curricula zu verwirklichen. Die Ausweitung der zu berücksichtigenden Gesichtspunkte in der Curriculumevaluation in den letzten Jahren hat noch nicht die anstehenden Probleme gelöst, vielleicht jedoch ihre Lösung ermöglicht. Manche der dargestellten Ansätze sind es wert, aufgegriffen und weiterentwickelt zu werden.

Die Entwicklung von Theorie und Methodologie der Evaluation kann nur in engem Zusammenhang mit der tatsächlichen Evaluation von Curricula erfolgen. Das bedeutet: es muß wenigstens bei der Planung einiger Curricula das Schwergewicht auf ihre Evaluation gelegt werden mit dem Ziel, die Probleme der Evaluation zu erforschen. Dabei gilt es zugleich, die Aufmerksamkeit auf die Ausbildung von Wissenschaftlern zu richten, die sich der vielschichtigen Probleme der Evaluation bewußt sind und ihre zahlreichen Methoden beherrschen [18]. Im Rahmen einer zu entwickelnden Theorie und zu erarbeitender Methoden der Curriculumevaluation sind vor allem folgende Probleme zu lösen:

1. Welches ist die gesellschaftliche Funktion der Evaluation?
2. Wie wird die Beziehung zwischen Curriculum und Evaluation bestimmt?
3. Wie werden Kriterien für die Evaluation von Curricula gewonnen?
4. Wie wird das Verhältnis von Beschreibung und Werturteilen in der Evaluation bestimmt?
5. Inwieweit ist es notwendig, das Vorfeld, das Aktionsfeld und das Wirkungsfeld in den verschiedenen Phasen der Curriculumentwicklung zu evaluieren, und welche Methoden sind angemessen?
6. Welche Funktion hat Evaluation für die verschiedenen Ebenen der Entscheidung in den unterschiedlichen Phasen der Curriculumentwicklung?
7. Welchen Beitrag kann Evaluation zur Auswahl und Revision von Lernzielen, zur Entwicklung von Curricula, zur Entwicklung spezifischer Lernsituationen und -methoden, zur Optimierung des Lernens von spezifischen Schülergruppen und Individuen leisten?
8. Wie kann eine Technologie der Evaluation entwickelt werden, und welche Instrumente und Methoden muß sie ausarbeiten?

II Grundfragen der Evaluation

Einführung

Im Rahmen der umfangreichen Anstrengungen zur Bildungsreform sind der Evaluation neue Aufgaben gestellt worden, denen die überkommenen Vorstellungen von Evaluation, wie sie z. T. in Anlehnung an die zu seiner Zeit bahnbrechenden Arbeiten Ralph Tylers entwickelt worden sind, längst nicht mehr gerecht werden konnten. Im folgenden Abschnitt wird der deutsche Leser mit den Beiträgen vertraut gemacht, die dieses neue Verständnis von Evaluation grundlegend bestimmt haben. Dabei ist es von entscheidender Bedeutung, daß diese Beiträge zu einem Zeitpunkt Berücksichtigung finden, in dem Evaluation im Zusammenhang mit den gegenwärtigen Reformen im Bildungswesen zu einem zentralen Bereich pädagogischer Forschung in der BRD wird. Die im Rahmen der amerikanischen Bildungsreform der letzten Jahre entstandenen Beiträge können wesentlich zu einem besseren Verständnis der Aufgabenbestimmung von »Begleitforschung« bzw. Evaluation beitragen. Die hier ausgewählten Aufsätze haben als einen gemeinsamen Hintergrund die Reformanstrengungen der sechziger Jahre, thematisieren aber innerhalb dieses Bezugsrahmens sehr unterschiedliche Aspekte der Evaluation.

Ausgehend von der Überzeugung, daß Evaluation nach Abschluß der Curriculumentwicklung kaum noch etwas zur Verbesserung des Curriculum beitragen könne, fordert Cronbach die Integration der Evaluation in den Prozeß der Curriculumentwicklung. Nur so kann das Potential der Evaluation voll genutzt werden, da zu diesem Zeitpunkt noch eine Verbesserung des Curriculum auf Grund der gewonnenen Daten möglich ist. Um die für diese Form der Evaluation benötigten Daten zu gewinnen, empfiehlt Cronbach die Aufgabe des klassischen »research design« mit Versuchs- und Kontrollgruppen und an seiner Stelle die genaue Untersuchung einzelner ausgewählter Versuchsgruppen. Dabei sollen in den einzelnen Gruppen unterschiedliche Testaufgaben aus einer umfangreichen Aufgabensammlung benutzt werden, da man auf diese Weise mehr In-

formationen über das Curriculum erhalten kann als bei Verwendung eines für alle Gruppen gemeinsamen Fragebogens.

Scriven greift den Gedanken der die Curriculumentwicklung und jede pädagogische Innovation begleitenden Evaluation auf und nennt sie »formative Evaluation«. Er betont aber im Unterschied zu Cronbach auch die Wichtigkeit einer »summativen Evaluation« nach Abschluß der Curriculumentwicklung oder des Schulversuchs. In ihr müsse eine Bewertung des Bildungsprogramms erfolgen, die es dem Adressaten der Innovation erlauben würde, sie im Vergleich zu anderen Projekten zu sehen, so daß etwa im Falle von Curriculummaterialien die Schulen die besten von vergleichbaren Materialien für sich auswählen könnten. Neben der Unterscheidung zwischen diesen beiden Formen der Evaluation erfolgen weitere für die Konzeptualisierung von Evaluation wichtige Differenzierungen, wie z. B. zwischen Evaluation und Überprüfung des Erreichens von Lernzielen, zwischen intrinsischer Evaluation und Ergebnisevaluation.

Stake greift in seinem Beitrag einige der Gedankengänge Cronbachs und Scrivens auf und integriert Beschreibung und Beurteilung (Bewertung) als Dimensionen der Evaluation in sein Evaluationsmodell. Über Cronbach und Scriven hinausgehend betont er die Notwendigkeit, neben den Prozessen und Ergebnissen auch die Voraussetzungen einer Evaluation zu untersuchen, um Reformen angemessen, d. h. in einem Bezug auf die Bedürfnisse der Adressaten, planen und entwickeln zu können.

Ähnlich umfassend ist Evaluation für Stufflebeam, der Kontext, Input, Prozeß und Ergebnis eines Bildungsprogramms der Evaluation unterziehen will. Für ihn besteht die zentrale Aufgabe der Evaluation darin, den Entscheidungsträgern in Schule, Schulverwaltung, Bundesministerium und Parlament die Informationen zur Verfügung zu stellen, die sie benötigen, um rationale Entscheidungen treffen zu können. Stufflebeam entwickelt zu diesem Zweck ein umfangreiches Evaluationssystem und entsprechende Evaluationspläne, die viele vorher angesprochene Aspekte der Evaluation integrieren.

Alkins Beitrag bringt einen weiteren wichtigen Aspekt der Evaluation in die Diskussion, der im Zusammenhang mit der Arbeit des Bildungsrats und Wissenschaftsrats einer breiteren pädagogischen Öffentlichkeit bewußt geworden ist. Sein Aufsatz weist auf die Notwendigkeit hin, die ökonomischen Voraussetzungen von Bildungsreformen nicht nur in makroökonomischen Analysen, sondern auch bei der Entwicklung einzelner Innovationen in Form von mikroökonomischen Aufwands-Effektivitäts-Analysen (cost-effectiveness analysis) zu berücksichtigen. Unseres Wissens liegen dazu im deutschsprachigen Bereich bisher keine ähnlichen Ansätze vor. Wenn man auch einige Einwände gegen Einzelaspekte des

Modells vorbringen kann, muß man sich als Pädagoge durchaus mit dem Gedanken vertraut machen, daß bei der Begrenztheit der Ressourcen Bildungsreformen und schulische Innovationen *auch* eine ökonomische Dimension haben und die Öffentlichkeit auch in dieser Hinsicht einen Effektivitätsnachweis der Reformen verlangen kann.

Glass versucht, den Stand der Diskussion in bezug auf die konzeptuelle Entwicklung der Evaluation und ihrer Methoden zusammenzufassen und die ungeklärten Fragen aufzudecken. Dazu unterzieht er einige der wichtigen Evaluationsmodelle einer kritischen Analyse und entwickelt in Anlehnung an Scriven das Zielkomplex-Modell, in dessen Zentrum die Aufgabe der Bewertung von Innovationen liegt und in dem er ein Evaluationsmodell sieht, das einer weiteren Entwicklung wert ist.

Die Beiträge in diesem Abschnitt sind so ausgewählt, daß sie den gegenwärtigen Stand der Diskussion in der Evaluation wiedergeben. Dabei werden die zentralen Probleme der *Beschreibung*, *Bewertung* und *Entscheidungsvorbereitung* in den unterschiedlichen Phasen eines Bildungsprogramms bzw. Schulversuchs von verschiedenen Standpunkten aus diskutiert und erhellt.

Lee J. Cronbach

Evaluation zur Verbesserung von Curricula

Das weit verbreitete Interesse an der Verbesserung des Bildungswesens gab den Anstoß für einige wichtige Projekte, die die Verbesserung von Curricula, besonders von Curricula der Sekundarstufe, zum Ziel hatten. Auf Tagungen für Leiter von Projekten, die zur Verbesserung von Curricula führen sollten und die von der National Science Foundation finanziert wurden, standen häufig Probleme der Evaluation zur Diskussion [1]. Die Motive, sich mit der Evaluation zu befassen, reichen von reinem wissenschaftlichen Interesse am Unterrichtsgeschehen bis hin zu dem Anliegen, einem Geldgeber Sicherheit für die Richtigkeit seiner Investitionen zu geben. Curriculumentwickler sind sicherlich ernsthaft daran interessiert, die Spezialkenntnisse der Evaluationsexperten für ihre Arbeit zu benutzen. Ich möchte aber bezweifeln, ob sie eine genaue Vorstellung darüber haben, was Evaluation leisten kann oder leisten sollte. Andererseits komme ich immer mehr zu der Überzeugung, daß einige Verfahren und Denkgewohnheiten der Evaluatoren für die gegenwärtigen Curriculumuntersuchungen nur in geringem Maß anwendbar sind. Welche Theorien und welche Methoden der Evaluation sind für die Durchführung dieser Untersuchungen erforderlich, und inwieweit müssen wir uns von den herkömmlichen Lehrmeinungen und festgefahrenen Vorgehensweisen der traditionellen Testanwendung lösen?

Die Funktion der Evaluation in Entscheidungsprozessen

Um die Fülle der Aufgaben der Evaluationsforschung in den Griff zu bekommen, definieren wir »Evaluation« *als Sammlung von Informationen und ihre Verarbeitung mit dem Ziel, Entscheidungen über ein Curriculum zu fällen.* Das kann eine Materialsammlung für den Unterricht, die Unterrichtsaktivitäten einer einzelnen Schule oder auch die Lernerfahrungen eines einzelnen Schülers betreffen. Viele Arten von Entscheidun-

gen müssen getroffen werden; dazu sind viele verschiedene Informationen von Nutzen. Bereits hier zeigt es sich deutlich, daß Evaluation ein komplexer Vorgang ist und daß nicht eine bestimmte Vorgehensweise allen Situationen gerecht werden kann. Aber die Testkonstrukteure haben sich so sehr auf ein Verfahren – nämlich auf die Herstellung von Papier- und Bleistifttests zur Beurteilung einzelner Schüler – konzentriert, daß die Regeln, die mit diesem Verfahren verbunden sind, gleichsam als *die* Prinzipien der Evaluation angesehen werden. Tests, so wurde gesagt, sollten den Inhalt des Curriculum repräsentieren, und nur solche Evaluationsverfahren sollten benutzt werden, die einen gültigen Testwert erwarten lassen. Diese und ähnliche Prinzipien sind für eine Evaluation zur Curriculumverbesserung nicht ganz geeignet. Bevor wir dazu übergehen, diese Behauptung zu stützen, möchte ich zwischen den Zielen der Evaluation unterscheiden und sie zu der Geschichte der Test- und Curriculumentwicklung in Beziehung setzen.

Man kann drei Arten von Entscheidungen, für die Evaluation notwendig ist, unterscheiden:

1. Curriculumverbesserung
   Entscheidungen über die Angemessenheit von Unterrichtsmaterial und Unterrichtsmethoden und über notwendige Änderungen.
2. Entscheidungen über Individuen
   Erkennen der Bedürfnisse des Schülers, um seinen Unterricht entsprechend planen zu können; Beurteilung der Leistungen der Schüler, um eine Auswahl und Gruppierung vornehmen zu können; Vertrautwerden des Schülers mit seinen Leistungsfortschritten und -schwächen.
3. Administrative Regelungen
   Entscheidungen über die Qualität eines Schulsystems und über die Eignung einzelner Lehrer usw.

Die Verbesserung von Curricula wurde durch den dazu benötigten großen Zeitaufwand und die großen Entfernungen zwischen den Bezugsgruppen erschwert; denn zur Curriculumverbesserung gehört eine Änderung von häufig benutzten Unterrichtsmaterialien und Unterrichtsmethoden. Die Entwicklung einer Standardübung zur Behebung von Verständnisschwierigkeiten könnte als Curriculumverbesserung bezeichnet werden; bei der Entscheidung über die Teilnahme eines bestimmten Schülers an dieser Übung würde es sich jedoch um die Entscheidung über ein Individuum handeln. Eine administrative Regelung hat eine verhältnismäßig örtlich begrenzte Wirkung, während die Verbesserung eines Curriculum wahrscheinlich überall dort, wo es verwendet wird, Auswirkungen zeigt.

Für die Verbesserung von Curricula war die Einführung der systema-

tischen Evaluation von großer Bedeutung. Als Joseph Rice seinen aufsehenerregenden Rechtschreibtest in mehreren amerikanischen Schulen einsetzte und auf diese Weise den ersten Anstoß für die pädagogische Testbewegung gab, galt sein Interesse der Evaluation eines Curriculum. Rice wandte sich gegen den sich immer mehr ausbreitenden Drill in der Rechtschreibung, der in den Lehrplänen der Schulen im Vordergrund stand. Indem er seine Wertlosigkeit nachwies, rief er eine Revision der Curricula hervor. Als sich die Testbewegung entwickelte, übernahm sie jedoch eine andere Funktion.

Die stärkste Ausbreitung einer systematischen Leistungsmessung konnte in den zwanziger Jahren beobachtet werden. In dieser Zeit wurden die Inhalte der Curricula als weitgehend feststehend angesehen. Kritik wurde nicht geübt, von kleinen Veränderungen thematischer Schwerpunktbildung abgesehen. Auf Anordnung der Verwaltung wurden Standardtests, die sich auf die Curricula bezogen, ausgegeben, um die Effektivität des Lehrers oder des Schulsystems abzuschätzen. Da die administrative Testdurchführung unkritisch und unzulänglich gehandhabt wurde, verlor sie in den zwanziger und dreißiger Jahren an Bedeutung. Beamte der Schulverwaltung und der Schulaufsichtsbehörden griffen jedoch bei der Beurteilung der Qualität einer Schule wieder auf sie beschreibende Merkmale zurück. Anstatt unmittelbar Daten über pädagogische Auswirkungen zu sammeln, beurteilten sie die Schulen nach dem Budget, nach dem Lehrer-Schüler-Verhältnis, nach der Größe der Versuchsräume und nach den Qualifikationsnachweisen, die die Lehrer während ihrer Fortbildung erlangten. Das scheint sich nun zu ändern. An vielen Universitäten richten Schulverwaltungen Forschungszentren ein, um mehr über das Ergebnis ihrer Arbeit zu erfahren. Die Anwendung von Tests, die auf Qualitätskontrollen hinzielt, scheint sich auch an weniger guten Schulen durchzusetzen. Dies läßt sich sehr deutlich anhand des Erlasses der kalifornischen Legislative nachweisen, in dem Testdurchführung an allen Schulen Kaliforniens gefordert wird.

Etwa nach 1930 wurden Tests fast ausschließlich zur Beurteilung von Einzelpersonen eingesetzt: Um Schüler für einen Kurs mit höherem Niveau auszuwählen, um Noten in einer Klasse festzusetzen und um Leistungsstärken bzw. -schwächen des einzelnen festzustellen. Für alle diese Entscheidungen benötigte man genaue und gültige Vergleiche zwischen einem Individuum und anderen oder zwischen einem Individuum und einer Norm. Ein großer Teil der Testtheorie und Testtechnologie befaßte sich mit der Präzisierung der Messungen. Obwohl für die meisten Entscheidungen, die über Individuen getroffen werden, Genauigkeit sehr wesentlich ist, möchte ich doch Gründe dafür anführen, daß es für die

Curriculumevaluation nicht erforderlich ist, genaue Testwerte für Einzelpersonen zu erhalten.

Während die Testkonstrukteure mit ihren üblichen Verfahren zur Bestimmung genauer Testwerte zufrieden waren, waren sie es weit weniger mit den Verfahren, mit denen sie die Gültigkeit der Testwerte nachzuweisen versuchten. Noch vor 1935 wurde meist das Faktenwissen des Schülers und die Bewältigung grundlegender Fertigkeiten geprüft. Forschungsarbeiten und Veröffentlichungen von Tyler aus diesen Jahren weckten das Bewußtsein, daß höhere geistige Denkabläufe nicht durch einfache Wissenstests hervorgerufen und darum auch nicht festgestellt werden können und daß der Unterricht, der Faktenwissen fördert, nicht notwendigerweise auch andere wichtigere pädagogische Ergebnisse begünstigt, sondern daß er im Gegenteil mit ihnen in Konflikt geraten kann. Tyler, Lindquist und ihre Schüler konnten zeigen, daß man auch Tests entwickeln kann, um allgemeine pädagogische Auswirkungen zu messen, wie z. B. die Fähigkeit, eine wissenschaftliche Methode zu verstehen. Während sich ein Schüler für einen Wissenstest nur durch einen Lehrgang vorbereiten kann, der die getesteten Fakten vermittelt, können viele verschiedene Lehrgänge dieselben *allgemeinen* Fähigkeiten und dieselben Einstellungen fördern. Wenn man heute neue Curricula evaluieren will, ist es selbstverständlich wichtig, abzuschätzen, welchen allgemeinen Bildungsstand der Schüler erreicht hat, da die Curriculumentwickler behaupten, daß der allgemeine Bildungsstand wichtiger sei als die Bewältigung bestimmter Unterrichtseinheiten. Es sei daran erinnert, daß z. B. die Biological Sciences Curriculum Study drei Fassungen eines Curriculum mit fachspezifisch unterschiedlichem Inhalt als alternative Möglichkeiten anbietet, um am Ende die gleichen Ziele zu erreichen.

Obwohl einige etwa um 1930 entwickelte Meßverfahren dazu geeignet sind, allgemeine Auswirkungen der Schulbildung zu messen, fanden sie keine weite Verbreitung. Die vorherrschende Auffassung über die Funktion von Curricula, besonders unter den »Progressiven«, besteht in der Forderung, ein Programm zu entwickeln, das auf lokale Erfordernisse abgestimmt ist und die Fähigkeiten und Erfahrungen der Schüler, die an dem betreffenden Ort leben, besonders berücksichtigt. Das Vertrauen, das man um 1920 in ein »Standard«-Curriculum gesetzt hatte, wurde durch die Erkenntnis ersetzt, daß die beste Lernerfahrung das Ergebnis gemeinsamer Unterrichtsplanung von Lehrer und Schüler sei. Da jeder Lehrer bzw. jede Klasse verschiedene Inhalte und auch unterschiedliche Lernziele wählen konnte, ließ diese Auffassung wenig Raum für standardisierte Testverfahren.

Viele Evaluationsexperten sahen in der Entwicklung von Tests eine

Strategie für die Lehrerweiterbildung, so daß die Testentwicklung an sich höher bewertet wurde als der daraus resultierende Test selbst oder die entsprechenden Testergebnisse. Folgende Ausführungen von Bloom (1961) stehen stellvertretend für eine bestimmte Denkrichtung (vgl. auch Tyler 1951):

»Das Kriterium für die Bestimmung der Qualität einer Schule oder ihrer pädagogischen Funktionen sollte die Erreichung der Ziele sein, die sie sich selbst gesetzt hat... Unsere Erfahrungen geben zu der Vermutung Anlaß, daß die Wahrscheinlichkeit, etwas für die Realisierung der Ziele der Schule getan zu haben, gering ist, wenn die Schule ihre Ziele nicht in spezielle operationale Definitionen übersetzt hat. Diese Ziele bleiben sonst fromme Hoffnungen und unverbindliche Äußerungen... Die Teilnahme des Lehrerkollegiums an der Auswahl und an der Entwicklung der Evaluationsverfahren hat einmal zu verbesserten Verfahren und zum anderen zur Klärung der Unterrichtsziele beigetragen. Es gelang hiermit auch, die Ziele für die Lehrer greifbarer und sinnvoller erscheinen zu lassen... Nach der aktiven Teilnahme der Lehrer an der Definition der Ziele und der Auswahl oder Entwicklung der Evaluationsverfahren wandten sie sich wieder mit mehr Energie und großem Einfallsreichtum den alltäglichen Unterrichtsproblemen zu. ... Lehrer, die sich für eine Reihe pädagogischer Ziele, die sie gut verstehen, engagiert haben, versuchen zahlreiche Erfahrungen zu vermitteln, die so verschieden und komplex sind, wie sie die jeweilige Situation verlangt.«

So wird Evaluation zu einer jeweils an einen Schulbezirk gebundenen sinnvollen Aktivität der Lehrerbildung. Der daraus resultierende Gewinn besteht in dem Nachdenken darüber, welche Informationen überhaupt gesammelt werden sollen. Über die wirkliche Verwendung der Testergebnisse wird wenig gesagt; man hat den Eindruck, daß der Test selbst vergessen wird, sobald die Testentwicklung abgeschlossen ist. Sicher hat man ein geringes Interesse daran, die Tests so zu überarbeiten, daß sie auch in anderen Schulen benutzt werden können; denn in diesem Fall würde man den Lehrern die Möglichkeit nehmen, an der Ausarbeitung ihrer Ziele und Verfahren selbst mitzuarbeiten.

Bloom und Tyler fassen die Curriculumentwicklung und die Evaluation als integrierende Bestandteile eines dezentralisierten Unterrichts auf. Diese Funktion der Evaluation ist von derjenigen zur Verbesserung eines Curriculum zu unterscheiden. Die gegenwärtigen großen Curriculumprojekte gehen davon aus, daß die Curriculumentwicklung zentralisiert werden kann. Sie bereiten Materialien vor, die von Lehrern überall in gleicher Weise angewandt werden sollen. Man nimmt an, daß die Materialien, die von Fachleuten entworfen und nach Vorversuchen überarbeitet wurden, zu einem besseren Unterrichtsablauf beitragen können als die Materialien, die der Lehrer aufgrund der örtlichen Gegebenheiten entwerfen könnte. In diesem Zusammenhang scheint es völlig angebracht, wenn

man die meisten Tests von einem zentral arbeitenden Team entwickeln läßt. Die Testergebnisse müssen dem Team wieder zur Verfügung gestellt werden, damit es das Curriculum weiter verbessern kann.

Stellt man die Evaluation in den Dienst der Curriculumverbesserung, so ist es das Hauptanliegen, die Auswirkungen des Curriculum und die Veränderungen zu ermitteln, die es bei den Schülern bewirkt. Es geht hier aber nicht nur um die Frage, ob das Curriculum effektiv ist oder nicht. Die Ergebnisse des Unterrichts sind multidimensional determiniert; eine gute Untersuchung muß die Wirkungen eines Curriculum hinsichtlich dieser verschiedenen Dimensionen aufzeigen können. Es ist falsch, unterschiedliche Leistungen, die erst nach der Arbeit mit dem Curriculum geprüft werden, in einem einzigen Meßwert zusammenzufassen, da ein Versagen bei der Erreichung eines Lernziels z. B. durch den Erfolg bei der Erreichung eines anderen Lernziels verdeckt werden kann. Da ein Gesamttestwert Beurteilungen über die Bedeutung der verschiedenen Einzelergebnisse beinhaltet und gewöhnlich keine Aufschlüsse über die Beurteilungen der Einzelergebnisse gibt, kann für die Pädagogen, die verschiedene Werthierarchien haben, demnach nur ein Bericht von Nutzen sein, der die Ergebnisse getrennt voneinander auswertet.

Der größte Beitrag, den die Evaluation leisten kann, liegt darin, die Aspekte des Curriculum herauszuarbeiten, für die eine Neubearbeitung erforderlich ist. Die für die Curriculumentwicklung Verantwortlichen würden gerne die Effektivität ihres Curriculum beweisen. Der Gedanke an eine »unabhängige Testinstitution«, die das Ergebnis ihrer Arbeit beurteilt, ist für sie sehr reizvoll. Wenn man den Evaluator lediglich nach Beendigung der Curriculumentwicklung hinzuzieht, um ihn bestätigen zu lassen, was bereits getan wurde, würde das bedeuten, daß man von den Fähigkeiten eines Evaluators einen nur begrenzten Gebrauch macht und seine Rolle unterschätzt. Um aber die Verbesserung von Curricula zu erreichen, sollten die Ergebnisse während der Curriculumentwicklung zur Verfügung stehen und nicht erst dann, wenn der Curriculumentwickler nicht mehr daran interessiert ist, eine von ihm als beendet betrachtete Sammlung von Materialien und Techniken erneut zu diskutieren. Evaluation, die auf die Verbesserung von noch in der Entwicklung befindlichen Curricula zielt, trägt mehr zur Verbesserung des Unterrichts bei als die Evaluation, die nur dazu dient, Produkte zu bewerten, die bereits auf dem Markt sind.

Evaluation sollte soweit wie möglich dazu beitragen, das Verständnis für die Art der Wirkungen des Curriculum und für die Variablen, die seine Effektivität beeinflussen, zu erweitern. Es ist z. B. zu beachten, daß das Ergebnis programmierten Unterrichts von der Einstellung des Lehrers abhängig ist; das dürfte wichtiger sein als die Feststellung, daß dieser Un-

terricht im Durchschnitt etwas bessere oder schlechtere Ergebnisse erzielt als der konventionelle Unterricht.

Hoffentlich sieht man die Aufgabe von Evaluationsuntersuchungen nicht nur darin, über das eine oder andere Curriculum einen Bericht abzugeben, sondern dazu beizutragen, Erziehungs- und Lernprozesse besser zu verstehen. Solche Einsichten tragen schließlich außer zur Entwicklung des Curriculum, dessen Lehrerfolge mit Hilfe von Tests nachgeprüft werden, auch zum Verständnis der allgemeinen Probleme der Curriculumentwicklung bei. In einigen neuen Curricula liegen Ergebnisse vor, die vermuten lassen, daß die Fähigkeiten der Schüler mit der Leistung am Ende eines Curriculum in geringerem Maße korrelieren als mit der Leistung in früheren Einheiten des Curriculum (vgl. Ferris 1962). Dieser Befund ist nicht gut abgesichert. Wenn er sich jedoch als richtig herausstellen sollte, dann käme ihm große Bedeutung zu. Auch wenn dies nur für die neuen Curricula zutreffend ist, hat das bereits Konsequenzen; wenn derselbe Effekt bei den herkömmlichen Curricula auftritt, so hat das einen anderen Stellenwert. In beiden Fällen ist das jedoch für Lehrer, Schulpsychologen und Erziehungswissenschaftler ein Grund zum Nachdenken. Evaluationsuntersuchungen sollten dazu beitragen, Erkenntnisse über die Merkmale von Fähigkeiten zu ermitteln, die zur Erreichung pädagogischer Ziele notwendig sind. Zwanzig Jahre nach der Eight-Year-Study der Progressive Education Association sind ihre Testverfahren noch immer von Bedeutung; aber wir wissen sehr wenig darüber, was diese Testverfahren eigentlich messen. Man denke z. B. an die »Anwendung wissenschaftlicher Prinzipien in den Naturwissenschaften«. Kann man hier in irgendeiner Hinsicht von einer einheitlichen Fähigkeit sprechen? Oder ist es dem guten Schüler nur gelungen, allmählich einige Prinzipien zu beherrschen? Ist die Fähigkeit, die in einem solchen Test geprüft wird, von größerem Voraussagewert für zukünftige Leistungen als Faktenwissen? Man sollte solchen Fragen große Bedeutung beimessen, obwohl sie für die Curriculumentwickler nur von begrenztem Interesse sind.

Das Ziel, Curricula miteinander zu vergleichen, sollte nicht die Pläne für die Evaluation bestimmen. Entscheidungsträger müssen zwischen mehreren Curricula wählen; dabei bleibt es nicht aus, daß alle Evaluationsberichte z. T. vergleichend interpretiert werden. Aber als Experiment geplante Untersuchungen, in denen man ein Curriculum mit einem anderen vergleicht, sind selten aussagekräftig genug, um den finanziellen Aufwand zu rechtfertigen. Die Unterschiede zwischen Durchschnittstestwerten, die das Ergebnis verschiedener Curricula darstellen, sind in der Regel gering im Vergleich zu den großen Unterschieden zwischen und in den Klassen, die mit demselben Curriculum unterrichtet worden sind. Bestenfalls kann

ein solcher Versuch zwei bereits bestehende Curricula miteinander vergleichen. Wenn man sich sehr bemüht, das schlechtere Curriculum zu optimieren, führt dies wahrscheinlich zur Umkehrung des Urteils über den Versuch.

Man gefährdet die Interpretation eines Versuchs, wenn man die Klassen nicht parallelisiert, die zu vergleichende Curricula benutzen. Leider sind solche Fehler fast unvermeidbar. Wenn man ein Medikament testet, ist man sich darüber klar, daß gültige Ergebnisse nur mit Hilfe eines Doppelblindversuchs gewonnen werden können. Im Doppelblindversuch bekommt die Hälfte der Probanden anstelle des Medikaments ein unwirksames Placebo. Placebo und Medikament sehen genauso aus, so daß weder Arzt noch Patient wissen, wer von den Patienten das Medikament bekommt. Ohne eine solche Kontrolle sind die Ergebnisse wertlos, selbst wenn der Zustand des Patienten anhand völlig objektiver Anzeichen überprüft wurde. In einem pädagogischen Versuch ist es schwer, die Schüler über ihre Rolle als Versuchsgruppe im unklaren zu lassen. Die Fehlerquellen, die durch die Person des Lehrers bedingt sind, können kaum so gut kontrolliert werden, wie die des Arztes im Doppelblindversuch. Infolgedessen kann man nicht mit Sicherheit sagen, ob ein beobachteter Gewinn der pädagogischen Innovation an sich zuzuschreiben ist oder dem größeren Engagement von Lehrern und Schülern bei einem Versuch mit einer neuen Methode. Man hat behauptet, daß alle Curricula, die besten nicht ausgenommen, viel von ihrer Anziehungskraft verlieren, sobald sie aufgrund ihres Erfolgs die Rolle des herkömmlichen Unterrichts übernehmen (vgl. Modell 1963).

Da die Ergebnisse von Gruppenvergleichen sehr fragwürdig sind, sollte man meiner Meinung nach eine gute Untersuchung in erster Linie so planen, daß sie es erlaubt, die Leistungen einer genau umschriebenen Gruppe am Ende eines Curriculum im Hinblick auf die wesentlichen Ziele und Nebenwirkungen zu bestimmen. Unser Problem ist mit dem eines Ingenieurs, der ein neues Auto überprüft, vergleichbar. Er kann sich die Aufgabe stellen, die Leistungsfähigkeit und Zuverlässigkeit des Autos genau zu bestimmen. Es würde an dem Problem vorbeiführen, wenn er sich die Frage stellen würde: Ist dieses Auto besser oder schlechter als die konkurrierende Automarke? In einem Versuch jedoch, in dem sich die verglichenen Curricula in zahlreicher Hinsicht unterscheiden, kann man keine neuen Erkenntnisse aufgrund des höheren Punktwertes des neuen Curriculum erwarten. Man kann nicht sagen, welche der Variablen für diesen Punktgewinn verantwortlich ist. Stärker analytische Versuche sind viel nützlicher als Feldversuche, die sehr unterschiedliche Curricula verschiedenen Gruppen zuteilen. Klein angelegte, gut kontrollierte Untersuchungen können zum

Vergleich alternativer Fassungen des gleichen Curriculum erfolgreich eingesetzt werden; in einer solchen Untersuchung sind die Unterschiede zwischen den Varianten des Curriculum gering und gut genug definiert, so daß die Ergebnisse zur Klärung des Problems beitragen.

Für die drei Ziele, Curricula zu verbessern, Entscheidungen über Einzelpersonen zu fällen und administrative Regelungen zu treffen, werden Meßverfahren von verschiedener Art benötigt. Wenn ein Test dazu benutzt werden soll, über den einzelnen Lehrer ein administratives Urteil zu fällen, dann ist eine gründliche und unparteiische Untersuchung zu fordern; die dafür benötigten Testverfahren sind extrem zeitraubend, wenn sie nicht nur ein Einzelergebnis erbringen sollen. Bei der Beurteilung eines Curriculum jedoch kann man zu zufriedenstellenden Interpretationen kommen, wenn die gesammelten Ergebnisse auf einer Stichprobe beruhen; in diesem Fall ist der Anspruch, die Leistungen jeder Klasse sorgfältig gemessen zu haben, nicht angebracht. Ähnliches gilt auch für die Testanwendung, wenn es um Entscheidungen über Einzelpersonen geht. Individualtests müssen außerordentlich gerecht sein und umfassend genug, wenn man für jedes Individuum einen verläßlichen Punktwert gewinnen will. Wenn aber die Leistung das Geschick des Individuums nicht beeinflußt, können wir es darum bitten, Aufgaben auszuführen, für die es durch das Curriculum nicht ausdrücklich vorbereitet wurde; wir können ferner Verfahren einsetzen, die, wenn man für jedes Individuum einen zuverlässigen Testwert erhalten will, bei sorgfältiger Anwendung sehr kostspielig wären.

Methoden der Evaluation

Spektrum der Methoden

Evaluation ist zu oft mit der Durchführung etwa einstündiger formaler Tests am Ende eines Curriculum gleichgesetzt worden. Es gibt aber noch viele andere Methoden zur Überprüfung von Schülerleistungen; doch auch die Schülerleistungen sind nicht die einzige Basis zur Bewertung eines Curriculum.

Es erscheint auch sinnvoll, Wissenschaftler zu befragen, ob ein Curriculum dem neuesten Stand des Wissens entspricht. Dies ist ein geeignetes und notwendiges Verfahren. Man kann ferner die pädagogische Konzeption des neuen Curriculum mit Hilfe von Meinungsumfragen evaluieren; doch kann dieses Vorgehen recht zufällige Ergebnisse erbringen. Wenn die Meinungen auf einigen Vorurteilen über eine Lehrmethode beruhen, so werden die Urteile widersprüchlich ausfallen und sehr wahrscheinlich zu

Fehlinterpretationen verführen. Es gibt keine pädagogischen Theorien, die so abgesichert sind, daß sie – ohne Vorversuche – Voraussagen über pädagogische Wirkungen zulassen.

Man kann von der Notwendigkeit einer pragmatischen Untersuchung des Curriculum überzeugt sein und dennoch Umfrageergebnisse als zusätzlich unterstützende Faktoren hinzuziehen. In den Versuchsstadien der Curriculumentwicklung verläßt man sich sehr auf die Berichte der Lehrer, die über die Schülerleistungen abgegeben werden: »Hier hatten sie Schwierigkeiten.« »Dies fanden sie langweilig.« »Hier wäre nur die Hälfte der vorgesehenen Übungen notwendig«, usw. Dabei handelt es sich um Verhaltensbeobachtung, die, auch wenn sie unsystematisch erfolgt, sehr wertvoll ist. Für einen Übergang zur systematischen Beobachtung spricht, daß sie gerechter, besser nachprüfbar und manchmal auch gründlicher ist. Wenn es um die Beurteilung der Qualität von Curriculuminhalten geht, vertraue ich z. B. den Fachkenntnissen des Historikers oder Mathematikers. Hingegen stimme ich nicht mit der Ansicht überein, daß Geschichts- oder Mathematiklehrer, die ein Curriculum ausprobieren, seine Effektivität am besten beurteilen können. Wissenschaftler haben sich zu oft über ihre Fähigkeit als Lehrer getäuscht, vor allem da sie das Nachplappern von Wörtern als Beweis von Verständnis gewertet haben, als daß man ihrem ungeschulten Urteilsvermögen vertrauen könnte. Systematische Beobachtung ist finanziell aufwendig; außerdem bringt sie eine zeitliche Verzögerung zwischen dem Unterrichtsgeschehen und der Rückmeldung der Ergebnisse mit sich. Daher wird die systematische Beobachtung für den Curriculumentwickler niemals die einzige Informationsquelle sein. Nachdem man sich mit den offensichtlich schwerwiegenden Unzulänglichkeiten eines Curriculum in früheren Entwürfen bereits auseinandergesetzt hatte, wird die systematische Datensammlung in den Zwischenstadien der Curriculumentwicklung von Nutzen sein.

Zu den Verfahren der Evaluation zählen Prozeßuntersuchungen (process studies), Leistungsuntersuchungen, Einstellungsuntersuchungen und Längsschnittuntersuchungen (follow-up studies). Eine Prozeßuntersuchung befaßt sich mit dem Unterrichtsgeschehen, Leistungs- und Einstellungsmessungen befassen sich mit beobachteten Veränderungen der Schüler, und Längsschnittuntersuchungen verfolgen den späteren Berufserfolg der Schüler, die mit einem bestimmten Curriculum gearbeitet haben.

Längsschnittuntersuchungen können die bleibenden pädagogischen Nach- oder Auswirkungen des Curriculum noch am ehesten erfassen. Der Abschluß einer solchen Untersuchung ist jedoch zeitlich so weit vom Unterricht entfernt, daß die Untersuchung für die Verbesserung des Curriculum oder für die Erklärung seiner Auswirkungen nur von geringem

Wert ist. In einer Hinsicht unterscheiden sich Längsschnittuntersuchungen deutlich von den anderen Arten der Evaluation. Wie bereits erwähnt, sollte sich Evaluation in erster Linie mit den Auswirkungen des untersuchten Curriculum befassen, weniger mit dem Vergleich von Curricula. D. h., ich würde besonders die Diskrepanz zwischen den Ergebnissen und den Zielvorstellungen, die Unterschiede in der Effektivität verschiedener Teile des Curriculum und die Unterschiede zwischen den einzelnen Testaufgaben herausarbeiten; hier sind Ansatzpunkte für die Verbesserung von Curricula zu finden. Aber diese Gesichtspunkte können nicht auf eine Längsschnittuntersuchung übertragen werden, die die Auswirkungen des Curriculum insgesamt bewertet und die nur von geringer Aussagekraft ist, wenn man nicht die Ergebnisse auf einer einheitlichen Basis vergleichen kann. Angenommen, 65 Prozent der Schüler lassen sich nach erfolgreichem Abschluß eines Curriculum in naturwissenschaftlichen und technischen Fächern einer Hochschule immatrikulieren, dann kann man nicht beurteilen, ob dies ein hoher oder niedriger Prozentsatz ist, es sei denn, man vergleicht den Prozentsatz dieser Schüler mit dem prozentualen Anteil derjenigen, die nicht nur in diesem Curriculum unterrichtet worden sind. In einer Längsschnittuntersuchung muß man Daten einer Kontrollgruppe erhalten, die wenigstens in groben Umrissen mit der Versuchsgruppe in bezug auf eindeutige demographische Variablen parallelisiert wurde.

Obwohl die Parallelisierung solcher Gruppen schwierig ist und die Daten einer Längsschnittuntersuchung nicht viel darüber aussagen, wie ein Curriculum verbessert werden kann, sollten solche Untersuchungen dennoch durchgeführt werden. Denn die vielen großen Stichproben der neuen Curricula eignen sich gut dazu, wichtige Fragen weiterzuverfolgen. Eine bekannte Form der Längsschnittuntersuchung besteht darin, den Erfolg des Studenten in einem Curriculum der Hochschule, das auf ein Curriculum der Sekundarschule aufbaut, zu ermitteln. Man kann die Noten des Schülers untersuchen oder ihn fragen, für welche Themen des Hochschulcurriculum er sich schlecht vorbereitet glaubte. Hoffentlich werden einige der neuen naturwissenschaftlichen und mathematischen Curricula unter Mädchen größeres Interesse als bisher hervorrufen; ob diese Hoffnung berechtigt ist, kann man nachprüfen, indem man untersucht, welche Haupt- und Nebenfächer die ehemaligen Schülerinnen im College gewählt haben. Ebenso verdient die Berufswahl Beachtung. Einige Befürworter der neuen Curricula würden es begrüßen, wenn mehr Begabte sich statt für technologische Disziplinen für die Grundwissenschaften entscheiden würden. Andere wiederum halten dies für möglicherweise verhängnisvoll; aber keiner würde Daten über eine solche Veränderung für unwichtig halten.

Für die Curriculumentwickler sind unter den Ergebnissen des Curricu-

lum Einstellungsänderungen von besonderer Bedeutung. Einstellungen sind Meinungen oder Überzeugungen und nicht nur Ausdruck von Zustimmung oder Ablehnung. Die Einstellung eines Menschen gegenüber den Naturwissenschaften enthält Vorstellungen über Sachverhalte, in denen ein Wissenschaftler eine Autorität sein kann; sie wird aber auch durch die Erforschung des Mondes, durch Untersuchungen über Affenmütter und die Ausbeutung von Naturschätzen geprägt. Ebenso wichtig ist die Frage nach der Übereinstimmung zwischen dem Selbstkonzept und dem Umweltverständnis, etwa: Welche Möglichkeiten kann die Wissenschaft mir bieten? Würde ich einen Wissenschaftler heiraten wollen? Jede Lernaktivität trägt zu Einstellungen bei, die weit über das Fachliche hinausreichen, so wie die Einstellung des Schülers über sein eigenes Können und seine Lernbereitschaft hinausreicht.

Einstellungen können auf sehr verschiedene Weise gemessen werden; die Fächer- und Berufswahl, die durch Längsschnittuntersuchungen aufgedeckt wird, kommt z. B. dafür in Betracht. Aber gewöhnlich wird die Messung in Form von direkter oder indirekter Befragung durchgeführt. Interviews, Fragebogen und ähnliche Verfahren sind durchaus wertvoll, solange man ihnen nicht blind vertraut. Sicherlich sollten wir auch alle *unerwünschten* Meinungsäußerungen, die von einem großen Teil der Absolventen eines Curriculum zum Ausdruck gebracht werden, ernst nehmen (z. B. die Meinung, ein Wissenschaftler könne mit besonderer Autorität über politische und ethische Fragen sprechen, oder die Ansicht, die Mathematik habe bereits die Grenzen ihrer Möglichkeiten erreicht).

Einstellungsfragebogen sind heftig kritisiert worden, weil sie leicht zu Verfälschungen führen, vor allem wenn ein Schüler durch weniger Offenheit zu einem besseren Testergebnis zu kommen hofft. Die Antworten sind wahrscheinlich eher zuverlässig, wenn die Fragen in einem Zusammenhang gestellt werden, der sich sehr von den Inhalten des Versuchscurriculum unterscheidet. So kann z. B. ein allgemeiner Fragebogen, der im Zusammenhang mit dem obligatorischen Englischunterricht ausgegeben wird, auch Fragen über die Neigung für verschiedene Fächer und Tätigkeiten enthalten; dieselben Fragen würden weniger zuverlässige Ergebnisse über die Einstellung gegenüber Mathematik ergeben, wenn sie von einem Mathematiklehrer verteilt worden wären. Obwohl die Schüler entgegen ihren wahren Anschauungen eher »günstige« Antworten geben, ist diese Verzerrung jedoch in einem Jahr nicht größer als im anderen und bei den Schülern nicht größer, die im Unterschied zu anderen an einem Versuchscurriculum teilgenommen haben. Im Gruppendurchschnitt gleichen sich viele Verfälschungen wieder aus. Die Fragebogen, die für das Testen einzelner Personen eine nicht hinreichende Gültigkeit besitzen, können je-

doch zur Evaluation von Curricula benutzt werden. Denn der Schüler wird hier nicht motiviert sein, Ergebnisse zu verfälschen, und der Evaluator wendet sie nur zum Vergleich von Mittelwerten und nicht zum Vergleich von Individuen an.

Um Leistungen messen zu können, benötigt man ebenfalls verschiedene Verfahren. Standardisierte Tests sind nützlich. Aber für die Curriculumevaluation erscheint es sinnvoll, verschiedenen Schülern *unterschiedliche* Fragen vorzulegen. Wenn man jedem Schüler in einer Grundgesamtheit von 500 Schülern den gleichen Test mit 50 Fragen gibt, so wird dieser Test für den Curriculumentwickler weniger informativ sein, als wenn man jedem Schüler 50 Fragen aus einer Sammlung von etwa 700 Testaufgaben zuteilt. Letzteres Verfahren bestimmt den durchschnittlichen Erfolg von etwa 75 repräsentativ ausgewählten Schülern in bezug auf jede dieser 700 Testaufgaben, das zuerst genannte Verfahren jedoch nur für 50 Testaufgaben (vgl. Lord 1962). Aufsatztests und offene Fragen, die für viele Formen der Evaluation im allgemeinen zu teuer sind, können zur Beurteilung bestimmter Fähigkeiten mit Gewinn eingesetzt werden. Man kann auch darüber hinaus Individuen oder Gruppen unter kontrollierten Bedingungen dabei beobachten, wie sie ein Forschungsproblem angehen und wie sie sich mit anderen umfassenden Problemen auseinandersetzen. Da man nur eine repräsentative Stichprobe von Schülern testen muß, stellt die Kostenfrage nicht ein so großes Problem dar wie bei der gewohnten Art der Testdurchführung. Weitere Gesichtspunkte zur Anwendung von Leistungstests sollen später noch berücksichtigt werden.

Der besondere Wert von Prozeßuntersuchungen (process measures), die das Unterrichtsgeschehen untersuchen, liegt darin, aufzudecken, wie ein Curriculum verbessert werden kann. Bei der Entwicklung von programmiertem Unterrichtsmaterial werden z. B. Aufzeichnungen gesammelt, aus denen zu ersehen ist, wie viele Schüler die einzelnen Testaufgaben jeweils nicht lösen konnten. Jede Häufung von Fehlern erfordert eine bessere Erklärung oder einen stärker gestuften Aufbau eines schwierigen Unterrichtsinhaltes. Kurz nach der Darbietung eines Lehrfilms kann man die Schüler z. B. um die Beschreibung eines Photos aus dem Film bitten. Mißverständliche Darstellungen und Inhalte, die unklar geblieben sind, können durch solche Methoden herausgefunden werden. Entsprechend können Interviews aufdecken, welchen Gewinn die Schüler vom Unterricht im Labor oder von einer Diskussion haben. Eine Prozeßuntersuchung kann sich auch auf das Unterrichtsverhalten des Lehrers richten. Für die Curricula, die eine Wahl der Themen zulassen, lohnt es sich, herauszufinden, welche Themen gewählt wurden und wieviel Zeit für jedes Thema zur Verfügung stand. Eine Aufzeichnung des Unterrichtsgeschehens, die eher ein Schüler als ein Leh-

rer erstellen sollte, kann zeigen, welche der für einen Fortbildungskursus empfohlenen Unterrichtstechniken wirklich verwendet wurden und welche Verfahren des neuen Curriculum nur in der Phantasie des Curriculumentwicklers existieren.

Leistungsmessung

Wie bereits ausgeführt, halte ich die Ergebnisse einzelner Testaufgaben für wichtiger als Gesamttestwerte. Aufgrund des Gesamttestwertes kann ein Curriculum positiv oder negativ bewertet werden; aber der Gesamttestwert sagt sehr wenig darüber aus, wie das Curriculum weiter verbessert werden kann. Ferris wies bereits 1962 darauf hin, daß solche Testwerte sehr leicht fehl- oder überinterpretiert werden. Die Frage, wie ein Curriculum zu verbessern ist, ist mit Hilfe des Testwertes einer einzelnen Testaufgabe oder einer Problemlösungsaufgabe, die mehrere Antworten hintereinander erfordert, eher als mit Hilfe eines Gesamttestwertes zu beantworten. Wenn wir die Testwerte der einzelnen Testaufgaben als aussagekräftig ansehen, darf man Evaluation nicht länger als punktuelles Ereignis am Ende eines Schuljahrs betrachten. Leistungen können zu jeder Zeit unter Berücksichtigung der Testaufgaben gemessen werden, die den engsten Bezug zu den letzten Unterrichtseinheiten haben. Dagegen hat es sich als sinnvoll erwiesen, Testaufgaben, die allgemeine Fähigkeiten erfassen, wiederholt während der Arbeit mit dem Curriculum einzusetzen (vielleicht bei verschiedenen Zufallsstichproben von Schülern), um zu ermitteln, wann und aufgrund welcher Erfahrungen sich diese Fähigkeiten verändern.

In der Curriculumevaluation braucht man sich nicht zu sehr darum zu bemühen, die Meßverfahren dem Curriculum anzupassen. Wie überraschend das auch immer ist und wie sehr das auch im Gegensatz zu den Prinzipien der Evaluation für andere Zwecke steht, so gilt das dennoch, wenn wir wissen wollen, welche Veränderungen ein Curriculum bei einem Schüler verursacht. Eine optimale Evaluation würde alle Arten der Leistungen miteinbeziehen, die für ein bestimmtes Problem relevant sind, und nicht nur die ausgewählten Ergebnisse, auf die das Curriculum sich konzentriert. Wenn man jedoch nur wissen will, wie gut ein Curriculum *seine* Ziele erreicht, dann muß der Test das Curriculum inhaltlich repräsentieren; wenn man aber wissen will, welchen Wert das Curriculum für die Gesellschaft hat, muß man alle Auswirkungen messen, für die es sich einzusetzen lohnt. In einem der neuen Mathematikcurricula könnte etwa numerische Trigonometrie oder elektronische Datenverarbeitung als Inhalt abgelehnt werden. Dennoch kann man zu Recht danach fragen, wie gut

die Absolventen des Curriculum diese Operationen durchführen können. Selbst wenn die Curriculumentwickler behaupten würden, daß elektronische Datenverarbeitung kein angemessenes Ziel des Sekundarschulunterrichts ist, werden einige Pädagogen diese Ansicht nicht teilen. Wenn man aber nachweisen kann, daß Schüler, die man im Rahmen des neuen Curriculum in diesen Fähigkeiten nicht ausdrücklich unterrichtet hatte, dennoch bei der elektronischen Datenverarbeitung einiges leisten, wird man auch die Kritiker zufriedenstellen können. Wenn jedoch keine Leistung erbracht wird, ist das der Nachweis, daß etwas versäumt worden ist. Ähnliches gilt für alternative Curricula der Biologen, die den Schwerpunkt auf Mikro-Biologie bzw. auf Ökologie legen. Auch hier ist die Frage berechtigt, wie gut die Absolventen des einen Curriculum die im anderen Curriculum behandelten Probleme verstehen. Eine optimale Evaluation, z. B. in Mathematik, wird Nachweise für alle Fähigkeiten sammeln, die in einem Mathematikcurriculum sinnvoll angestrebt werden können, das entsprechende gilt für andere Fachbereiche.

Ferris behauptet, daß der Anderson Chemistry Test (ACS), so gut er auch konstruiert sein mag, für die Evaluation des neuen Chemical Bond Approach Project (CBA) und der neuen Chemical Education Material Study (CHEM) ungeeignet ist, weil er ihre Lernziele nicht prüft.

Man kann mit dieser Behauptung übereinstimmen, ohne die Verwendung des ACS-Tests im Zusammenhang mit diesen Curricula für unangemessen zu halten. Dieser Test darf jedoch nicht *allein* zur Evaluation verwendet werden. Er kann wertvolle Aufschlüsse darüber geben, wieviel Allgemeinwissen das neue Curriculum vermittelt. Die Curriculumentwickler haben bewußt auf einige der konventionellen Leistungsanforderungen verzichtet. Sie haben bei fachkundiger Interpretation von diesen Testergebnissen nichts zu befürchten, besonders wenn die Ergebnisse für jede Testaufgabe einzeln untersucht werden.

Die Forderung, daß Tests sich auf die Ziele eines Curriculum beziehen sollen, spiegelt die Tatsache wieder, daß herkömmliche Prüfungen bestimmen, was gelehrt wird. Wenn die Fragen im voraus bekannt sind, konzentrieren sich die Schüler mehr auf das Lernen ihrer Antworten als auf das Lernen anderer Teile des Curriculum. Das muß jedoch kein Nachteil sein. Wenn es darauf ankommt, bestimmte Inhalte zu bewältigen, von denen man weiß, daß sie getestet werden, bewirkt das eine hohe Anstrengungsbereitschaft. Andererseits besteht ein erheblicher Unterschied zwischen dem Lernen von Antworten auf eine Reihe von Fragen und dem Verständnis der Inhalte, auf die sich die Fragen beziehen. Vielleicht besteht deshalb in der Verwendung »sicherer« Tests ein Vorteil für die Curriculumevaluation. Sicherheit kann nur dadurch erreicht werden, daß man je-

des Jahr neue Tests entwickelt und auch keine Vor- und Nachvergleiche mit denselben Testaufgaben durchführt. Die Verwendung unterschiedlicher Testaufgaben bei verschiedenen Schülern und die Tatsache, daß weniger Anreiz zum Auswendiglernen der Testaufgaben besteht, wenn Schüler und Lehrer nicht beurteilt werden, würde die »Sicherheit« zu einem weniger wichtigen Problem werden lassen.

Die Unterscheidung zwischen Wissenstests und Tests für komplexere Denkprozesse, wie sie z. B. in der *Taxonomy of Educational Objectives* getroffen wurde, ist für die Planung von Tests wertvoll, obwohl die Klassifikation von Testaufgaben »zur Erfassung von Wissen«, »Anwendung« (application), »Problemlösungsverhalten« usw. schwierig und oft unmöglich ist. Ob eine gegebene Antwort Auswendiggelerntes oder eine vernünftige Denkleistung widerspiegelt, hängt davon ab, wie der Schüler unterrichtet wurde, und nicht allein von der gestellten Testaufgabe. Man kann z. B. eine biologische Umwelt beschreiben und nach Voraussagen über die Wirkung eines bestimmten Eingriffs fragen. Schüler, die sich niemals mit ökologischen Sachverhalten befaßt haben, würden entweder aufgrund ihrer allgemeinen Fähigkeit, über komplexe Vorgänge nachdenken zu können, erfolgreich sein, oder sie versagen; Schüler, die in ökologischer Biologie unterrichtet worden sind, würden mit größerer Wahrscheinlichkeit Erfolg haben, da sie in ihrem Denken bestimmte Prinzipien der Ökologie verwenden können. Schüler, die in einer solchen Umwelt gelebt oder darüber gelesen haben, müßten aufgrund ihrer Erinnerung erfolgreich antworten. Deshalb sollte man nur selten testen, ob ein Schüler bestimmte Inhalte kennt oder nicht kennt. Es kommt vielmehr auf das Ausmaß des Wissens und seine Anwendbarkeit an. Zwei Personen können mit denselben Tatsachen oder Prinzipien vertraut sein, aber dennoch wird einer sie besser verstehen und besser in der Lage sein, mit widersprüchlichen Daten, irrelevanten Aspekten eines Problems und offensichtlichen Ausnahmen von der Regel umzugehen. Um kognitive Fähigkeiten zu messen, muß man die Tiefe, die Kohärenz und die Anwendbarkeit des Wissens messen.

Testaufgaben sind zu oft curriculumspezifisch und so formuliert, daß man sie nur dann beantworten kann, wenn man durch den Unterricht darauf vorbereitet wurde, die gestellten Fragen zu verstehen. Solche Fragen können im allgemeinen daran erkannt werden, daß sie in einer Fachsprache formuliert sind. Manchmal sind einzelne Elemente dieser Fachsprache allgemein bekannt, und wir können annehmen, daß alle getesteten Schüler mit ihnen vertraut sind. Ein Biologietest aber, in dem ein Stoffwechselvorgang mit Hilfe einer Formel bezeichnet wird, stellt für die Schüler eine Schwierigkeit dar, die zwar die wissenschaftliche Frage über den Stoffwechselhaushalt durchdenken können, aber die Formel nicht kennen. Ein

trigonometrisches Problem, das die Benutzung einer trigonometrischen Tabelle erfordert, ist allein dann angebracht, wenn man die Vertrautheit mit den Bezeichnungen der Funktionen testen will. Dieselbe Frage in numerischer Trigonometrie kann auch in einer Form gestellt werden, die für den Durchschnittsschüler beim *Eintritt* in die Sekundarstufe klar und verständlich ist; wenn nötig, können den Schülern die Tabellen der Funktionen zusammen mit einer verständlichen Erklärung gegeben werden. In dieser Form ist die Fragestellung curriculumunabhängig. Man kann zu Recht fragen, ob die Absolventen eines Versuchscurriculum auch Probleme lösen können, mit denen sie vorher nicht konfrontiert wurden, während es jedoch sinnlos ist, danach zu fragen, ob sie Fragen beantworten können, deren Sprache für sie unverständlich ist. Ohne Zweifel ist die Kenntnis einer bestimmten Terminologie ein wichtiges Unterrichtsziel; aber für die Curriculumevaluation sollte das Testen der Terminologie nach Möglichkeit von dem Testen anderer Formen des Verstehens getrennt werden. Um das Verständnis von Prozessen und Relationen einzuschätzen, ist eine Frage dann gut, wenn sie für einen Schüler verständlich ist, der nicht an dem Curriculum teilgenommen hat. Das bedeutet nicht, daß er die Antwort oder das zur Beantwortung der Frage angebrachte Vorgehen kennen muß, aber er sollte wenigstens verstehen, was die Frage beinhaltet. Solche curriculumunabhängigen Fragen können wie standardisierte Verfahren zur Untersuchung jedes Curriculum benutzt werden.

Schüler, die sich nicht mit einem Thema befaßt haben, werden es in der Regel schwerer haben als solche, die sich damit auseinandergesetzt haben. Die Absolventen meines hypothetischen Mathematikcurriculum werden mehr Zeit zur Lösung trigonometrischer Aufgaben benötigen als Schüler, die Trigonometrie gelernt haben. Aber Schnelligkeit und Qualität der Lösung dürfen nicht miteinander verwechselt werden; im kognitiven Bereich ist die Qualität der Leistung stets von größerer Bedeutung. Wenn das Curriculum dem Schüler ermöglicht, sich mit einem Inhalt, mit dem er sich nicht beschäftigt hat, richtig, wenn auch nur langsam auseinanderzusetzen, dann kann man von ihm erwarten, daß er später nach wiederholter Konfrontation mühelos mit dem Inhalt umgehen kann.

Das wichtigste Ziel vieler neuer Curricula scheint in der Förderung der Fähigkeit zu liegen, neue Aufgaben innerhalb desselben Fachbereichs besser zu bewältigen. Ein Biologiecurriculum kann nicht alle wichtigen biologischen Inhalte behandeln; es kann jedoch durchaus darauf abzielen, den Schüler in die Lage zu versetzen, Beschreibungen ihm unbekannter Organismen und eine neue Theorie und deren Hintergründe zu verstehen und einen Versuch zur Überprüfung neuer Hypothesen zu planen. Dies ist ein Beispiel für den Transfer des Gelernten. Man hat bislang kaum erkannt,

daß es zwei Arten des Transfer gibt. Sie befinden sich auf einem Kontinuum, dessen einer Pol durch einen unmittelbar wirksamen und dessen anderer durch einen langfristig wirksamen Transfereffekt gekennzeichnet ist. Den unmittelbar wirksamen Transfereffekt kann man als anwendbaren Transfer (applicational transfer) bezeichnen, den langfristig wirksamen Transfereffekt als Zuwachs an Fähigkeit (vgl. Ferguson 1954).

In fast der gesamten pädagogischen Transfer-Forschung hat man die unmittelbar sich zeigende Leistung an einer teilweise neuen Aufgabe getestet. Wir lehren die Schüler, Gleichungen mit der Unbekannten x zu lösen und fordern im Test Lösungen von Gleichungen mit a oder z. Wir lehren die Prinzipien des ökologischen Gleichgewichts am Beispiel der Wälder und fragen in einem Transfertest nach der Wirkung der Umweltverschmutzung auf die Population eines Sees. Wir beschreiben einen nicht im Test dargestellten Versuch und fordern die Schüler auf, mögliche Interpretationen und benötigte Kontrollen zu erörtern. Alle diese Tests können kurzfristig gehandhabt werden, aber die wichtigere Art des Transfer ist die steigende Lernfähigkeit auf einem bestimmten Gebiet. Wahrscheinlich besteht ein bedeutsamer Unterschied zwischen der Fähigkeit, Folgerungen aus einem sorgfältig beendeten Versuch zu ziehen, und der Fähigkeit, Erkenntnis aus ungeordneten und sich widersprechenden Beobachtungen zu gewinnen, die im Laufe kontinuierlicher Versuchsarbeit an einem Problem auftauchen. Der Schüler, der mit einem guten Biologie-Curriculum unterrichtet wird, kann bestimmte Arten von Theorien und Daten besser verstehen, so daß er bei der Beschäftigung mit Ethnologie im folgenden Jahr einen größeren Gewinn hat; dieser Gewinn kann nicht gemessen werden, indem man das Verständnis des Schülers anhand kurzer Abschnitte aus der Ethnologie prüft. Selten hat man die Fähigkeit bewertet, eine Problemsituation oder einen komplexen Wissensbereich über einen Zeitraum von Tagen oder Monaten zu bearbeiten. Trotz der praktischen Schwierigkeiten, die dem Versuch entgegenstehen, die Wirkungen eines Curriculum auf das spätere Lernen einer Person zu messen, ist das »Lernen zu lernen« so wichtig, daß ernsthafte Anstrengungen unternommen werden sollten, um solche Wirkungen aufzudecken und ihre Entwicklung zu fördern.

Die Methode des programmierten Unterrichts kann dazu dienen, die Lernfähigkeit eines Schülers abzuschätzen. Man kann z. B. die Schnelligkeit messen, mit der ein Schüler eine in sich selbständige programmierte Einheit über das physikalische Problem der Hitze oder über ein anderes Thema bewältigt, mit dem er sich nicht beschäftigt hat. Ist das Programm in sich abgeschlossen, dann kann es jeder Schüler bewältigen; der Schüler mit dem größeren naturwissenschaftlichen Verständnis wird voraussichtlich jedoch weniger Fehler machen und schnellere Fortschritte erzielen. Das Pro-

gramm sollte in mehreren logisch vollständigen Fassungen hergestellt werden, wobei diese von einer Fassung mit sehr kleinen Schritten bis hin zu einer mit sehr wenigen internen Wiederholungen (internal redundancy) reichen sollten; dem liegt die Hypothese zugrunde, daß der bessere Schüler das weniger redundante Programm bewältigen kann und vielleicht auch mehr von der größeren Eleganz des Programms angesprochen wird.

Zusammenfassung

Alte Denkgewohnheiten und schon lange etablierte Methoden eignen sich nicht für die Evaluation, die zur Curriculumverbesserung erforderlich ist. In der Vergangenheit zielte pädagogisches Testen vorwiegend auf die Gewinnung gerechter und genauer Testwerte, um Einzelpersonen miteinander zu vergleichen. In pädagogischen Experimenten befaßte man sich vorwiegend mit dem Vergleich der Testmittelwerte konkurrierender Curricula. Aber Curriculumevaluation erfordert die Beschreibung der Ergebnisse. Diese Beschreibung sollte auf einer möglichst breiten Skala erfolgen, selbst unter Aufgabe vordergründiger Objektivität und Genauigkeit.

Curriculumevaluation sollte die von einem Curriculum bewirkten Veränderungen feststellen und die Aspekte des Curriculum identifizieren, die einer Verbesserung bedürfen. Die beobachteten Ergebnisse sollten allgemeine Ergebnisse berücksichtigen, die weit über die Inhalte des Curriculum selbst hinausreichen: Einstellungen, Berufswahl, allgemeine Verständnisfähigkeit und die Fähigkeit, weiter zu lernen. Die Analyse der Schülerleistung bei einzelnen Testaufgaben oder bestimmten Problemarten liefert mehr Informationen als die Analyse von Gesamttestwerten. Es empfiehlt sich nicht, allen Schülern denselben Test zu geben; statt dessen sollten aus einer Sammlung von möglichst vielen Testaufgaben Gruppen verschiedener Testaufgaben zusammengestellt werden, die jeweils verschiedenen kleineren Schülerstichproben gegeben werden sollten. Aufwendige Methoden wie Interviews und Aufsatztests können bei Schülerstichproben erfolgreich eingesetzt werden, während dagegen das Testen der Grundgesamtheit nicht in Frage kommt. Richtige Fragestellungen zu pädagogischen Ergebnissen können zur Verbesserung pädagogischer Effektivität viel beitragen. Selbst wenn die richtigen Daten gesammelt werden, wird die Funktion der Evaluation nur sehr begrenzt sein, wenn sie sich lediglich auf die positive bzw. negative Bewertung der Curricula beschränkt. Evaluation ist ein grundlegender Bestandteil der Curriculumentwicklung. Ihre Aufgabe besteht darin, Daten zu sammeln, die der Curriculumentwickler zur besseren Erfüllung seiner Aufgabe verwenden kann und die ein besseres Verständnis der pädagogischen Prozesse ermöglichen.

Michael Scriven

Die Methodologie der Evaluation

Einführung

Die bisherigen Konzeptionen der Evaluation sind in Theorie und Praxis noch unzureichend. Mit diesem Beitrag soll versucht werden, einige Unzulänglichkeiten aufzudecken und zu verringern. Geistiger Fortschritt ist nur möglich, weil junge Wissenschaftler auf den Arbeiten von Koryphäen aufbauen können. Dazu gehört jedoch auch, daß beide Seiten die Leistungen der anderen Seite respektieren. Verpflichtet bin ich Lee Cronbachs Aufsatz von 1963, weiterführenden Gesprächen mit den Mitarbeitern des Center for Instructional Research and Curriculum Evaluation (CIRCE) an der Universität von Illinois, Urbana, und anregendem Schriftwechsel mit mehreren Kollegen, insbesondere James Shaver und Ray Barglow.

Überblick

Der Schwerpunkt dieses Beitrags liegt auf der Curriculumevaluation; fast alle Überlegungen können jedoch ohne weiteres auf andere Arten der Evaluation übertragen werden. Die Überschriften der einzelnen Abschnitte sind aus sich selbst heraus verständlich und erscheinen in folgender Reihenfolge:

1. Überblick
2. Ziele und Rollen der Evaluation; die formative und summative Rolle der Evaluation
3. Professionelle und Amateur-Evaluation
4. Evaluationsuntersuchungen und Prozeßuntersuchungen
5. Evaluation und Überprüfung der Zielerreichung
6. Intrinsische Evaluation und Ergebnisevaluation
7. Praktische Vorschläge für eine Mischform der Evaluation (Hybrid Evaluation)
8. Das Für und Wider einer reinen Ergebnisevaluation
9. Vergleichende und nicht-vergleichende Evaluation
10. Praktische Verfahrensweisen für die Evaluation mit Kontrollgruppen.

Ziele der Evaluation und Rollen der Evaluation; die formative und die summative Rolle der Evaluation

Die Funktion der Evaluation kann von zwei Seiten her begriffen werden. Auf der methodischen Ebene können wir von den *Zielen* der Evaluation sprechen; darüber hinaus lassen sich in einem bestimmten soziologischen oder pädagogischen Kontext verschiedene *Rollen* der Evaluation unterscheiden.

In bezug auf die Ziele der Evaluation kann man davon ausgehen, daß Evaluation bestimmte *Fragen* über bestimmte *Einheiten* zu beantworten versucht. Die Einheiten sind die verschiedenen pädagogischen Instrumente (Prozesse, Personal, Verfahrensweisen, Programme usw.) Zu den Fragen gehören solche über die Form: *Wie gut* funktioniert dieses Instrument (in bezug auf diese oder jene Kriterien)? Funktioniert es *besser* als ein anderes Instrument? *Was* leistet dieses Instrument, d. h. welche Variablen der uns interessierenden Gruppe werden von seiner Anwendung signifikant beeinflußt? *Rechtfertigt* der Gebrauch dieses Instruments seine Kosten? Evaluation an sich ist ein methodisches Vorgehen, das im Grunde genommen gleich ist, unabhängig davon, ob man Kaffeemaschinen, Lehrmaschinen, Pläne für ein Haus oder ein Curriculum zu evaluieren versucht. Es besteht einfach im Sammeln und Kombinieren von Verhaltensdaten mit einem gewichteten Satz von Skalen, mit denen entweder vergleichende oder numerische Beurteilungen erlangt werden sollen, und in der Rechtfertigung (a) der Datensammlungsinstrumente, (b) der Gewichtungen, (c) der Kriterienauswahl.

In einem bestimmten pädagogischen Kontext kann die *Rolle* der Evaluation jedoch sehr unterschiedlich sein. Evaluation kann Bestandteil der Lehrerbildung, der Curriculumentwicklung, eines Feldversuchs zur Verbesserung einer Lerntheorie oder einer Untersuchung von Unterrichtsmaterial vor einer Kaufentscheidung sein. Evaluation kann ferner zu einer Datensammlung führen, mit deren Hilfe eine Steuererhöhung oder ein Forschungsvorhaben unterstützt wird; sie kann aber auch z. B. in einem Trainingsprogramm, in einem Gefängnis oder in einer Schulklasse als Mittel zu einer positiven oder negativen Rückmeldung (feedback) dienen. Häufig hat man versäumt, diese wichtige Unterscheidung zwischen Rollen und Zielen der Evaluation zu treffen. Das hat u. a. zu einem starken Substanzverlust in der Evaluation geführt, so daß sie nicht länger zur Beantwortung von Fragen über den Wert von Bildungsprogrammen beiträgt, obwohl gerade darin ihre Aufgabe liegen sollte. Man darf Evaluation nur ablehnen, wenn man begründen kann, daß man sich nicht um eine Antwort auf Fragen nach dem Wert pädagogischer Instrumente bemühen sollte. Dazu wür-

de jedoch auch der kaum zu erbringende Nachweis gehören, daß diese Fragen bei überhaupt *keinen* Aktivitäten gestellt werden dürfen. Aus der Tatsache, daß der Evaluation manchmal eine unangemessene Rolle zugewiesen wird, darf man nicht schließen, daß man Fragen nach dem Ziel der Evaluation niemals beantworten sollte. Die besonders bei Lehrern oder Schülern häufig anzutreffende Furcht vor Evaluation ist eine oft zu Unrecht verallgemeinerte Reaktion auf die berechtigte Ablehnung einer Situation, in der Evaluation eine Rolle zugewiesen wurde, die nicht mehr in ihren Geltungsbereich fällt und der sie auch nicht gerecht werden kann.

Zu Recht verweist man häufig auf die Rolle der Evaluation im Prozeß der Curriculum*entwicklung*, die natürlich die Evaluation des *Endergebnisses* dieses Prozesses nicht ersetzen kann. In der Regel kann und muß Evaluation verschiedenen Rollen gerecht werden. Aus der Behandlung der Evaluation in einer Reihe neuerer Veröffentlichungen und Forschungsprojekte läßt sich jedoch erkennen, daß man die Anforderungen an Evaluation bereits zu erfüllen meint, wenn man an *irgendeiner* Stelle im Laufe des Projekts evaluiert. Evaluation kann jedoch im Rahmen eines pädagogischen Projekts nicht nur verschiedene Rollen übernehmen; sie kann auch innerhalb jeder Rolle mehrere spezifische Ziele haben. So kann Evaluation bei der Verbesserung des Curriculum eine Rolle spielen. Um ihr gerecht zu werden, kann man z. B. folgende Fragen stellen: Gelingt es dem Curriculum wirklich, den Unterschied zwischen Vorurteil und politischer Einstellung zu vermitteln? Benötigt es dafür zuviel Unterrichtszeit? Diese Form der Evaluation kann man als *formative* Evaluation bezeichnen. Ihre Ergebnisse bleiben innerhalb der die Curriculumentwicklung tragenden Institution und dienen zur Verbesserung des Programms.

Der Evaluation kommt eine andere Rolle zu, wenn sie den Beamten der Schulverwaltung bei der Entscheidung helfen soll, ob das fertiggestellte Curriculum den anderen Alternativen so weit überlegen ist, daß sich die Ausgaben für seine Einführung in das Schulsystem rechtfertigen lassen. Man kann diese Form der Evaluation als *summative* Evaluation bezeichnen. Ihre Ergebnisse werden Interessenten außerhalb der die Curriculum tragenden Institution zur Verfügung gestellt und dienen zum besseren *Verständnis* und zur besseren *Verwendung* des Programms.

Die häufige Verwechslung und fehlende Unterscheidung zwischen Rollen und Zielen der Evaluation liegt zum Teil in dem wohlgemeinten Versuch begründet, die Beunruhigung der Lehrer durch Evaluation zu verringern. Indem man jedoch die konstruktive Funktion der Evaluation so stark betont, übergeht man stillschweigend, daß zu den Zielen der Evaluation auch die Beurteilung von Leistung, Bedeutung, Wert usw. gehört, die in einer anderen Rolle der Evaluation zu Entscheidungen über die

Förderung von Personen und Curricula beitragen kann. Man sollte die Furcht vor Evaluation nicht dadurch zu verringern suchen, daß man diese wichtige Funktion der Evaluation nicht genügend berücksichtigt und die Darstellung der Evaluationsergebnisse verfälscht. Die nachteiligen Folgen für das Erziehungswesen sind zu groß. Die Wirtschaft kann es sich auch nicht erlauben, Fabriken zu unterhalten oder leitende Angestellte zu beschäftigen, die ganz offensichtlich keine gute Arbeit leisten. Ebenso sollte die Gesellschaft – solange gute Leistungen erbracht werden können – keine schlechten Bücher und Kurse verwenden und unzulängliche Lehrer und Schulaufsichtsbeamte beschäftigen. Um der Furcht vor Evaluation entsprechend zu begegnen, muß man für die Personen, deren Position oder Ansehen in Gefahr ist, Aufgaben schaffen, für die sie besser geeignet sind. Wenn man die Schülerleistung nicht beurteilt, führt das zu unzulänglichen Leistungen in der Schulklasse. Scheut man sich, die Lehrerleistung zu evaluieren, bewirkt das inkompetenten Unterricht. Daher dürfen wir, wenn wir der gesellschaftlichen Verantwortung für das Erziehungswesen gerecht werden wollen, uns gegenüber dem einzelnen nicht immer so zaghaft verhalten. Vielleicht trifft es zu, daß »der größte Beitrag, den Evaluation leisten kann, darin liegt, die Aspekte des Curriculum herauszuarbeiten, für die eine Neubearbeitung erforderlich ist« (Cronbach 1963, 41, *46*). Jedoch gibt es ohne Zweifel auch im Rahmen der meisten Curriculumprojekte und Innovationen gleich wichtige andere Funktionen der Evaluation. Es gibt z. B. viele Situationen, in denen man durch die abschließende Evaluation eines Projekts oder einer Person seiner Verantwortung gegenüber der Person, dem Projekt oder dem Steuerzahler nachkommt. Wenn man mit Hilfe einer umfassenden Ergebnisevaluation deutlich machen kann, daß ein teures Schulbuch nicht besser als ein anderes Schulbuch ist, mit dem es verglichen wurde, oder aber daß das teure Schulbuch bei weitem besser als alle anderen ist, so sollte man doch nicht wie Cronbach diesen Beitrag der Evaluation für unbedeutend erklären. Das heißt: Wenn man z. B. zeigen kann, daß ein bestimmtes Verfahren, Mathematik zu unterrichten, in keiner von Mathematikern für wichtig gehaltenen Dimension signifikant bessere Schülerleistungen erzielt, dann kann dieses Ergebnis Zeit und Geld sparen helfen und somit einen Beitrag zur Entwicklung des Bildungswesens leisten. Das gleiche gilt natürlich auch, wenn mit Hilfe des Verfahrens signifikant bessere Leistungen erzielt werden können. Demnach müssen zunächst einige erhebliche Einschränkungen erfolgen, bevor man Cronbach zustimmen und der formativen Evaluation größere Bedeutung zuschreiben kann als der summativen: »Evaluation, die auf die Verbesserung von noch in der Entwicklung befindlichen Curricula zielt, trägt mehr zur Verbesserung des Unterrichts bei als die Evaluation, die nur dazu dient, Produkte

zu bewerten, die bereits auf dem Markt sind« (Cronbach 1963, 41 46). Das beste Gegenbeispiel bilden die erfolgreichen, jedoch rassistischen Grundschulbücher der späten fünfziger Jahre. Es bedurfte einer summativen Evaluation mit negativem Ergebnis, um sie aus den Schulen zu verdrängen und durch andere bessere Bücher zu ersetzen. Doch zum Glück braucht man sich nicht für eine der beiden Rollen der Evaluation zu entscheiden. Bei pädagogischen, besonders jedoch bei curricularen Projekten muß man versuchen, beide Rollen der Evaluation zu erfüllen.

Wahrscheinlich hat jeder Curriculumentwickler seine Aufgabe übernommen, weil er das gegenwärtige Curriculum aufgrund vorläufiger summativer Evaluation für unzureichend hält. Während er das neue Curriculummaterial entwickelt, evaluiert er es laufend, indem er es besser als das bereits vorhandene Material zu machen versucht. Wenn er sich auch nur ein wenig der Begrenztheit seines Urteils über die eigene Arbeit bewußt ist, wird er das Curriculum noch während seiner Entwicklung in der Schule testen. Dadurch erhält der Curriculumentwickler eine Rückmeldung, auf deren Basis er es revidieren kann. Dieses Vorgehen ist eine formative Evaluation; wenn diese Felduntersuchung gut ausgeführt wird, kann sie sogar zu einer summativen Evaluation der *frühen Formen* des neuen Curriculum werden. Im allgemeinen arbeitet der Curriculumentwickler mit Lehrern oder anderen Kollegen zusammen, die das Material fortwährend kommentieren und beurteilen. Auch das ist eine Form der Evaluation, mit deren Hilfe das Curriculum verbessert werden kann.

Wenn formative Evaluation sinnvoll durchgeführt werden soll, dann sollte möglichst ein *professioneller Evaluator* in dem Curriculumprojekt mitarbeiten. Im allgemeinen werden dabei die Vorteile überwiegen; dennoch deuten einige Erfahrungen in der Praxis darauf hin, daß die Mitarbeit eines professionellen Evaluators auch Nachteile haben kann. Diese Frage berührt natürlich nicht die Frage nach der summativen Evaluation und der Rolle des professionellen Evaluators in ihr. Beide Fragen sollen in einem Teil des nächsten Abschnitts weiter erörtert werden.

Professionelle und Amateur-Evaluation

Der Evaluator ist zwar in seinem Gebiet ein Fachmann, selten aber in dem für das Curriculum inhaltlich relevanten Bereich. Im Unterschied zu den Curriculumentwicklern steht er dem Projekt im allgemeinen distanzierter gegenüber. Diese unterschiedliche Einstellung führt nicht selten zu Spannungen und Zwistigkeiten, die jedem Projektleiter nur zu vertraut sind.

Die unzulängliche Kommunikation zwischen Evaluatoren, Lehrern und

Curriculumentwicklern hat leider zu zwei extremen Reaktionen geführt. Es entstand einmal eine starre Anti-Evaluations-Haltung; sie ist oft nur eine Rationalisierung der Furcht, die durch die Gegenwart eines externen Beurteilers ausgelöst wird, der sich mit den Zielen des Projekts nicht identifiziert hat und der ihnen auch nicht verpflichtet ist. Das andere ebenso unerfreuliche Extrem besteht in dem rigiden Evaluator, der nur Operationalisierungen gelten läßt und der deshalb oft dem Sinn nach sagen dürfte: »Wenn Sie mir nicht in operationalisierter Form sagen, welche Variablen Sie beeinflussen wollen, kann ich keinen entsprechenden Test entwikkeln, und solange die Variablen nicht getestet worden sind, dürfen Sie nicht annehmen, daß Sie die Variablen erfolgreich beeinflußt haben.«

Zur Präzisierung dieser beiden Positionen wollen wir den Unterschied zwischen einem großen Curriculumprojekt der Gegenwart und einem in den späten dreißiger Jahren von zwei oder drei Lehrern gemeinsam verfaßten Algebra-Text deutlich herausarbeiten.

Erstens werden die heutigen Projekte häufig mit umfangreichen öffentlichen Geldern finanziert. Um diese Ausgaben zu rechtfertigen, muß ein objektiver Nachweis über den Wert des Produkts erbracht werden. Sodann ist für die *weitere* Finanzierung der Arbeit in diesem Bereich oder anderer Projekte derselben Curriculumentwickler ein objektiver Leistungsnachweis erforderlich. Da die Finanzen nicht ausreichen, alle Antragsteller zu unterstützen, muß man den Wert der Projekte aufgrund eines Vergleichs beurteilen. Objektive Grundlagen dafür sind natürlich den persönlichen Meinungsäußerungen von Kollegen überlegen. Schließlich werden durch die hohen Kosten für die *Einführung* solcher Curriculummaterialien in ein Schulsystem weitere Steuergelder verbraucht; diese Ausgaben sollten nur dann erfolgen, wenn sie sich auf Grund ausreichender Daten rechtfertigen lassen. Daher müssen Projektleiter, finanzierende Institutionen und Schulen auf die Durchführung einer summativen Evaluation drängen. Da formative Evaluation zu jedem rationalen Versuch gehört, gute Ergebnisse in der summativen Evaluation zu erzielen, muß sie auch von Anfang an erfolgen; nach unserem Verständnis ist sie sogar bis zu einem gewissen Grad durch den Prozeß der Curriculumentwicklung selbst bedingt. Davon unabhängig ist die Frage, ob und wie man professionelle Evaluatoren zum Projekt hinzuziehen soll. Die Beantwortung der Frage hängt davon ab, inwieweit formative Evaluation zur Verbesserung des Curriculum beitragen, bzw. seine Entwicklung behindern kann; und es gibt durchaus eine Reihe von Situationen, in denen sie den Prozeß der Curriculumentwicklung eher negativ beeinflußt.

Professionelle Evaluatoren können z. B. so kritisch sein, daß sie die Arbeit einer produktiven Gruppe ernsthaft gefährden. Auch wenn sie der

Gruppe, im ganzen gesehen, in der Regel behilflich sind, stellen sie z. B. oft so hohe Anforderungen an die operationale Formulierung der Ziele, daß zuviel Zeit für eine im Grunde sekundäre Tätigkeit verwendet wird. Daher muß ein Kompromiß geschlossen werden. Der Evaluator muß einen Teil *seiner* Aufgabe darin sehen, einen Satz von testbaren Kriterien für das Curriculum zu entwickeln. Dabei kann ihm die Tatsache helfen, daß sich das Projektteam für bestimmte Ziele ausdrücklich entschieden und andere verworfen hat. Für die Formulierung der Kriterien wird ihm außerdem die Kritik des Teams nützlich sein. Die Kommunikation zwischen Evaluator und Curriculumentwickler muß jedoch in beiden Richtungen erfolgen. Den Curriculumentwicklern unterlaufen häufig Fehler; sie sind oft voreingenommen oder zu sehr von ihrem Projekt begeistert. Evaluatoren wiederum haben, solange sie noch nicht mit den Themen und Zielen der Curriculumentwickler vertraut sind, nur begrenzte Wirkungsmöglichkeiten. Wenn sie sich aber mit diesen Zielen und dem Gesamtprojekt identifizieren, verlieren sie leicht die für eine objektive Evaluation wichtige Unabhängigkeit. Daher sollte man die formativen Evaluatoren nach Möglichkeit auch deutlich von den summativen Evaluatoren unterscheiden, denen sie natürlich bei der Entwicklung eines summativen Evaluationsplans behilflich sein können. Wenn man zwischen formativen und summativen Evaluatoren unterscheidet, kann man die Vorteile objektiver professioneller Evaluation wahrnehmen, ohne eine Störung der Kooperation im Team zu riskieren.

Bei der Evaluation im Bildungswesen und der Verwendung eines Evaluators im Prozeß der Curriculumentwicklung ergeben sich noch viele weitere Probleme. Auf mehrere hat J. Myron Atkin (1963) hingewiesen. Einige von ihnen sollen später in diesem Beitrag behandelt werden; auf zwei Probleme sei jedoch bereits hier aufmerksam gemacht. Eines besteht darin, daß das Testen bestimmter differenzierter Begriffe dadurch eine negative Auswirkung haben kann, daß es dem Schüler die Rolle eines Begriffs zu früh bewußt werden läßt und dadurch die natürliche Entwicklung des Begriffsverständnisses verhindert. Das zweite Problem besteht darin, daß manchmal bei einigen Begriffen die Verständnisfähigkeit eines Kindes im Verlauf der Arbeit mit einem Curriculum oder während eines bestimmten Schuljahrs nur wenig zunimmt. Dennoch kann dieser geringe Zuwachs für die langfristige Entwicklung der Verständnisfähigkeit von großer Bedeutung sein. Der Verständniszuwachs würde sich jedoch in Tests nicht niederschlagen und könnte sogar durch die Verwendung von Tests beeinträchtigt werden; dennoch muß er potentiell im Curriculum enthalten sein, um das gewünschte Endprodukt hervorzubringen. In einem solchen Fall wäre eine Evaluation jedoch unergiebig und vielleicht sogar hinderlich.

Wenn ein Curriculumteam seine Arbeit mit den Lehrern des gegenwärtigen Curriculum diskutiert, bringt das trotz möglicher Vorteile auch Nachteile mit sich. Das gilt auch für eine frühzeitige Hinzuziehung eines Evaluators. Ein einfallsreicher Projektleiter kann eine solche Situation mit verschiedenen Möglichkeiten zum Besten des Projekts nutzen. Er kann z. B. dem Evaluator lediglich die Curriculummaterialien geben, ohne daß dieser die Curriculumentwickler selber kennenlernt. Die Kommentare des Evaluators werden dann dem Projektleiter zugeleitet, der vielleicht zunächst nur die grundsätzliche und ernsthafte Kritik an das Team weitergibt und die anderen kritischen Anmerkungen bis zu dem Zeitpunkt zurückhält, an dem eine umfassende Revision erfolgen soll. Das sind jedoch Überlegungen für die Praxis. Es bleiben zwei grundsätzliche Einwände, die kurz erwähnt werden sollen und von denen der erste sich unmittelbar auf Atkins Befürchtungen bezieht.

Jeder, der ein neues Curriculum in der Schule getestet hat, weiß, daß dieses Curriculum sehr unterschiedliche Wirkungen auf die Schüler haben kann, die sich häufig aus ihren vorherigen Leistungen nicht voraussagen lassen. So findet z. B. ein Kind, das sich bereits für die Beobachtung von Vögeln interessiert, einen entsprechenden Zugang zur Biologie vielleicht attraktiver als einen anderen. Bei einigen Kindern hängt ihr Interesse davon ab, wie weit das Unterrichtsmaterial für die ihnen bereits vertrauten Probleme relevant ist; für andere hingegen sind z. B. die Eigenschaften der Möbiusschen Fläche [1] sofort faszinierend. Im allgemeinen kann die Unterrichtsorganisation die Motivation der Schüler in unterschiedlicher Weise beeinflussen. Der nicht-direktive Erziehungsstil, der wegen seines vermuteten Zusammenhangs mit der induktiven Unterrichtsmethode häufig bevorzugt wird, ist für diejenigen Kinder nicht geeignet, die stärker durch eine aggressive, rivalitätsgeladene, kritische Interaktion zur Aktivität herausgefordert werden. Trotz dieser Unterschiede greift man noch immer auf Tests für die ganze Klasse als undifferenzierte Evaluationsinstrumente zurück. Doch selbst wenn man die Testergebnisse in bezug auf individuellen Leistungszuwachs aufschlüsselt, hat man die Möglichkeiten des Materials noch nicht voll ausgenutzt. Sie würden sich erst dann zeigen, wenn man das richtige Material *und* die richtige Unterrichtstechnik für jedes Kind mit seinen entsprechenden Einstellungen, Interessen und Fähigkeiten auswählt. Wenn jemand der Evaluation skeptisch gegenübersteht, wird er vielleicht vorschlagen, man solle seine Hoffnung auf den kreativen und wissenschaftlich kompetenten Curriculumentwickler setzen und die Felduntersuchungen nur als Nachweis dafür ansehen, daß man Schüler unter geeigneten Bedingungen für das Curriculummaterial interessieren und mit ihm unterrichten könne. Das bedeutet, unser Kriterium sollte der deutliche

Leistungszuwachs bei *einigen* oder *mehreren* Schülern, nicht aber der Leistungszuwachs der ganzen Klasse sein. Dem muß der Evaluator mit dem Einwand begegnen, man dürfe nicht übersehen, daß z. B. ein unzulängliches Verständnis vieler Schüler und eine deutliche relative Verschlechterung der Leistungen mehrerer Schüler den Leistungszuwachs bei einigen Schülern aufhebt. Auch dürfe man nicht ausschließen, daß das pädagogische Geschick oder das Engagement des Lehrers und nicht die Materialien für den Erfolg bei der Felduntersuchung verantwortlich sind. Die Curriculummaterialien müssen daher auch anderen Lehrern gegeben werden, damit sie feststellen können, ob sie ihrer Meinung nach brauchbar sind. Um diese Fragen beantworten zu können, benötigt man professionelle Evaluation.

Aus dieser Kritik läßt sich jedoch eine wichtige Anregung ableiten. Auf jeden Fall muß man nämlich die Ansichten und Urteile der Fachwissenschaftler über die Qualität von Curriculuminhalten gewissenhaft berücksichtigen. Manchmal wird man zwar kaum Informationen für ihre Weiterentwicklung erhalten können; in einigen Fällen werden sie jedoch für bestimmte Entscheidungen durchaus genügen. Auf jeden Fall sollten diese Urteile sorgfältig bedacht und ausdrücklich in der Evaluation berücksichtigt werden; denn eine *fehlende* Unterstützung durch das Urteil von Fachwissenschaftlern ist oft bereits ein ausreichender Grund für eine vollständige Ablehnung des Curriculummaterials.

Schließlich trifft man in vielen Diskussionen auf die Ansicht, daß Evaluation Werturteile erfordert und daß diese Werturteile im Grunde genommen subjektiv und nicht wissenschaftlich sind. Dies ist genau so abwegig wie die Ansicht, daß Aussagen einer Person über sich selbst im Grunde genommen subjektiv sind und damit nicht rational vorgebracht werden können. Einige Werturteile sind im wesentlichen Äußerungen von wichtigen persönlichen Präferenzen (Geschmacksfragen) und als solche faktische Aussagen, die durch die üblichen Verfahren der psychologischen Forschung nachgewiesen werden können. Der Nachweis solcher Urteile gibt keine Auskunft darüber, ob es für jemanden richtig oder falsch ist, solche Wertvorstellungen zu haben. Man erfährt lediglich, ob jemand diese Wertvorstellungen hat oder nicht. Eine andere Form eines Werturteils ist die Einschätzung (assessment) der Leistung oder relativen Leistung einer curricularen Einheit in einem klar definierten Kontext, die schließlich zu dem Urteil führt, die Leistung dieser Einheit sei in bezug auf deutlich identifizierbare und deutlich gewichtete Kriteriumsvariablen so gut wie oder gar besser als die einer anderen curricularen Einheit. Bei Werturteilen dieser Art läßt sich allerdings nicht nur feststellen, ob die Personen, die sie abgeben, zu ihnen stehen oder nicht, sondern man kann auch feststellen, ob es richtig oder falsch ist, diese Werturteile zu haben. Sie sind lediglich komplexe

Gesamturteile aufgrund der Einstufung und Gewichtung verschiedener Leistungen. In diesem Sinne können wir also genau feststellen, daß die Palek Quartz zur Zeit die beste Armbanduhr ist oder daß ein bestimmtes Wörterbuch für Benutzer mit umfangreichen wissenschaftlichen Interessen am besten geeignet ist. Schließlich gibt es noch Werturteile, bei denen die Kriterien selbst umstritten sind; diese Werturteile sind, philosophisch gesehen, die wichtigsten; ihre Strittigkeit verweist darauf, daß wichtige Probleme meist nur schwer lösbar sind. Ein Beispiel für ein solches Urteil ist die Behauptung, daß die wichtigste Rolle der Evaluation im Prozeß der Curriculumentwicklung liegt, daß der Intelligenztest ein überholtes Untersuchungsinstrument ist oder daß die Kopenhagener Interpretation der Quantenphysik allen bekannten Alternativen überlegen ist.

In allen diesen Fällen streitet man sich darüber, was als gut gelten soll, und argumentiert dabei kaum mit den »wirklichen Fakten« der Situation. Dennoch darf man solche Urteile nicht außer acht lassen. Vielmehr sollte man auf jeden Fall die Gründe untersuchen, die für solche Urteile angeführt werden, und dann erst entscheiden, ob und wie diese Fragen rational diskutiert werden können. Nach einer häufig vertretenen Auffassung müssen wir im Umgang mit Menschen, also etwa bei der Erziehung, *ethische* Werturteile fällen, die im Grunde genommen subjektiv sind. Aber erstens sind Werturteile über Menschen keineswegs notwendigerweise ethisch, weil sie sich auch auf ihre Gesundheit, ihre Intelligenz oder ihre Leistungen beziehen können. Zweitens, selbst wenn sie ethisch sind, sind wir alle wohl einem ethischen Prinzip, der Rechtsgleichheit für alle Menschen, verpflichtet. Auf dieser Voraussetzung und einem entsprechenden Bezugsrahmen beruht der größte Teil der öffentlichen Auseinandersetzungen über ethische Fragen. Wenn man nicht dieses Axiom in Frage stellen und rationale Argumente für eine Alternative beibringen will, sind selbst ethische Werturteile rationaler Diskussion zugänglich. Was auch immer das Ergebnis einer solchen Diskussion ist, die Tatsache, daß Evaluation manchmal eine ethische Evaluation ist und daß ethische Evaluation zum Teil kontrovers ist, läßt bei weitem nicht den Schluß zu, daß Curriculumevaluation nicht ein wichtiger Bereich der angewandten Wissenschaft ist, zu der man sonst auch die Ingenieurswissenschaften und die Medizin nicht zählen dürfte.

Evaluationsuntersuchungen und Prozeßuntersuchungen

Wenn man den Begriff Evaluation zu erklären versucht, sollte man sich vor Simplifizierung in acht nehmen. Obwohl Evaluation im allgemeinen auf

Urteile über Leistung und Wert zielt, ist eine analytische Beschreibung und Interpretation des Prozesses notwendig, in dem jemand eine *Situation* oder die *Auswirkung* bestimmter Materialien evaluiert. In diesem Sinne kann man einige Formen von Prozeßforschung als Evaluation begreifen. Prozeßforschung überschneidet sich jedoch mit Evaluation nur teilweise und sollte nicht unter Evaluation subsumiert werden. Nach Cronbach lassen sich drei Arten der Prozeßforschung unterscheiden.

(1) Die erste Art der Prozeßforschung besteht in der deskriptiven Untersuchung der wirklichen Unterrichtsgeschehnisse. Vielleicht kann man sie am ehesten als eine Untersuchung des Lehr- und Lernprozesses kennzeichnen. Man kann z. B. erforschen, wie lange der Lehrer in einer Unterrichtsstunde spricht, wieviel Zeit die Schüler pro Unterrichtsstunde für Hausarbeiten aufwenden oder wieviel Gesprächszeit z. B. für Erklärungen, Definitionen und ähnliches benötigt wird (Meux/Smith 1961). Bei einigen dieser Untersuchungen wird man nur schwer ihren Wert einsichtig machen können. Denn nicht selten sind sie nur Forschungen um ihrer selbst willen. Deshalb muß auch die Arbeit von Smith und Meux besonders erwähnt werden, da sie wirklich originell und sehr erfolgversprechend ist. Dennoch darf man im ganzen davon ausgehen, daß der größte Teil dieser Art der Prozeßforschung in der Erziehung und der Psychotherapie weder für die Theorie noch für die Praxis fruchtbar ist.

(2) Die zweite Art von Prozeßforschung zielt auf die Erforschung der kausalen Beziehungen zwischen den Prozeßelementen (»dynamische Hypothesen«). Hier will man z. B. erforschen, ob ein größerer Zeitaufwand für eine an den curricularen Zielen orientierte Diskussion, die auf Kosten der Zeit für Übungsaufgaben geführt wird, zu einem besseren Verständnis für Algebra oder Geographie führt.

In einer anderen Variante dieser Art der Prozeßforschung versucht man Fragen zu beantworten wie: Wird durch die Betonung des Lehrer-Schüler-Dialogs die Bildung von Untergruppen und die Identifikation mit dem Lehrer gefördert? Diese Untergruppe von Prozeßhypothesen unterscheidet sich von Evaluationshypothesen dadurch, daß die unabhängigen Variablen entweder gar nicht unter den Kriterien einer summativen Evaluationsuntersuchung auftauchen oder daß sie nur eine Untergruppe der summativen Kriterien bilden würden. In beiden Fällen versucht man jedoch nicht, aufgrund von Korrelationsuntersuchungen die Vorzüge zu bestimmen.

Prozeßhypothesen dieser zweiten Art sind im allgemeinen genauso schwierig zu konkretisieren wie Hypothesen über Ergebnisse. Tatsächlich lassen sie sich manchmal sogar noch schwerer konkretisieren, weil sie vielleicht nur die Messung einer einzigen unter mehreren unabhängigen Va-

riablen erfordern und die gebräuchlichen Verfahren der Parallelisierung sich nur schwer dazu verwenden lassen, die anderen Variablen zu kontrollieren. Einige summative Evaluationsuntersuchungen haben den Vorteil, daß sie sich nur mit der Evaluation der Gesamtauswirkungen eines von Lehrern unterrichteten Curriculum befassen und daher nicht die spezifischen Elemente aufdecken müssen, die für die Verbesserung oder Verschlechterung der Ergebnisse verantwortlich sind. Dieser Vorteil wird jedoch hinfällig, wenn wir herausfinden wollen, welche nur Auswirkungen auf das Curriculum und nicht auf den Lehrer zurückgehen.

(3) Formative Evaluation. Diese Art der Forschung wird oft Prozeßforschung genannt, aber sie ist natürlich nur eine Ergebnisevaluation in einem Zwischenstadium der Curriculumentwicklung. Zwischen formativer Evaluation und der oben beschriebenen zweiten Art der Prozeßforschung gibt es zwei Unterschiede. Der eine Unterschied liegt in den Rollen. Die Rolle der formativen Evaluation besteht darin, die Schwächen und Stärken in der vorläufigen Fassung eines neuen Curriculum zu entdecken. Die Rolle der Forschung bei dieser zweiten Art der Prozeßforschung besteht aus spezifisch eigenen Aufgaben. Sie soll wichtige Fragen über Unterrichtsmechanismen zu beantworten versuchen. Der zweite Unterschied zur formativen Evaluation besteht in der unterschiedlichen Bedeutung, die der Frage zukommt, inwieweit die angewandten Kriterien den Zielen des Curriculum entsprechen. Im Unterschied zur formativen Evaluation braucht die Prozeßforschung der zweiten Art, die sich auf die Erforschung der kausalen Beziehungen zwischen Prozeßelementen richtet, die curricularen Ziele nicht zu berücksichtigen. Diese zwei Arten der Prozeßforschung lassen sich jedoch nicht immer scharf voneinander trennen; sie sind beide für die Curriculumforschung von großer Bedeutung.

Natürlich sollte man, wenn man einen Schulversuch durchführt, seine *Ergebnisse* evaluieren. Im allgemeinen ist ein Versuch sogar so angelegt, daß die Verfahren für die Evaluation der Ergebnisse mitgeplant werden. Das bedeutet jedoch nicht, daß der größte Teil der Forschung Evaluationsforschung ist. Sogar Prozeßforschung ist nicht immer Evaluationsforschung. Daß die Interpretation von Daten als Evaluation von Ergebnissen beschrieben werden kann, bedeutet noch nicht, daß die Interpretation und die Erklärungen sich auf die *Leistung* eines Curriculum beziehen. Sie können sich z. B. auch auf die zeitliche Dauer seiner verschiedenen Elemente richten. Darin liegt ein deutlicher Unterschied; allerdings verrät ein erheblicher Teil der Diskussion pro und contra Evaluationsforschung beträchtliche Unkenntnis über die Grenzen der Evaluation.

Evaluation und Überprüfung der Zielerreichung

Eine Reaktion auf die Verunsicherung durch Evaluation und vielleicht auch auf die Verwendung zu wenig sensibler Evaluationsverfahren besteht in der extremen Relativierung der Evaluationsforschung. In ihrem Verlauf wird die Frage, wie gut ein Curriculum seine Ziele erreicht, anstelle der Frage, wie gut ein Curriculum ist, die zentrale Frage der Evaluation. Es ist jedoch recht unwichtig, wie gut man Ziele erreicht, wenn sie überhaupt nicht wert sind, erreicht zu werden. Dieser Relativismus im Bereich der Evaluation konnte nur dadurch entstehen, daß man davon ausging, Urteile über Ziele seien subjektive, nicht auf rationaler Begründung beruhende Werturteile. Das verhält sich zweifellos oft so; jedoch bedeutet es nicht, daß es in diesem Bereich keine Objektivität geben kann. So könnte z. B. von einem Curriculum über amerikanische Geschichte, das nur auf das Auswendiglernen von Namen und Daten zielt, auf keinen Fall behauptet werden, es sei ein gutes Curriculum, auch dann nicht, wenn es seine Ziele gut erreicht. Genau so unzulänglich wäre jedoch auch ein Curriculum, bei dem überhaupt keine Namen und historische Daten gelernt werden. So wäre auch ein Curriculum in moderner Mathematik, in dem die Mehrzahl der Sekundarschulabgänger nicht gelernt hat, zuverlässig zu addieren und zu multiplizieren, völlig unzulänglich, unabhängig davon, was es sonst noch vermittelt. Solche Werturteile über Ziele werden durchaus abgegeben. Dafür, daß sie unterbleiben sollten, hat noch keiner *gute* Argumente angeführt. Denn dies sind gut begründete Werturteile, die auch spezifisch genug sind.

So gehören zu einem angemessenen Verständnis von Evaluation neben der Leistungsmessung in bezug auf die Ziele auch Verfahren zur Evaluation dieser Ziele. In den nächsten beiden Abschnitten werden wir Evaluationsverfahren erörtern, die sich auf Ziele beziehen und Verfahren einschließen, die solche Beziehungen zu umgehen versuchen. Zuerst soll dargelegt werden, daß Urteile über curriculare Ziele Teil der Evaluation sind, d. h. z. B., daß man nicht einfach beliebige Ziele akzeptieren darf. Das wiederum bedeutet jedoch nicht, daß diese Ziele für jede Schule, jeden Schulbezirk, jeden Lehrer, jede Altersstufe gleich sind. Eine Schule, in der die Mehrzahl der Schulabgänger direkt in den Beruf geht, sollte andere Ziele als eine Schule haben, die 95 Prozent der Schulabgänger zur Hochschule entläßt. Das heißt natürlich nicht, daß die Lehrer, Schulleiter oder Curriculumentwickler bei der Auswahl der Ziele nicht kritisiert werden dürfen. Ein großer Teil der Energie in den gegenwärtigen Bemühungen um Curriculumreform geht unmittelbar auf die Überzeugung zurück, daß die bisherigen Ziele grundsätzlich falsch waren, daß z. B. Lebensanpassung als

Erziehungsziel viel zu stark betont wurde. Nun in die entgegengesetzte Richtung einzuschwenken ist nur allzu leicht und in keiner Weise besser.

Der Prozeß der Relativierung hat jedoch nicht nur zu übertriebener Toleranz für zu restriktive Ziele geführt, sondern hat auch zu inkompetenter Evaluation des Ausmaßes, in dem diese Ziele erreicht wurden, beigetragen. Wie man sich auch zur Evaluation stellt, es läßt sich leicht nachweisen, daß es gegenwärtig in den USA nur sehr wenige professionell kompetente Evaluatoren gibt. Der Gedanke, daß jedes Schulsystem oder jeder Lehrer seine Leistung sinnvoll evaluieren kann, ist genauso abwegig wie die Ansicht, daß jeder Psychotherapeut dazu fähig ist, seine Arbeit mit den Patienten zu evaluieren. Gewiß können sie durch die sorgfältige Untersuchung ihrer eigenen Arbeit viel lernen; sie können ohne Zweifel darin einige gute und schlechte Aspekte identifizieren. Aber wenn man die wichtigen Fragen über den Prozeß oder das Ergebnis beantwortet haben will, braucht man Fertigkeiten und Mittel, die nur sehr schwer zu finden sind.

Intrinsische Evaluation und Ergebnisevaluation

Für die Evaluation eines Unterrichtsinstruments lassen sich zwei Ansätze unterscheiden, die beide auch in der Literatur häufig einander gegenübergestellt werden. Wenn man ein Werkzeug, etwa eine Axt evaluieren will, kann man einmal die Konstruktion der Schneide, die Gewichtsverteilung, die Stahllegierung, die Güte des Ahorngriffs untersuchen; man kann aber auch die Art und Geschwindigkeit der Hiebe untersuchen, die sie in der Hand eines guten Holzfällers macht. In beiden Fällen kann die Evaluation summativ oder formativ sein; denn diese beiden Begriffe kennzeichnen Rollen, nicht unterschiedliche Verfahrensweisen der Evaluation.

Der erste Ansatz zielt auf eine Bewertung des Instruments selbst; in unserem Zusammenhang würde er z. B. der Evaluation der Inhalte, Ziele, Zensierungsverfahren, Lehrereinstellungen entsprechen. Wir nennen diesen Ansatz *intrinsische Evaluation.* Seine Kriterien sind im allgemeinen nicht operational formuliert; sie beziehen sich auf das Instrument selbst und nur indirekt auf seine pädagogischen Auswirkungen. Der zweite Ansatz zielt ausschließlich auf die Untersuchung der Wirkungen des Unterrichtsinstruments auf den Schüler und spezifiziert diese gewöhnlich operational, wobei die Auswirkungen auf Lehrer, Eltern usw. auch berücksichtigt werden können. Zu diesem Ansatz gehört z. B. eine Bewertung der Unterschiede zwischen Vor- und Nachtests, zwischen den Tests der Versuchsgruppe und der Kontrollgruppe in bezug auf eine Anzahl von Kriterien. Wir nennen diese Form der Evaluation *Ergebnisevaluation.* Ihre Befür-

worter würden behaupten, daß nur von Bedeutung ist, welche Wirkungen das Curriculum auf die Schüler hat, und daß die Evaluation der Ziele und Inhalte nur insofern sinnvoll ist, als ihre Ergebnisse mit der Ergebnisevaluation korrelieren. Im Gegensatz dazu könnte der intrinsische Evaluator darauf hinweisen, daß viele wichtige Werte und Charakteristika des Curriculum sich in der reinen Ergebnisevaluation nicht niederschlagen. Um seine Auffassungen zu veranschaulichen, braucht der intrinsische Evaluator nur auf Qualitäten wie Eleganz, Modernität, Struktur eines Curriculum zu verweisen, die sich am besten durch eine unmittelbare Analyse des gesamten Materials erfassen und durch Berücksichtigung von Aspekten des Unterrichts wie Schülerbeteiligung und Unterrichtsklima beurteilen lassen.

Im vorherigen Abschnitt wurde die Behauptung vertreten, daß die bloße Überprüfung des Erreichens von Zielen ein schlechter Ersatz für summative Evaluation sei, da man damit der Grundproblematik der Evaluation ausweiche. Wenn man unter Berücksichtigung der Ziele evaluieren will, muß man die Ziele selbst einer Evaluation unterziehen. Dabei liegt eine Schwierigkeit darin, daß die intrinsische Evaluation sekundäre Ziele und Kriterien benötigt und deshalb sofort auch die Frage nach dem Wert dieser Kriterien unter Bezug auf die primären Ergebniskriterien stellen muß. Im Rahmen der Evaluation ist ein sinnvoller Kompromiß möglich, wenn man einige intrinsische Kriterien und einige Ergebniskriterien berücksichtigt. Zweifellos läßt sie sich in zahlreichen Evaluationssituationen anwenden. Doch bevor wir weiter zu beurteilen versuchen, welches Verhältnis zwischen beiden Arten der Kriterien im Rahmen der Evaluation angemessen ist, wollen wir die entsprechenden praktischen Verfahren etwas näher untersuchen.

Praktische Vorschläge für eine Mischform bei Evaluationsuntersuchungen (Hybrid Evaluation)

Zu Beginn liegen allen Curriculumprojekten allgemeine Zielvorstellungen zugrunde. Selbst wenn sie nur einen interessanteren oder moderneren Unterricht bewirken sollten, wurden sie begonnen, weil man mit dem gegenwärtigen Curriculum in bestimmten Punkten nicht zufrieden war und mit Hilfe des Projekts eine Verbesserung der Situation herbeiführen wollte. Im allgemeinen werden die Ziele und Vorstellungen während der Planungsdiskussion stärker spezifiziert. Wenn drei gleich vertretbare Ziele formuliert werden können, die zu unvereinbaren Anforderungen an das Curriculum führen, kann man sich nach einer gründlichen Diskussion der

Projektziele z. B. dazu entschließen, ein dreigliedriges, auf die unterschiedlichen Lehrer- und Schülerinteressen zielendes Curriculum zu entwickeln. Oder man entschließt sich dazu, dasselbe Ziel – in sehr allgemeinem Sinn – mit drei verschiedenen, gleichwertigen Curriculumvarianten zu erreichen. Diese Varianten werden dann zu den sekundären Kriterien für das Curriculum. Daß diese Varianten oft als miteinander unvereinbar angesehen werden, macht deutlich, daß sie ziemlich wichtigen Inhalts sind.

Ein anderes wichtiges sekundäres Kriterium bezieht sich auf den Geltungsbereich; von Anfang an weiß man, daß wenigstens bestimmte Themen behandelt werden sollten. Wenn das nicht möglich ist, muß eine Abdeckung durch andere Themen erfolgen. Im allgemeinen enthält ein Projekt wenigstens einige abstrakt formulierte Kriterien für primäre und sekundäre Eigenschaften, z. B. für die Verhaltensziele und die intrinsischen Qualitäten eines Instruments. In diesem Fall sollte man eine Mischform der Evaluation wählen, in der beide Kriterienarten berücksichtigt werden, um den Erfolg eines Curriculum festzustellen. In diesem frühen Stadium curricularer Planung sollten einige Projektmitglieder die Aufgabe übernehmen, die Ziele zu formulieren. Die häufig dagegen geäußerten Bedenken sind eine Reaktion auf die rigide Forderung nach präziser Formulierung der Ziele in dieser Phase der curricularen Planung. Alle vom Projektteam akzeptierten Ziele – wie abstrakt oder spezifisch sie auch formuliert sein mögen – einschließlich der Ziele, die nur als vorläufige oder mögliche Ziele angesehen werden, sollten schon in diesem Entwicklungsstadium in einer Liste zusammengestellt werden. Keines der Ziele sollte als absolut verbindlich angesehen werden, da sie lediglich eine Hilfsfunktion haben. Dabei sollte man durchaus auch an das negative Beispiel von Projekten denken, bei denen das kreative Engagement der Mitarbeiter dazu geführt hat, die Bedingungen und Erfordernisse der Realität zu vernachlässigen. Daher sollte man von Anfang an beachten, daß bei zu großer Abweichung vom traditionellen Curriculum, die Implementation des neuen Curriculum in der Schule sehr schwierig wird. Wenn eine umfassende Implementation eine zentrale Zielsetzung des Curriculum ist, sollte sie zusammen mit den kognitiven und affektiven Zielen genannt werden. Eine solche Zielsetzung ist durchaus angemessen, da man das Bildungswesen ja nicht mit Curricula reformieren kann, die niemals in die Schule gelangen. Bereits in einem frühen Entwicklungsstadium empfiehlt es sich, unter Bezug auf diese Ziele die Inhalte auszuwählen.

Im Verlauf der Projektentwicklung sollten drei mit der Zielformulierung zusammenhängende Dinge gesehen werden. Erstens sollten alle formulierten Ziele regelmäßig überprüft und unter Berücksichtigung der im Verlauf der Curriculumentwicklung entstandenen Divergenzen an den

Stellen modifiziert werden, an denen diese Veränderungen zu anderen, wertvolleren Zielen geführt haben. Doch selbst wenn keine Modifikation der Ziele erfolgt, dient die Überprüfung der Ziele dazu, die Curriculumentwickler an die übergreifenden Ziele des Projekts zu erinnern.

Zweitens sollte man möglichst rechtzeitig mit der Konstruktion einer Sammlung von Testaufgaben beginnen. Aus den Leistungstests können Testaufgaben in diese Sammlung aufgenommen werden. Mit ihrer Entwicklung soll eine operationale Fassung der Ziele erstellt werden. Deshalb ist ihre Überprüfung gleichzeitig eine Überprüfung der allgemeiner formulierten Ziele. Schon wenn das Projekt bei der Entwicklung der ersten Einheit eines auf zehn Einheiten angelegten Curriculum ist, empfiehlt es sich, die Testaufgaben so zu formulieren, daß sie in der Schlußprüfung bei der letzten Einheit oder in einem ein Jahr später eingesetzten Test verwendet werden können. Bekanntlich verändert sich die Konzeption der Curriculumziele bei der Formulierung solcher Aufgaben. Man sollte nicht zuviel Zeit dafür aufwenden, und doch sollte man im Zusammenhang mit den Zielen darüber nachdenken, welche Testaufgaben eine bestimmte Lernleistung oder eine Veränderung der Motivation in der abschließenden Prüfung oder in einer späteren Untersuchung erfassen. Manchmal werden sich keine Testaufgaben entwickeln lassen, da nicht alle Werte eines Curriculum in der abschließenden oder in einer später erfolgenden Prüfung direkt in Erscheinung treten. Wo sie sich nicht zeigen, sollte wenigstens angegeben werden, wann und wie sie sich etwa in der Berufswahl, in den Einstellungen von Erwachsenen oder im Unterrichtsverhalten ausdrücken.

Drittens sollte man in einem mittleren Stadium der Curriculumentwicklung versuchen, einige externe Beurteilungen über den Zusammenhang zwischen den angegebenen Zielen, den wirklichen curricularen Inhalten und den gesammelten Testaufgaben zu erhalten. Denn ohne solche Beurteilungen dürfte die Validität der Tests und der praktische Nutzen des Curriculum wohl eingeschränkt sein. Um diese Aufgaben zu erfüllen, muß der einzelne Beurteiler nicht unbedingt ein professioneller Evaluator sein. Professionelle Evaluatoren sind sogar oft wenig geeignet. Ein Fachwissenschaftler, ein pädagogischer Psychologe oder ein Curriculumfachmann kann diese Aufgaben besser erfüllen. Die dazu benötigten Qualitäten sind nicht mit professionellen Fähigkeiten identisch. Sie bestehen in der Fähigkeit zur »Konsistenzanalyse«. Dies ist ein Gebiet, für das man Mitarbeiter nicht ohne Probezeit einstellen sollte. Vielleicht sollte der Wissenschaftler, der die Konsistenzanalyse macht, wenigstens in der Probezeit nicht persönlich mit dem Projektteam zusammenarbeiten. Zu diesem Zeitpunkt genügt vielleicht ein kurzer schriftlicher Bericht, in dem die vorhandenen In-

formationen zur Verfügung gestellt werden. In einem späteren Stadium jedoch, wenn große Divergenzen zwischen (a) verbalisierten, (b) impliziten und (c) getesteten Zielen vermieden werden sollen, ist die Konsistenzanalyse sehr wichtig. Ein Wissenschaftler kann mit einer guten Konsistenzanalyse nicht nur verhindern, daß das Projekt durch den Übereifer seiner Mitarbeiter oder durch Fehleinschätzungen seiner tatsächlichen Auswirkungen in Sackgassen gerät, sondern er kann auch wertvolle Anregungen zur Entwicklung des Projekts in eine neue Richtung liefern. Er muß auf fehlende und überflüssige Testaufgaben in der entsprechenden Sammlung und auf fehlende und irrelevante Ziele in der entsprechenden Liste achten. Schließlich läßt sich auch die Psychotherapie nicht dadurch rechtfertigen, daß der Psychotherapeut *meint*, er würde dem Patienten helfen, sondern nur dadurch, daß er es tatsächlich tut; entsprechendes gilt für die Curriculumforschung.

Daher braucht man ein besser entwickeltes, wenn auch ähnliches Verfahren, um die Diskrepanz zwischen den curricularen Materialien, Zielen, Tests und den Normen eines solchen Curriculum zu identifizieren. Die Curriculumhersteller neigen leicht zu der Annahme, daß diese Größen kongruent sind. Daher bedarf es eines externen Evaluators. Wenn er gut ist und über hinreichende Erfahrungen verfügt, wird er Nebenwirkungen und Diskrepanzen entdecken, die für die finanzierenden Ministerien oder Stiftungen und die Adressaten aufschlußreich sind. Der Beweis dafür, daß der Evaluator nicht nur seine eigenen Vorurteile zum Maßstab seiner Evaluation macht, muß in seinen Argumenten liegen, die in vielen Fällen die Curriculumhersteller durchaus überzeugen können.

Wenn man während der ganzen Curriculumentwicklung in der beschriebenen Weise vorgeht, wird man am Ende eine große Testaufgabensammlung haben. Die Antworten auf diese Testaufgaben können dazu dienen, jedes angestrebte Ergebnis des Curriculum zu überprüfen; das Ergebnis dürfte dann im allgemeinen wirklich nur auf das Curriculum zurückzuführen sein. Diese Sammlung hat mehrere erhebliche Vorteile. Sie ist eine operationale Fassung der Curriculumziele, die erstens den Schülern eine Vorstellung von den an sie gestellten Erwartungen vermitteln kann, die zweitens ein wertvolles Hilfsmittel für die Konstruktion der abschließenden Prüfung ist und die drittens dem Curriculumentwickler ein detailliertes Bild seines eigenen Erfolgs bietet. Um sich darüber Gewißheit zu verschaffen, kann er jedem Schüler eine jeweils unterschiedliche Zufallsauswahl von Testaufgaben im Rahmen einer formativen Evaluation vorlegen, anstatt jedem Schüler eine bestimmte Zufallsauswahl zur Beantwortung zu geben, wie man es vielleicht gerechterweise in einer abschließenden Prüfung machen müßte.

Bisher wurden die Grundzüge einer Evaluation beschrieben, die unter Zuhilfenahme sekundärer Kriterien erfolgt. Dabei haben wir vor allem auf inhaltliche Charakteristika als eine Form der Zielangabe hingewiesen, da Curriculumteams oft behaupten, daß ein Vorteil ihres Curriculum in der Modernität und Aktualität seiner Inhalte liegt. Um diese Behauptung zu verifizieren, braucht man das Curriculum nur von einigen qualifizierten Fachwissenschaftlern analysieren zu lassen. Dabei ergeben sich jedoch besondere Schwierigkeiten. Bestenfalls können wir in Erfahrung bringen, ob das Curriculum starke Verzerrungen oder Mängel hinsichtlich der wichtigsten derzeitigen Kenntnisse und Anschauungen enthält. Offen bleibt bei diesem Verfahren die Frage – worauf vor allem der Befürworter der Ergebnisevaluation hinweisen würde –, inwieweit die Ziele und Inhalte des Curriculummaterials den Schülern wirklich vermittelt werden. Selbst wenn ein Curriculum im Hinblick auf seine fachliche Qualität für wissenschaftliche Experten nicht ganz zufriedenstellend ist, kann es von den fachlichen Inhalten manchmal durchaus eine bessere Vorstellung vermitteln als ein nur nach fachwissenschaftlichen Kriterien entwickeltes anderes Curriculumprojekt. Der Vorteil der geschilderten Methode besteht darin, ein Verfahren bereitzustellen, die Lücke zwischen intrinsischer Evaluation und Ergebnisevaluation, zwischen bloßem Messen, ob die Lernziele erreicht worden sind, und vollständiger Evaluation auszufüllen.

Weitere Verbesserungen der obigen Ausführungen sind erforderlich und in jeder guten Untersuchung unerläßlich. Sie hängen mit der Rolle der Konsistenzanalyse zusammen und sind für formative Evaluationsuntersuchungen noch wichtiger als für summative, da sie die Gründe für schlechte Ergebnisse zu entdecken helfen. Man muß in Erfahrung bringen, ob es gelungen ist, eine Entsprechung zwischen drei zusammenhängenden Problemen herzustellen:

1. die Entsprechung von Zielen und Curriculuminhalten,
2. die Entsprechung von Zielen und Prüfungsinhalten,
3. die Entsprechung von Curriculuminhalten und Testinhalten.

Im Grunde genommen, brauchte man nur zwei der Probleme zu lösen, um auch das dritte evaluieren zu können. Aber in der Praxis empfiehlt es sich, um Irrtümer möglichst auszuschließen, jedes Problem unabhängig vom anderen zu behandeln. Aufgrund dieser Überlegungen könnte man den Eindruck gewinnen, als könnte eine Person oder Gruppe alle Entsprechungen abschätzen. Es empfiehlt sich jedoch, alle Einschätzungen unabhängig voneinander durchführen zu lassen und sogar von nicht an dem Projekt beteiligten Personen wiederholen zu lassen. Nur so kann man wahrscheinlich die wirklichen Gründe für enttäuschende Ergebnisse herausfinden. Sogar das Curriculumprojekt des Physical Science Study Committee, das so sorg-

fältig wie die meisten derzeitigen Curriculumprojekte getestet wurde, hat an keiner Stelle die hier als notwendig bezeichneten Verfahren der Analyse angewandt.

Das schwierigste Problem der Testtheorie und der Herstellung von Tests liegt in der Konstruktvalidität, die durch das angesprochene Problem vor allem berührt wird. Man darf die mit ihr verbundenen Probleme nur vernachlässigen, wenn man dafür in Kauf zu nehmen bereit ist, (1) daß die intendierten Ziele nicht im Curriculum realisiert werden oder (2) daß die Prüfungen nicht testen, was das Curriculum lehrt, oder (3) daß die Prüfungen nicht die Werte und Materialien testen, die das Curriculum vermitteln soll. Es gibt in der Praxis viele Möglichkeiten, mit denen man die hier beschriebenen Vergleiche durchführen kann: die Anwendung von Q- und R-Techniken (Q-sorts, R-sorts), von parallelisierten Tests und projektiven Tests für die Analyse usw. Die Aufgabe muß jedoch im Rahmen der Evaluation in irgendeiner Form gelöst werden.

Das Für und Wider einer reinen Ergebnisevaluation

Der »reine« Ergebnisevaluator betrachtet die sich bei der beschriebenen experimentellen Planung ergebenden Schwierigkeiten mit Skepsis. Nach seiner Auffassung ist die Berücksichtigung von Ziel- und Inhaltsbewertung oder einer anderen sekundären Bewertung im Rahmen der Curriculumevaluation nicht nur irrelevant, sondern auch unzuverlässig. Seiner Ansicht nach braucht man weder zu untersuchen, was ein Lehrer zu tun behauptet oder nach Aussage seiner Schüler tut, noch was er im Unterricht sagt und was die Schüler in ihren Schulbüchern lesen. Wichtig ist lediglich, was der Schüler nach Beendigung seiner Arbeit mit dem Curriculum sagt und was er nicht gesagt hätte, wenn er nicht mit diesem Curriculum gearbeitet hätte. Nach seiner Auffassung kommt es also nur darauf an, die Auswirkungen eines Curriculum festzustellen, und nicht, ob ihm gute Intentionen zugrunde liegen.

Für den »reinen« Ergebnisevaluator gibt es jedoch auch erhebliche Schwierigkeiten. Er kann das Problem der Konstruktvalidität nicht ganz umgehen, d. h. er kann den Schwierigkeiten nicht ausweichen, die in dem Versuch liegen, die Lernleistung der Schüler *mit einem sinnvollen Grad an Allgemeinheit* zu beschreiben. Es ist zwar einfach, die Testergebnisse so darzustellen, daß ersichtlich wird, wieviel Schüler (in Prozenten) die einzelnen Testaufgaben gelöst haben; man muß jedoch wissen, ob man aufgrund dieser Lösung sagen kann, daß sie bestimmte Elemente der Astronomie oder den ökologischen Ansatz in der Biologie *besser verstehen.* Doch von

Daten über die Lösung spezifischer Testaufgaben zu derartigen Schlußfolgerungen ist es ein weiter Weg. Der Ergebnisevaluator hat zwar recht mit der Behauptung, daß man für solche Schlußfolgerungen nicht unbedingt eine Diskussion der Ziele benötigt. *Ohne* eine solche Zieldiskussion verfügt man jedoch nicht über die benötigten Daten, um eine Entscheidung über die Angemessenheit unterschiedlicher Erklärungen des Erfolgs oder Versagens bestimmter Aspekte des Curriculum zu fällen. Wenn man z. B. den Ansatz der reinen Ergebnisevaluation wählt und dann entdeckt, daß die von den Schülern erinnerten und reproduzierten Inhalte von Fachwissenschaftlern als inadäquat bezeichnet werden, dann weiß man noch nicht, ob dies auf die Unzulänglichkeit der Intentionen, der curricularen Realisierung der Ziele oder der curriculumspezifischen Prüfungen zurückzuführen ist.

Das soll weiter erläutert werden: Wenn man eine reine Ergebnisevaluation durchführen will und die Schülerleistung nur am Ende des Curriculum von einem externen Beurteiler beurteilen lassen will, muß man jemanden für die Beurteilung auswählen. Dabei zeigt es sich, daß die Auswahl von bestimmten Interessen und bestimmten Zielen abhängt, die man genausogut explizit machen könnte. Genauso muß der Evaluator die Schülerleistung auf ein *bestimmtes* Kriterium beziehen. Das kann seine Auffassung von einem angemessenen Fachverständnis des ganzen Bereichs oder seine Ansicht über die angemessene Verständnisfähigkeit eines Schülers der zehnten Klasse sein. Der Ergebnisevaluator behauptet zu Recht, daß man von jeder Zieldiskussion absehen und doch genau feststellen kann, was Schüler gelernt haben. Mit dem gleichen Recht geht er davon aus, daß letzteres die wichtigste Variable überhaupt ist. Aber er irrt sich in der Annahme, daß man ohne weiteres die Ergebnisse des Lernens so beschreiben kann, daß sie für unsere Zwecke nützlich sind, oder daß man das Curriculum ohne Bezug auf allgemeine Ziele rechtfertigen kann. Deshalb ist eine reine Ergebnisevaluation etwas oberflächlich, so daß beim augenblicklichen Diskussionsstand eine Mischform der Evalution vorzuziehen ist.

Der Ergebnisevaluator weist konsequenterweise zu Recht darauf hin, daß es unverantwortlich ist, »elegante«, »moderne«, »präzise« Curricula zu entwickeln, wenn ihre Qualitäten nicht bis zu den Schülern gelangen. Solange es sich dabei nur um sekundäre Qualitäten handelt, reicht es aus, ihre Existenz lediglich anzunehmen; sobald man sie aber für wesentlich hält, darf man sich damit nicht begnügen. Deshalb muß man einen wissenschaftlichen Evaluator berufen, dessen Aufgabe darin besteht, nicht nur die Curriculummaterialien oder die Sammlung der Testaufgaben, sondern auch die exakte Leistung der Klasse bei jeder Testaufgabe zu untersuchen. Mit Hilfe dieser Ergebnisse soll er abschätzen, inwieweit das Cur-

riculum die Inhalte angemessen vermittelt. Trotzdem fehlt uns dann immer noch die Diagnose der Ursache für die Mängel. Deshalb ist dies ein schlechtes Verfahren für formative Evaluation; doch wir können summative Evaluation durch dieses Verfahren vereinfachen. Deshalb müssen wir unseren umfassenden Plan durch eine genaue Analyse der *Ergebnisse* der Schülertests und nicht nur des Curriculum und der Testinhalte ergänzen. Es lohnt sich nicht, viel Mühe auf die Aufstellung und wechselseitige Analyse der Ziele, Tests und Inhalte eines Curriculum zu verwenden, wenn man lediglich versucht, eine Prozentangabe in bezug auf die maximal möglichen Punkte als Index für das Ausmaß zu benutzen, in dem die Ziele erreicht worden sind – es sei denn, diese Angabe liegt zufällig ziemlich dicht bei 100 oder 0 Prozent. Die Leistungen der Schüler in den Tests der mittleren Stadien müssen analysiert werden, um exakt festzustellen, wo z. B. ein ausreichendes Verständnis grundlegender Fakten und die Übung wichtiger Fertigkeiten fehlen. Prozentangaben sind dabei nicht so wichtig. Es ist vielmehr die *Art* der Fehler, die für die Evaluation und für die Neufassung des Curriculum wichtig ist. Daher braucht man für die formative und die summative Evaluation eine klare Beschreibung der Stärken und Schwächen des Curriculum. Die umfangreiche Sammlung von Testaufgaben ist ein bewährtes Verfahren, um die Unzulänglichkeiten im Curriculum zu lokalisieren. Aber es kann nur dann voll ausgenutzt werden, wenn die Ergebnisse adäquat evaluiert werden. Dazu muß man in gleicher Weise unabhängige Beurteiler, Hypothesenentwicklung, Testen der Art der Fehler, Längsschnittanalysen von Leistungsunterschieden bei gleichen Schülern verwenden. Daher ist eine angemessene Evaluation von Curriculummaterialien sehr schwierig. Zudem ist die Verwendung von Aufsätzen, die Entwicklung und Anwendung von neuen Instrumenten, die Verwendung der Berichte der Versuchsleiter, die Übertragung dieses gesamten Materials in ein speziell entwickeltes Beurteilungsschema kostspielig und zeitraubend. Doch ist diese Art auch nicht zeitraubender als gute Forschungs- und Entwicklungsarbeit in der Technik. In diesem Zusammenhang soll eine weitere Unterscheidung zwischen zwei Methoden getroffen werden.

Vergleichende und nicht-vergleichende Evaluation

Die Ergebnisse der Evaluation neuerer Curricula sind oft erstaunlich ähnlich. Bei dem Vergleich zwischen Schülern, die nach dem alten Curriculum unterrichtet werden, und Schülern, die nach dem neuen unterrichtet werden, schneiden letztere gewöhnlich besser bei den für ihr Curriculum kon-

struierten Tests ab als erstere und schlechter bei den für das alte Curriculum konstruierten Tests; für die Schüler, die nach dem alten Curriculum unterrichtet werden, sind die Ergebnisse entsprechend umgekehrt. Es fehlt im allgemeinen ein größerer Leistungszuwachs für dieselben Kriterien. Leicht hat man den Eindruck, daß ein solches Ergebnis kaum wirklich relevant ist; denn daß das Ergebnis positiv und nicht negativ ist, hängt ausschließlich von den der Evaluation zugrunde gelegten Kriterien, d. h. von den benutzten Tests ab. Aufgrund dieses Sachverhalts erhebt sich die berechtigte Frage, ob man nicht den von den Fachwissenschaftlern den Inhalten und Zielen zugeteilten Wert viel schwerer gewichten sollte als die geringen Unterschiede im Leistungsniveau in bezug auf ungewichtete Kriterien. Wenn man sich dazu entschließt, werden relativ unbedeutende Leistungsverbesserungen hinsichtlich der richtigen Ziele sehr wertvoll, und, so gesehen, schneidet das neue Curriculum bei dem Vergleich wesentlich besser ab. Ob diese Veränderung der Gewichtung sich wirklich rechtfertigen läßt, muß gründlich untersucht werden. Dazu muß z. B. die wirkliche Bedeutung der zum Verständnis moderner Physik im alten Curriculum fehlenden Elemente analysiert werden. Denn nur zu leicht ist man in der Versuchung, die neue Gewichtung für richtig zu halten, da man ja von der Überlegenheit des neuen Curriculum fest überzeugt ist.

Ferner muß man sich fragen, ob die Tests bei den nach dem neuen Curriculum unterrichteten Schülern wirklich das Verständnis erfassen. In diesem Zusammenhang empfiehlt es sich, eine umfangreiche Sammlung von Testaufgaben zu verwenden. Cronbach schlägt eine Sammlung von 700 Testaufgaben vor. Bei einer gründlichen Evaluation eines ein- oder zweijährigen Curriculum ist diese Größenordnung durchaus sinnvoll. In dieser Sammlung sollte man keine Testaufgaben aufnehmen, die nur auf die terminologischen Unterschiede des neuen Curriculum zielen. Wenn die Sammlung hauptsächlich solche Aufgaben enthält, werden die Schüler des neuen Curriculum natürlich viel besser abschneiden, obwohl eine echte Überlegenheit nicht besteht. Cronbach weist daher mit Recht auf die Unzulänglichkeit curriculumabhängiger Terminologie hin, obwohl er mit der Unterscheidung und Trennung zwischen Verständnis und Terminologie zu weit geht. Deshalb sollte man auch hier externe Evaluatoren zur Konstruktion und Beurteilung der Aufgabensammlung hinzuziehen.

Die Reaktionen auf diesen Sachverhalt reichen von der etwas naiven Vermutung, daß solche Resultate nur die Schwächen von Evaluationsverfahren zeigen, bis zur folgenden interessanten Überlegung Cronbachs: »Da die Ergebnisse von Gruppenvergleichen sehr fragwürdig sind, sollte man meiner Meinung nach eine gute Untersuchung in erster Linie so planen, daß sie es erlaubt, Leistungen einer genau umschriebenen Gruppe am

Ende eines Curriculum im Hinblick auf die wesentlichen Ziele und Nebenwirkungen zu bestimmen« (1963, 43, *48*). Cronbach schlägt offensichtlich ein Verfahren vor, bei dem wir zwar den Vergleich mit Lernzielen nicht vermeiden können, wohl aber den mit einer anderen Gruppe, die der Testgruppe in bezug auf relevante Variablen entspricht. Wie sieht nun eine nicht-vergleichende alternative Verfahrensweise für Evaluation aus? Cronbach führt dazu aus: »Unser Problem ist mit dem eines Ingenieurs, der ein neues Auto überprüft, vergleichbar. Er kann sich die Aufgabe stellen, die Leistungsfähigkeit und Zuverlässigkeit des Autos genau zu bestimmen. Es würde aber an dem Problem vorbeiführen, wenn er sich die Frage stellen würde: Ist dieses Auto besser oder schlechter als die konkurrierende Automarke?« (a. a. O.). Es ist richtig, daß der Ingenieur vielleicht nur an der Frage der Leistung und Verläßlichkeit des neuen Autos interessiert ist. Aber kein Ingenieur hat jemals nur dieses Interesse gehabt, und keiner wird es jemals haben. Ziele sind nur im Kontext einer praktischen Entscheidung wichtig. Unrealistische Lernziele z. B. sind nicht wichtig. Das Maß der Leistung und der Verläßlichkeit eines Autos und unser Interesse daran hat seinen Ursprung *ausschließlich* in dem Wissen darüber, was sich bisher innerhalb einer bestimmten Preisklasse mit bestimmtem Raum und bestimmtem Gesamtgewicht als möglich erwiesen hat. Die Anwendung von geeichten Instrumenten ist keine Alternative, sondern eine indirekte Form der vergleichenden Untersuchung. Was wir messen, hat eine absolute Qualität. Der Grund dafür, daß wir sie messen, liegt darin, daß wir sie für eine wichtige Variable im Rahmen eines Vergleichs halten. Wenn man genau wüßte, daß alle Autos die Eigenschaft P haben, dann brauchte man sie nicht zu messen. Aber im allgemeinen ist P eine stärker oder geringer bewertete Variable, der unterschiedliche Bedeutung zugemessen wird und die wir messen, weil sie eine Grundlage für den Vergleich bildet.

Dasselbe gilt für den Bereich der Curriculumentwicklung. Es gibt bereits Curricula für fast jedes Thema, und es besteht wahrhaftig kein Interesse daran, Curricula um ihrer selbst willen zu produzieren. Man ist an neuen Curricula interessiert, weil sie vielleicht in wichtigen Aspekten besser als die vorhandenen sind. Man kann jemanden beauftragen, ein Curriculum in bezug auf bestimmte Variablen zu bewerten, ohne gleichzeitig zu fordern, daß er die Leistung anderer Curricula bezüglich dieser Variablen feststellt. Aber wenn man das Curriculum – im Unterschied zur Beschreibung seiner Leistung – *evaluiert*, dann ist man unweigerlich mit der Frage seiner Über- oder Unterlegenheit in bezug auf andere Curricula konfrontiert. Die Behauptung, ein Curriculum sei ein »wertvoller Beitrag«, ein »wünschenswertes«, »nützliches« oder »gutes« Curriculum, bedeutet bereits, ihm einen relativen Wert zuzuordnen. Tatsächlich sind die Skalen,

die wir zur Leistungsmessung von Curricula anwenden, oft Prozentskalen oder andere Skalen mit implizitem Vergleich.

Es gibt sogar wichtige Gründe, die Frage sofort in der Form eines Vergleichs zu formulieren. Vergleichende Evaluation ist oft viel einfacher als nicht-vergleichende zu handhaben, weil man häufig Tests benutzen kann, die Unterschiede erbringen, anstatt eine absolute Skala konstruieren zu müssen, um dann schließlich die absoluten Testwerte (scores) zu vergleichen. Handelt es sich z. B. um Curricula für das Lernen von Schach, kann man zwei Gruppen in bezug auf Hintergrundvariablen parallelisieren, unterrichtet sie nach verschiedenen Curricula und läßt sie dann in einem Turnier gegeneinander spielen. Ein absolutes Maß für eine Fertigkeit aufzustellen wäre außerordentlich schwierig; durch die Art der vergleichenden Evaluation jedoch können wir leicht konsistente und signifikante Unterschiede erhalten. Cronbach macht nicht den Fehler der reinen Ergebnisevaluation, den Bezug auf allgemeine Ziele zu leugnen; aber er schlägt ein Verfahren vor, das das implizit vergleichende Element in jedem Bereich des »Social Engineering« einschließlich der Curriculumevaluation unterschätzt, so wie Ergebnisevaluation die implizite Aussagekraft abstrakter Kriterien unterschätzt.

Sodann entwickelt Cronbach in diesem Abschnitt einen Gedankengang, über den es keine Meinungsverschiedenheiten gibt. Er weist darauf hin, daß in allen Vergleichen zwischen sehr unterschiedlichen Unterrichtsinstrumenten kein wirkliches Verständnis der Gründe für eine Leistungsdifferenz dadurch gewonnen wird, daß man die Überlegenheit eines Unterrichtsinstruments gegenüber den anderen entdeckt. Denn niemand kennt einzelne Elemente, die für die Überlegenheit verantwortlich sind. Aber ein Verständnis der Leistungsdifferenz zwischen Curricula ist nicht das *einzige* Ziel der Evaluation. Sie richtet sich ebenso auf die Fragen der Unterstützung, Annahme, Anerkennung, Verbesserung von Curricula usw. Diese äußerst wichtigen Fragen können schon – wenn auch nicht immer vollständig – durch die Feststellung der Überlegenheit des Curriculum beantwortet werden. Wie wir in einem früheren Abschnitt ausgeführt haben, ist die reine Ergebnisevaluation der zielbezogenen Evaluation darin unterlegen, daß zu ihren Ergebnissen nicht die Daten gehören, die uns helfen, die Ursachen der Schwierigkeiten usw. aufzudecken. Nach Cronbachs Auffassung kann hier eine nicht-vergleichende Methode mit größerer Wahrscheinlichkeit die Daten liefern, die für spätere Verbesserungen benötigt werden. Das ist jedoch nicht ein Vorteil der nicht-vergleichenden Methode als solcher. Es ist lediglich der Vorteil von Methoden, bei denen eine größere Zahl von Variablen genauer untersucht wird. Wenn man aber die Gründe für die Unterschiede zwischen den Curricula feststellen will, kann man »gut

kontrollierte Untersuchungen kleinerer Größenordnung mit Gewinn dazu benutzen, alternative Fassungen desselben Curriculum zu vergleichen«, während der umfassende Vergleich großen Ausmaßes weniger wertvoll ist. Das bedeutet aber nicht, daß keine vergleichenden Untersuchungen im Lauf des Evaluationsverfahrens benötigt werden. Cronbachs Argument besagt lediglich, daß man, um *Erklärungen* zu erhalten, mehr Kontrollgruppen und kurzfristige Untersuchungen braucht, als für summative *Evaluation* notwendig sind. Das ist unbestreitbar, aber beweist nicht, daß man für die umfassende Evaluation einen umfassenden Vergleich vermeiden sollte.

Man kann den entscheidenden Punkt in Form folgender Analogie fassen: In der Geschichte der Konstruktion von Automotoren ist es häufiger vorgekommen, daß ein Konstrukteur einen Motor entwarf, der der Konkurrenz ganz überraschend überlegen war. Vielleicht hatte man etwa 30 Variablen bei Konstruktion des neuen Motors verändert; nachdem der Motor in die Produktion gegangen war, wußte man noch lange nicht, der Konstrukteur eingeschlossen, welche von diesen Variablen vor allem für die Verbesserung verantwortlich war. Aber die Entscheidung, den Motor in die Produktion zu geben, die Entscheidung, weitere Forschung in den Motor zu investieren, machte es möglich, den Grund des Erfolgs herauszufinden. Tatsächlich war für den Beginn einer neuen Ära der Konstruktion von Motoren wirklich vor allem die *vergleichende* Evaluation notwendig. Man setzt ein großes Team ein und hofft, daß es Gold entdecken wird, muß jedoch auf das Metall stoßen, bevor das Kapital investiert wird, das man benötigt, um die Lage von Flözen genau festzustellen. So müssen wir in jedem Bereich arbeiten, wo wir zu viele Variablen und zu wenig Zeit haben.

Praktische Verfahrensweisen für die Evaluation mit Kontrollgruppen

In einer seiner Hauptthesen behauptet Cronbach, daß Vergleiche mit Kontrollgruppen für die Curriculumentwicklung nicht sehr nützlich sind. Nach unseren Ausführungen bietet sein Versuch, eine brauchbare Alternative bereitzustellen, im Kontext einer typischen Evaluation keine heuristische Möglichkeit. Man muß deshalb einige seiner Einwände gegen Kontrollgruppen zu entkräften versuchen, die sich unserer Meinung nach für die Evaluation durchaus eignen.

Der These, daß grobe Vergleiche nur geringe Unterschiede erbringen, kann man zunächst dadurch begegnen, daß man die Präzision der Untersuchungsmethoden verbessert. Das bedeutet die Entwicklung einer grö-

ßeren Zahl verschiedenartiger Testaufgaben, die Vergrößerung der Gruppe, um differenziertere Unterschiede zu gewinnen, und die Entwicklung neuerer und besserer Tests. Wenn wir einen besonders relevanten Faktor entdecken, versuchen wir, diesen bei der Neukonstruktion des Curriculum stärker zu berücksichtigen, um seinen Erfolg zu vergrößern. Doch bleibt die Tatsache bestehen, daß man die Verbesserung des Curriculum wahrscheinlich nur in geringen Testunterschieden messen kann und daß das Streben nach größeren Unterschieden im allgemeinen eine Methode mit mehreren Angriffspunkten erfordert; daher muß sie nicht nur auf das Curriculum zielen, sondern muß auch die Verfahren der Gruppeneinteilung, die Stoffdarbietung des Lehrers, die Verteilung der Unterrichtszeit berücksichtigen. Darüber hinaus muß sie versuchen, Langzeiteffekte, z. B. ein allgemeines Anwachsen des Interesses, zu erforschen, die von Verbesserungen in jedem Bereich des Schulcurriculum bewirkt werden. Darin liegt eine wichtige Aufgabe der Evaluation. Eine genaue Parallele finden wir im Bereich der Psychotherapie, in der wir menschliches Verhalten dadurch zu ändern versuchen, daß wir Menschen mehrere Jahre lang wöchentlich für einige Stunden behandeln. Vielleicht sind wir allzusehr an die Erfindung von Wunderdrogen oder an technologische Durchbrüche im Bereich der Raumfahrt gewöhnt, daß wir diesen plötzlichen Fortschritt nicht mehr als ungewöhnliche Ausnahme begreifen. Selbst im Bereich der Konstruktion von Autos, um bei Cronbachs Beispiel zu bleiben, deuten zahlreiche Erfahrungen darauf hin, daß die Weiterentwicklung eines bereits erprobten Modells bessere Resultate erbringt als die Einführung eines vielversprechenden, aber radikal neuen Modells. Realistisch gesehen, kann man keine großen Sprünge, sondern nur eine langsame und stetige Verbesserung als Ergebnis erwarten, wobei sich natürlich manchmal Sackgassen nicht vermeiden lassen. Das systematisch geplante Experiment, das ein Curriculum einem anderen gegenüberstellt, sollte genügend eindeutige Ergebnisse haben, um die zu ihrer Gewinnung erforderlichen Kosten zu rechtfertigen. Cronbach berücksichtigt jedoch dabei nicht genügend, daß das Ausbleiben klarer Unterschiede oft gerade das Ergebnis ist, das man benötigt. Wenn man sich wirklich davon überzeugt hat, daß man mit guten Tests die wichtigsten Kriteriumsvariablen erfaßt, dann *ist* es äußerst informativ, zu sehen, daß die Ergebnisse gleich sind, denn »kein Unterschied« bedeutet keineswegs »kein Wissen«.

Natürlich können wir aus einem Null-Resultat nicht schließen, daß alle in einem neuen Curriculum enthaltenen Verfahren wertlos sind. Wir müssen mit Mikro-Untersuchungen fortfahren, aus denen wir entnehmen können, ob eine dieser Techniken etwas wert ist. Die Durchführung einer groben vergleichenden Untersuchung kostet stets dasselbe, unabhängig davon,

welche Resultate sie erbringt; hinzu kommt, daß man sie früher oder später doch durchführen muß. Es ist also falsch aufzuhören, wenn man nichtsignifikante Unterschiede entdeckt hat; vielmehr muß man weitere analytische Forschung der von Cronbach empfohlenen Art betreiben. Wenn auch Cronbach in seinem Beitrag den Untersuchungen mit Kontrollgruppen eine geringe Bedeutung zumißt, kann man diese jedoch lediglich dann als unangemessen bezeichnen, wenn man sie als *einzige* Evaluationsmethode für den *gesamten* curricularen Bereich ansieht. Wir wollen hier versuchen, einige praktische Vorschläge für die Anlage von Versuchen zu machen, die mehr als eine grob vergleichende Evaluation ergeben.

Ein wichtiger Grund für Cronbachs Ablehnung von vergleichenden Untersuchungen liegt in der Überzeugung, daß man keine Doppelblindversuche durchführen könne. »In einem pädagogischen Versuch ist es schwer, die Schüler über ihre Rolle als Versuchsgruppe im unklaren zu lassen. Die Fehlerquellen, die durch die Person des Lehrers bedingt sind, können kaum so gut kontrolliert werden wie die durch den Arzt im Doppelblindversuch bedingten. Infolgedessen kann man nicht mit Sicherheit sagen, ob ein beobachteter Gewinn der pädagogischen Innovation an sich zuzuschreiben ist oder dem größeren Engagement von Lehrer und Schülern bei einem Versuch mit einer neuen Methode« (Cronbach 1963, 42, *43*). Cronbachs Schlußfolgerungen sind jedoch übereilt. Im medizinischen Bereich bilden nicht die Untersuchungen über Medikamente die Analogie, bei denen wir Doppelblindbedingungen ohne weiteres herstellen können, sondern psychotherapeutische Untersuchungen, bei denen der Therapeut die Behandlung engagiert durchführt und der Patient nicht darüber in Ungewißheit gelassen werden kann, daß er behandelt wird. Wenn Cronbachs Argumentation richtig ist, wäre es nicht möglich, eine adäquate Untersuchung der Ergebnisse einer psychotherapeutischen Behandlung zu planen. Das *ist* jedoch möglich, und die entsprechende Methode besteht darin, mehrere Vergleichsgruppen zu benutzen (vgl. Scriven 1959). Wenn wir nur eine Kontrollgruppe benutzen, können wir keine Aussage darüber machen, ob das Engagement oder die Versuchstechnik den Unterschied erklärt. Wenn wir aber mehrere Versuchsgruppen haben, können wir die Auswirkungen des Engagements einschätzen. Wir vergleichen mehrere Therapiegruppen, in denen der Therapeut jeweils engagiert ist, in denen aber die Therapiemethode jeweils verschieden ist. Nach Möglichkeit sollte man Therapiemethoden mit sehr verschiedenen Verfahrensweisen anwenden. Die Patienten, die nach *einer* Methode behandelt werden, sollten jedoch soweit wie möglich vergleichbar sein. Es gibt eine Anzahl von Therapien, die die erste Bedingung in mehreren Dimensionen erfüllen, und es ist leicht, Pseudo-Therapien zu entwickeln, die vielversprechend genug sind, um bei eini-

gen praktizierenden Ärzten Engagement zu wecken. Die Feststellung der Unterschiede in Verbindung mit der Kovarianzanalyse ermöglicht zu entscheiden, ob Engagement der einzige oder ein Hauptfaktor beim therapeutischen Erfolg ist, wenn Doppelblindbedingungen nicht erreicht werden können. Dies ist auch nicht der einzige Forschungsplan, mit dem das erreicht werden kann; andere Methoden sind verfügbar, und fähige Wissenschaftler werden zweifellos noch weitere Methoden entwickeln können, die es uns ermöglichen, dieses Forschungsproblem zu bewältigen. Eine Doppelblinduntersuchung ist also nicht unbedingt notwendig.

Im curricularen Bereich sind die Fragen noch etwas schwieriger als im Bereich der Psychotherapie, weil es sehr schwer ist, gemeinsame Elemente aus den verschiedenen Vergleichsgruppen auszusondern. Zwar wird der durchschnittlich intelligente Patient aufgrund der vielen unzulänglichen Ärzte und aufgrund des Wunsches, geheilt zu werden, fast alles als Form von Therapie akzeptieren, doch ist es längst nicht so einfach, Schüler und Lehrer davon zu überzeugen, daß sie einen bestimmten Unterricht in Geometrie bekommen oder geben sollen, es sei denn, die Art der Geometrie erscheint ihnen sinnvoll. Und wenn es sich so verhält, dann ist die Interpretation jedes der möglichen Ergebnisse mehrdeutig, d. h., wenn mehrere Gruppen ungefähr gleich gut abschneiden, kann es *entweder* sein, daß das Engagement sich so auswirkt, *oder*, daß die gemeinsamen Inhalte effizient sind. Trotzdem ist vergleichende Evaluation lohnend; denn wenn wir einen *deutlichen* Unterschied zwischen den Gruppen feststellen und Engagement bei Lehrern und Schülern in beiden Fällen vorhanden ist, können wir einigermaßen sicher sein, daß der Unterschied auf die Curriculuminhalte zurückzuführen ist. Die Sequenz der Darbietung, die Methoden, die Schwierigkeiten, die Beispiele usw. können sicherlich genügend variiert werden, so daß nicht unterscheidbare Resultate unwahrscheinlich sind.

Es ist nicht sehr schwierig, entsprechendes Engagement in den Gruppen zu erreichen. In Analogie zu den scharf kalkulierten Vergleichsgruppen der »neuen Therapie«, bei denen die Therapieverfahren in ein oder zwei Tagen freier Assoziation durch ein »Brainstorming« der Wissenschaftler entstehen, konstruieren wir auf folgende Weise einige »neue Curricula«:

Zuerst holen wir uns zwei intelligente Studenten höherer Semester, z. B. aus den Wirtschaftswissenschaften, geben ihnen eine Liste mit wirtschaftswissenschaftlichen Fachausdrücken für die zehnte Klasse und zahlen ihnen 500 Dollar für die Übersetzung eines Kapitels aus Samuelsons Wirtschaftslehre in die Sprache der zehnten Klasse. Wir ermutigen sie, ihre Originalität zu zeigen und neue Ideen zu entwickeln. Sie können den Text wahrscheinlich in einem Sommer bearbeiten; so haben wir für einige tausend Dollar, einschließlich der Kosten für die Reproduktion des Versuchs-

materials, ein Curriculum, das wir einem der teuren wirtschaftswissenschaftlichen Curricula entgegensetzen können, die mit großer finanzieller Unterstützung auf kostspieligen Felduntersuchungen aufbauen. Dann suchen wir einige intelligente jüngere Studenten aus verschiedenen Hochschulen, die Wirtschaftswissenschaften studieren. Sie haben zu diesem Zeitpunkt Erfahrungen beim Absolvieren von Einführungskursen in Wirtschaftwissenschaften gesammelt und ein Problembewußtsein in bezug auf das Begriffsverständnis in diesem Bereich erworben. Wir geben ihnen einen Sommer Zeit für die Entwicklung eines Curriculum zur Einführung in die Wirtschaftswissenschaften für die zehnte Klasse, das nicht einen bestimmten Text in den Mittelpunkt rücken soll.

Für eine dritte Vergleichgruppe suchen wir einige Lehrer aus, die eine hohe Meinung von einem in den Sekundarschulen verwendeten Text für »Wirtschaftswissenschaften« haben. Dann lassen wir sie zusammen mit den Autoren eine Revision erarbeiten und dabei einige Beispiele der Reaktionen ihrer Kollegen auf den im Unterricht benutzten Text berücksichtigen. Da wir vor allem Curriculumentwickler unterrichten lassen, setzen wir sie in nur grob parallelisierten Vergleichsgruppen in Schulsystemen ein, die geographisch weit von denen entfernt sind, in denen wir die teuren Curricula testen. Wir setzen, um ihre Initiative zu wecken, eine vorher angekündigte Geldprämie für diese Gruppe aus, wenn sie nicht von dem großen Curriculum signifikant übertroffen wird. Wenn wir *trotzdem* einen erheblichen Unterschied zugunsten des großen Curriculum bekommen, können wir zu Recht annehmen, daß wir die Engagement-Variable beachtet haben. Darüber hinaus brauchen wir dies nicht für jedes Fach durchzuführen, da Engagement unabhängig vom Fachbereich in seinen Auswirkungen ziemlich konstant ist. Auf jeden Fall sollte eine kleinere Stichprobe genügen, um dies zu überprüfen.

Diese Art der vergleichenden Untersuchung hat einen besonderen Vorzug. Selbst wenn wir nur unbedeutende Unterschiede und damit ein mehrdeutiges Resultat erhalten, das uns darüber in Zweifel geraten läßt, ob ein allgemeines Engagement dafür verantwortlich ist oder ob alle Curricula der Wirtschaftswissenschaften ungefähr gleichen Unterricht bewirken, ersparen wir uns große Unkosten. Wenn wir mit wenig Kapital neue Curricula entwickeln können, die gute Ergebnisse erbringen, so ist das um so besser. Wir können das häufiger praktizieren und uns dadurch die Unterstützung engagierter Projektleiter erhalten und die Chancen vergrößern, einen Newton der Curriculumreform zu finden, der einen grundlegend neuen Ansatz entdeckt.

Weiterhin können wir, ohne daß neue Kosten entstehen, selbst im Fall einer Verbindung zwischen den verschiedenen Curricula die Engagement-

Frage recht schnell lösen, indem wir die Curricula einigen *negativ* und einigen *neutral* eingestellten Lehrern für den Unterricht während des nächsten Jahres oder der nächsten zwei Jahre geben. Andererseits bilden die von Anfang an beteiligten Curriculumentwickler eine Gruppe gut ausgesuchter, innovationswilliger Lehrer sorgfältig für die gleiche Arbeit aus. Vergleiche zwischen der Leistung dieser drei neuen Gruppen und der Leistung der alten sollten es uns ermöglichen, die Rolle des Engagements und zusätzlich die Unabhängigkeit der verschiedenen Curricula gegenüber fehlendem Engagement, die unzweifelhaft eine Variable darstellt, recht genau zu erfassen.

Offensichtlich müssen einige der oben dargestellten Verfahrensweisen in einer tatsächlich durchgeführten Untersuchung erweitert werden, z. B. die Möglichkeit für die neuen Curriculumentwickler, einige Nachmittage auf die Felduntersuchung der ersten Abschnitte ihres neuen Curriculum zu verwenden, um ihnen ein »Gefühl« für die Schnelligkeit zu vermitteln, mit der Schüler dieser Altersstufe die neuen Begriffe aufnehmen können, und um die Lehrer in bezug auf ihre konservative Einstellung, ihre Abneigung oder Lethargie mit Hilfe von Selbsteinschätzung und Einschätzung von Kollegen in Verbindung mit Einstellungsskalen sorgfältig auszusuchen.

Die »Schwierigkeit« mit der Engagement-Variablen ist ein Beispiel für die Auswirkungen der Versuchssituation. Andere Beispiele sind der Placebo-Effekt in der Medizin und der Hawthorne-Effekt in der Betriebs- und Sozialpsychologie. In jedem dieser Fälle sind wir daran interessiert, die Wirkungen eines bestimmten Faktors festzustellen, aber wir können den Faktor nicht in die experimentelle Situation einführen, ohne eine Störung hervorzurufen, die ihrerseits für die beobachteten Veränderungen verantwortlich sein kann. Im medizinischen Bereich besteht die Störung darin, dem Patienten etwas zu geben, was er für ein Medikament hält. Weil das für ihn kein gewöhnlicher Vorgang ist, kann dieser, ganz abgesehen von den intrinsischen Wirkungen des Medikaments, eigene Wirkungen hervorrufen. Beim Hawthorne-Effekt besteht die Störung beispielsweise in der Änderung von Arbeitsbedingungen, die den Arbeiter möglicherweise vermuten läßt, daß er Gegenstand einer speziellen Untersuchung und eines speziellen Interesses ist, und *dies* mag mehr zu einem verbesserten Arbeitsergebnis führen als die physikalischen Änderungen in der Umgebung, die die zu untersuchenden Kontrollvariablen darstellen. Die bisher erwähnten Fälle sind solche, bei denen die Überzeugung der Versuchspersonen der intervenierende Faktor zwischen Störung und mehrdeutiger Wirkung ist. Dies ist im Bereich der Psychologie charakteristisch, aber die Situation ist nicht grundlegend verschieden von der, die in der naturwissenschaftlichen Forschung auftaucht. Dort treffen wir auf Probleme wie die Absorp-

tion von Hitze durch ein Thermometer, wodurch die Temperatur, die gemessen werden soll, geändert wird. Das heißt, einige der beobachteten Wirkungen sind auf die Tatsache zurückzuführen, daß man das, was man messen will, ändern muß, um überhaupt eine Messung zu erhalten. Der Meßprozeß bringt ein anderes physikalisches Objekt in die Nähe des gemessenen Objektes. Das Instrument selbst hat eine bestimmte Wärmekapazität – ein Faktor, an dessen Einfluß man nicht interessiert ist. Dennoch muß man seine Größe abschätzen, um das Gewünschte herausfinden zu können. Das optimale Doppelblindmodell ist nur bei bestimmten Gegebenheiten angemessen, und es ist nur eine von vielen Methoden, mit denen wir diese Wirkungen umgehen können. Cronbachs Annahme, daß die Unmöglichkeit einer Doppelblinduntersuchung in der Curriculumentwicklung vergleichende Evaluation nicht zuläßt, erscheint deshalb zu pessimistisch. Tatsächlich stimmt er der Wichtigkeit vergleichender Arbeit zu, soweit er Längsschnittuntersuchungen erörtert.

Die Schlußfolgerung erscheint zwingend, daß vergleichende Evaluation die richtige Methode für die Behandlung der Probleme der Evaluation ist.

Robert E. Stake

Verschiedene Aspekte pädagogischer Evaluation

Präsident Johnson, Präsident Conant, Mrs. Hull (Saras Lehrer) und Herr Tykoziner (der Mann nebenan) ähneln sich in ihrem Vertrauen auf Erziehung. Aber sie haben recht unterschiedliche Ideen darüber, was Erziehung ist. Der Wert, den sie der Erziehung beimessen, gibt keine Aufschlüsse über die Art, wie sie Erziehung bewerten. Genauso unterschiedliche Auffassungen haben Pädagogen über den Inhalt und Wert eines Bildungsprogramms. Die vielen Möglichkeiten in den Zielen und Methoden der Evaluation erlauben es jedem, seine eigene Perspektive zu behalten. Da viele Pädagogen ein zu begrenztes Verständnis von Evaluation haben, sehen nur wenige ihre eigenen Programme in voller Komplexität. Um den eigenen Unterricht besser zu verstehen und zur Verbesserung der Wissenschaft vom Unterricht beizutragen, sollte jeder Pädagoge sich sämtliche Möglichkeiten der Evaluation vergegenwärtigen. Pädagogische Evaluation hat ihre formalen und informalen Seiten. Informale Evaluation ist durch gelegentliche Beobachtungen, implizite Ziele, intuitive Normen und subjektive Urteile gekennzeichnet. Weil diese vielleicht auch im täglichen Leben üblich sind, kommt informale Evaluation zu Ergebnissen, die selten in Frage gestellt werden. Sorgfältige Untersuchungen zeigen, daß informale pädagogische Evaluation von unterschiedlicher Qualität ist; manchmal ist sie scharfsinnig und einsichtsvoll, manchmal oberflächlich und irreführend.

Formale pädagogische Evaluation ist durch Strichlisten, strukturierte Unterrichtsbeobachtungen, Vergleiche mit Kontrollgruppen und Untersuchungen von Schülern mit standardisierten Tests gekennzeichnet. Einige dieser Verfahren haben sich seit langem bewährt. Bei der Planung von Evaluation denken leider nur wenige Pädagogen an diese vier Verfahren. Es ist viel gebräuchlicher, informal zu evaluieren: den Lehrer nach seiner Meinung zu fragen, über die Logik des Programms nachzudenken oder das Ansehen seiner Befürworter in Betracht zu ziehen. Selten sucht man nach relevanten Forschungsberichten oder Verhaltensdaten, die das Ergebnis curricularer Entscheidungen sind.

Die Unzufriedenheit mit dem formalen Ansatz hat gute Gründe. Es gibt wenige wirklich relevante, lesbare Forschungsberichte. Die pädagogischen Zeitschriften sind nicht bereit, Evaluationsuntersuchungen zu veröffentlichen. Verhaltensdaten zu gewinnen, ist kostspielig; auch sie geben oft nicht die gesuchten Antworten. Zu vielen Pädagogen, die Unterrichtsbesichtigungen machen, fehlt eine entsprechende Schulung oder Erfahrung in Evaluation. Viele Strichlisten sind ungenau; einige betonen die äußeren Bedingungen einer Schule zu sehr. Psychometrische Tests werden eher dazu entwickelt, zwischen Schülern mit etwa gleicher Bildung zu differenzieren als die Auswirkungen des Unterrichts auf den Erwerb von Fertigkeiten und Verständnis zu erfassen. Ein moderner Pädagoge kann sich wenig auf formale Evaluation verlassen, weil ihre Ergebnisse selten die von ihm gestellten Fragen beantworten.

Der mögliche Beitrag formaler Evaluation

Die Skepsis des Pädagogen gegenüber formaler Evaluation resultiert zum Teil auch aus seiner Empfindlichkeit gegen Kritik. Häufig versteckt er sich hinter Begriffen wie »Innovationsphase« und »akademische Freiheit«, um Evaluation zu vermeiden. Die »Politik« der Evaluation ist ein interessantes Problem, das in diesem Zusammenhang jedoch nicht erörtert werden soll. Das Thema unserer Ausführungen ist der *mögliche* Beitrag formaler Evaluation zur Erziehung. Pädagogen sehen heute kaum, welche Hilfe formale Evaluation ihnen leisten könnte. Sie sollten Testkonstrukteure bitten, eine Methodologie zu entwickeln, die den Reichtum, die Komplexität und die Wichtigkeit ihrer Programme berücksichtigt. Das geschieht jedoch bisher noch nicht.

Wenn man die gegenwärtigen Bemühungen um formale Evaluation in der Pädagogik untersucht, findet man geringe Anstrengungen, die Voraussetzungen (antecedent conditions) und die Unterrichtsprozesse (transactions) – einige werden von Beobachter-Teams aufgezeichnet – zu erforschen; auch findet man zu wenig Versuche, sie mit den verschiedenen Ergebnissen – einige werden in konventionellen Testwerten ausgedrückt – in Verbindung zu bringen. Man hat selten versucht, das Verhältnis zwischen dem, was ein Pädagoge zu tun beabsichtigt, und dem, was er wirklich tut, zu erfassen. Das traditionelle Bemühen der Testkonstrukteure um die Reliabilität der Punktwerte individueller Schüler und die Voraussage-Validität (vgl. Lindquist 1951) ist ein zweifelhaftes Mittel. Bei der Evaluation von Curricula sollte man, anstatt die individuellen Unterschiede zwischen den Schülern zu betonen, besser die Kontingenzen zwischen den

Voraussetzungen, den Unterrichtsaktivitäten und den schulischen Ergebnissen beachten.

In diesem Beitrag soll nicht erörtert werden, was oder wie man messen sollte; es soll ein Hintergrund für die Entwicklung eines Evaluationsplans gegeben werden. Was und wie evaluiert wird, muß später entschieden werden. Mir geht es hier eher um Bildungsprogramme als um die Ergebnisse der Bildung. Ich setze voraus, daß der Wert eines Bildungsergebnisses auf dem verwendeten Programm beruht. Die Evaluation eines Programms schließt die Evaluation seiner Materialien ein.

Dieses Verständnis von pädagogischer Evaluation scheint sich zu verändern. Auf den folgenden Seiten will ich zeigen, welches Verständnis von Evaluation sich m. E. empfiehlt. Ich werde eine Konzeptualisierung der Evaluation zu entwickeln versuchen, die sich an dem komplexen und dynamischen Charakter der Erziehung orientiert und die die verschiedenen Zielsetzungen und Urteile des Praktikers angemessen berücksichtigt.

Ein großer Teil des neuerlichen Interesses an Curriculumevaluation liegt in den gegenwärtig umfangreichen Bemühungen um Curriculuminnovation begründet; aber die Ausführungen in diesem Beitrag gelten für herkömmliche und neue Curricula gleichermaßen. Sie betreffen z. B. Titel I- und Titel III-Projekte, die im Rahmen des Elementary and Secondary Education Act von 1965 finanziert worden sind. Die Erörterungen sind für alle Curricula relevant, unabhängig davon, ob sie sich mehr an den fachspezifischen Inhalten oder mehr an den Interessen der Schüler orientieren. Dabei ist es gleichgültig, ob das Curriculum allgemeine Zielsetzungen oder spezielle Förderungsaufgaben hat.

Ziele und Verfahren pädagogischer Evaluation sind von Fall zu Fall verschieden. Was für eine Schule angemessen ist, mag für eine andere weniger zweckmäßig sein. Manchmal empfehlen sich standardisierte Leistungstests und manchmal nicht. In einem Fall stehen geringe, im anderen Fall umfangreiche finanzielle Mittel zur Verfügung. Wie unterscheiden sich Ziele und Verfahren der Evaluation? Was sind die grundlegenden Charakteristika der Evaluation? In den folgenden Ausführungen werden sie als Evaluationshandlungen, Datenquellen, Kongruenz und Kontingenzen, Normen und Verwendungsarten der Evaluation identifiziert. Doch zuerst soll zwischen Beschreibung und Beurteilung in der Evaluation unterschieden werden.

Die Erwartung, die der Pädagoge an die Evaluation stellt, ist nicht gleich der Auffassung des Evaluators. Dieser erblickt seine Aufgabe in der Beschreibung von Einstellungen, Umwelt und Leistungen. Der Lehrer und der Beamte der Schulverwaltung jedoch erwarten vom Evaluator, daß er etwas oder jemanden nach seiner Leistung bewertet. Sodann erwarten

sie, daß er Dinge nach äußeren Normen beurteilt, vielleicht mit Kriterien, die nur eine geringe Beziehung zu den Mitteln und Zielen der örtlichen Schule haben.

Keiner begreift Evaluation umfassend genug. Beschreibung und Beurteilung sind erforderlich; sie sind in der Tat die beiden grundlegenden Evaluationshandlungen. Ein Evaluator kann versuchen, sich des Urteils oder der Sammlung von Urteilen anderer Personen zu enthalten. Ein anderer Evaluator kann ausschließlich darauf bedacht sein, den Wert des Programms deutlich zu machen. Aber die Evaluation beider ist unvollständig. Um vollständig verstanden zu werden, muß das Erziehungsprogramm vollständig beschrieben und beurteilt werden.

Auf dem Weg zu einer vollständigen Beschreibung

Der Evaluator scheint in zunehmendem Maße die Wichtigkeit einer vollständigen Beschreibung zu betonen. Seit vielen Jahren evaluiert er vor allem dadurch, daß er feststellt, inwieweit Lernziele von Schülern erreicht worden sind. Diese Lernziele wurden gewöhnlich mit den traditionellen Disziplinen, z. B. Mathematik, Englisch und Politischer Bildung (social studies) gleichgesetzt. Standardisierte oder vom Lehrer hergestellte Leistungstests hielt man für nützlich, um das Ausmaß zu beschreiben, in dem einige curriculare Lernziele von einzelnen Schülern in einem speziellen Kurs erreicht wurden. In diesem frühen Stadium war Evaluation für viele Evaluatoren und Pädagogen nichts anderes als der Einsatz und die normative Interpretation von Leistungstests.

In den letzten Jahren haben darüber hinaus einige Evaluatoren abzuschätzen versucht, inwieweit einzelne Schüler bestimmte interdisziplinäre und extracurriculare Ziele erreicht haben. Dabei bestand ihr Ziel vor allem darin, die Integration von Verhaltensweisen in Individuen festzustellen, das Verständnis der Beziehungen zwischen den wissenschaftlichen Disziplinen zu erfassen und die Entwicklung von Haltungen, Fertigkeiten und Einstellungen zu untersuchen, die ein Individuum dazu befähigen, ein Handwerker oder Wissenschaftler zu sein. Für die beschreibende Evaluation solcher Ergebnisse hat die Eight-Year-Study (Smith/Tyler 1942) als Modell gedient. Das National Assessment Program wird vielleicht – wie aus den in einem Zwischenbericht erschienenen folgenden Ausführungen hervorgeht – zu einem anderen Modell werden: »... Alle Kommissionen arbeiteten innerhalb der folgenden allgemeinen Definition des National Assessment.

1. Um in etwa die Ziele der Erziehung in den USA zu reflektieren, sollte

das National Assessment traditionelle und moderne Curricula ins Auge fassen, alle Zielsetzungen berücksichtigen, die die Schulen für die Entwicklung von Einstellungen, Motivation, Kenntnissen und Fertigkeiten haben« (Educational Testing Service 1965).

In seinem Beitrag *Evaluation zur Verbesserung von Curricula* forderte Lee Cronbach (1963) eine möglichst umfangreiche Einbeziehung verhaltenswissenschaftlicher Variablen, um die Ursachen und Auswirkungen von gutem Unterricht zu untersuchen. Nach seinem Vorschlag liegt das Hauptziel der Evaluation darin, solche dauerhaften Beziehungen zu entdecken, die für die Entwicklung zukünftiger Bildungsprogramme relevant sind. Die traditionelle Beschreibung der Schülerleistung ergänzen wir durch die Beschreibung des Unterrichts und die Beschreibung der Beziehungen zwischen ihnen. Wie der Bildungsforscher versucht nach unserer Auffassung auch der Evaluator, Generalisationen über pädagogische Praktiken zu entwickeln. Viele Evaluatoren von Curriculumprojekten haben sich diese Definition von Evaluation zu eigen gemacht.

Die Rolle des Urteils

Beschreibung ist eine Sache, Beurteilung eine andere. Die meisten Evaluatoren haben sich entschlossen, keine Urteile abzugeben. In seiner kürzlich erschienenen *Methodologie der Evaluation* hat Michael Scriven (1967) jedoch den Evaluatoren die Aufgabe zugeschrieben, Urteile über den Wert einer pädagogischen Handlung zu fällen. Er hat vom Evaluator verlangt, die Erwartungen zu erfüllen, die Pädagogen an ihn stellen. Nach Scrivens Ansicht findet Evaluation nicht statt, bevor nicht Beurteilung erfolgt. Wenn der Evaluator sich dessen bewußt ist, ist er am besten zur Abgabe von Urteilen qualifiziert.

Aufgrund seiner zahlreichen Erfahrungen und Kenntnisse in diesem Bereich der Forschung und pädagogischen Praxis ist der Evaluator wenigstens teilweise dazu befähigt, Urteile abzugeben. Aber soll er wirklich diese Aufgabe übernehmen? Selbst zur Zeit, da wenige Evaluatoren bereit sind, Urteile zu fällen, wehren Pädagogen sich gegen formale Evaluation. Wenn man die Funktion der Evaluatoren *häufiger* mit der Aufgabe identifiziert, Urteile über den Unterschied zwischen schlechteren und besseren Programmen, über die Gewährung von Unterstützung und über die Formulierung von Kritik abzugeben, würde sich der Zugang der Evaluatoren zu Daten wahrscheinlich erschweren. Evaluatoren arbeiten mit anderen Sozialwissenschaftlern und Verhaltensforschern zusammen. Alle Forscher, die keine Urteile fällen wollen, bedauern die Übernahme dieser Aufgabe durch ihre

Kollegen. Sie sind der Überzeugung, daß viele Praktiker noch mehr Einwände als bisher gegen die Sozialwissenschaften und die verhaltenswissenschaftliche Forschung erheben würden.

Viele Evaluatoren glauben, sie seien nicht in der Lage – wozu ihrer Meinung nach ein Beurteiler fähig sein sollte –, den eindimensionalen *Wert* alternativer Programme wahrzunehmen. Sie erwarten z. B. folgendes Dilemma: Curriculum I hat als Ergebnis drei Fertigkeiten und zehn Erkenntnisse, Curriculum II vier Fertigkeiten und acht Erkenntnisse. Sie scheuen sich, zu beurteilen, ob der Gewinn einer Fertigkeit den Verlust von zwei Erkenntnissen wert ist. So bestärkt der Evaluator, sei es aus Ängstlichkeit, Desinteresse oder aufgrund rationaler Entscheidung, häufig die Entscheidungen der Gemeinden, ihr Recht, eigene Normen aufzustellen und den Wert ihres Bildungssystems selbst zu beurteilen. Er setzt voraus, daß das, was für eine Gemeinde gut ist, auch für eine andere gut sein muß; er traut sich nicht zu, eine Entscheidung darüber zu fällen, was für eine ihm erst seit geraumer Zeit bekannte Gemeinde am besten ist.

Scriven macht darauf aufmerksam, daß gegenwärtig wenige und in Zukunft noch weniger Evaluatoren komplexe Curricula beurteilen können. Verschiedene Entscheidungen müssen getroffen werden, z. B. ob das Physical Science Study Committee Program oder das Harvard Physical Program unterrichtet werden soll. Sie sollen jedoch nicht aufgrund trivialer Kriterien – beispielsweise Erwähnung in der Presse, Persönlichkeit des Vertreters des Projekts, administrative Bequemlichkeit oder pädagogischer Mythos – getroffen werden. Wer soll Urteile fällen? Scriven findet die Antwort z. T. so leicht, weil er zwischen Schüler und Curriculum wenige Interaktionen erwartet. D. h.: er geht davon aus, daß das, was für einen Schüler – wenigstens in groben Umrissen – am besten ist, auch für andere am besten sein muß. Er setzt ferner voraus, daß, wenn die Interessen einer Gemeinde sich nicht mit denen der Gesamtgesellschaft decken, erstere denen der Gesamtgesellschaft abträglich sind, so daß daher das freie Entscheidungsrecht eingeschränkt werden muß. – Nach Scriven muß der Evaluator Urteile fällen.

Ob die Evaluatoren Scrivens Aufforderung berücksichtigen oder nicht, bleibt abzuwarten. Wahrscheinlich werden jedoch Beurteilungen einen zunehmend größeren Teil des Evaluationsberichts ausmachen. Die Evaluatoren werden sich darum bemühen, die Ansichten von qualifizierten Personen aufzuzeichnen. Obwohl diese Auffassungen subjektiv sind, können sie sehr nützlich sein und objektiv gesammelt werden, d. h. unabhängig von denen, die diese Ansicht vertreten. Der Evaluator kann sich eher der Aufgabe unterziehen, Urteile bei der Evaluation zu verwerten als sie selber abzugeben.

Taylor und Maguire (1966) haben fünf Gruppen genannt, deren Ansichten über Erziehung wichtig sind: Sprecher der Gesamtgesellschaft, Fachwissenschaftler, Lehrer, Eltern und Schüler. Die Urteile der Vertreter dieser und anderer Gruppen sollten gehört werden. Oberflächliche Umfragen, Briefe an den Herausgeber und andere beiläufig geäußerte Urteile sind unzureichend. Die Evaluation sollte den Wert und die Unzulänglichkeiten eines Schulprogramms nach dem Urteil gut informierter Gruppen mit Hilfe systematisch gesammelter und verarbeiteter Daten deutlich machen. D. h. also: Urteilsdaten und Beschreibungsdaten sind gleichermaßen für die Evaluation von Bildungsprogrammen erforderlich.

Datenmatrizen

Um evaluieren zu können, muß ein Pädagoge bestimmte Daten sammeln. Sie werden wahrscheinlich in zahlreichen heterogenen Bereichen mit mehreren verschiedenartigen Mitteln gewonnen. Unabhängig davon, ob das unmittelbare Ziel Beschreibung oder Beurteilung ist, sollten Informationen in drei Bereichen gesammelt werden. Im Evaluationsbericht empfiehlt es sich, zwischen *Voraussetzungs*daten, *Prozeß*daten und *Ergebnis*daten zu unterscheiden.

Eine Voraussetzung ist jede Bedingung, die vor dem Unterrichten und Lernen besteht und die Einfluß auf die Ergebnisse haben kann. Die Ausgangssituation eines Schülers vor dem Unterricht, z. B. seine Fähigkeit, vorherige Erfahrung, Interesse und Bereitschaft, ist eine komplexe Voraussetzung. Beim Programmierten Unterricht nennt man einige Voraussetzungen »Eingangsverhalten«. Die »akkreditierende Institution des Staates« wiederum richtet ihre besondere Aufmerksamkeit auf die Investition der Ressourcen der Gemeinde. Dies sind Beispiele für die Voraussetzungen, die ein Evaluator beschreiben kann.

Prozesse sind die zahllosen Begegnungen zwischen Schülern und Lehrern, Schülern und Schülern, Autoren und Lesern, Eltern und Schulpsychologen – das Aufeinanderfolgen von pädagogischen Handlungen, das den Prozeß der Erziehung ausmacht. Beispiele sind die Vorführung eines Films, eine Unterrichtsdiskussion, die Lösung einer Hausaufgabe, eine Erklärung am Rand einer Prüfungsarbeit und der Einsatz eines Tests. Smith und Meux (1962) haben solche Prozesse genau untersucht und dazu ein 18 Kategorien umfassendes Klassifikationssystem aufgestellt. Auf eine bestimmte Art von Prozessen wurde durch die Förderung der Entwicklung audiovisueller Medien im Rahmen des National Defense Education Act besonderer Wert gelegt.

Während Voraussetzungen und Ergebnisse relativ statisch sind, sind Prozesse dynamisch. Die Grenzen zwischen den Bereichen sind nicht deutlich. Wir können z. B. während eines Prozesses bestimmte Ergebnisse identifizieren, die als Rückmeldung Voraussetzung für nachfolgendes Lernen sind. Die Abgrenzung zwischen den Bereichen braucht nicht exakt zu erfolgen. Die Kategorien sollen eher dazu dienen, eine umfangreiche Sammlung von Daten anzuregen, als sie in Gruppen zu unterteilen.

In der Vergangenheit konzentrierte man sich bei formaler Evaluation vorwiegend auf Ergebnisse wie Fähigkeiten, Leistungen, Einstellungen und Erwartungen der Schüler, die sie aufgrund einer pädagogischen Erfahrung gewonnen hatten. Versteht man unter Ergebnissen aber den gesamten Bereich der dazu gehörenden Informationen, müßte man auch die Auswirkungen des Unterrichts auf Lehrer, Verwaltungsbeamte, Schulpsychologen und andere untersuchen. Hierzu gehören auch Daten über die Abnutzung der Ausstattung, den Einfluß der Lernbedingungen und die Kosten. Bei der Evaluation müssen außer den nachweisbaren oder sogar deutlich greifbaren Ergebnissen auch die Anwendung des Gelernten, der Transfer und die Auswirkungen wiederholenden Lernens, die sich vielleicht erst viel später messen lassen, berücksichtigt werden. Die Beschreibung der Ergebnisse von Fahrschulunterricht z. B. könnte sinnvollerweise Berichte darüber enthalten, inwieweit jemand im Laufe seines Lebens

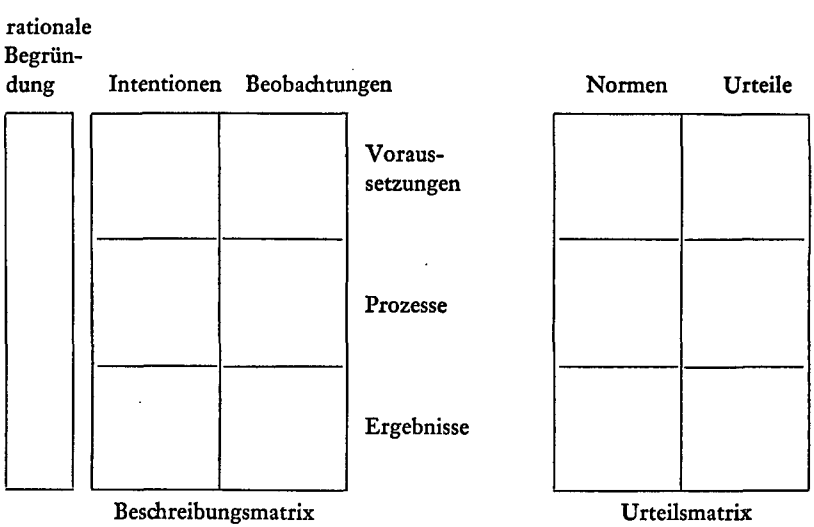


Abb. 1: Eine Matrix für Daten, die vom Evaluator eines Bildungsprogramms gesammelt werden sollen

Unfälle vermieden hat. Ergebnisse sind also, kurz gesagt, die unmittelbaren und langfristigen kognitiven und affektiven, persönlichen und gesellschaftlichen Folgen der Erziehung.

Voraussetzungen, Prozesse und Ergebnisse sind Elemente der Evaluationsmatrix und müssen – wie Abbildung 1 zeigt – bei der Beschreibung und Beurteilung berücksichtigt werden. Um diese Matrix auszufüllen, sammelt der Evaluator Urteile und Beschreibungen, z. B. über Vorurteile in der Gemeinde, über Stile des Problemlösens und die Persönlichkeit des Lehrers. Aus Abbildung 1 geht auch hervor, daß Urteile entweder als allgemeine Qualitätsnormen oder als spezifische Urteile über ein gegebenes Programm klassifiziert werden. Beschreibende Daten werden als Intentionen und Beobachtungen klassifiziert. Der Evaluator kann die Sammlung seiner Daten entsprechend den Kategorien der Abbildung 1 organisieren.

Der Evaluator kann aufzeichnen, was Pädagogen beabsichtigen und Beobachter wahrnehmen, was die verantwortlichen Geldgeber im allgemeinen erwarten und wie Beurteiler das gegenwärtige Programm bewerten. Die Aufzeichnung kann versuchen, Voraussetzungen, Prozesse und Ergebnisse getrennt innerhalb der vier Gruppen als *Intentionen, Beobachtungen, Normen und Urteile* zu identifizieren (vgl. Abb. 1). Die folgenden Ausführungen liefern ein Beispiel für zwölf Daten, die in die zwölf Felder eingetragen werden können. Sie beginnen mit einer intendierten Voraussetzung und gehen jede Spalte hinunter, bis ein Urteil über die Ergebnisse gefällt worden ist.

Im Wissen, daß

(1) Kapitel 11 als Aufgabe aufgegeben worden ist, und daß er beabsichtigt (2), am Mittwoch über das Thema eine Vorlesung zu halten, gibt ein Professor an (3), was die Studenten bis zum Freitag können sollen – z. T. dadurch, daß er einen Fragebogen über das Thema bearbeiten läßt. Er beobachtet, (4) daß einige Studenten am Mittwoch abwesend waren, (5) daß er wegen der langen Diskussion nicht die Vorlesung beenden konnte und (6) daß einen wichtigen Begriff im Fragebogen nur zwei Drittel der Hörer zu verstehen schienen. Im allgemeinen erwartet er (7), daß einige abwesend sind, aber daß das Versäumnis durch die für den Fragebogen aufgewandte Zeit aufgeholt wird; er erwartet (8), daß seine Vorlesungen für etwa 90 % der Zuhörer so klar sind, daß sie ohne Schwierigkeiten verstehen können; und er weiß (9), daß seine Kollegen erwarten, daß nur einer von zehn Studenten alle wichtigen Begriffe in solchen Vorlesungen versteht. Nach seinem Urteil bot (10) die aufgegebene Lektüre keine ausreichenden Hintergrundinformationen für seine Vorlesung; Studenten äußerten (11), daß die Vorlesung provokativ war; der Hilfsassistent, der die Fragebogen las, sagte (12), daß eine entmutigend große Zahl der Studenten einen wichtigen Begriff mit einem anderen zu verwechseln schien.

Nicht einmal für die ferne Zukunft erwarten Pädagogen und Evaluatoren,

daß Daten so genau aufgezeichnet werden. Meine Absicht war es, hier zwölf Beispiele für Daten zu geben, die den zwölf verschiedenen Feldern in der Matrix zugeordnet werden können. Im folgenden möchte ich die Matrix für die Beschreibungsdaten erläutern.

Ziele und Intentionen

Seit vielen Jahren sind Unterrichtstechnologen und Testkonstrukteure für eine stärker explizite Formulierung der pädagogischen Ziele eingetreten. Nach meiner Auffassung sind Ziele, Lernziele und Intentionen synonym. Ich benutze als Bezeichnung der Kategorie *Intentionen*, weil heute viele Pädagogen »Ziele« und »Lernziele« mit »intendiertem Schülerverhalten« gleichsetzen. In diesem Beitrag umfassen Intentionen die intendierten Bedingungen der Umwelt, die geplanten Demonstrationen, die beabsichtigte Behandlung von bestimmten fachspezifischen Inhalten und das angestrebte Schülerverhalten. Zu den drei Feldern dieser Reihe gehören erwünschte, erhoffte, erwartete und sogar befürchtete Auswirkungen. Diese Datengruppe enthält die Ziele und Pläne anderer Personen, vor allem jedoch die der Schüler. (Man sollte bedenken, daß Pädagogen nicht das Recht haben, die Untersuchung einer Variablen dadurch auszuschließen, daß sie nicht als ein Lernziel angesehen wird. Der Evaluator sollte die Variable und ihre Ablehnung erfassen.) Die daraus sich ergebende Sammlung der *Intentionen* ist eine – nach Priorität geordnete – Aufzeichnung aller möglicherweise eintretenden Ereignisse.

Die Tatsache, daß viele Pädagogen heute die »Ziele« mit den »intendierten Schülerverhaltensweisen« gleichsetzen, geht auf die Behavioristen, vor allem jedoch auf die Vertreter des Programmierten Unterrichts zurück. Indem sie das Schwergewicht auf die spezifischen Unterrichtshandlungen und -übungen gelegt haben, die zur Verbesserung der Schülerantworten beitragen, haben sie eine gewisse Reform des Unterrichts bewirkt. Das American Association for the Advancement of Science Elementary Project (A. A. A. S. S.) hat z. B. sein Curriculum erfolgreich mit Hilfe von Verhaltenszielen entwickelt. Einige innovative Curriculumprojekte haben jedoch festgestellt, daß die Betonung behavioristischer Ergebnisse für kreativen Unterricht hinderlich ist (vgl. Atkin 1963). Der pädagogische Evaluator sollte Ziele nicht nur als erwartetes Schülerverhalten formulieren. Um ein Bildungsprogramm zu *evaluieren*, muß man untersuchen, was gelehrt und gelernt werden soll. (Viele Voraussetzungen und viele Unterrichtsprozesse können auf Wunsch behavioristisch formuliert werden.) Wie Intentionen formuliert werden, ist kein Kriterium für ihre Berücksichtigung bei der

Evaluation. Intentionen können die allgemeinen Ziele der »Educational Policies Commission« oder die detaillierten Ziele der Hersteller von Programmen sein (vgl. Mager 1962). Taxonomische, mechanistische, humanistische, biblische – alle noch so verschiedenartigen Zielformulierungen müssen bei der Evaluation berücksichtigt werden.

Bei dem Versuch, die Ziele des Pädagogen aufzuzeichnen, stößt ein Evaluator gegenwärtig auf Schwierigkeiten. Zu Beginn seiner Arbeit fordert er den Pädagogen auf, seine Lernziele so darzulegen, daß Verfahren für das Testen der Ergebnisse entwickelt werden können. Dabei erfährt er, daß der Pädagoge sich entweder sträubt oder unfähig ist, seine Ziele zu verbalisieren. Obwohl der Evaluator das Formulieren von Verhaltenszielen für die Aufgabe des Pädagogen hält, hilft er ihm sorgfältig und gern dabei. Nach unserer Auffassung ist es jedoch nicht Aufgabe des Pädagogen, Verhaltensziele zu formulieren. In Übereinstimmung mit Scrivens Ausführungen liegt u. E. die Beschreibung curricularer Lernziele beim Evaluator. Er ist mit der Terminologie des Verhaltens und seiner Ausdrucksformen vertraut. So wie es seine Aufgabe ist, die Verhaltensweisen eines Lehrers und die Antworten eines Schülers in Daten umzuformen, muß er auch die Intentionen und Erwartungen eines Pädagogen in Daten transformieren. Wiederholt muß der Evaluator den Pädagogen bitten, seine Intentionen zu äußern. Er sollte versuchen, die Zahl der Antworten durch Fragen zu erhöhen, wie: »Kann man es auch so sagen? Ist das ein Beispiel für das, was Sie meinen?« Natürlich kann der Evaluator den interessierten Pädagogen über Verhaltensziele unterrichten. Das kann seine Arbeit erleichtern. Darauf zu bestehen, daß jeder Pädagoge Verhaltensziele verwendet, ist jedoch falsch.

Authentische Formulierungen der Intentionen zu erhalten, ist für den Evaluator eine schwierige Aufgabe. Die benötigte Methodologie muß noch entwickelt werden. Im weiteren soll nun die zweite Reihe der Datenfelder behandelt werden.

Die Auswahl von Methoden der Beschreibung

Die meisten deskriptiven Daten, die am Anfang des vorherigen Abschnitts erwähnt wurden, werden als *Beobachtungen* klassifiziert. Wenn der Evaluator[1] die Voraussetzungen, Prozesse und die daraus sich ergebenden Folgen beschreibt, gibt er (nach Abb. 1) seine Beobachtungen wieder. Manchmal macht er die Beobachtungen direkt und persönlich, manchmal benutzt er Instrumente. Zu seinen Instrumenten gehören Inventurverzeichnisse, Listen mit bibliographischen Daten, Routine-Interviews, Strichlisten, Fragebogen zur Erforschung von Meinungen und alle Arten psy-

chometrischer Tests. Der erfahrene Evaluator konzentriert seine Aufmerksamkeit auf das Messen der Schülerleistungen; aber er beobachtet auch die anderen Ergebnisse, Voraussetzungen und unterrichtlichen Prozesse.

Viele Pädagogen fürchten, daß der von außen kommende Evaluator nicht die Merkmale berücksichtigt, die nach dem Urteil des Lehrerkollegiums die wichtigsten sind. Dies trifft manchmal zu; oft aber richten Evaluatoren *zuviel* Aufmerksamkeit auf das, was sie beobachten sollen, und zuwenig Aufmerksamkeit auf andere Dinge. Bei der Auswahl der Variablen für die Evaluation muß der Evaluator eine subjektive Entscheidung treffen. Selbstverständlich muß er für eine Untersuchung die Zahl der Elemente begrenzen. Alle Elemente, die nicht berücksichtigt werden, tragen nach seinem Urteil nicht zum Verständnis des pädagogischen Geschehens bei. Der Evaluator sollte die Variablen besonders beachten, die durch die Lernziele des Pädagogen angegeben werden. Darüber hinaus muß er aber auch zusätzliche Variablen beobachten und die ungewollten Nebenwirkungen und zufälligen Ergebnisse untersuchen. Der Evaluator hat die Beobachtungsgegenstände und Meßverfahren auszuwählen.

Ohne die rationale Begründung (rationale) des Programms darzulegen, ist eine Evaluation nicht vollständig. Sie muß gesondert berücksichtigt werden (vgl. Abb. 1). Jedes Programm enthält eine allerdings oft nur implizite Begründung. Sie macht den philosophischen Hintergrund und die grundlegenden Ziele des Programms deutlich. Berlack (1966) hat dargelegt, wie wichtig die rationale Begründung für die Evaluation ist. Sie soll eine Grundlage für die Evaluation der Intentionen bieten. Der Evaluator muß sich oder anderen Beurteilern die Frage stellen, ob der von den Pädagogen entwickelte Plan einen logischen Schritt zur Implementation der grundlegenden Ziele darstellt. Die Begründung ist auch für die Wahl der Personen wichtig, die das Programm verwenden sollen, z. B. für die Mathematiker und Mathematiklehrer, die später verschiedene Aspekte des Programms beurteilen sollen.

Eine Formulierung der Begründung zu erhalten ist oft schwer. Häufig ist ein effektiver Lehrer beim Formulieren der Begründung für sein pädagogisches Handeln recht uneffektiv. Wenn er gedrängt wird, kann er schließlich vielleicht das sagen, was man von ihm erwartet. Die Begründung sollte jedoch in der Sprache des Pädagogen formuliert werden. Die Vorschläge des Evaluators können leicht hinderlich werden, da sie vielleicht übernommen werden, weil sie attraktiv sind, ohne jedoch die wirklichen Gründe für die Handlungen des Pädagogen anzugeben.

Die Urteilsmatrix bedarf weiterer Erläuterung. Aber ich verschiebe das bis nach der Behandlung der Grundlagen für die Verarbeitung deskriptiver Daten.

Kontingenz und Kongruenz

Um deskriptive Evaluationsdaten zu verarbeiten, gibt es für jedes Bildungsprogramm zwei wichtige Verfahren. Man muß die Kontingenzen zwischen den Voraussetzungen, Prozessen und Ergebnissen und die Kongruenz zwischen den Intentionen und Beobachtungen finden. Die Verarbeitung der Urteile folgt einem anderen Modell. Die ersten beiden Spalten der Datenmatrix in Abbildung 1 enthalten die deskriptiven Daten. Das Schema für die Verarbeitung dieser Daten ist in Abbildung 2 dargestellt. Wenn das, was intendiert ist, wirklich geschieht, sind die Daten für ein Curriculum *kongruent*. Um vollständig kongruent zu sein, müssen die intendierten Voraussetzungen, Prozesse und Ergebnisse eintreten. (Das geschieht selten – und oft sollte es nicht geschehen.) Innerhalb einer Reihe der Datenmatrix sollte der Evaluator die Felder, die Intentionen und Beobachtungen enthalten, vergleichen, um Diskrepanzen festzustellen und den Grad der Kongruenz in dieser Reihe zu beschreiben (die Wichtigkeit der Kongruenz der Ergebnisse wurde in dem von Taylor/Maguire (1966) erarbeiteten Evaluationsmodell hervorgehoben). Die Kongruenz gibt keinen Hinweis darauf, ob die Ergebnisse reliabel oder valide sind, sondern lediglich darauf, daß das Intendierte eintrat.

Ähnlich dem Gestaltpsychologen, der in dem Ganzen mehr findet als die Summe seiner Teile, findet der Evaluator bei der Untersuchung der Variablen von beliebigen zwei der drei Bereiche in einer Spalte der Datenmatrix mehr zu beschreiben als die Variablen selbst. Die Beziehungen oder *Kontingenzen* zwischen den Variablen verdienen zusätzliche Aufmerksamkeit. Insofern als Evaluation die Suche nach Beziehungen ist, die die Verbesserung der Erziehung ermöglichen, ist es die Aufgabe des Evaluators, die Ergebnisse zu identifizieren, die mit bestimmten Voraussetzungen und Unterrichtsprozessen kontingent sind.

Unterrichtsplanung und Curriculumrevision haben in den letzten Jahren auf dem Vertrauen in bestimmte Kontingenzen beruht. Täglich organisiert der gute Lehrer seinen Unterricht und wählt seine Curriculummaterialien entsprechend seinen unterrichtlichen Zielen aus. Für ihn sind die Kontingenzen in der Hauptsache logisch intuitiv, die durch zahlreiche befriedigende und bestätigende Erfahrungen unterstützt werden. Sogar der erfahrene und zweifellos der weniger erfahrene Lehrer müssen ihre intuitiv erwarteten Kontingenzen der Überprüfung durch Evaluatoren unterziehen.

Als erster Schritt in der Evaluation ist es wichtig, die Kontingenzen aufzuzeichnen. Ein Film über eine Überschwemmung kann dazu dienen (intendierter Prozeß), Schülern einen Hintergrund für eine entsprechende Schutzgesetzgebung (intendiertes Ergebnis) zu geben. Denen, die die

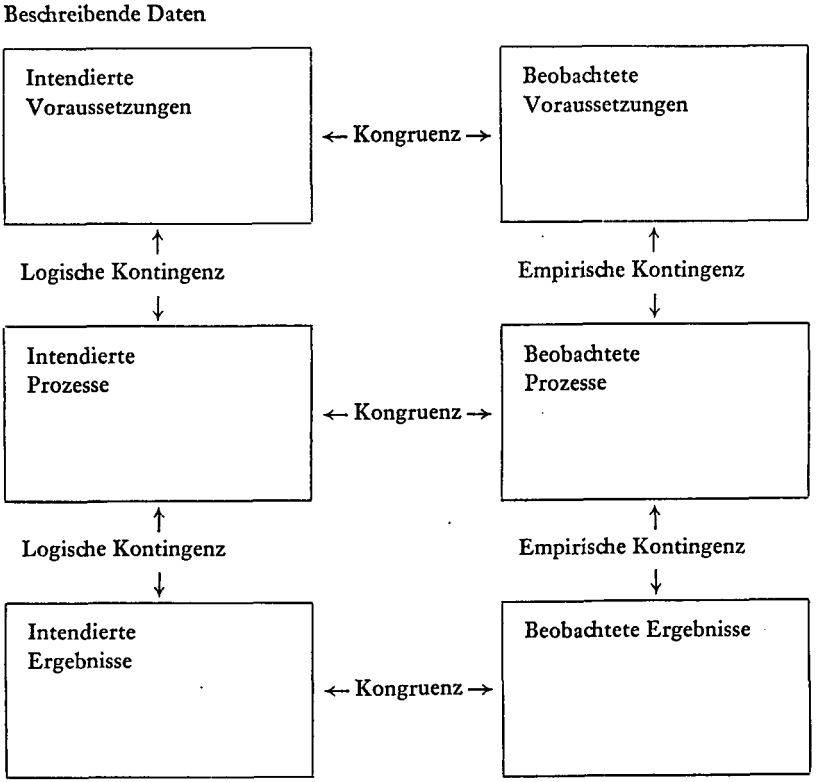


Abb. 2: Eine Darstellung des Prozesses der Verarbeitung von beschreibenden Daten

Sach- und die pädagogischen Probleme kennen, kann man folgende Frage stellen: »Gibt es eine logische Verbindung zwischen diesem Prozeß und dieser Zielsetzung?« Wenn es sie gibt, existiert eine logische Kontingenz zwischen diesen beiden Intentionen. Die Aufzeichnung sollte sie deutlich machen. Bei der Evaluation von Intentionen ist das Kontingenzkriterium immer ein Kriterium der Logik. Um die Logik einer pädagogischen Kontingenz zu testen, greifen die Evaluatoren auf ihre vorherigen Erfahrungen und vielleicht auch auf Forschungserfahrungen mit ähnlichen Erscheinungen zurück. Keine unmittelbare Beobachtung dieser Variablen ist jedoch erforderlich, um die Stärke der Kontingenzen zwischen Intentionen zu testen.

Die Evaluation von Beobachtungskontingenzen hängt von der empirischen Evidenz ab. Um sagen zu können: Diese Klasse macht im Rechnen

schnelle Fortschritte, weil der Lehrer gute, aber nicht zu differenzierte Kenntnisse in Mathematik hat, braucht man empirische Daten aus der Evaluation oder aus der Forschungsliteratur (vgl. Bassam 1962). Aus der Evaluation eines einzigen Programms allein kann man nicht die Daten erhalten, die für die Formulierung einer Kontingenz notwendig sind. Dazu bilden auch frühere Erfahrungen mit ähnlichen Erscheinungen eine grundlegende Qualifikation für den Evaluator.

Die Kontingenzen und Kongruenzen, die die Evaluatoren identifiziert haben, müssen genauso wie einheitliche Beobachtungsdaten der Beurteilung von Experten und Teilnehmern unterzogen werden. Das Auftreten einer *Nicht-Kongruenz* wird entsprechend den unterschiedlichen Standpunkten verschieden bewertet werden. So sind vielleicht ein Schulrat und ein Schulpsychologe verschiedener Ansicht darüber, welche Bedeutung der Streichung von Stunden für die im Stundenplan vorgesehene Sexualhygiene zukommt. Für die Beurteilung von Kontingenzen bietet sich als Beispiel das Ausmaß an, in dem die Lehrfähigkeit eines Lehrers während eines ganzen Schultages kontingent ist; darin kann ein Beurteiler einen ausreichenden Grund dafür sehen, auf eine Unterrichtsstunde am frühen Morgen zu verzichten, ein anderer jedoch nicht. Die Vorstellungen, die über die Bedeutung von Kongruenz und Kontingenz bestehen, müssen vom Evaluator sorgfältig untersucht werden.

Normen und Urteile

Nach allgemeinem Konsens besteht das Ziel der Erziehung in der optimalen Bildung der Schüler. Wie und unter welchen Umständen die Schüler sie jedoch erreichen, wird immer umstritten sein. Unabhängig davon, ob die Ziele von den örtlichen Gemeinden aufgestellt werden und nur für sie gelten oder ob sie für das ganze Land gelten sollen, erfordert die Evaluation des Bildungsniveaus eher explizite als implizite Normen (standards). Die gegenwärtigen Bildungsprogramme werden keiner normorientierten Evaluation unterzogen. Das bedeutet nicht, daß sich die Schulen nicht anstrengen oder keine Erfolge erzielen; es bedeutet lediglich, daß Normen – allgemeingültige Verhaltensformen – nicht überall gebräuchlich sind. Selbst wenn Schulen in allen Teilen des Landes die gleichen Evaluationsbogen verwenden [2], wird die Interpretation der gewonnenen Daten mit unklaren, individuell gebrauchten Begriffen erfolgen. Sogar im Rahmen informaler Evaluation kann keine Schule die Auswirkungen ihres Curriculum evaluieren, ohne zu wissen, wie andere Schulen ähnliche Lernziele zu erreichen versuchen. Leider wehren sich viele Pädagogen gegen die sy-

stematische Sammlung solcher Kenntnisse (vgl. Hand 1965 u. Tyler 1965). Über die Qualität der Erziehung eines Schülers weiß man gegenwärtig wenig. Die Schulzensuren beruhen auf den privaten Kriterien und Normen eines einzelnen Lehrers. Die meisten Werte in standardisierten Tests geben eher Auskunft darüber, wo ein Schüler bei der Lösung psychometrisch brauchbarer Aufgaben im Verhältnis zu seiner Bezugsgruppe steht, als über das Ausmaß an Kompetenz, mit der er wesentliche schulische Aufgaben erfüllt. Obwohl die meisten Lehrer in der Lage sind, ihre Fächer zu unterrichten und Lernschwierigkeiten zu erkennen, haben nur wenige die Fähigkeit, zu *beschreiben,* wie ein Schüler sich mit seiner geistigen Umwelt auseinandersetzt. Weder Schulzensuren noch Punktwerte in standardisierten Tests, noch die Ansichten der Lehrer enthalten genügend Informationen über das Bildungsniveau der Schüler.

Selbst wenn die Meßwerte erfolgreich interpretiert werden, ist Evaluation aufgrund der zahlreichen Normen schwierig. Die Normen unterscheiden sich von Schüler zu Schüler, Lehrer zu Lehrer und Bezugsgruppe zu Bezugsgruppe, und das ist auch richtig so. In einer pluralistischen Gesellschaft haben verschiedene Gruppen unterschiedliche Normen. Die Aufgabe der Evaluation besteht zum Teil darin, deutlich zu machen, wer welche Normen hat.

Es wurde bereits dargelegt, daß im Verlauf eines längeren Zeitraums die *Intentionen* eines Pädagogen sich ändern. Das bedeutet, daß sich während des Unterrichts die Kriterien und die Normen des Pädagogen wandeln. Während der Entwicklung und Dissemination eines Curriculum ändern sich sogar die Hauptgruppen der Kriterien. In einer umfassenden Analyse des Prozesses, in dem neue Curricula an die speziellen Bedingungen der einzelnen Schulen adaptiert werden, identifizierten Clark und Guba (1965) acht Stadien der Veränderung. Für jedes Stadium erarbeiteten sie spezifische Kriterien (jedes mit seinen eigenen Normen), aufgrund derer das Curriculum evaluiert werden soll, bevor man zum nächsten Stadium fortschreitet. Alle ihre Kriterien bedürfen weiterer Ausführung; hier soll nur angedeutet werden, daß es in jedem aufeinanderfolgenden Stadium der Curriculumentwicklung recht unterschiedliche Kriterien gibt. In der informalen Evaluation werden die Kriterien oft unspezifiziert gelassen. Formale Evaluation ist spezifischer. Je sorgfältiger die Evaluation ist, desto weniger Kriterien scheint es zu geben; je sorgfältiger die Kriterien spezifiziert sind, desto weniger wird die Angemessenheit der ihnen zugrunde liegenden Normen beachtet. Leider haben die am besten ausgebildeten Evaluatoren die Erziehung mit einem Mikroskop anstatt mit einem Weitwinkelsucher untersucht.

Es gibt keine genauen Kenntnisse darüber, was Schulen und Curriculum-

projekte gegenwärtig leisten; z. T. liegt es daran, daß die Methoden für die Verarbeitung von Urteilsdaten unzulänglich sind. Bei dem gegenwärtig geringen Ausmaß an formaler Evaluation berücksichtigt man zu wenig Kriterien, ist zu tolerant für implizite Normen und kümmert sich nicht um die Vorteile relativer Vergleiche. Es bedarf weiterer Ausführungen über relative und absolute Normen.

Vergleichen und Urteilen

Die Beurteilung der Charakteristika eines Bildungsprogramms kann erfolgen (1) in bezug auf absolute Normen, wie sie sich in persönlichen Urteilen äußern, und (2) in bezug auf relative Normen, wie sie in den Charakteristika alternativer Curricula zum Ausdruck kommen. Man kann das School Mathematics Study Group Project in bezug auf persönliche Ansichten darüber, was ein Mathematikcurriculum sein soll, oder in bezug auf andere Mathematikcurricula evaluieren. Die Vergleiche und Beurteilungen des Evaluators sind in Abbildung 3 dargestellt. Der obere linke Teil der Abbildung entspricht der Datenmatrix in Abbildung 2. Auf der oberen rechten Seite sind Normengruppen dargestellt, mit denen ein Curriculum absolut beurteilt werden kann. Da es zahlreiche Bezugsgruppen oder Gesichtspunkte geben kann, gibt es viele unterschiedliche Normengruppen. Die verschiedenen Matrizen auf der unteren linken Seite stellen verschiedene alternative Curricula dar, mit denen das Curriculum, das evaluiert wird, verglichen werden kann.

Wenn alle absoluten Normengruppen formalisiert werden, würden sie angemessene und wertvolle Bezugsebenen für die Voraussetzungen, Prozesse und Ergebnisse bilden. Bislang ging es nur um das Aufstellen von Normen, nicht um ihre Beurteilung. Bevor der Evaluator ein Urteil fällt, muß er bestimmen, ob alle Normen getroffen werden. Wenn Normen nicht vorhanden sind, müssen sie gesetzt werden. Der Urteilsakt selbst entscheidet, welche Normengruppe berücksichtigt wird. Genauer gesagt, Urteile fällen heißt: jeder Normengruppe eine bestimmte Bedeutung zuordnen. Rationales Urteilen in der pädagogischen Evaluation ist eine Entscheidung darüber, wieviel Beachtung den Normen jeder Bezugsgruppe bei der Entscheidung darüber zukommt, ob eine administrative Handlung erfolgen oder nicht erfolgen soll.

Der relative Vergleich wird ähnlich durchgeführt, wobei allerdings die Normen aus der Beschreibung anderer Bildungsprogramme stammen. Es ist nicht sehr schwierig, ein Urteil darüber zu fällen, ob ein Curriculum in einem Charakteristikum besser ist als ein anderes, aber es gibt viele Cha-

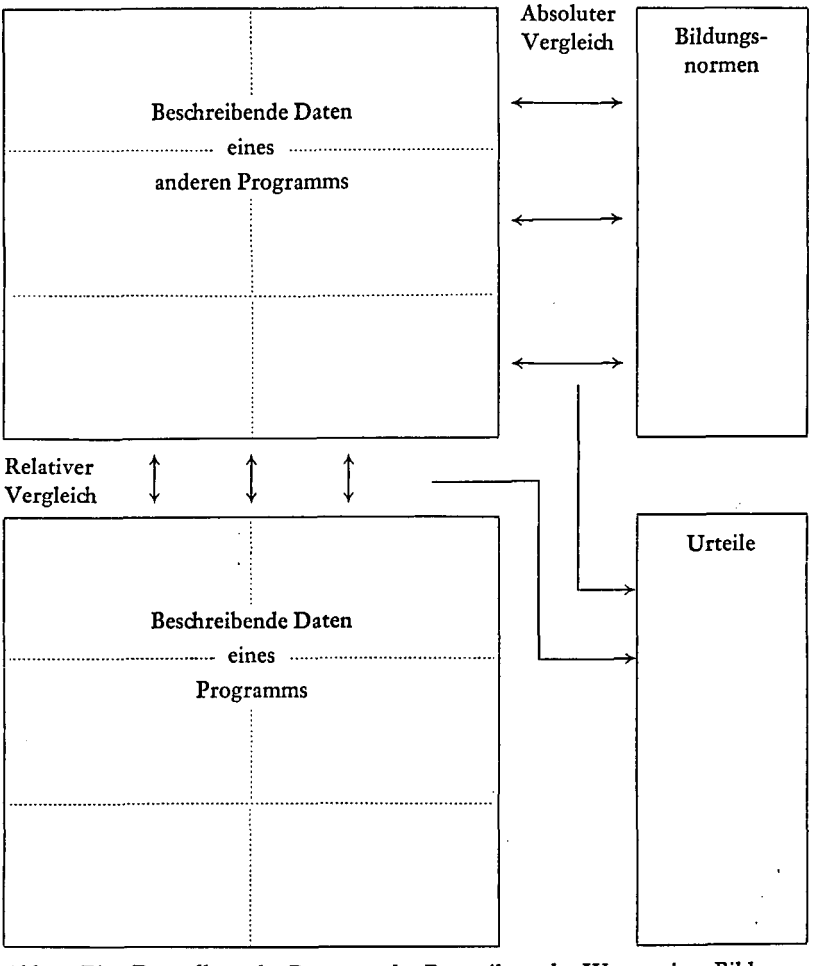


Abb. 3: Eine Darstellung des Prozesses der Beurteilung des Wertes eines Bildungsprogramms

rakteristika, die verschieden wichtig sind. Der Evaluator wählt aus, welche Charakteristika zu berücksichtigen und mit welchen Bildungsprogrammen sie zu vergleichen sind. Mit Hilfe der relativen und der absoluten Beurteilung eines Bildungsprogramms kann man ein Gesamturteil über seine Qualität (vielleicht mit einigen modifizierenden Aussagen) erhalten, das die Grundlage für eine Bildungsentscheidung sein kann. Aufgrund dieser abschließenden Beurteilung kann dann eine Empfehlung ausgesprochen werden.

Absolute und relative Evaluation

Ob absoluter oder relativer Evaluation der Vorzug zu geben ist, ist zwischen Scriven und Cronbach umstritten geblieben. Cronbach (1963) vertritt die Auffassung, es sei sehr gewagt, aus den Ergebnissen curricularer Vergleichsuntersuchungen Generalisationen zu machen, die auf eine örtliche Schulsituation zutreffen sollen – selbst wenn diese Untersuchungen umfangreich, gut angelegt und richtig kontrolliert worden sind –, so daß in Vergleichsuntersuchungen eine unzulängliche Forschungsinvestition besteht. Darüber hinaus ist wahrscheinlich der Unterschied in den Zielsetzungen der verglichenen Curricula so groß, daß die Ergebnisse nicht interpretierbar sind, es sei denn, ein Curriculum ist einem anderen weit überlegen. Da Cronbach diesen Fall selten erwartet, tritt er dafür ein, weniger Vergleichsuntersuchungen, dafür aber mehr intensive Prozeß- und Einzeluntersuchungen von Curricula mit umfangreichen Messungen und sorgfältiger Beschreibung zu machen.

Scriven (1967) andererseits vertritt die Auffassung, daß der Pädagoge vor allem wissen will, ob ein Curriculum besser ist als ein anderes, und daß das beste Verfahren zur Beantwortung dieser Frage der direkte Vergleich ist. Er weist darauf hin, wie schwierig die begrifflich klare Beschreibung der Ergebnisse komplexer Lernprozesse in bezug auf absolute Normen ist, wenn man sie mit der Beobachtung der relativen Ergebnisse zweier Bildungsprogramme vergleicht. Inwieweit Scrivens Ausführungen überzeugend sind, hängt wahrscheinlich vom Adressaten ab. Ein Pädagoge, der über die Adaptation eines Programms entscheiden muß, wird seine Überlegungen eher plausibel finden als ein Curriculuminnovator und Unterrichtstechnologe.

Die von Scriven getroffene Unterscheidung zwischen *formativer* und *summativer* Evaluation ist eine sehr wertvolle Differenzierung im Rahmen der Evaluation. Seine Verwendung der Begriffe bezieht sich vor allem auf das Stadium der Entwicklung von curricularem Material. Solange das Material noch nicht so fertiggestellt ist, daß es an die Lehrer verteilt werden kann, ist Evaluation formativ; nach Abschluß seiner Entwicklung ist Evaluation summativ. Wahrscheinlich empfiehlt es sich eher, zu unterscheiden zwischen einer Evaluation, die sich an den Kriterien und Normen der Curriculumentwickler, Autoren und Verleger orientiert, und einer Evaluation, die sich an den Kriterien und Normen der Schüler, Beamten der Schulverwaltung und Lehrer orientiert. Die Unterscheidung zwischen formativer und summativer Evaluation könnte so definiert werden, und ich will die Begriffe so verwenden. Die Kommission eines Lehrerkollegiums, die ein Curriculum für die Schule auswählen soll, stellt Fragen nach seiner Qua-

lität und Geeignetheit. Der Curriculumentwickler, der Cronbachs Rat befolgt, fragt danach, wie das Curriculum verbessert werden kann. (Keiner befaßt sich mit den individuellen Unterschieden zwischen den Schülern.) Um diese Fragen zu beantworten, muß der Evaluator verschiedene Daten untersuchen und sich auf verschiedene Normen beziehen.

Der Evaluator, der seine Aufgabe eher in der summativen als in der formativen Evaluation sieht, muß die Adressaten über die Qualität des Curriculum informieren. Sein Ziel besteht darin (vgl. Abb. 3), zu Urteilen zu gelangen. Wahrscheinlich wird er die Schulsituationen zu beschreiben versuchen, in denen die Verfahren bzw. Materialien benutzt werden können. Seine Aufgabe kann er darin sehen, herauszufinden, wie gut ein Curriculum in ein bereits bestehendes Schulprogramm paßt. Er muß in Erfahrung bringen, ob die für das Curriculum intendierten Voraussetzungen, Prozesse und Ergebnisse sich mit den finanziellen Mitteln, Normen und Zielen der Schule vereinbaren lassen. Dazu kann es erforderlich sein, der Schule ebensoviel Aufmerksamkeit wie dem Curriculum zuzuwenden.

Der formative Evaluator andererseits interessiert sich stärker für Kontingenzen, wie sie in Abbildung 2 dargestellt worden sind. Er wird in der Evaluationsuntersuchung und in Querschnittsuntersuchungen (across studies) nach gemeinsamen Veränderungen suchen, um auf ihrer Grundlage die Entwicklung gegenwärtiger oder zukünftiger Curricula zu steuern.

Für größere Evaluationsuntersuchungen verfügt ein Evaluator allein natürlich nicht über die vielen benötigten Fähigkeiten und Kenntnisse. Ein Team von Sozialwissenschaftlern ist für viele Aufgaben erforderlich. Solche Teams werden sicherlich aus Experten in der Unterrichtstechnologie, der Psychometrie, in den Skalierungsmethoden, in der Forschungsplanung (research design) und der Dissemination von Informationen bestehen. Curriculare Innovationen haben ohne Zweifel tiefe und umfassende Wirkungen auf unsere Gesellschaft; daher empfiehlt es sich auch, zu manchen Evaluationsteams einen Kulturanthropologen hinzuzuziehen. Auch Wirtschaftswissenschaftler und Philosophen können einen wesentlichen Beitrag zur Evaluation leisten. Experten für die Untersuchung von Wertvorstellungen, für Stichprobenerhebungen und statistische Methoden werden benötigt.

Der Pädagoge, der sich vor der Teilnahme an einer Evaluation scheut, wird erst recht davon unangenehm berührt, daß ein Evaluationsteam in seiner Schule arbeitet. Darüber hinaus erhebt sich für ihn die Frage, wie Evaluatoren die wirkliche Beschaffenheit der Erziehung, die durch ihre Gegenwart beeinflußt wird, beobachten und beschreiben können. Die Bedenken des Pädagogen sind berechtigt; die Evaluation – ja die bloße Gegenwart der Evaluatoren – hat manchmal einen positiven und manchmal einen negativen Einfluß auf die Erziehung. In beiden Fällen trägt sie je-

doch zum atypischen Charakter des Unterrichts bei. Einige Wissenschaftler nehmen an (Webb/Campbell/Schwartz/Sechritz 1966), daß die Verfahren der Evaluation eines Tages soweit entwickelt sein werden, daß sie die Evaluation nicht mehr beeinträchtigen.

Abschließend möchte ich den Leser darauf aufmerksam machen, daß zur Zeit eine der größten Investitionen im Bildungswesen in der Entwicklung neuer Bildungsprogramme besteht. Schulaufsichtsbeamte können ein Curriculum noch nicht mit Hilfe rationaler Begründungen revidieren, da es die dazu notwendige Evaluation nicht gibt. Wie können die Erfahrungen aus den Anstrengungen der Innovatoren der sechziger Jahre genutzt werden, wenn in den sechziger Jahren keine Evaluationsaufzeichnungen vorhanden sind? Für die Innovatoren und Lehrer der siebziger Jahre sind solche Informationen notwendig, denn die bisherigen Kenntnisse können nicht die Sammlung ausreichender Kenntnisse ersetzen. In unseren Datensammlungen sollten wir die Auswirkungen und ihre Gründe, die Kongruenz zwischen Intentionen und Ergebnissen und die verschiedenen Urteile der Adressaten festhalten. Solche Aufzeichnungen sollten gesammelt werden, um das pädagogische Handeln zu fördern, nicht um es zu hemmen. Evaluation sollte ihre Aufgabe darin sehen, Daten für Entscheidungen zu sammeln, und nicht darin, Unruhe in der Schule zu stiften.

Pädagogen sollten ihre eigenen Evaluationen sorgfältiger machen und stärker formalisieren. Alle, die in ihren Klassen oder in überregionalen Kommissionen daran interessiert sind, können sich vielleicht ihre Aufgabe durch die Beantwortung folgender Fragen verdeutlichen:

(1) Ist diese Evaluation vorwiegend deskriptiv oder vorwiegend beurteilend oder deskriptiv und beurteilend zugleich?
(2) Soll diese Evaluation die Voraussetzungen, die Prozesse oder die Ergebnisse allein oder eine Kombination dieser oder aber ihre funktionalen Kontingenzen betonen?
(3) Soll diese Evaluation die Kongruenz zwischen den Intentionen und den Ergebnissen angeben?
(4) Richtet sich diese Evaluation auf ein Programm allein, oder findet sie als Vergleich zwischen zwei oder mehreren Programmen statt?
(5) Soll diese Evaluation eher zur Weiterentwicklung von Curricula dienen, oder soll sie zwischen vorhandenen Curricula auszuwählen helfen?

Mit der Beantwortung dieser Fragen werden restriktive Auswirkungen unvollständiger und unangemessener Auffassungen von Evaluation leichter vermieden.

DANIEL L. STUFFLEBEAM

Evaluation als Entscheidungshilfe

In den letzten zweieinhalb Jahren habe ich mit Mitarbeitern aus Schulen, Erziehungsministerien mehrerer Bundesstaaten und dem amerikanischen Erziehungsministerium intensiv an Problemen der Evaluation gearbeitet. In den meisten Fällen galt es, Projekte zu evaluieren, die aus Titel I und Titel III des Elementary and Secondary Education Act von 1965 finanziert wurden. Der vorliegende Beitrag beruht auf diesen Erfahrungen und stellt den Versuch dar, einige meiner Vorstellungen über die Aufgaben der Evaluation im Rahmen der gegenwärtigen innovativen Bildungsprogramme darzustellen.

Der Aufsatz umfaßt zwei Teile. Im ersten Teil soll der gegenwärtige Stand der Evaluation im Bildungswesen dargestellt werden. Es gilt die Aufgaben der Evaluation zu beschreiben und zu zeigen, daß Pädagogen bislang bei dem Versuch, diesen Aufgaben gerecht zu werden, ineffektiv waren. Sodann sollen einige mögliche Gründe für das unzulängliche Ausmaß an Evaluation im Bildungswesen genannt werden. Im zweiten Teil des Beitrags versuche ich dann einige alternative Ansätze der pädagogischen Evaluation zu entwickeln. Es gilt Evaluation zu definieren, vier m. E. für Innovationen im Bildungsbereich besonders wichtige Evaluationsstrategien zu entwickeln und die Struktur von Evaluationsplänen zu erklären.

I. Der gegenwärtige Stand der Evaluation im Bildungswesen

Erziehung wird in immer stärkerem Maße als ein Mittel zur Befriedigung der sozialen, wirtschaftlichen und geistigen Bedürfnisse der Gesellschaft angesehen. Um dieser ständig schwieriger werdenden Rolle gerecht zu werden, müssen Pädagogen sich mit den wichtigen gesellschaftlichen Problemen befassen. Zu ihnen gehören die Fragen der Chancenungleichheit der verschiedenen Rassen, der De-facto-Segregation, der Aufstände in den Städten, der Desillusionierung der Jugend und der zahlreichen vorzeitigen

Schulabgänge. Den hier sich zeigenden Tendenzen und ihrer Ausweitung muß im Interesse unserer Gesellschaft entgegengearbeitet werden. Der Erziehung kommt dabei eine wichtige und schwierige Aufgabe zu, zu deren Bewältigung zahlreiche Reformen erfolgen müssen.

Voraussetzungen

Um den Pädagogen die Erfüllung ihrer neuen Aufgaben zu ermöglichen, stellt die Gesellschaft in allen Bildungsbereichen jährlich viele Milliarden Dollar im Rahmen der einzelstaatlichen und bundesstaatlichen Programme und mit Hilfe von Stiftungen zur Verfügung. Ein Beispiel für eine derartige Investition im Bildungswesen sind der Elementary and Secondary Education Act von 1965, das Head Start Program, der Education Professions Act und das Experienced Teacher Fellowship Program. Außerdem entwikkelten viele Industriezweige Bildungsinitiativen, so daß es voraussichtlich bald viele von der Industrie finanzierte Bildungsprojekte geben wird. Aufgrund der neuen Aufgaben gibt es im Bildungswesen noch nie dagewesene Möglichkeiten für die Entwicklung innovativer Programme.

Diese Situation hat jedoch auch dazu geführt, die Evaluation neuer Bildungspläne und Bildungsprogramme zu verlangen. Das gilt vor allem für die aus Bundesmitteln finanzierten Programme wie Titel I und Titel III des Elementary and Secondary Education Act. Hier fordert das Gesetz ausdrücklich, daß die finanzierten Projekte mindestens einmal jährlich einen Evaluationsbericht einreichen. Deshalb müssen viele Pädagogen in allen Sektoren des Bildungswesens erstmals eine formale Evaluation durchführen.

Die Vorschrift, solche Evaluationsuntersuchungen durchzuführen, ist sinnvoll und meiner Meinung nach längst fällig gewesen. Geldgeber und Öffentlichkeit haben das Recht, zu erfahren, ob ihre hohen Bildungsausgaben die erwünschten Erfolge erzielen. Noch stärker benötigen die Pädagogen selbst evaluative Informationen, um eine rationale Grundlage für Entscheidungen über alternative Pläne und Verfahren zu haben. Die Forderung nach der Evaluation von Bildungsprogrammen bewirkt jedoch noch nicht ihre Operationalisierung. Pädagogen müssen die Notwendigkeit der Evaluation einsehen und wirksame Evaluationsuntersuchungen durchführen.

Die Notwendigkeit besserer Evaluation im Bildungswesen

Ohne Zweifel sind viele Pädagogen davon überzeugt, daß Bildungsprogramme evaluiert werden müssen. Die Mehrzahl der gegenwärtig vorhan-

denen Evaluationsberichte der Schulen, Erziehungsministerien und Regional Educational Laboratories lassen erkennen, daß Pädagogen viel Zeit, Anstrengung und Geld für die Evaluation ihrer Programme aufwenden. Das hat jedoch noch nicht dazu geführt, eine wirkungsvolle Evaluation durchzuführen. Denn obwohl Pädagogen zahlreiche Evaluationsuntersuchungen gemacht haben, haben ihre Bemühungen nicht dazu beigetragen, die Informationen zu gewinnen, die als Grundlage für Entscheidungen über die evaluierten Programme notwendig sind.

Viele Evaluationsberichte enthalten nur bruchstückhafte Informationen. Obwohl solche Informationen für die Entscheidungsträger wichtig sein können, fehlt ihnen jedoch im allgemeinen das Ausmaß an Zuverlässigkeit, das Entscheidungsträger brauchen, um ihre Entscheidungen zu rechtfertigen, so daß derartige Informationen für wichtige Entscheidungen nur selten nützlich sind. Ein Beispiel dafür ist der erste Jahresbericht für den Titel I des Elementary and Secondary Education Act [1]. Dieser Bericht war sehr wichtig, da er die vielen tausend Projekte des Titel I umfaßte. Sein Wert wurde jedoch dadurch erheblich eingeschränkt, daß er fast keine empirisch gewonnenen Daten enthielt. Statt dessen bot er viele anekdotische Berichte, in denen Projektleiter darlegten, daß ihrer Meinung nach ihr Programm erfolgreich sei; viele von ihnen dachten sogar darüber nach, welches die Gründe für die angeblichen Erfolge sein könnten. Obwohl diese Anekdoten manchmal Fragen anschnitten, die für die Verbesserung des Titel I Program von Bedeutung waren, konnten die Entscheidungsträger im Kongreß, im amerikanischen Erziehungsministerium, in den Erziehungsministerien der Bundesstaaten und in den örtlichen Schulbezirken wichtige Entscheidungen kaum auf solche Beweisstücke gründen.

Beim Titel III des Elementary and Secondary Education Act ist die Situation kaum anders. Die für den Titel III verantwortlichen Beamten im amerikanischen Erziehungsministerium bestimmten die Qualität der Titel-III-Anträge immer aufgrund von 15 Kriterien mit Hilfe einer Fünf-Punkte-Skala [2]. Die Beurteilung im Kriterium, das sich auf die Evaluation bezog, lag in der Regel im negativen Bereich der Skala und war schlechter als bei dreizehn der anderen Kriterien; Ausnahme war das Kriterium, das sich auf die Dissemination bezog. Guba hat überzeugend dargelegt, daß die Evaluationspläne in den Anträgen zu Titel-III-Projekten unzureichend sind [3]. Aufgrund einer Analyse von 32 Anträgen zu Titel-III-Projekten kam Guba zu folgendem Ergebnis: »Es ist fraglich, ob die Ergebnisse dieser Evaluationsuntersuchungen überhaupt nützlich sind. Sie entsprechen wahrscheinlich der unter Pädagogen verbreiteten stereotypen Auffassung von Evaluation als etwas, das von oben gefordert wird und zu dessen Herstellung Zeit und Mühe erforderlich ist, das aber für Handlungsabläufe nur wenig re-

levant ist.«[4] Im Unterschied zu den bereits erwähnten Evaluationsuntersuchungen von Titel I und Titel III enthalten einige andere empirisch gewonnene Daten. So wurden z. B. für den Evaluationsbericht des New York City Higher Horizons Program exakte Forschungsverfahren benutzt (Wrightstone o. J.), um die Leistungen einer Versuchsgruppe, die im Rahmen dieses Programms unterrichtet wurde, mit denen einer Kontrollgruppe zu vergleichen, die in einigen Punkten mit der Versuchsgruppe parallelisiert worden war. Die Ergebnisse dieses fast 300 Seiten umfassenden Berichts waren für die Ergebnisse genauer Evaluationsuntersuchungen typisch. Es gab keine signifikanten Unterschiede. Im deutlichen Unterschied dazu stellte der Bericht jedoch auch fest, daß nach der Überzeugung der an dem Programm beteiligten Lehrer und Schulleiter die Unterschiede so stark waren, daß man das Programm auf keinen Fall wieder aufgeben dürfe.

Obwohl die Evaluationsuntersuchungen der Projekte der Titel I und III sich von der Evaluation des Higher Horizons Program in ihrer Genauigkeit unterschieden, waren sie in einer Hinsicht jedoch gleich. Keine von ihnen half den Entscheidungsträgern, die evaluierten Programme zu verbessern. Obwohl ich nur drei Beispiele für Unzulänglichkeiten bei gegenwärtigen Evaluationsuntersuchungen angeführt habe, sind sie meiner Meinung nach genügend aussagekräftig, um meinen Standpunkt zu veranschaulichen. In vielen Fällen helfen Evaluationsberichte den Entscheidungsträgern nur wenig oder gar nicht, so daß Entscheidungen im Bildungswesen zu treffen intuitiv und riskant bleibt.

Probleme der Evaluation im Bildungswesen

Wie läßt sich diese Situation erklären? Warum können Pädagogen nicht Evaluationsuntersuchungen durchführen, die zugleich nützlich und wissenschaftlich einwandfrei sind? Warum gewinnt man aus Evaluationsuntersuchungen mit Hilfe klassischer Forschungsmethoden nur Informationen, die für Entscheidungen über Bildungsprogramme von begrenztem Wert sind? Warum stehen die Ergebnisse vieler Evaluationsuntersuchungen, die keinen signifikanten Unterschied aufweisen können, in Widerspruch zu den Erfahrungen der an dem Programm Beteiligten?

Man kann diesen Fragen nicht einfach mit dem Argument begegnen, daß die Praxis der Evaluation zu weit hinter den Ansprüchen der Theorie zurückbleibt oder daß die Pädagogen sich nicht genügend bemühen, ihr Programm zu evaluieren. Auch sollte man nicht die evaluativen Äußerungen der beteiligten Personen als unglaubwürdig hinstellen oder behaupten, daß Ergebnisse mit nicht signifikanten Unterschieden typisch sind,

weil in der Pädagogik alle Anstrengungen kaum jemals einen Unterschied bewirken. Nach meiner Meinung besteht der Mangel an angemessenen Evaluationsdaten, weil verschiedene grundlegende Probleme erst gelöst werden müssen, bevor die Evaluationsuntersuchungen sich verbessern lassen. Zu diesen Problemen gehört das Fehlen ausgebildeter Evaluatoren, adäquater Evaluationsinstrumente und Evaluationsverfahren und einer angemessenen Theorie der Evaluation. Nach meiner Auffassung liegt das größte Problem im Fehlen einer für die Evaluation von Bildungsprogrammen geeigneten Konzeptualisierung oder Theorie.

Die konzeptuellen Grundlagen sind für Evaluationsuntersuchungen von grundlegender Bedeutung. Wenn die Konzeptionen falsch sind, dann sind die auf ihnen beruhenden Evaluationsergebnisse auch falsch. Daher ist es wichtig, die Qualität der Konzeptualisierungen zu untersuchen, die den gegenwärtigen Anforderungen an Evaluation zugrunde liegen. Es empfiehlt sich, die Konzeptualisierungen in folgende drei Klassen einzuteilen und jede getrennt zu betrachten:

1. Konzeptionen von der Beschaffenheit der Bildungsprogramme, für die Evaluationsuntersuchungen gebraucht werden, z. B. von den Entscheidungsprozessen und den entsprechenden Informationsbedürfnissen, die die Evaluationsuntersuchungen befriedigen sollen
2. Konzeptionen vom Wesen der Evaluation und ihrer Beziehung zu einzelnen Klassen von Bildungsprogrammen
3. Konzeptionen von der Struktur der Evaluationspläne, die für die Durchführung pädagogischer Evaluationsuntersuchungen gebraucht werden.

Probleme bei der Bestimmung der Anforderungen für Evaluation im Bildungswesen

Zunächst wollen wir einmal die Probleme untersuchen, die sich bei der Bestimmung der Aufgabe und Funktion pädagogischer Evaluationsuntersuchungen ergeben. Um eine Evaluation machen zu können, muß man natürlich erst wissen, was evaluiert werden soll. Das zu wissen, ist jedoch gegenwärtig eine außerordentlich schwierige Aufgabe. Die augenblicklichen Bemühungen um Evaluation sind aufgrund der neuen pädagogischen Programme und Aktivitäten entstanden. Zu diesen Aktivitäten gehören auch die erst in letzter Zeit für die Pädagogen entstandenen neuen Aufgaben, die neuen Verhältnisse in den verschiedenen Bereichen des Bildungswesens und das Anliegen zahlreicher Institutionen, zu gemeinsamen Bildungsentscheidungen zu kommen. Daher sollte man sich nicht dadurch beirren lassen, daß die bislang für das Bildungswesen gültige Theorie der Evaluation nicht länger ausreichte, um die Informationen, die für die Entwicklung

der neuen Bildungsprogramme notwendig sind, zu gewinnen. Viele neue Bildungsprogramme unterscheiden sich von den bisherigen so sehr, daß unsere Evaluationsuntersuchungen Fragen beantworten müssen, die sich von denen der Vergangenheit stark unterscheiden.

Meiner Meinung nach brauchen wir Konzeptualisierungen, die den Entscheidungsprozessen und Informationsbedürfnissen bei den neuen Bildungsprogrammen gerecht werden. Solche Programme, die zur Verbesserung des Bildungswesens beitragen sollen, sind von zahlreichen unterschiedlichen Entscheidungen abhängig; um diese Entscheidungen zu fällen, werden Informationen benötigt. Evaluatoren, die diese Informationen beschaffen sollen, müssen die relevanten Entscheidungsprozesse und entsprechende Informationsbedürfnisse kennen, bevor sie angemessene Evaluationsuntersuchungen planen können. Sie müssen den Ort, den Schwerpunkt, den Zeitpunkt und die kritische Reflektiertheit der Entscheidungen kennen, die sie vorbereiten sollen. Gegenwärtig gibt es weder eine adäquate Kenntnis der Entscheidungsprozesse und der entsprechenden Informationsbedürfnisse bei Bildungsprogrammen, noch gibt es ein systematisches Programm, diese Kenntnisse zu gewinnen. Kurz gesagt, es gibt keine angemessenen Konzeptualisierungen der Entscheidungen und der entsprechenden Informationsbedürfnisse; es gibt auch keine Vorstellungen über Programme, mit deren Hilfe sie gewonnen werden können.

Probleme bei der Definition von Evaluation im Bildungswesen

Als nächstes wollen wir Fragen nach der Bedeutung der Evaluation im Bildungswesen erörtern. Im allgemeinen haben Pädagogen es als die Aufgabe der Evaluation angesehen, das Ausmaß zu bestimmen, in dem Lernziele erreicht worden sind. Der erste Schritt zur Operationalisierung dieser Definition besteht darin, Programmziele in Verhaltensbegriffen zu formulieren. Um eine Beziehung zwischen Ergebnissen und Zielen herzustellen, muß man sodann Kriterien definieren und operationalisieren. Zur Operationalisierung dieser Kriterien gehört auch die Spezifikation der Instrumente, mit deren Hilfe die Ergebnisse und die Normen gemessen werden, die zur Bewertung der Ergebnisse dienen sollen.

Normen sind entweder absolut oder relativ. Eine absolute Norm könnte z. B. in einer bestimmten Punktzahl bestehen, die alle Schüler als Mindestdurchschnitt in einem ausgewählten Leistungstest erreichen sollen. Eine relative Norm könnte z. B. dadurch gebildet werden, daß eine Schülergruppe, die mit einem neuen Programm arbeitet, in einem ausgewählten Leistungstest im Durchschnitt höhere Punktzahlen erreichen soll als eine entsprechende Schülergruppe, die mit einem herkömmlichen Programm ar-

beitet. Ungeachtet der verwendeten Evaluationsnorm werden bei einer solchen Untersuchung die Daten erst nach einem vollständigen Ablauf des Programms analysiert, um zu bestimmen, in welchem Ausmaß die Ziele erreicht worden sind.

Evaluationsuntersuchungen, die nach diesem Modell durchgeführt werden, liefern nach Ablauf des Programms Daten über seine Gesamtwirkung und helfen Entscheidungen über das Programm zu fällen. Sie unterstützen aber den Pädagogen nicht bei der Anfangsplanung und der Realisierung der Programme. Solche Evaluationsuntersuchungen bieten daher nur eine ungenügende Hilfe für die Lösung der Probleme der Pädagogen, die innovative Programme planen und durchführen müssen.

Die Unzulänglichkeit der vorhandenen Evaluationskonzepte wird durch den folgenden Auszug aus den Ausführungen über die Evaluationsuntersuchungen des Titel I belegt, die eine Gruppe New Yorker Bürger vor einer Kommission des Kongresses machte: »Wir bitten um eine Verbesserung der gesetzlichen Bestimmungen über die Evaluation von Titel-I-Projekten, damit die Ergebnisse der Evaluation besser genutzt werden. Das Gesetz bestimmt lediglich, daß Evaluationsuntersuchungen gemacht werden müssen, nicht jedoch, daß ihre Ergebnisse für eine zukünftige Planung verwendet werden sollen. In New York City wurden in diesem Jahr Projekte wiederholt durchgeführt, ohne daß die Evaluationsergebnisse des vergangenen Jahres vorlagen. Um Evaluationsuntersuchungen wirksamer zu machen, sollten sie auch Alternativen und Empfehlungen des Evaluators enthalten. Was bislang nur eine kostspielige, für die Programmentwicklung sekundäre Aktivität war, sollte eine echte Aufgabe des Evaluators werden, damit den örtlichen Schulverwaltungen geholfen wird, ihre Entscheidungen auf Erfahrungen und empirische Daten zu gründen. Die amerikanische Wirtschaft wäre auch nicht funktionsfähig, wenn nicht ihre Berater, sobald sie die Wirksamkeit bestimmter Programme untersucht haben, das Management mit Alternativen versorgen würden.«[5]

Nach dieser Auffassung sind die aufgrund der gegenwärtigen Evaluationsprogramme verfaßten Berichte weder spezifisch noch rechtzeitig genug verfügbar, um Bildungsprogramme beeinflussen zu können. Evaluationsuntersuchungen, die diesen beiden Kriterien jedoch nicht gerecht werden, sind nur von geringem Nutzen.

Probleme der Planung von Evaluation im Bildungswesen

Schließlich sollen Probleme der Methodologie der Evaluation erörtert werden. Wenn die gegenwärtigen Konzeptionen der Evaluation sich nicht für die Evaluation heutiger pädagogischer Aktivitäten eignen, können auch

die entsprechenden Evaluationspläne nicht angemessen sein. Die bestehenden Verfahren der Evaluation wurden entwickelt, um die Aufgaben der Evaluation so zu erfüllen, wie sie in der Vergangenheit bestimmt worden waren.

Die Unzulänglichkeit der vorhandenen Evaluationsmethodologie wird deutlich, wenn man untersucht, welche Pläne Pädagogen für die Evaluation ihrer Programme verwenden. Wenn sie überhaupt einen Evaluationsplan entwickelt haben, ist es im allgemeinen ein experimenteller Versuchsplan. Sein wichtigstes Ziel besteht darin, Daten so zu erheben, daß sie eine innere Validität (internal validity) haben. Dazu müssen einige Bedingungen erfüllt werden. Die Einheiten des Bildungsprogramms, die untersucht werden sollen, müssen nach dem Zufallstichproben-Verfahren den Versuchs- und den Kontrollbedingungen zugeteilt werden. Eine Reihe von Schülern kann z. B. nach dem Zufallstichproben-Verfahren in zwei Gruppen geteilt werden; eine davon arbeitet mit dem neuen Programm, die andere mit dem herkömmlichen Programm der Schule. Sodann müssen die Versuchs- und Kontrollbedingungen geschaffen werden, die während des ganzen Versuchs konstant gehalten werden und der Anfangsdefinition der Bedingungen entsprechen müssen. Die Bedingungen des neuen Programms dürfen im Verlauf des Untersuchungsprozesses nicht verändert werden, da man sonst nicht weiß, was wirklich evaluiert worden ist.

Alle an dem Experiment teilnehmenden Schüler müssen in gleichem Ausmaß den Bedingungen ihrer Gruppe ausgesetzt werden; es muß darauf geachtet werden, daß die Schüler der Versuchsgruppe von denen der Kontrollgruppe deutlich getrennt sind. Denn wenn eine Kontaktaufnahme zwischen den Gruppen stattfindet, kann man nach Abschluß des Projekts nicht mehr feststellen, welche Ergebnisse durch welche Bedingungen verursacht worden sind. Daher muß man bis nach Abschluß des Experiments der Versuchung widerstehen, die erfolgreichen Aktivitäten der Versuchs- *oder* der Kontrollgruppe den an diesem Versuch in einer der beiden Gruppen teilnehmenden Schülern zugute kommen zu lassen, selbst wenn die Aktivitäten in einer Gruppe der Schüler sehr unzureichend sind.

Schließlich muß ein Instrument, das für ein bestimmtes Kriterium valide und reliabel ist, nach einer gewissen Zeitspanne – im allgemeinen nach einem ganzen Programmablauf – Versuchspersonen aus beiden Gruppen des Experiments erneut vorgelegt werden. Wenn alle diese Bedingungen erfüllt würden, könnte man mit Hilfe statistischer Verfahren und Entscheidungsregeln unzweideutig bestimmen, ob es zwischen den Versuchs- und den Kontrollgruppen im Hinblick auf die interessierenden Variablen signifikante Unterschiede gegeben hat oder nicht.

Auf den ersten Blick scheint sich die Anwendung eines experimentellen

Versuchsplans auf Evaluationsprobleme zu empfehlen, da bisher experimentelle Forschung und Evaluation dazu verwendet worden sind, Hypothesen über die Auswirkungen bestimmter Programme zu überprüfen. In dieser Aufgabenbestimmung sind jedoch vier wichtige Probleme enthalten:

Erstens gerät die Anwendung eines experimentellen Versuchsplans auf Evaluationsprobleme mit der Aufgabe der Evaluation, zur kontinuierlichen Verbesserung eines Bildungsprogramms beizutragen, in Konflikt. Ein experimenteller Versuchsplan verhindert die Modifikation der Versuchsbedingungen eher, als daß er sie fördert, weil die Versuchs- und Kontrollbedingungen im Verlauf des Versuchs nicht verändert werden dürfen, ohne daß die Daten über die Unterschiede zwischen den Versuchs- und den Kontrollgruppen verfälscht werden. Somit richten sich die Versuchs- und die Kontrollbedingungen nach dem Evaluationsplan anstatt umgekehrt; der Versuchsplan verhindert also eher Veränderungen in den Bedingungen, als daß er sie fördert.

Man kann nicht erwarten, daß die Leiter von Innovationsprojekten sich den Bedingungen eines experimentellen Versuchsplans unterwerfen. Sie können die Entwicklung ihres Projekts nicht im Anfangsstadium lassen, nur um am Jahresende Evaluationsdaten mit innerer Validität zu haben. Die Projektleiter müssen vielmehr alle erhältlichen Daten verwenden, um den Projektplan und seine Implementation kontinuierlich zu verbessern oder ihn in einigen Fällen von Grund auf zu verändern. Deshalb braucht man Konzeptionen von Evaluationsprogrammen, die eine dynamische Entwicklung der Bildungsprogramme fördern und nicht hemmen.

Eine zweite Unzulänglichkeit des experimentellen Versuchsplans besteht darin, daß er zwar nach Abschluß eines Projekts Daten für Entscheidungen zur Verfügung stellt, daß er jedoch für Entscheidungen während der Planung und Implementation eines Projekts fast nutzlos ist. Er liefert nach Abschluß des Versuchs Daten über die relative Wirkung der Programme in den Versuchs- und Kontrollgruppen. Solche Daten sind jedoch weder genügend spezifisch und umfassend, noch stehen sie zur rechten Zeit zur Verfügung, um dem Entscheidungsträger bei der Bestimmung der Projektziele und des Projektplans oder bei der Modifikation seiner Implementation behilflich zu sein. Bestenfalls zeigen experimentelle Versuchspläne hinterher, ob ein Projekt seine Ziele erreicht hat. Dann ist es jedoch zu spät, Entscheidungen über Pläne und Verfahren zu treffen, die bereits weitgehend den Erfolg oder Mißerfolg eines Projekts bestimmt haben.

Guba hat auf ein drittes mit dem experimentellen Versuchsplan verbundenes Problem hingewiesen: Dieser Plan eignet sich eher für die antiseptischen Bedingungen des Laboratoriums als für die septischen Bedingungen

der Schule[6]. Die potentiellen Störvariablen (confounding variables) müssen entweder kontrolliert oder durch Randomisierung eliminiert werden, wenn die Ergebnisse eine inhaltliche Validität haben sollen. Im Bereich der Erziehung gelingt dies jedoch fast nie. Untersuchen wir z. B. das folgende Zitat aus einem Evaluationsbericht von Julian Stanley (1966):

»Selbst wenn das Programm einen hohen positiven Einfluß auf die Berufslaufbahn einer Person hat, tritt dieser vielleicht nur langsam in Erscheinung und kann so mit anderen Einflüssen verbunden sein, daß ihn selbst die betreffende Person nicht deutlich erkennen kann. Dennoch müssen wir alle Daten zu der Entscheidung darüber heranziehen, ob sich die wiederholte Benutzung von bestimmten Bildungsprogrammen empfiehlt, bzw. welche ihrer Teile modifiziert werden müssen, um ihre Wirkung zu verbessern. Bei für einen experimentellen Versuchsplan optimalen Bedingungen müßten wir das Programm als kontrolliertes Experiment mit einer parallelisierten Kontrollgruppe, die nicht an dem Sommerkurs teilnimmt, durchführen, müßten dann beide Gruppen über mehrere Jahre hinweg weiter verfolgen, um die Unterschiede zwischen ihnen bestimmen zu können. Wenn die Auswahl der Teilnehmer für den Sommerkurs rechtzeitig beginnt und die Gruppe der Bewerber so groß ist, daß aus ihr zwei Gruppen mit genügend hohem Niveau gebildet werden können, kann der experimentelle Versuchsplan verwendet werden. Dennoch können auch in diesem Fall die Reaktionen der abgewiesenen Bewerber und die fehlende Möglichkeit, ihre Aktivitäten während des Sommerkurses zu kontrollieren, eine unerwünschte Auswirkung auf das Ergebnis des Experiments haben. Die Teilnahme an einem angesehenen Programm bestätigt zu bekommen, dürfte bereits eine wirkungsvolle Hilfe sein. ... Unser wichtigstes Mittel für die Evaluation des Programms waren die Berichte der Projektmitarbeiter und vor allem der Teilnehmer.«

In diesen Ausführungen hat Stanley viele Gründe dafür genannt, warum ein experimenteller Versuchsplan sich nicht für die Evaluation im Bildungswesen eignet. In zahlreichen innovativen Programmen gibt es zu viele Störfaktoren, die sich nicht wirksam kontrollieren lassen.

Die von Stanley z. B. erwähnten Störfaktoren weisen auf ein viertes Problem bei der Anwendung des experimentellen Versuchsplans hin. *Während die innere Validität (internal validity) durch die Kontrolle äußerer Variablen (extraneous variables) erreicht werden kann, wird dieser Gewinn allerdings auf Kosten der äußeren Validität (external validity) erreicht.* Wenn die äußeren Variablen streng kontrolliert werden, sind die Ergebnisse so lange zuverlässig, wie eine Innovation unter kontrollierten Bedingungen stattfindet. Die Ergebnisse solcher Untersuchungen können jedoch nicht auf die Schulwirklichkeit übertragen werden, in der die äußeren Variablen sich nicht kontrollieren lassen. Deshalb gilt es in Erfahrung zu bringen, wie pädagogische Innovationen sich in der Schulwirklichkeit bewähren.

Bislang habe ich in diesem Beitrag den gegenwärtigen Stand der Be-

mühungen um Evaluation im Bildungswesen darzustellen versucht. Am Anfang meiner Ausführungen legte ich dar, daß Pädagogen mit vielen neuen und unterschiedlichen Forderungen nach Evaluation konfrontiert werden. Sodann versuchte ich nachzuweisen, daß die Anstrengungen der Pädagogen, diesen Ansprüchen zu genügen, bislang nicht ausreichen. Schließlich habe ich auf drei Unzulänglichkeiten hingewiesen, die Pädagogen daran hinderten, ergiebige Evaluationsuntersuchungen zu machen:

(1) Fehlende Kenntnis der Entscheidungsprozesse und Informationen, die bei der Entwicklung innovativer Bildungsprogramme benötigt werden,
(2) Fehlen einer Definition der Evaluation, die den sich abzeichnenden Forderungen nach Evaluation im Bildungswesen gerecht wird,
(3) Fehlen angemessener Evaluationspläne

II. Das Wesen der Evaluation

Da dies ein Arbeitsbericht ist, sollte ich nicht länger die gegenwärtigen Bedürfnisse und Probleme der Evaluation behandeln. Man kann meine Ausführungen überprüfen, modifizieren oder ablehnen. Sobald man zu einem Konsens darüber gekommen ist, welches die wirklichen Probleme der Evaluation sind, könnte man relevante Lösungen entwickeln. Hier sollte ich einige Vorstellungen zur Lösung der gegenwärtigen Schwierigkeiten vortragen. Daher werde ich im Verlauf dieses Beitrags einige alternative Konzepte der pädagogischen Evaluation entwickeln.

Der folgende Teil dieses Beitrags gliedert sich in vier Abschnitte. Im ersten soll Evaluation ganz allgemein definiert werden. Im zweiten werden neuere innovative Programme analysiert und die Arten der Entscheidungen identifiziert, für die in diesen Programmen Evaluationsuntersuchungen benötigt werden. Der dritte Teil enthält einen Überblick über vier Strategien zur Evaluation von Bildungsprogrammen. Der Beitrag schließt im vierten Teil mit dem Versuch, die Struktur von Evaluationsplänen zu entwickeln.

Merkmale der Evaluation
Eine rationale Begründung (Rationale)

Wenn Entscheidungsträger ihre Möglichkeiten maximal ausnutzen wollen, müssen sie vernünftige Entscheidungen über vorliegende Alternativen treffen. Dazu müssen sie jedoch zunächst wissen, welche Alternativen ihnen zur Verfügung stehen. Sie müssen außerdem in der Lage sein, begründete Urteile über den relativen Wert der Alternativen abzugeben. Dies erfor-

dert jedoch relevante Informationen. Die Entscheidungsträger sollten daher über wirksame Mittel verfügen, mit deren Hilfe evaluative Informationen gewonnen werden können. Anderenfalls dürften ihre Entscheidungen von vielen unerwünschten Elementen abhängen. Günstigstenfalls sind die Urteile lediglich von Sympathien, Vorurteilen und Interessen abhängig. Häufig gibt es dabei eine Tendenz, persönliche Erfahrungen, Gerüchte und die Ansicht einer Autorität überzubewerten; so werden gewiß zuviel Entscheidungen getroffen, ohne daß die möglichen Alternativen bekannt sind.

Die Qualität von Programmen hängt von der Qualität der Entscheidungen in den Programmen und über die Programme ab; die Qualität der Entscheidungen wird durch die Fähigkeit der Entscheidungsträger bestimmt, die Alternativen zu identifizieren, die in Entscheidungssituationen auftreten, und entsprechend vernünftige Urteile über sie zu fällen; vernünftige Urteile bedürfen valider und reliabler Informationen über die Alternativen; um solche Informationen zu erhalten und den Entscheidungsträgern zur Verfügung zu stellen, braucht man systematische Verfahren. Die Prozesse, mit deren Hilfe die für die Entscheidungen erforderlichen Informationen gewonnen werden, müssen Teil des Evaluationskonzepts sein. Auf diesen Ausführungen aufbauend, möchte ich eine Definition von Evaluation entwickeln.

Definition der Evaluation

Im allgemeinen bedeutet Evaluation die Gewinnung von Informationen durch formale Mittel wie Kriterien, Messungen und statistische Verfahren mit dem Ziel, eine rationale Grundlage für das Fällen von Urteilen in Entscheidungssituationen zu erhalten. Zur Erläuterung dieser Definition sollen einige zentrale Begriffe erklärt werden:

Eine Entscheidung ist eine Wahl zwischen Alternativen.

Eine Entscheidungssituation besteht aus einer Reihe von Alternativen.

Ein Urteil fällen bedeutet, die Alternativen zu bewerten.

Ein Kriterium ist ein Maßstab, aufgrund dessen die Alternativen bewertet werden; im Idealfall umfaßt ein Kriterium die Spezifikation von Variablen, Messungen und Normen für die Beurteilung des Untersuchungsgegenstandes.

Statistik ist die Wissenschaft von der Analyse und Interpretation einer Reihe von Meßwerten.

Unter Messung wird die Übertragung von Zahlen auf Einheiten aufgrund bestimmter Regeln verstanden; nach solchen Regeln erfolgt im allgemeinen die Spezifikation der Stichprobenelemente, der Meßverfahren und der Bedingungen für die Durchführung und Beurteilung der Meßverfahren.

Vereinfacht gesagt, ist Evaluation also die Wissenschaft, mit deren Hilfe Informationen für Entscheidungsprozesse zur Verfügung gestellt werden.

Zur Methodologie der Evaluation gehören vier Funktionen: *Sammlung, Organisation, Analyse* und *Bericht* von Informationen. Zu den Kriterien für die Einschätzung der Angemessenheit der Evaluation gehören *Validität* (Ist die Information diejenige, die der Entscheidungsträger braucht?), *Reliabilität* (Ist die Information reproduzierbar?), *Rechtzeitigkeit* (Steht die Information für den Entscheidungsträger rechtzeitig zur Verfügung?), *Verfügbarkeit* (Erreicht die Information alle Entscheidungsträger, die sie brauchen?) und *Zuverlässigkeit* (Vertrauen die Entscheidungsträger der Information?).

Evaluation außerhalb des Bildungswesens

Das bisher entwickelte Evaluationskonzept ist sehr allgemein gefaßt, da das Bewerten von Alternativen in allen Bereichen des menschlichen Lebens üblich ist und da Menschen immer bestrebt sind, rational vertretbare Grundlagen für ihre Urteile zu erhalten. Es lassen sich jedoch zahlreiche Arten der Evaluation, die alle die Bedingungen der genannten Definition erfüllen, voneinander unterscheiden. So ist z. B. auch für Marktforschung, Kosten-Nutzen-Analyse (cost-benefit analysis), experimentelle Versuchsplanung, objektive Testverfahren, militärwissenschaftliche Forschung, Planungsforschung, Program Evaluation and Review Technique (PERT), Planning Programming and Budgeting System (PPBS), Qualitätskontrolle und Systemanalyse die erwähnte allgemeine Definition der Evaluation gültig.

Für jedes dieser Forschungsverfahren ist die Anwendung systematischer Verfahren zur Bewertung von Alternativen in Entscheidungssituationen charakteristisch. Diese verschiedenen Arten der Evaluation lassen sich nach Entscheidungssituationen, Entscheidungsbedingungen, Art der verwendeten Instrumente und Verfahren, Ausmaß der Präzision bei der Sammlung und Analyse von Informationen und den methodischen Fähigkeiten der Evaluatoren und ihrer Adressaten unterscheiden. Diese inhaltlichen und methodischen Unterschiede erklären wahrscheinlich, warum die verschiedenen Formen der Evaluation unterschiedliche Namen haben. Das wird z. B. auch aus folgenden Ausführungen Quades (1967,4) deutlich: »Evaluationsuntersuchungen, die Entscheidungsträgern bei der Wahl zwischen Systemen und bei der Ermittlung ihrer Effektivität im Hinblick auf ihre Ziele oder bei der Entwicklung eines Bezugsrahmens für ihre Erforschung helfen sollen, können selbstverständlich Systemanalysen genannt werden.«

Obwohl Quade behauptet, daß Systemanalyse eine Form der Evaluation

ist, erkennt er zugleich auch, daß die Bezeichnung Systemanalyse wegen der spezifischen Beschaffenheit dieser Art der Evaluation gewählt worden ist.

Betrachtet man die Entstehung der genannten Formen der Evaluation, wird deutlich, daß alle für spezifische Anwendungsbereiche entwickelt worden sind. Program Evaluation and Review Technique (PERT) wurde entwickelt, um dem Militär bei der Entscheidung über die Entwicklung komplexer Waffensysteme zu helfen. Systemanalyse entstand, um dem Militär die Entscheidung über die Entwicklung und Durchführung militärischer Operationen zu erleichtern. Objektive Testverfahren werden beim Militär vor allem zur Auswahl für den Wehrdienst eingesetzt.

Diese Formen der Evaluation wurden rasch entwickelt, da ein großes Bedürfnis nach begründbaren Entscheidungen bestand; sie entsprechen daher auch der Art der erforderlichen Entscheidungen und den Entscheidungsbedingungen. Neue Ansätze der Evaluation wurden entwickelt, weil die bestehenden für die Entscheidungsprozesse nicht genügend Informationen liefern und einmal getroffene falsche Entscheidungen ernsthafte Konsequenzen haben konnten. Militärische Entscheidungen können den Ausgang eines Krieges beeinflussen; also wurden entsprechende Verfahren der militärwissenschaftlichen Forschung, Systemanalyse usw. entwickelt. Wirtschaftliche Entscheidungen können zum Gewinn, Verlust oder Bankrott von Tausenden von Aktionären führen; also wurde die Kosten-Nutzen-Analyse entwickelt.

Evaluation im Bildungswesen

Bislang hatten Entscheidungen im Bildungswesen weniger deutliche Auswirkungen als Entscheidungen in Wirtschaft, Landwirtschaft und Militär. Deshalb fanden auch im Bildungswesen geringere Anstrengungen statt, hochspezialisierte Formen der Evaluation zu entwickeln, um die zahlreichen unterschiedlichen Bildungsentscheidungen zu unterstützen. Die meisten Pädagogen hatten erhebliche Schwierigkeiten, die wichtigsten pädagogischen Entscheidungssituationen zu identifizieren, die im Rahmen der Evaluation besonderer Behandlung bedürfen. Man darf daraus jedoch nicht schließen, daß es in der Pädagogik bislang keine Evaluationsverfahren gegeben hat. Standardisierte Tests wurden bei Entscheidungen über Hochschulzulassungen zur Hilfe herangezogen; sie dienten als Grundlage zur Notengebung, Einstufung der Schüler in das Curriculum und Vergabe von Diplomen. Die Buros Mental Measurement Yearbooks (1965) wurden herausgegeben, um den Pädagogen bei der Auswahl und bei der Anwendung von Tests zu helfen. Kürzlich wurde der Educational Product Information

Exchange (EPIE) eingerichtet[7], um Pädagogen bei der Auswahl alternativer Unterrichtsmaterialien behilflich zu sein. Abgesehen davon wurden im Bildungswesen bislang jedoch keine speziellen Verfahren entwickelt, die bei Entscheidungen über Bildungsprogramme behilflich sein könnten.

Im Bildungswesen war es durchaus üblich, auch andere Bereiche zu beachten, in denen ähnliche Probleme in Angriff genommen und gelöst wurden. Deshalb haben auch die Pädagogen den experimentellen Versuchsplan als einen Evaluationsplan adaptiert. Dabei wird allerdings ein Verfahren, das zunächst den Bauern bei der Unterscheidung zwischen verschiedenen Arten von Düngemitteln und Saaten helfen sollte, im Erziehungswesen benutzt, um eine Auswahl zwischen alternativen Bildungsinnovationen zu treffen. Offensichtlich ist die Ähnlichkeit zwischen pädagogischen Innovationen und Düngemitteln jedoch außerordentlich gering.

In letzter Zeit hat man die Program Evaluation and Review Technique (PERT), die Systemanalyse und das Planning Programming and Budgeting System (PPBS) im Erziehungswesen angewandt. Obwohl ausgewählte Verfahren aus anderen Bereichen den Pädagogen helfen können, Zeit und Mühen zu sparen, möchte ich aber auch davor warnen, Verfahren aus anderen Bereichen unkritisch zu übernehmen. Anderenfalls könnte es zu einer unangemessenen Anwendung solcher Verfahren auf pädagogische Probleme kommen. Meiner Ansicht nach ist die Anwendung des experimentellen Versuchsplans zur Evaluation innovativer Programme ein Beispiel für die unkritische Übernahme in anderem Zusammenhang entwickelter Verfahren. Die Verwendung des experimentellen Versuchsplans in diesem Kontext hat Pädagogen viel Zeit und Anstrengung gekostet, ohne ihnen bei den Entscheidungsprozessen sehr geholfen zu haben.

Wie bereits dargelegt, braucht meiner Meinung nach das Bildungswesen eine neue Konzeptualisierung, um eine Theorie und Methodologie der Evaluation entwickeln zu können, die für die pädagogischen Probleme relevant ist. Bislang habe ich nur eine allgemeine Begründung und Definition von Evaluation gegeben; im weiteren möchte ich eine rationale Begründung und Definition für Evaluation im Bildungswesen entwickeln.

Eine rationale Begründung der Evaluation im Bildungswesen

Die Programme der Titel I und III des Elementary and Secondary Education Act von 1965 bilden für die Entwicklung einer rationalen Begründung einer pädagogischen Evaluation einen komplexen Kontext. Fast alle Schulbezirke sind an einem oder an beiden Programmen beteiligt. Die Ziele dieser Programme bestehen darin, schulische Leistungen, Erfahrungen und Möglichkeiten sozial benachteiligter Schüler zu verbessern und das Aus-

maß und die Qualität der Innovationen in zahlreichen Bildungsinstitutionen zu erhöhen. Beide Programme finden in allen Teilen der USA Anwendung und sind entsprechend konzipiert. Sie werden auf der Ebene der Bundesstaaten koordiniert, kontrolliert und in den örtlichen Schulbezirken implementiert. Insgesamt stellen beide Programme den örtlichen Schulbezirken jährlich mehr als eine Milliarde Dollar zur Verfügung.

Abbildung 1 stellt eine Konzeptualisierung des Prozesses und der Funktion der Evaluation für Entscheidungsabläufe dar, wie sie in den bundesstaatlichen Programmen bestehen könnten. Eine Reihe von Kontrollschleifen veranschaulicht die Beziehungen zwischen den örtlichen, einzel- und bundesstaatlichen Evaluationsuntersuchungen, denen die Projekte der beiden Programme unterzogen werden. Die Schleife an der rechten Seite veranschaulicht örtliche, die mittlere einzelstaatliche und die linke bundesstaatliche Aktivitäten. Jede Schleife enthält eine Reihe von Blöcken, die die wichtigsten Evaluationsfunktionen repräsentieren.

Block 1 stellt das Bildungsprogramm eines Schulbezirks dar. Es bildet den Kontext, aus dem die Bedürfnisse nach pädagogischen Innovationen entstehen und in dem die Reformen schließlich realisiert werden müssen. Es enthält die *Inputs* des Systems, d. h. Schule, Curriculum, Lehrkörper, Organisation, Politik, Finanzen, schulische Anlagen, Beziehungen zwischen Schule und Gemeinde, und die *Outputs* des Systems, d. h. das kognitive, psychische, physische und soziale Befinden der Schüler und späteren Erwachsenen.

Das erste Segment des Umfangs rechts von Block 1 veranschaulicht die Informationssammlung. Sie findet auf der Ebene der örtlichen Schulbezirke als systematische Sammlung aller Informationen statt, die für spätere Entscheidungen auf der örtlichen, einzelstaatlichen und bundesstaatlichen Ebene benötigt werden.

Block 2 repräsentiert die Organisation der Informationen. Dabei werden die Informationen nach vorher bestimmten Kategorien kodiert, bearbeitet, systematisch gespeichert und im Bedarfsfall abgerufen.

In Block 3 werden die in Block 2 organisierten Informationen unter dem Aspekt der Vorbereitung von Entscheidungsprozessen auf örtlicher, einzelstaatlicher und bundesstaatlicher Ebene analysiert und den örtlichen und einzelstaatlichen Entscheidungsträgern berichtet.

Block 4 stellt die Programmentscheidungen dar, die auf örtlicher Ebene getroffen werden. Zu den örtlichen Entscheidungsträgern, denen die Ergebnisse der Evaluation zur Verfügung gestellt werden, gehören das Board of Education, die Schulverwaltung, der Projektleiter, die Lehrer und der Schulleiter.

Die Entscheidungen, die in Block 4 fallen, werden in Block 5 durchge-

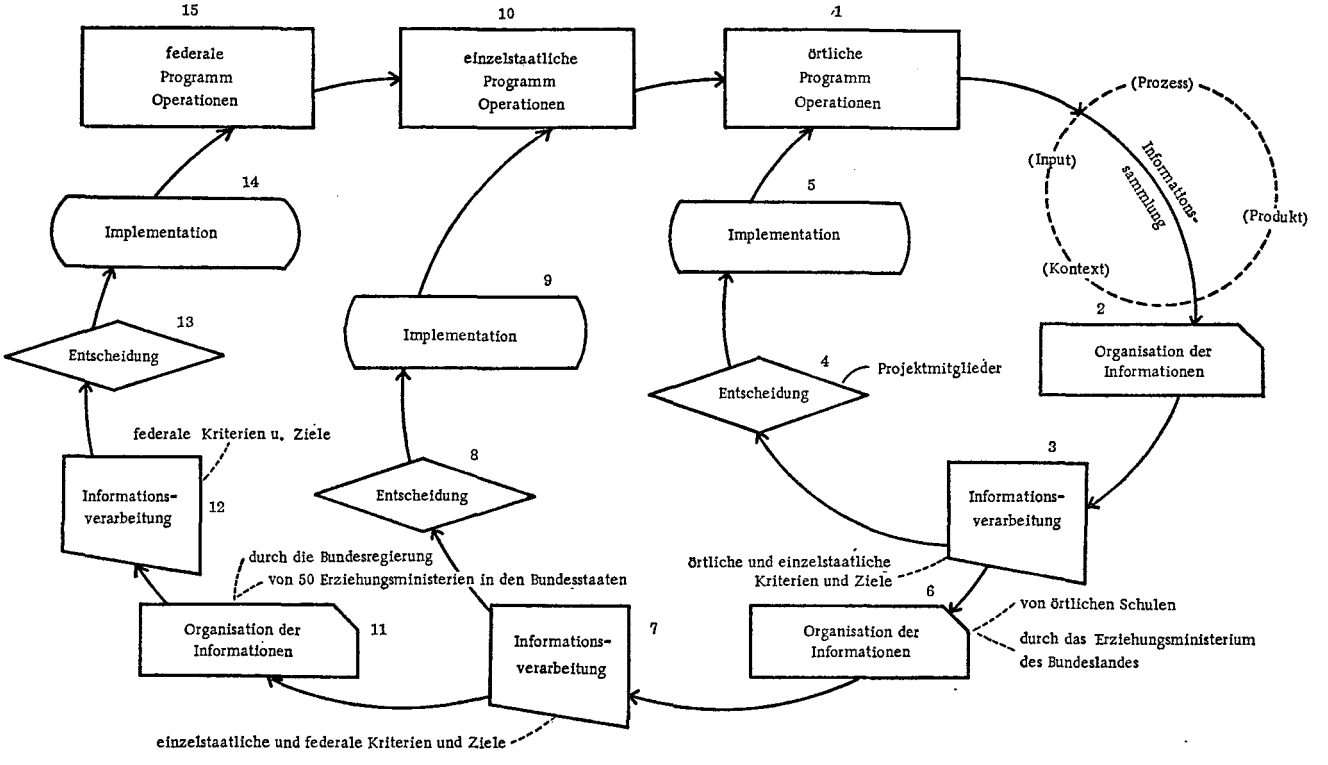


Abb. 1: Evaluation in vom Bund unterstützten Bildungsprogrammen (Stufflebeam 1967)

führt. So beginnt der Zyklus mit zahlreichen Modifikationen des Schulprogramms in Block 1 wieder von neuem.

In Block 3 wird dargestellt – um darauf zurückzukommen –, daß Evaluationsberichte für staatliche Erziehungsministerien jährlich von allen örtlichen Schulbezirken angefertigt werden sollen. In Block 6 würde das Erziehungsministerium eines Bundesstaates dann diese Berichte nach der Art der Projekte organisieren und zu den Informationen über ähnliche Projekte in Beziehung setzen. Diese Informationen würden dann in Block 7 analysiert werden, um die Stärken und Schwächen des Programms im ganzen Bundesstaat zu bestimmen. Die für das Programm in einem Bundesstaat verantwortlichen Beamten würden diese Informationen dazu nutzen, um die pädagogischen Bedürfnisse und Probleme in diesem Staat abzuschätzen, um dann Entscheidungen über Programmschwerpunkte und über die Kontrolle in Block 8 zu treffen. Entscheidungen, die in Block 8 gefällt werden, würden in Block 9 ausgeführt und würden wiederum das Bildungsprogramm des Einzelstaates in Block 10 berühren und den Zyklus in Block 1 wieder von neuem anfangen lassen.

In Block 7 würden jährlich Evaluationsberichte von 50 Staaten an die verantwortliche Bundesinstitution gesandt werden. Die Organisation der Informationen erfolgt dann in Block 11, so daß die wichtigsten Programmansätze aus allen Teilen des Landes in Block 12 auf der bundesstaatlichen Ebene überprüft und analysiert und die Berichte für den Associate Commissioner, der für den Elementary and Secondary Education Act verantwortlich ist, für den Minister, den Präsidenten und den Kongreß vorbereitet werden können. Entscheidungen über Programmschwerpunkte und Finanzen würden auf bundesstaatlicher Ebene in Block 13 gefällt; die Implementation solcher Entscheidungen in Block 14 oder das bundesstaatliche Programm in Block 15 berühren das einzelstaatliche Programm in Block 10 und die örtlichen Projekte der Schulen in Block 1. Somit würde der Zyklus von neuem beginnen.

Zusammengefaßt stellt die Abbildung 1 folgenden Prozeß dar:

(a) Die Informationen auf bundesstaatlicher, einzelstaatlicher und örtlicher Ebene werden weitgehend in den örtlichen Schulbezirken gesammelt.
(b) Diese Informationen stellen die Grundlage für bundesstaatliche, einzelstaatliche und örtliche Entscheidungen dar, die schließlich das Handeln in den örtlichen Schulbezirken beeinflussen.
(c) Evaluationspläne müssen auf der bundesstaatlichen, einzelstaatlichen und örtlichen Ebene entwickelt, verbreitet und koordiniert werden, wenn die Informationen angemessen sind, die die Schulen für die Unterstützung des Entscheidungsprozesses auf allen drei Ebenen zur Verfügung stellen.

Um ein angemessenes Evaluationssystem für Programme wie Titel I und Titel III zu entwickeln, braucht man zunächst einige Kenntnisse der gegebenen Entscheidungssituationen. Die Kenntnis dieser Entscheidungssituationen sollte die Beantwortung einer Reihe von Fragen ermöglichen. Erstens sollte man die Stelle, bzw. *Ebene* der Entscheidungen identifizieren, auf der die Autorität und Verantwortung für die Entscheidungen liegt, d. h. man muß bestimmen, ob die Entscheidungen auf der Ebene der Schulen, der Einzelstaaten oder des Bundes erfolgt. Zweitens sollte man den *Schwerpunkt* (focus) der Entscheidungen bestimmen, also z. B. die Frage, inwieweit sich die Entscheidungen auf die Ziele der Forschung, Entwicklung, Lehrerbildung und Implementation beziehen. Drittens muß man den *Inhalt* der Entscheidungen kennen und wissen, ob sie sich auf Mathematik, Sprachen, Kunst u. a. beziehen und welche Alternativen es in jeder Entscheidungssituation gibt. Viertens muß man die *Funktion* der Entscheidungen kennen und wissen, ob sie sich auf die Planung, das Programm, die Implementation oder die wiederholte Verwendung des Programms beziehen. Fünftens muß man mit dem *Gegenstand* der Entscheidungen (z. B. Personen, Ereignisse oder Dinge) vertraut sein. Sechstens muß man den *Zeitpunkt* der Entscheidungen genau kennen. Siebentens muß man das Ausmaß an *kritischer Reflektiertheit* der Entscheidungen identifizieren.

Wenn man alle genannten Entscheidungsvariablen berücksichtigt, lassen sich viele verschiedene Entscheidungssituationen im Bildungswesen identifizieren. Deshalb kann man auch mehrere Arten der Evaluation unterscheiden. Aus diesem Grunde sollte man ein Klassifikationssystem für die verschiedenen Arten der pädagogischen Evaluation entwickeln, das die allgemeine konzeptuelle Evaluationsdefinition mit den zahlreichen spezifischen Arten der Evaluation in Beziehung setzt, die sich in einer detaillierten Analyse und Klassifikation pädagogischer Entscheidungssituationen aus einer Berücksichtigung der genannten Variablen gewinnen lassen. Sodann gilt es schließlich, für die identifizierten Klassen pädagogischer Evaluation brauchbare Begriffe zu finden.

Um ein Klassifikationssystem für pädagogische Entscheidungssituationen und Programme zu erarbeiten, konzentrierte ich mich anfangs ausschließlich auf die Funktion der Entscheidungen (Stufflebeam 1967). Meiner Ansicht nach lassen sich die Funktionen von Entscheidungen im Bildungswesen als *Planung*, *Programmgestaltung*, *Implementation* und *modifizierte Programmwiederholung* klassifizieren. *Planungsentscheidungen* richten sich auf erforderliche Reformen und präzisieren ihren Bereich und ihre allgemeinen und spezifischen Ziele. *Programmentscheidungen* richten sich auf die Verfahren, die beteiligten Personen und die zeitlichen und finanziellen Bedingungen für die Implementation der geplanten Akti-

vitäten. *Implementationsentscheidungen* richten sich auf die im Programm intendierten Handlungen. Zu den Entscheidungen, die mit der Frage der *wiederholten, bzw. modifizierten Verwendung des Programms* zusammenhängen, gehören diejenigen über die Beendigung, Weiterführung, Entwicklung oder Veränderung des Programms.

Vier Strategien zur Evaluation von Bildungsprogrammen

In Entsprechung zu diesen vier Arten von Bildungsentscheidungen gibt es auch vier Arten von Evaluation. Sie werden in Tabelle 1 als Kontext-, Input-, Prozeß- und Produktevaluation bezeichnet. *Kontextevaluation* findet während der ersten Phase der Projektplanung statt. *Inputevaluation* erfolgt gleich danach bei der spezifischen Planung des Programms. *Prozeßevaluation* findet während der Implementation des Projekts statt. *Produktevaluation* erfolgt im allgemeinen nach der Beendigung des Projekts. Diese vier Formen der Evaluation sollen im folgenden weiter entwickelt werden.

Kontextevaluation

Das Hauptziel der Kontextevaluation besteht darin, die Voraussetzungen, unter denen eine Reform erfolgt, und die unbefriedigten Bedürfnisse der Umwelt und die mit ihnen verbundenen Probleme zu bestimmen. Die Umwelt kann z. B. aus den Grundschulen des Zentrums einer großen Stadt bestehen. Die Erforschung dieser Bedingungen könnte ergeben, daß die wirklichen Leseleistungen der Schüler in diesem Schulbezirk weit unter den Erwartungen des Schulsystems liegen. Damit wäre ein Mangel identifiziert, d. h., die Kontextevaluation hätte ergeben, daß die Leseleistungen der Schüler verbessert werden müssen.

Als ersten Schritt in der Kontextevaluation müßten die Schulen die Gründe für die mangelnden Leseleistungen zu erkennen versuchen. Ist der Unterricht der Schüler angemessen? Entspricht das Unterrichtsmaterial ihren Bedürfnissen? Gibt es Sprachbarrieren? Bleiben die Schüler dem Unterricht fern? Sind die Erwartungen der Schulen an diese Schüler erfüllbar? Dies sind meiner Ansicht nach mögliche Probleme und Schwierigkeiten, die verhindern, daß die angestrebten Ziele erreicht werden, und die dadurch dazu führen, daß solche Unzulänglichkeiten entstehen.

Bei der Kontextevaluation beginnt man mit einer konzeptuellen Analyse, um den Untersuchungsbereich mit seinen wichtigsten Teilbereichen zu identifizieren und zu begrenzen. Sodann werden empirische Untersuchungen mit Stichprobenerhebungen, Umfragen und standardisierten Tests durchgeführt. Das Ziel der Kontextevaluation besteht darin, die Diskre-

Tabelle 1: Das CIP-Evaluationsmodell – Ein Klassifikationsschema der Strategien zur Evaluation pädagogischer Reformen (Stufflebeam 1967)

	Die Strategien			
	Kontextevaluation	*Inputevaluation*	*Prozeßevaluation*	*Produktevaluation*
Ziel	Definition des *Programmkontexts*, Identifikation und Einschätzung der *Bedürfnisse* in dem Kontext und Identifikation und Beschreibung der *Probleme*, die mit den Bedürfnissen verbunden sind.	Identifikation und Abschätzung der Systemmöglichkeiten, der verfügbaren Input-*Strategien* und der *Pläne* zur Implementation der Strategien.	Identifikation oder Voraussage, der Unzulänglichkeiten des den Prozeß steuernden Plans oder seiner Implementation und die Aufzeichnung von *Ereignissen und Aktivitäten des Prozesses.*	In-Beziehung-setzen der *Ergebnisse* mit den Lernzielen, dem Kontext, dem Input und dem Prozeß.
Methode	Individuelle Beschreibung der wichtigsten Teilsysteme des Kontexts unter relevanten Gesichtspunkten; Vergleich wirklicher und beabsichtigter Inputs und Outputs der Teilsysteme; Analyse möglicher Gründe für die Diskrepanz zwischen Wirklichkeit und Intention.	Beschreibung und Analyse der verfügbaren menschlichen und materiellen Ressourcen, Lösungsstrategien und Verfahrenspläne in bezug auf Relevanz, *Durchführbarkeit* und Wirtschaftlichkeit während der Durchführung.	Beachtung der möglichen im Prozeß auftretenden Hindernisse, der Aktivitäten und *ständige* Aufmerksamkeit gegenüber *unerwarteten* Hindernissen.	Operationale Definition und Messung der mit den Zielen verbundenen Kriterien durch Vergleich der Meßwerte mit im voraus bestimmten Normen und durch Interpretation des Ergebnisses in bezug auf die aufgezeichneten Input- und Prozeßinformationen.
Beziehung zum Fällen von Entscheidungen im Reformprozeß	Entscheidungen über die *Ausgangsbedingungen*, die *Ziele*, die zur Verbesserung der Situation dienen sollen, und die *Lernziele*, die zur Problemlösung, d. h. zur Planung der benötigten Reformen bestimmt sind.	Auswahl der *Finanzierungsquellen*, der *Lösungsstrategien* und *Verfahrenspläne*, d. h. systematische Planung der Reformaktivitäten.	*Implementation und Verbesserung des Programmplans* und des Verfahrens, um z. B. den Verlauf wirksam zu kontrollieren.	Entscheidung über die *Weiterentwicklung, Beendigung, Modifikation oder Schwerpunktverlagerung* einer Reformaktivität und Verbindung der Aktivität mit anderen wichtigen Phasen des Reformprozesses, z. B. neu in Erscheinung tretenden Reformaktivitäten.

panzen zwischen intendierten und wirklichen Situationen für alle Teilbereiche des untersuchten Gesamtbereichs darzulegen und somit die benötigten Reformen zu identifizieren. Schließlich gehören empirische und konzeptuelle Analysen, Theorien und Ansichten von Autoritäten zur Kontextevaluation, um die mit dem Mangel und den Unzulänglichkeiten verbundenen Probleme zu beurteilen.

Zu den Entscheidungen, für die die Kontextevaluation Daten zur Verfügung stellen soll, gehören Entscheidungen über die Ziele, die sich aus den Bedürfnissen ergeben und mit deren Hilfe die Probleme gelöst werden sollen. Derartige Entscheidungen werden im allgemeinen in einleitenden Abschnitten von Projektanträgen an Ministerien oder Stiftungen sichtbar.

Inputevaluation

Um über die Verteilung von Ressourcen zur Realisierung von Programmzielen zu entscheiden, bedarf es einer Inputevaluation. Ihr Ziel besteht darin, die relevanten Möglichkeiten des Antragstellers, die Strategien und Pläne zur Realisierung der Programmziele und der entsprechenden Lernziele zu identifizieren und zu beurteilen. Das Ergebnis einer Inputevaluation ist die Analyse von alternativen Verfahrensplänen im Hinblick auf mögliche Kosten und möglichen Gewinn.

Insbesondere werden die alternativen Pläne unter Bezug auf die folgenden Aspekte beurteilt: Ressourcen, Zeit, eventuelle Unzulänglichkeiten bei den intendierten Realisierungsverfahren, die Möglichkeiten und Kosten ihrer Überwindung, die Relevanz von Plänen im Hinblick auf Programmziele und die Gesamtmöglichkeiten des Plans für die Realisierung der Programmziele. Im allgemeinen liefert Inputevaluation Informationen für eine Entscheidung darüber, ob für die Realisierung der Ziele äußere Hilfe in Anspruch genommen werden soll und welche Strategien dabei gewählt werden müssen, ob also z. B. bereits verfügbaren Bildungsmöglichkeiten oder der Entwicklung neuer Strategien der Vorzug zu geben ist, und welcher Plan oder welche Vorgehensweise für die Implementation der ausgewählten Strategien verwendet werden soll.

Bislang fehlen im Bildungswesen Methoden der Inputevaluation. Zu den bisher verbreiteten Praktiken gehören Kommissionsberatungen, eine Berücksichtigung der Fachliteratur und die Befragung von Experten. In einigen Bereichen gibt es bereits formale Instrumente, um den Entscheidungsträgern bei Inputentscheidungen behilflich zu sein. Bei der Planung von Testprogrammen kann man in Buros Mental Measurements Yearbooks (1949) eine wesentliche Hilfe finden.

Der pädagogische Forscher, der einen experimentellen Versuchsplan

wählen möchte, kann für die Identifikation und Beurteilung alternativer experimenteller Pläne in dem Kapitel über experimentelle Versuchspläne in Gages Handbook of Research on Teaching (1963) erhebliche Hilfen finden. In diesem Kapitel werden für den Forscher, der vor der Entscheidung über einen experimentellen Versuchsplan steht, die relevanten Alternativen experimenteller Forschung ausführlich dargestellt. Alle diese experimentellen Versuchspläne werden in bezug auf die Kriterien innerer und äußerer Validität beurteilt. Sodann werden für alle erwähnten Versuchspläne mögliche Probleme und Schwierigkeiten in der Durchführung angegeben.

Entscheidungen aufgrund von Inputevaluation führen im allgemeinen in den Anträgen an die Ministerien und Stiftungen zu einer Spezifikation der Verfahren, Materialien, Zeitpläne, Stellenanforderungen und des Budgets. Die Anträge werden von den Geldgebern wieder einer Inputevaluation unterzogen, um danach über eine Finanzierung der vorgeschlagenen Projekte zu entscheiden. Stiftungen und Ministerien haben im allgemeinen für ihre Inputevaluation Experten als Berater und Beurteiler.

Prozeßevaluation

Wenn die Richtung des Vorgehens bestimmt worden ist und die Implementation des Plans begonnen hat, brauchen Projektleiter und die anderen Verantwortlichen eine systematische Prozeßevaluation, um dadurch eine kontinuierliche Kontrolle und Verbesserung der Pläne und Verfahren bewirken zu können. Die Aufgabe der Prozeßevaluation besteht darin, während der einzelnen Stadien der Implementation Unzulänglichkeiten im Verfahrensplan oder in seiner Durchführung zu entdecken oder vorauszusagen. Die Gesamtstrategie zielt darauf, die möglichen Ursachen für Fehlschläge in einem Projekt zu identifizieren: Zu diesen möglichen Ursachen gehören die interpersonellen Beziehungen zwischen den Mitarbeitern, die Kommunikationsstrukturen, das Verständnis und die Unterstützung der Intentionen des Programms durch die Programmentwickler und die Adressaten sowie die Angemessenheit der Ressourcen, der schulischen Anlagen, der zeitlichen Planung und die Eignung der Mitarbeiter.

Im Unterschied zur Evaluation mit einem experimentellen Versuchsplan erfordert Prozeßevaluation weder eine kontrollierte Zuordnung der am Versuch beteiligten Personen in Versuchs- und Kontrollgruppen noch konstant gehaltene Versuchsbedingungen. Ihre Aufgabe besteht darin, den Mitarbeitern des Projekts dabei zu helfen, ihre alltäglichen Entscheidungen ein wenig rationaler zu fällen, um so die Qualität der Programme zu verbessern. In der Prozeßevaluation ist der Evaluator bereit, an dem Pro-

gramm in seiner augenblicklichen und modifizierten Form mitzuarbeiten und die Gesamtsituation so gut wie möglich zu berücksichtigen. Dabei sollte er sich bemühen, bei den wichtigsten Aspekten des Projekts möglichst empfindliche und nicht intervenierende Verfahren der Datensammlung zu verwenden. Eine solche Evaluation ist multivariat; nicht alle wichtigen Variablen lassen sich vor dem Beginn des Projekts spezifizieren. Der Prozeßevaluator konzentriert sich vor allem auf die in der Theorie des Programms entwickelten Variablen, aber er muß auch bereit sein, seine Aufmerksamkeit auf unerwartete, aber wichtige Ereignisse zu richten. Bei einer Prozeßevaluation werden die Informationen täglich gesammelt, systematisch organisiert, periodisch, d. h. z. B. wöchentlich, analysiert und so oft vorgetragen, wie die Projektmitglieder solche Informationen anfordern.

Dadurch erhalten die Entscheidungsträger eines Projekts nicht nur die Informationen, die sie für die Antizipation und Überwindung prozeßbedingter Schwierigkeiten brauchen, sondern auch einen Bericht über den Prozeß der Implementation, der auch für die spätere Interpretation der Projektergebnisse verwendet werden kann.

Produktevaluation

Produktevaluation dient dazu, nach Beendigung des Projekts seine Wirksamkeit festzustellen. Ihre Aufgabe besteht darin, die Ergebnisse auf die Ziele, den Kontext, den Input und den Prozeß zu beziehen, d. h. die Ergebnisse zu messen und entsprechend zu interpretieren. Dazu muß man die Kriterien, die zu den Intentionen einer Handlung gehören, operational festlegen und messen, die Meßwerte mit den im voraus bestimmten absoluten oder relativen Normen vergleichen und die Ergebnisse mit Hilfe der aufgezeichneten Kontext-, Input- und Prozeßinformationen rational interpretieren. Die Kriterien für die Produktevaluation können entweder primäre oder sekundäre sein (vgl. Scriven 1967). Die sekundären Kriterien beziehen sich auf die Ergebnisse eines Programms, die zur Erreichung der Verhaltensziele beitragen. Clark und Guba haben für pädagogische Innovationen 1965 eine Taxonomie von Zielen mit den dazu gehörenden Kriterien entwickelt, deren Schema ich in Tabelle 2 adaptiert habe. Die primären Kriterien beziehen sich vor allem auf Verhaltensziele, für deren Identifikation Blooms Taxonomy of Educational Objectives (1954) nützlich ist.

Im Reformprozeß bietet die Produktevaluation Informationen, mit deren Hilfe über die Weiterentwicklung, Beendigung, Modifikation einer Reform entschieden und aufgrund derer diese Innovation mit anderen Phasen des Reformprozesses verbunden werden soll. So kann z. B. die Produktevalua-

tion eines Programms, mit dem die Lernbereitschaft von Schülern aus sozio-kulturell benachteiligten Familien angeregt werden sollte, zeigen, daß die Programmziele gut erreicht worden sind, und daß das entwickelte innovative Programm auch auf andere Schulen übertragen werden kann.

Nachdem ich diese vier Formen der Evaluation dargestellt habe, soll die Methodologie ihrer Implementation im nächsten Abschnitt dieses Beitrags entwickelt werden.

Die Struktur von Evaluationsplänen

Wenn ein Evaluator eine Evaluationsstrategie, d. h. z. B. Kontext-, Input-, Prozeß- oder Produktevaluation, gewählt hat, muß er einen Plan (design) für ihre Durchführung auswählen oder entwickeln. Das ist eine schwierige Aufgabe, weil es nur wenige generalisierbare Evaluationspläne gibt, die den Erfordernissen des Bildungswesens genügen. Daher müssen Pädagogen Evaluationspläne im allgemeinen ganz neu entwickeln.

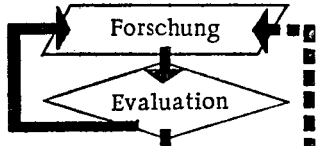
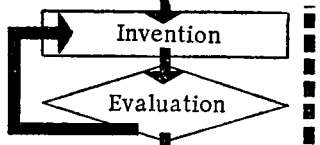
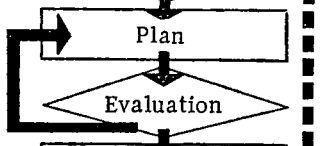
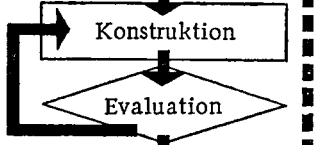
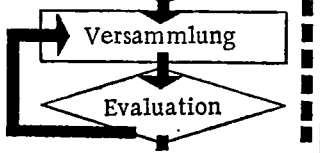
In dem folgenden Abschnitt dieses Beitrags sollen einige allgemeine Richtlinien für die Entwicklung von Evaluationsplänen behandelt werden. Dabei möchte ich die Struktur von Evaluationsplänen im Bildungswesen darzustellen versuchen. Hoffentlich werden diese allgemeinen Ausführungen den Pädagogen bei ihren Versuchen, Evaluationspläne zu entwickeln, behilflich sein, und hoffentlich werden die folgenden Ausführungen Experten in Methodenfragen dazu anregen, generalisierbare Pläne für die Kontext-, Input-, Prozeß- und Produktevaluation zu entwickeln.

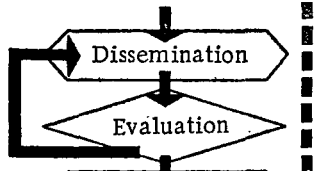
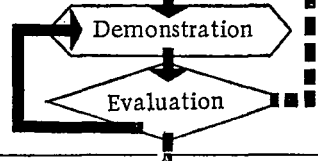
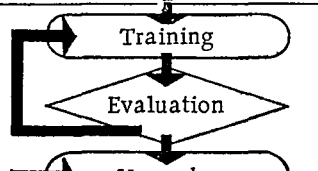
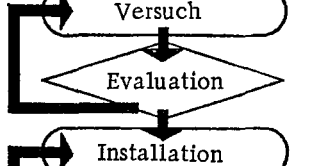
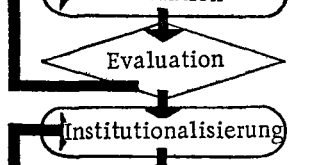
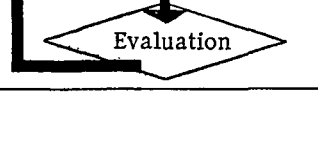
Definition des Plans

Im allgemeinen dient ein Plan zur Vorbereitung einiger Entscheidungssituationen, die zur Realisierung bestimmter Ziele führen sollen. Diese Definition besagt dreierlei:

1) Man muß die Ziele bestimmen, die durch die Realisierung des Plans erreicht werden sollen. In einer Produktevaluation könnte ein solches Ziel z. B. in der Untersuchung darüber bestehen, ob alle Schüler in einem Leseprogramm bestimmte Leseleistungen und Lesefertigkeiten erreichen.
2) Man muß die Entscheidungssituationen während der Realisierung des Evaluationsziels identifizieren. In dem Beispiel vom Leseprogramm müßte man die Meßverfahren bestimmen, die sich für die Einschätzungen der Lesefertigkeit eignen.
3) Der Evaluator muß in allen identifizierten Entscheidungssituationen zwischen möglichen Alternativen wählen.

Tabelle 2. Eine Prozeßtabelle, die die Rolle der Evaluation im Prozeß der Bildungsreform darstellt. Sie beruht auf D. L. Clark und E. G. Guba (1965) und ist abgedruckt aus: D. L. Stufflebeam (1966)

	Institution	Ziel	Prozeß	Kriterien	Beziehung zur Innovation
FORSCHUNG	Universitäten, Forschungs- und Entwicklungszentren und Bildungszentren	Verbesserung des Wissens, d. h. darstellen, in Beziehung setzen, konzeptualisieren und überprüfen.		Validität (innere und äußere).	Liefert eine Basis für eine Innovation.
ENTWICKLUNG	Universitäten, Forschungs- und Entwicklungszentren, Bildungszentren.	Formulieren einer neuen Lösung eines oder mehrerer Probleme, d. h. innovieren.		Augenscheinvalidität (face validity); geschätzte Funktionsfähigkeit; Einfluß (relativer Beitrag).	Erzeugt die Innovation.
		Entwurf eines Plans zur Konstruktion der Innovation.		*Durchführbarkeit* (Produktion und Benutzung); Handlichkeit (leicht zu benutzen, zu kontrollieren und in der Verwendung zu unterweisen).	*Bewirkt, daß die* Innovation den Merkmalen der Zielsituation angemessen ist.
		Entwicklung der Komponenten, d. h. Konstruktion.		Spezifikation des Plans; individuelle Verhaltensweisen.	Schafft die für die Implementation des Plans notwendigen Komponenten.
		Integration der Komponenten in ein funktionierendes System, z. B. Beendigung der Entwicklung der Innovation für den Verkauf.		Spezifikation des Plans; Präzisierung aller Verhaltensweisen, Funktionsfähigkeit, Leistungsfähigkeit.	Erzeugt das koordinierte funktionierende System.

IMPLEMENTATION	Regierung, Universitäten und Bildungszentren.	Schafft verbreitete Kenntnis der Innovation bei den Praktikern, d. h. informiert.		Verständlichkeit, Zuverlässigkeit, Überzeugungskraft; Einfluß (Ausmaß, in dem Hauptziele erreicht werden).	Informiert über die Innovation.
		Möglichkeit, die Qualität der Reform zu überprüfen und zu beurteilen, d. h. sie überzeugend machen.		Glaubwürdigkeit; Angemessenheit; Bewertung.	Überzeugt von der Innovation.
ADAPTATION	Universitäten, Bildungszentren und Schulen.	Ausbildung der Lehrer in den Schulbezirken, mit der Innovation umzugehen und sie durchzuführen, d. h. Personalausbildung.		Quantität, Kontinuität, Angemessenheit, Motivation und Tüchtigkeit des ausgebildeten Personals.	Errichtet und erhält die Funktionsfähigkeit für die Durchführung der Innovation.
		Vertrautmachen mit der Innovation und Schaffen einer Grundlage zur Einschätzung der Qualität, des Wertes, der Angemessenheit und der Verwendbarkeit der Innovation in einer bestimmten Institution.		Anwendbarkeit; Durchführbarkeit; Handlung.	Prüft die Innovation im Kontext einer bestimmten Situation.
		Die Charakteristika der Innovation und der sie implementierenden Institution aufeinander beziehen.		Wirksamkeit, Leistungsfähigkeit.	Operationalisiert die Innovation für die Verwendung in einer bestimmten Institution.
		Die Innovation als eine anerkannte Komponente des Systems assimilieren, d. h. etablieren.		Kontinuität, Bewertung; Unterstützung.	Etabliert die Innovation als einen Teil eines bestehenden Programms, überführt sie in eine »Nicht-Innovation«.

Somit würde ein vollständiger Evaluationsplan zahlreiche Entscheidungen über die Durchführung der Evaluation und die Wahl der verwendeten Instrumente enthalten.

Eine Liste mit den in vielen Evaluationsplänen gleichen Entscheidungssituationen, wäre für Evaluatoren außerordentlich nützlich. Sie würde es ihnen ermöglichen, die Probleme des Evaluationsplans systematisch in Angriff zu nehmen. Außerdem könnte sie für die Formulierung der Abschnitte über Evaluation in den Anträgen auf Forschungs- und Entwicklungsprojekte nützlich sein. Die Ministerien und Stiftungen könnten auch mit Hilfe dieser Liste ihre allgemeinen Richtlinien für Anforderungen im Bereich der Evaluation strukturieren. Sie könnte auch zur Bestimmung der Ausbildungsanforderungen wertvoll sein.

In Tabelle 3 wird eine Liste von allgemeinen Entscheidungssituationen für Evaluationspläne zusammengestellt. Sie beruht auf der Voraussetzung, daß die Struktur des Evaluationsplans für die Kontext-, Input-, Prozeß- und Produktevaluation gleich ist. Diese Struktur besteht nach meiner Auffassung aus sechs wichtigen Elementen:

a) Evaluationsschwerpunkt
b) Informationssammlung
c) Informationsorganisation
d) Informationsanalyse
e) Informationsbericht
f) Administration der Evaluation.

Alle sechs Elemente sollen im einzelnen kurz dargestellt werden.

Evaluationsschwerpunkt

Das erste Element der Struktur eines Evaluationsplans besteht im Evaluationsschwerpunkt. Aus ihm ergeben sich die Ziele der Evaluation und die mit ihrer Realisierung verbundenen Verfahrensfragen. Zu diesem Element des Evaluationsplans gehören vier Aspekte.

Erstens gilt es, die wichtigsten Entscheidungsebenen zu identifizieren, für die Informationen zur Verfügung gestellt werden sollen. So würden z. B. im Titel III des Elementary and Secondary Education Act von den einzelnen Schulen evaluative Informationen für die Ebene des Schulbezirks, des Einzelstaates und des Bundesministeriums benötigt. Bei der Entwicklung eines Evaluationsplans muß man alle relevanten Ebenen berücksichtigen, da man auf den einzelnen Ebenen unterschiedliche Informationen zu verschiedenen Zeitpunkten braucht.

Nachdem die wichtigsten der für die Evaluation relevanten Entscheidungsebenen genannt worden sind, müssen zweitens die Entscheidungs-

situationen auf jeder Ebene identifiziert werden. Bei unseren gegenwärtig geringen Kenntnissen über Entscheidungsprozesse im Bildungswesen liegt darin eine schwierige Aufgabe. Sie zu erfüllen ist jedoch außerordentlich wichtig; sie sollte deshalb sobald als möglich in Angriff genommen werden. Zunächst sollten Entscheidungssituationen im Hinblick auf die relevanten Entscheidungsträger wie Lehrer, Schulleiter, Schulverwaltung und Gesetzgeber bestimmt werden. Sodann sollten die wichtigsten Arten der Entscheidungssituationen, z. B. die Allokation von Mitteln und die Zustimmung zur Programmweiterentwicklung, festgelegt werden. Schließlich sollten diese Arten der Entscheidungssituationen z. B. als Forschung, Entwicklung, Dissemination oder Adaptation klassifiziert werden, wobei dieser Schritt vor allem für die Bestimmung relevanter Evaluationskriterien nützlich ist.

Die identifizierten Entscheidungssituationen sollten dann in bezug auf ihre kritische Reflektiertheit analysiert werden. Relativ weniger wichtige Entscheidungen, für die man die Evaluations-Ressourcen leicht aufbrauchen könnte, sollten nicht berücksichtigt werden. Ferner sollte man abschätzen, wann die ausgewählten Entscheidungssituationen eintreten, so daß man mit Hilfe der Evaluation gewonnene relevante Daten rechtzeitig bereitstellen kann. Schließlich sollte man versuchen, für jede wichtige Entscheidungssituation die Alternativen, die im Verlauf des Entscheidungsprozesses in Frage kommen können, mitzuberücksichtigen.

Wenn die Entscheidungssituationen ausgearbeitet worden sind, müssen drittens die erforderlichen Informationen bestimmt werden. Insbesondere sollte man für jede Entscheidungssituation die Kriterien dadurch festlegen, daß man die Variablen für die Messung und die Normen für die Beurteilung der Alternativen spezifiziert.

Viertens müssen die Richtlinien der Evaluation bestimmt werden. Man muß z. B. entscheiden, ob eine Selbstevaluation oder eine Fremdevaluation erfolgen soll. Ferner gilt es, die Adressaten der Evaluationsberichte zu bestimmen. Schließlich muß man noch das Ausmaß der Datenerhebung durch das Evaluationsteam festsetzen.

Informationssammlung

Das zweite wichtige Element der Struktur von Evaluationsplänen besteht in der Planung und Sammlung von Informationen. Dieses Element muß in enger Beziehung zu den Kriterien, die im vorigen Abschnitt identifiziert wurden, gesehen werden.

Bei Verwendung dieser Kriterien sollte man zunächst einmal festlegen, welche Informationen gesammelt werden sollen. Dabei muß man vor allem zwei Aspekte berücksichtigen:

(1) den Ursprung der Informationen, z. B. Schüler, Lehrer, Schulleiter oder Eltern,
(2) die gegenwärtige Beschaffenheit der Informationen als Ergebnis zufälliger oder systematischer Aufzeichnungen.

Sodann sollte man Instrumente und Methoden zur Sammlung der erforderlichen Informationen, z. B. Leistungstests, Interviews und die relevante Fachliteratur, angeben. Metfessel und Michael (1967) haben eine umfassende Liste von Instrumenten erstellt, die für die Datensammlung im Rahmen der Evaluation relevant sein können.

Für jedes Instrument, das verwendet werden soll, sollte man zunächst das anzuwendende Stichprobenverfahren spezifizieren. Wenn möglich, sollte man einen Schüler nicht zu viele Instrumente bearbeiten lassen. So könnte ein zweckmäßiges Verfahren darin bestehen, eine Stichprobe für das Instrument A zu ziehen, die in dieser Stichprobe enthaltenen Individuen nicht in die Grundgesamtheit zurückzulegen und aus der verbleibenden Gesamtheit eine Stichprobe für das Instrument B zu ziehen usw. In ähnlicher Weise empfiehlt sich, wenn ein Gesamttestwert für den einzelnen Schüler nicht benötigt wird, ein komplexes Stichprobenverfahren, bei dem kein Schüler mehr als eine Stichprobe der Aufgaben eines Tests bearbeitet.

Schließlich sollte man einen Zeitplan für die Datensammlung entwickeln. Er sollte die Beziehungen zwischen der Auswahl der Stichproben und Instrumente und den Terminen für die Informationssammlung im einzelnen festlegen.

Informationsorganisation

In Evaluationsberichten wird häufig darüber geklagt, daß die Ressourcen nicht ausreichen, alle wichtigen Daten zu verarbeiten. Um diese Situation zu vermeiden, sollte man das dritte Element des Evaluationsplans, die Informationsorganisation, sorgfältig planen. Zur Organisation der Informationen gehört die Entwicklung eines Plans zur Klassifikation der Informationen und zur Bestimmung der Verfahren, der Kodierung des Übertragens auf Lochkarten und des Abrufens von Informationen.

Informationsanalyse

Das vierte wichtige Element des Evaluationsplans besteht in der Analyse der Informationen. Ihre Aufgabe ist es, für die deskriptive oder statistische Analyse der Informationen zu sorgen, die den Entscheidungsträgern zur Verfügung gestellt werden sollen. Dazu gehören auch Interpretationen und Empfehlungen. Wie bei der Organisation der Informationen muß der Evaluationsplan entsprechende Mittel für die Durchführung dieser Analysen

vorsehen. Diese Aufgabe sollte einem qualifizierten Mitglied des Evaluationsteams oder einem besonderen Team zugeteilt werden, das sich auf die Probleme statistischer Analysen spezialisiert hat. Die für die Analyse der Informationen verantwortlichen Mitarbeiter müssen auch an der Planung der Analyseverfahren beteiligt sein.

Informationsbericht

Das fünfte Element eines Evaluationsplans bildet der Bericht der Informationen. Er zielt darauf ab, den Entscheidungsträgern die benötigten Informationen rechtzeitig in benutzbarer Form zur Verfügung zu stellen. In Übereinstimmung mit den Grundsätzen und Richtlinien der Evaluation sollten die Adressaten identifiziert werden. Sodann sollten die Verfahren bestimmt werden, um jedem Adressaten die für ihn relevanten Informationen zur Verfügung zu stellen. Ferner muß das Ausmaß und die Form des Evaluationsberichts festgelegt werden.

Administration der Evaluation

Das letzte Element des Evaluationsplans besteht in der Administration der Evaluation. Ihre Aufgabe liegt darin, die Durchführung des Evaluationsplans zeitlich zu koordinieren. Erstens muß daher ein Gesamtzeitplan für den Ablauf der Evaluation entwickelt werden. Dazu empfehlen sich Verfahren wie die Program Evaluation and Review Technique (PERT). Zweitens muß man die Stellenanforderungen bestimmen. Drittens gilt es, die erforderlichen Mittel festzulegen, um den Grundsätzen und Richtlinien für die Durchführung der Evaluation gerecht zu werden. Viertens muß man untersuchen, inwieweit der Evaluationsplan valide, reliable, zuverlässige, aktuelle und überzeugende Informationen liefern kann. Fünftens gilt es, Verfahren zu entwickeln, um den Evaluationsplan regelmäßig auf den neuesten Stand zu bringen. Sechstens muß man einen Finanzierungsplan für die Evaluation ausarbeiten.

Tabelle 3: Entwicklung eines Evaluationsplans

Die logische Struktur eines Evaluationsplans ist für die Kontext-, Input-, Prozeß- oder Produktevaluation gleich. Sie enthält folgende Elemente:

A. *Evaluationsschwerpunkt*
1. Identifikation der wichtigsten Entscheidungsebenen (z. B. örtliche, einzelstaatliche und/oder bundesstaatliche)
2. Planung und Beschreibung aller Entscheidungssituationen auf jeder Entschei-

dungsebene in bezug auf ihren Schwerpunkt, die kritische Reflektiertheit, den Zeitpunkt und die Komposition der Alternativen
3. Bestimmung der Kriterien für jede Entscheidungssituation durch Spezifikation der Variablen für die Messungen und der Normen für die Beurteilung von Alternativen
4. Definition der Grundsätze und Richtlinien, innerhalb deren die Evaluation erfolgen soll

B. *Informationssammlung*
1. Spezifikation des Ursprungs der zu sammelnden Informationen
2. Bestimmung der Instrumente und Methoden für die Sammlung der erforderlichen Informationen
3. Spezifikation des anzuwendenden Stichprobenverfahrens
4. Spezifikation der Bedingungen und des Zeitplans für die Informationssammlung

C. *Informationsorganisation*
1. Erstellung eines Plans für die Informationen, die gesammelt werden sollen
2. Bestimmung der Mittel zur Kodierung, Organisation, Speicherung und zum Wiederabruf der Informationen

D. *Informationsanalyse*
1. Auswahl der analytischen Verfahren, die angewendet werden sollen
2. Bestimmung der Mittel zur Durchführung der Analyse
3. Spezifikation des Ausmaßes und der Form der Evaluationsberichte
4. Zeitplan des Informationsberichts

E. *Informationsbericht*
1. Definition der Adressatengruppe
2. Bestimmen der Mittel der Informationsvermittlung
3. Festlegen des Formats des Evaluationsberichts
4. Planung der Elemente für die Darstellung der Information

F. *Administration der Evaluation*
1. Zusammenfassung des Evaluationsplans
2. Bestimmung der für die Evaluation erforderlichen Mitarbeiterstellen und Finanzen
3. Spezifikation der Mittel, um die Evaluation gemäß ihren Grundsätzen und Richtlinien durchzuführen
4. Evaluation der Möglichkeiten des Evaluationsplans, valide, reliable, zuverlässige, aktuelle und überzeugende Informationen zu liefern
5. Spezifikation und zeitliche Planung der Mittel, um den Evaluationsplan regelmäßig auf den neuesten Stand zu bringen
6. Bereitstellung eines Budgets für das ganze Evaluationsprogramm

Ich bin am Ende meiner Ausführungen angelangt. Obwohl ich nur einen groben Überblick über einige Probleme der Evaluation im Bildungswesen gegeben habe, wird deutlich geworden sein, daß die Planung und Durchführung pädagogischer Evaluation ein höchst komplexes und schwieriges Unternehmen ist. Es bedarf einer erheblichen Anstrengung aller im Bereich pädagogischer Evaluation arbeitenden Wissenschaftler, um entsprechende Fortschritte zu erzielen. Bleiben sie aus, wird meiner Ansicht nach das Erziehungswesen darunter leiden, daß die für wichtige Entscheidungen erforderlichen Informationen fehlen.

Marvin C. Alkin

Die Aufwands-Effektivitäts-Evaluation von Unterrichtsprogrammen

Vergleich von Kosten-Nutzen-Evaluation (Cost-Benefit Evaluation) und Aufwands-Effektivitäts-Evaluation (Cost-Effectiveness Evaluation)

Was versteht man unter Kosten-Nutzen-Analyse? Auf was für Schwierigkeiten stößt man bei der Anwendung dieser Technik auf Entscheidungssituationen, in denen sich die meisten Pädagogen der jeweiligen Schul- oder Schulbezirksebene befinden? Und welche Evaluationstechnik ließe sich schließlich anstelle der Kosten-Nutzen-Analyse für die Evaluation von Bildungssystemen heranziehen?

Techniken wie die Kosten-Nutzen-Analyse sollen in erster Linie Entscheidungshilfen bei der Formulierung von Vorschlägen sein. Wenn man solche Verfahren anwenden will, muß man deshalb Angaben über die Wirklichkeit machen, die Grundlage für eine Handlungsdirektive der Entscheidungsträger sein können. Bei dieser Art von Betrachtung muß man verschiedene Handlungsabläufe bewerten, wobei man nicht nur die Ergebnisse oder Outputs des Prozesses betrachtet, sondern auch den jeweils damit verbundenen finanziellen Aufwand. Für die meisten Pädagogen ist es anscheinend wesentlich leichter, den Wert eines Programms an Erträgen bzw. Ergebnisgrößen als am Aufwand zu messen. Trotz der Vernachlässigung durch die Pädagogen ist der Aufwand von erheblicher Bedeutung und kann nur dann außer Betracht bleiben, wenn man sich in der glücklichen Lage befindet, über unbegrenzte Mittel zu verfügen, und zwar nicht nur in Form von materiellen Gütern und Dienstleistungen, sondern auch in Form von Zeit und Energie. Ein solcher Idealzustand entspricht gewiß nicht der heutigen Wirklichkeit.

Die Idee der Kosten-Nutzen-Analyse ist verblüffend einfach: Lediglich die jeweils mit unseren Alternativen verbundenen Kosten (Aufwand) und der Nutzen (Ertrag) müssen bestimmt werden. Sind Aufwand und Ertrag der einzelnen Alternativen erst einmal ermittelt, kann man leicht die Alternative herausfinden, die bei gegebenem Aufwand den größten Ertrag liefert oder die Alternative, die einen bestimmten Ertrag mit den geringsten Kosten erzielt. Die weitverbreitete Ansicht, daß die Kosten-Nutzen-Analyse zur gleichen Zeit eine Maximierung der Gewinne oder Erträge und eine

Minimierung des Aufwands anstrebe, ist nicht richtig; angenommen, sie sei zutreffend, ließe sich das Problem doch nicht lösen. Es wäre genau das gleiche, als wenn man von einem Geographen verlangte, den tiefsten See auf dem höchsten Berg ausfindig zu machen. Ganz gleich, welchen See er auswählte, es würde immer einen etwas seichteren See auf einem etwas höheren Berg geben; schließlich wäre er bei einem Wassertropfen auf dem Gipfel des Mount Everest angelangt. Wenn wir jedoch die Aufgabe so umformulieren, daß wir entweder die Tiefe des Sees oder die Höhe des Berges begrenzen, dann kann das Problem gelöst werden. Die gleichen Überlegungen gelten für die Kosten-Nutzen-Analyse. Es ist unmöglich, eine Strategie zu wählen, die gleichzeitig den Ertrag maximiert und den Aufwand minimiert. Eine derartige Strategie existiert nicht. Wenn wir zwei Strategien A und B vergleichen, so kann A zufällig einen größeren Ertrag aufweisen und doch weniger kosten als B. In diesem Fall ist A natürlich B überlegen. Die Strategie A minimiert jedoch nicht den Aufwand und maximiert gleichzeitig den Ertrag. Der maximale Ertrag ist unendlich groß, die minimalen Kosten betragen Null. Eine Strategie mit diesem Ergebnis werden wir nicht finden.

Damit wir die Kosten-Nutzen-Analyse sinnvoll anwenden können, müssen alle Kosten und Nutzen – unsere Entscheidungskriterien – spezifiziert werden können. Darüber hinaus müssen wir angeben können, welche Größen (Kosten- bzw. Nutzenarten) frei veränderlich und welche begrenzt oder beschränkt sind. Schließlich muß noch abgesteckt werden, innerhalb welcher Grenzen sich Kosten und Nutzen jeweils bewegen dürfen und in welchem Verhältnis Verluste in einer Dimension durch gleichzeitige Gewinne in einer anderen Dimension aufgewogen werden können (*Trade-offs*).

Die Kosten-Nutzen-Analyse ist in erster Linie eine ökonomische Analyse. Mit anderen Worten: Bei der Methode der Kosten-Nutzen-Analyse handelt es sich um ein Instrument des Ökonomen, das vornehmlich der Untersuchung von Wirtschaftseinheiten dient. *Eine der Hauptbedingungen der Kosten-Nutzen-Analyse ist, daß sowohl Inputs als auch Outputs in der gleichen Einheit, nämlich Dollar, gemessen werden können.* Dieses Konzept ist von Bedeutung, wenn es darauf ankommt, bestimmte Programme zu beurteilen. So mag sich etwa im privatwirtschaftlichen Sektor ein Unternehmen dazu entschließen, das Kapital zu erhöhen, um eines jener Programme auszuweiten, das ein günstiges Kosten-Nutzen-Verhältnis aufweist, d. h. aller Wahrscheinlichkeit nach einen in Geldeinheiten meßbaren Gewinn abwerfen wird.

Kosten-Nutzen-Analysen im öffentlichen Sektor wurden bislang vorwiegend im Bereich der Wasserwirtschaft und der Landesverteidigung ange-

wendet. In jedem Fall erfordert das Verfahren eine Darlegung der verschiedenen möglichen Ergebnisse, ausgedrückt in Dollar. So wird etwa der wichtigste direkte Ertrag eines Wasserkraftwerks der in Dollar ausgedrückte Wert der produzierten elektrischen Energie sein; daneben fallen noch indirekte Erträge, etwa durch Vermeidung von Schäden an Häusern, Besitztümern und Ernten wegen geringerer Überschwemmungsgefahr an. Auch die weniger leicht zu quantifizierenden (intangiblen) Erträge, wie z. B. Verbesserung des körperlichen und geistigen Wohlbefindens der Bewohner durch Beseitigung der Furcht vor Überschwemmungen, werden bewertet, und man ordnet ihnen Dollar-Beträge zu (McKean 1958).

Im Bildungswesen wurde die Kosten-Nutzen-Analyse gewöhnlich auf umfassende Systeme (z. B. von Regionen, Bundesländern, Staaten) angewendet. Das ist verständlich, denn auf diesen Ebenen sind Daten über Ergebnisse von Bildungsprozessen in Form von Dollar-Werten eher erhältlich. So konzentrierte Becker (1962) sich auf den gesellschaftlichen Nutzen der Hochschulbildung, den er an den Auswirkungen auf die gesamtwirtschaftliche Produktivität maß. Er kam u. a. zu dem Ergebnis, »daß die Rendite für den einzelnen bei einer Investition in Hochschulbildung höher ist als bei Anlage in einem Unternehmen« (Becker 1962). In einer anderen Untersuchung (Hansen 1963) wurde der interne Zinsfuß für aufeinanderfolgende Ausbildungsstufen berechnet, wobei der Ertrag aus Querschnittsdaten der Einkommen der Ausgebildeten – klassifiziert nach Alter und Ausbildungsstufe – errechnet wurde. Schließlich haben 1966 Hirsch und Marcus Aufwand und Ertrag einer allgemeinen Junior-College-Ausbildung mit der alternativen Verwendung der gleichen finanziellen Mittel für Sommerprogramme in Sekundarschulen verglichen.

Charakteristisch für diese wenigen Beispiele ist, daß in allen Fällen die Ergebnisgröße durch die Verwendung gebräuchlicher ökonomischer Indizes in Dollar-Beträge umgewandelt wurden. Da aber die Schulbezirke oft nicht mit anderen Verwaltungseinheiten übereinstimmen, sind ökonomische Daten auf der Ebene einzelner Schulen oder Schulbezirke nicht verfügbar. Selbst wenn sie verfügbar wären, müßte geprüft werden, ob die Kosten-Nutzen-Analyse wegen der Mobilität der Schüler über die Grenzen der Schulbezirke und wegen der Schwierigkeit, langfristige ökonomische Erträge für so kleine Bildungseinheiten wie einzelne Schulen ermitteln zu können, tatsächlich noch geeignet ist. Außerdem hilft uns die Kosten-Nutzen-Analyse nicht bei der Lösung des Problems für die hier zu erörternde Einheit. Kurz gesagt, soll sich das Interesse hier weniger auf die ökonomischen Auswirkungen bestimmter Entscheidungen über Investitionen in Bildung als vielmehr auf die Evaluation der Komponenten eines Systems im Hinblick auf die definierten Zielgrößen richten.

Im Gegensatz zur Kosten-Nutzen-Analyse, die keine direkte Anregung für den allgemeinen Entscheidungsprozeß eines politischen Systems bietet, soll in dieser Arbeit eine Entscheidungssituation der Wirklichkeit untersucht werden, in der nicht alle Ergebnisse in ökonomischen Größen ermittelt werden können.

Ich fasse zusammen: Wenn ich mich im Rahmen dieses Beitrags auf die Aufwands-Effektivitäts-Analyse beziehe, soll damit ein Modell gemeint sein, mit dessen Hilfe die relevanten Elemente von Bildungssystemen auf der Ebene einer einzelnen Schule oder eines einzelnen Schulbezirks untersucht werden können, um (a) die Ergebnisse des Bildungsprozesses bei verschiedenen Einheiten zu vergleichen, (b) die Auswirkungen unterschiedlichen finanziellen Aufwands festzustellen, (c) alternative Wege zur Erreichung bestimmter Bildungsziele auszuwählen.

Die Komponenten eines Aufwands-Effektivitäts-Modells

Welches sind nun die Komponenten eines Aufwands-Effektivitäts-Modells, mit dessen Hilfe Entscheidungsträger Bildungsprozesse evaluieren können? Dazu gilt es zunächst zu definieren, was hier unter einem Modell verstanden werden soll.

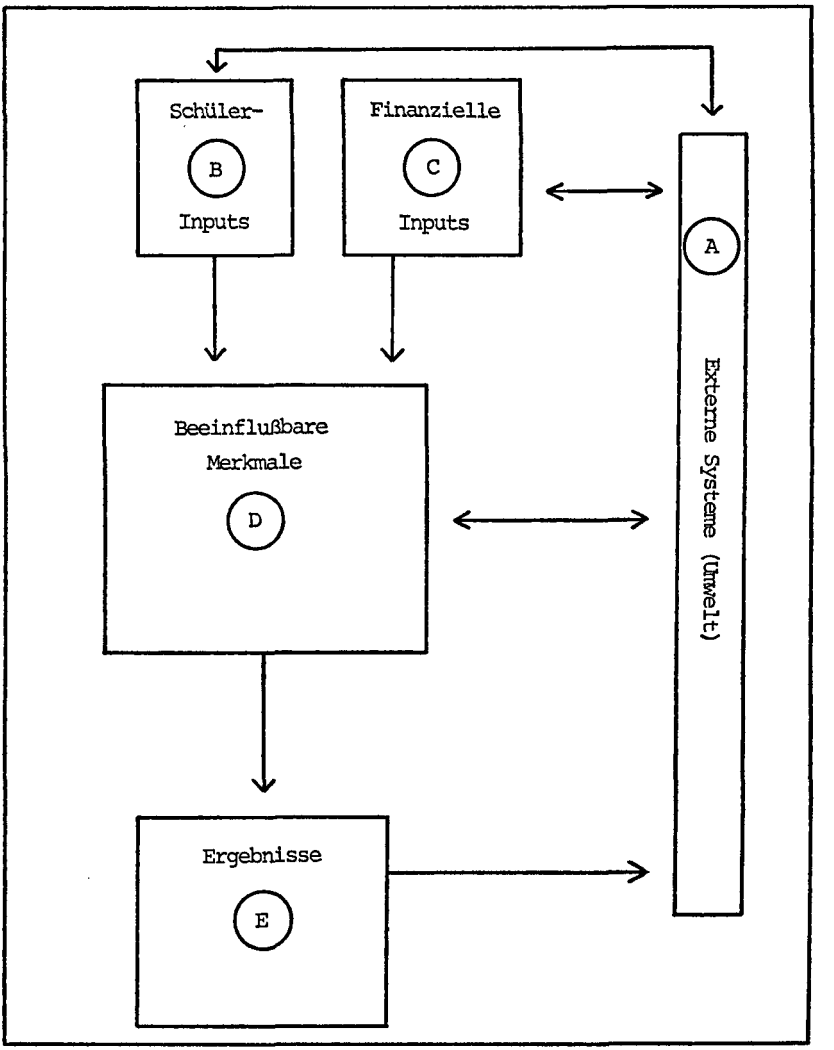
Kurz gesagt, ein Modell stellt einfach einen Versuch dar, die Hauptelemente einer Einheit oder eines Phänomens im Hinblick auf jeweilige Funktionen und gegenseitige Beziehungen zu klassifizieren, damit leichter beobachtet werden kann, wie die Elemente innerhalb der Einheit funktionieren, wie sie die Einheit funktionsfähig machen und wie sie gegenseitig aufeinander einwirken. Auf diese Weise können wir auch die Auswirkungen einer Veränderung der Elemente feststellen. Die meisten Modelle spiegeln die Neigung und Interessen derjenigen wider, die sie konstruiert haben. Auch dieses Modell bildet keine Ausnahme; unser Hauptinteresse gilt der Untersuchung administrativer und finanzieller Variablen im Bildungswesen, insbesondere wenn eine einzelne Schule oder ein Schulbezirk die zu analysierende Einheit ist. Gewiß ist ein Evaluationsmodell – wie jedes andere Modell – eine vereinfachte Darstellung oder Veranschaulichung komplexer Wechselbeziehungen. Eine solche Veranschaulichung hat lediglich den Zweck, demjenigen, der das Modell entwickelt hat, die für ihn wichtigen Tatbestände strukturieren zu helfen.

Aus welchen Elementen besteht unser Evaluationsmodell? (1) Die Schüler – d. h. in diesem Fall ihre Eigenschaften und Merkmale zu Beginn des zu evaluierenden Prozesses – sind eine Eingangsgröße (Input) des Modells. (2) Ergebnisse des Bildungsprozesses bilden eine Ausgangsgröße

(Output) des Modells. Damit meinen wir zwei Dinge: (a) kognitive und nicht-kognitive Veränderungen, die sich in den Schülern vollziehen, nachdem sie mit dem Unterrichtsprogramm konfrontiert worden sind, (b) die Auswirkung des Programms auf externe Systeme wie häusliche Verhältnisse, die Gemeinde, andere Programme usw. (3) Finanzielle Inputs – d. h. Mittel, die für die Durchführung des Programms aufgewendet werden – sind ein weiteres Element des Modells. (4) Die beeinflußbaren Größen oder Aktionsparameter (manipulatable characteristics) – z. B. Lehrkörper, Schulorganisation und Unterrichtsprogramme – geben an, wie die finanziellen Inputs in Verbindung mit den Schüler-Inputs innerhalb eines Programms verwendet werden. Schließlich (5) muß unser Evaluationsmodell externe Systeme berücksichtigen. Dieses Element betrifft den Rahmen gesellschaftlicher, politischer, gesetzlicher, ökonomischer und anderer außerhalb der Schule liegender formeller und informeller Systeme, d. h. die Umwelt, soweit sie Einfluß auf das Programm hat und ihrerseits von den Ergebnissen des Programms verändert wird.

Bei der Erörterung der beeinflußbaren Merkmale gehen wir von der Annahme aus, daß sie die einzigen administrativen beeinflußbaren Variablen sind. Im Rahmen unseres Modells wollen wir annehmen, daß (a) externe Systeme nicht sofort durch die Outputs des Systems verändert werden und (b) daß die schulischen Entscheidungsträger keine Macht über den Einfluß der Umwelt auf die Schule haben. Wenn wir davon ausgingen, daß die Rückkopplung das System sofort ändert, käme für diese Betrachtungen nur ein dynamisches Modell anstatt des hier verwendeten statischen Modells in Frage. Die zweite Annahme beinhaltet, daß kein Versuch unternommen wird, die Eigenschaften der in das System eingehenden Schüler zu verändern; d. h., wir machen uns im allgemeinen keine Gedanken über mögliche Veränderungen in der Gemeinde, die die Schüler-Inputs qualitativ ändern könnten. Wir gehen ferner davon aus, daß die Schüler-Inputs von außerhalb des Systems relativ unbeeinflußbar sind. Wir richten unser Augenmerk also nur auf die beeinflußbaren Größen innerhalb des Systems, d. h. die Aktionsparameter, die der Maximierung des Ergebnisses bei den Schülern dienen. Wir geben zu, daß eine gewisse Schwäche in dieser Annahme steckt und daß sich einige schulbezogene Manipulationsmöglichkeiten einrichten ließen, die die Eigenschaften der in das System eingehenden Schüler verändern würden. Mittel hierfür sind z. B. Veränderungen der Einzugsbereiche, der Einsatz von Schulbussen mit der Absicht, die Schüler-Inputs bestimmter Schulen zu manipulieren, schulische Maßnahmen der Gemeinden (wie etwa Sonderprogramme in sozial benachteiligten Gebieten) und Vorschulprogramme (wie z. B. das Headstart-Projekt). Die Annahme eines statischen Modells und nichtbeein-

Abbildung 1
Aufwands-Effektivitäts-Modell



flußbarer externer Systeme erscheint in diesem frühen Entwicklungsstadium des Modells notwendig.

Mit unserer Definition von Evaluation und unter Beachtung der genannten Unzulänglichkeiten kann nun das Evaluationsmodell erörtert werden.

Schüler-Inputs

Wir wollen den Schüler-Input als eine Beschreibung des Schülers zu Beginn des Prozesses betrachten oder bei einem umfassenderen Unterrichtsprogramm als eine aggregierte statistische Beschreibung der in das System eintretenden Schüler (vgl. Abb. 1). Im Idealfall wird den Schülern bei ihrem Eintritt in das System ein vollständiger Katalog aller gebräuchlichen Leistungs-, Intelligenz- und Persönlichkeitstests vorgelegt, des weiteren Fragebogen, die Informationen über die häuslichen Verhältnisse, den Status innerhalb der Gemeinde, den familiären Hintergrund, die Mitgliedschaft von Familienangehörigen in anderen gesellschaftlichen Systemen u. ä. liefern. Leider gibt es diesen Idealfall nicht; deshalb müssen wir eine Reihe von Näherungswerten für den Schüler-Input entwickeln. Häufig sind Werte von Intelligenztests für die eintretenden Schüler verfügbar; gewöhnlich liefert auch die Schülerkartei einige Familiendaten. Manchmal sind Leistungstests des vergangenen Jahres oder der letzten beiden Vorjahre als Maß für die Eingangsleistung der Schüler vorhanden. Die meisten der zusätzlich erwünschten Daten müssen jedoch entweder in den Schulen erhoben werden oder häufiger noch aus anderen besser zugänglichen Daten abgeleitet werden. Aus diesem Grunde wendet man sich oft dem Milieu und den Merkmalen der Gemeinde, der der Schüler entstammt, als einem Indikator für die Art der Schüler-Inputs des Systems zu.

Finanzielle Inputs

Eine zweite Gruppe von Eingangsgrößen des Systems sind die finanziellen Inputs. Wenn wir einen Schulbezirk als ein System betrachten, dann gehen nicht nur die Schüler als Input in das System ein, sondern auch finanzielle Mittel, die von Bund, Ländern und Gemeinden bereitgestellt werden und die teilweise der Realisierung von unterschiedlichen instrumentalen Faktorkombinationen innerhalb des Systems dienen. Vielleicht ist es wichtig, den Anteil von Bund, Ländern und Gemeinden an der gesamten Mittelaufbringung zu bestimmen. Unter Umständen sollte man auch aufzeigen, mit welcher Maßgabe Mittel aus Bundesquellen und besonderen Länderprogrammen vergeben werden, damit man sich der Auflagen und ihrer Folgen für die Mittelverwendung im System bewußt wird.

Wenn wir nur einen Teil des Systems evaluieren, etwa das Programm für den Mathematikunterricht oder die Schülerberatung, dann müßten wir Art und Umfang der finanziellen Inputs für dieses Teilsystem bestimmen. Leider liefern die derzeitigen Verfahren der Rechnungsführung in allen Ländern nur Daten in Form von verwaltungsmäßig gegliederten Aus-

gaben und nicht in Form von Programmausgaben; d. h. für eine Reihe von Faktoren, wie etwa für Verwaltung, Instandhaltung, laufenden Unterhalt, Unterricht und fixe Belastungen, sind Ausgabedaten verfügbar, die aber nicht nach Programmen gegliedert sind. Wollte man finanzielle Inputdaten in Evaluationsuntersuchungen einbeziehen, müßte man je nach der zu untersuchenden Ebene entweder die gegebenen Budgetdaten für unsere Ziele entsprechend neu gliedern oder aber mit neuen Verfahren der Rechnungsführung beginnen.

Externe Systeme

Die Schule wird von zahlreichen gesellschaftlichen Systemen umgeben (externer gesellschaftlicher Kontext). Im Falle einer einzelnen Schule gehören dazu z. B. die Gemeinde, der Schulbezirk und die Form seiner Verwaltung, andere Verwaltungssysteme, wie etwa die Stadt, der Landkreis sowie die Art des Gemeindelebens und die Teilnahme der Bürger daran. Jedes dieser externen Systeme stellt – entsprechend den verschiedenen von ihnen ausgeübten Funktionen – eine Reihe von Anforderungen und legt dem Bildungssystem (Schule) und dem einzelnen innerhalb des Systems Beschränkungen auf. Alle diese Systeme verfolgen bestimmte integrative, adaptive, zielorientierte und strukturerhaltende Funktionen im Makrosystem. Folglich ist es notwendig, die im Hinblick auf ihren Beitrag zur Erzielung des Bildungs-Outputs des Systems wichtigen Eigenschaften und Beziehungen dieser externen Systeme zu erkennen und zu quantifizieren.

In der Wirklichkeit stehen die externen Systeme mit dem Bildungssystem in Wechselbeziehungen. Während man einerseits davon ausgehen kann, daß jedes System seine eigenen Inputs, eine bestimmte Reihe von variablen Zwischengliedern und Outputs hat, ist andererseits jedes dieser Systeme gegenüber dem Bildungssystem wiederum ein externes System und *umgekehrt*. Folglich kann jedes außerhalb des Bildungsbereichs liegende System sowohl als Quelle von Inputs als auch als Empfänger von Outputs angesehen werden.

Beeinflußbare Merkmale (Aktionsparameter)

Eine vierte Gruppe von Elementen des Evaluationsmodells bezeichnen wir als beeinflußbare Merkmale. Es bieten sich zahlreiche Möglichkeiten für die Verwendung des finanziellen Inputs eines Systems. Wir können das zahlenmäßige Schüler-Lehrer-Verhältnis verringern, Normen festlegen, die die Einstellung von Lehrern mit bestimmten Eigenschaften sicherstellen,

andere Verwaltungsregelungen innerhalb der Schule treffen, mehr Bücher für die Schulbibliothek anschaffen, den Schülern direkt mehr Lehrbücher zur Verfügung stellen, andere Lehrpläne einführen, andere Unterrichtsverfahren anwenden oder zusätzliche Materialien beschaffen. Die Aktionsparameter sind also Veränderungen und Beeinflussungen durch die Entscheidungsträger auf allen Ebenen des Bildungswesens ausgesetzt. Uns fehlt jedoch ein eindeutiger Hinweis darauf, welche instrumentale Faktorkombination im Hinblick auf die Erreichung des Ziels der Schule, d. h. für die Erzielung des angestrebten Bildungs-Outputs, am wirkungsvollsten ist.

An dieser Stelle muß allerdings darauf hingewiesen werden, daß wir nicht unterstellen, daß alle den Bildungs-Output beeinflussenden Faktoren vom finanziellen Input abhängen. Die Durchführung von Veränderungen in der Schulumgebung oder im Lehrerverhalten kann z. B. relativ wenig finanziellen Aufwand erfordern. Häufig ist das vom Lehrer angewandte Unterrichtsverfahren (bzw. die Substitution eines Verfahrens durch ein anderes) mit geringen oder gar keinen zusätzlichen Kosten verbunden. Allerdings sind manche Veränderungen im System, wie z. B. neue Verwaltungsregelungen und der Einsatz neuerer technischer Mittel und Verfahren im Unterricht, außerordentlich teuer. Daher muß der durch die Änderung zu ermöglichende Output im Hinblick auf die damit verbundenen Kosten untersucht werden.

Der Standpunkt, daß mehr Geld für die Lehrerbesoldung aufgewendet werden sollte und daß auf diese Weise höchstwahrscheinlich das Bildungsprogramm verbessert werden würde, läßt sich leicht verteidigen. Es gibt Anzeichen dafür, daß eine Beziehung zwischen höheren Lehrergehältern und der Qualität der Bildung besteht. Die eigentliche Frage ist jedoch, inwieweit durch eine alternative Verwendung eines gegebenen Dollar-Inputs bestimmte Outputs des Bildungsprozesses gesteigert werden können. Dies ist ein Problem für die Aufwands-Effektivitäts-Analyse; es ist schließlich ein zentraler Bestandteil der Evaluation oder letztlich einer der Gründe, warum wir überhaupt evaluieren.

Wir wiesen bereits darauf hin, daß durch die Auswahl geeigneter instrumentaler Faktorkombinationen die Bildungs-Outputs eines Systems maximiert werden können. Gleichwohl muß angemerkt werden: Es existieren nicht nur verschiedene Setzungen von Aktionsparametern, die sich zur Produktion eines gegebenen Bildungs-Outputs eignen; bedeutsam ist vielmehr, daß diese Setzungen ganz verschiedene Bildungs-Outputs in unterschiedlichen Systemen oder für unterschiedliche Schülergruppen hervorbringen können. James Coleman beobachtete diesen Tatbestand in einer Untersuchung für die Civil Rights Commission mit dem Titel »Equality of

Educational Opportunity«, in der er hervorhob: »... es ist zu folgern, daß eine Verbesserung der schulischen Bedingungen eines zu einer (ethnischen) Minderheit gehörenden Schülers seine Leistung stärker anheben kann als eine ebensolche Verbesserung bei einem weißen Schüler.« Ähnlich kann die Leistung eines Durchschnittsschülers aus einer ethnischen Minderheit unter dem niedrigen Niveau einer Schule stärker leiden als die eines durchschnittlichen weißen Schülers. Er leitet hieraus den Schluß ab, daß »dies darauf hindeutet, daß für die am stärksten benachteiligten Kinder Verbesserungen in der Qualität der Schule die größten Leistungssteigerungen erbringen« (Coleman 1966). Die geeignete Festlegung der Aktionsparameter hängt deshalb nicht nur von den gewünschten Bildungs-Outputs ab, sondern ebenso von der Art der Schüler-Inputs und von dem gegebenen System.

Wie bereits früher erwähnt, gehen wir davon aus, daß die Glieder zwischen Input und Output die einzigen Aktionsparameter sind. Diese vereinfachende Annahme wurde von uns nicht zuletzt deshalb gemacht, damit wir statt mit einem komplexeren dynamischen mit einem statischen Modell arbeiten können. Die in dieser Annahme zum Ausdruck kommende Sichtweise ergibt sich auch aus dem von uns bei der Konstruktion des Modells verfolgten Hauptzweck, nämlich ein Entscheidungsmodell zu schaffen, mit dessen Hilfe Schulen und deren Tätigkeit evaluiert werden können.

Ergebnisse des Bildungsprozesses

Die erste Gruppe von Ergebnissen, die uns bei dem Modell beschäftigt, betrifft die im Schüler bewirkten Veränderungen, die von dem Zeitpunkt ihres Eintritts in das System bis zu dem Zeitpunkt ihres Austritts hervorgerufen wurden. Viele dieser Veränderungen werden durch die Art und Weise der finanziellen Aufwand erfordernden Zwischenglieder bewirkt. Hier zeigt sich erneut ein Problem, denn die Ergebnisse des Bildungsprozesses in einer Schule oder in einem Schulbezirk lassen sich nicht ausschließlich aufgrund der von den Schülern erzielten Ergebnisse in fachspezifischen Leistungstests messen[1]. Welches sind die nicht-kognitiven Aspekte des Ergebnisses oder des Outputs? Wie hat sich das Verhalten der Schüler geändert? Welcher Zusammenhang besteht zwischen den Aktivitäten, die in einem Schulbezirk oder in einer Schule stattfinden, und dem etwaigen Erfolg der Schüler in ihrem beruflichen Weiterkommen oder ihren zukünftigen Bildungsbemühungen? Welche Hilfe leisten die in der Schule gewonnenen Erfahrungen dem Schüler bei der Behandlung politischer Probleme und auf kulturellem Gebiet? In welchem Ausmaß beeinflußt die soziale Situation der Schule neben dem im Unterricht Gelernten

den Schüler? Dies sind nur einige der unbeantworteten Fragen, die mit der Identifikation der Ergebnisse des Bildungsprozesses zusammenhängen; beantworten lassen sie sich sicher nur durch weitere Forschung.

Während es zwei Input-Faktoren bei dem System gibt, nämlich die schülerbezogenen und die nicht-schülerbezogenen oder finanziellen Inputs, so wollen wir davon ausgehen, daß es keine finanziellen Ergebnisse gibt, es sei denn, wir wollten bestimmte Verhaltensänderungen in Geldeinheiten bewerten, oder aber die Ergebnisse bei den Schülern brächten finanzielle oder ökonomische Erträge individueller oder gesamtwirtschaftlicher Art mit sich [2].

Die zweite Gruppe von Ergebnissen im Modell sind die Outputs, die nicht beim Schüler anfallen. Die zwei Gruppen von Ergebnisgrößen (schülerbezogene und nicht-schülerbezogene) können als Rückkopplungsschleifen aufgefaßt werden, die bis zu einem gewissen Grade die Eigenschaften der zukünftigen Inputs des Systems verändern. Die Veränderungen in den Schülern haben u. a. soziale, politische und ökonomische Auswirkungen; damit ist gemeint: Die Struktur der externen Systeme wird durch die Schüler-Outputs verwandelt. Es gibt allerdings noch weitere Ergebnisse des Bildungsprozesses: Die im Zusammenhang mit den Aktionsparametern stehenden Bildungsentscheidungen haben Rückwirkungen auf die externen Systeme. Häufig berühren diese Outputs den einzelnen Schüler oder die Schülerergebnisse nur am Rande. Z. B. könnten viele Entscheidungen über die geeignete Verwendung der volkswirtschaftlichen Güter (Ressourcen) zahllose nicht unmittelbar schülerbezogene Bildungsergebnisse hervorbringen. Hier sei nur angeführt, daß Entscheidungen über die Anzahl und die Besoldung der Lehrer und des übrigen Personals in vieler Hinsicht die Struktur einiger externer Systeme ändern können, und zwar besonders dann, wenn die genannten Beschäftigten im Schulbezirk wohnen würden. In welchem Umfang sind unterschiedlich besoldete Lehrkräfte bereit, auf zusätzliche Einkünfte zu verzichten und statt dessen sich am Gemeindeleben und an Vereinigungen zu beteiligen? Wie verändert weiterhin die Entscheidung über eine bestimmte Kombination der Aktionsparameter im Bildungssystem, die höhere Bezüge für Lehrer vorsieht, diese externen Systeme? Ferner: Wie beeinflussen Art und Qualität der auszuwählenden Lehrer die sich ändernde Struktur der Gemeinde? Ein anderes Beispiel dürfte der Einfluß auf die Wirtschaft der Gemeinde sein, der durch die Auswahl solcher Aktionsparameter verursacht wird, die mit großen Sachinvestitionen oder großen Mengen am Ort eingekaufter Güter für den laufenden Schulbetrieb zusammenhängen? Inwieweit haben die Entscheidungen im Schulsystem über den Einsatz von Schulbussen, die Unterrichtszeit oder den Stundenplan – nicht nur im Hinblick auf die Schul-

stunden, sondern auch auf die Nutzung der schulischen Einrichtungen in Freizeit und Ferien – Folgen für die Arbeitsgestaltung und -gewohnheiten und die Freizeitgestaltung der Eltern? Und in welchem Umfang beeinflußt die Schule durch die Vermittlung von Fakten, Wissen und Gedanken die Einstellungen in der Gemeinde zu politischen, sozialen und kulturellen Angelegenheiten? Wenngleich die Liste noch weiter fortgesetzt werden könnte, wollen wir sie mit folgender Frage abschließen: Welcher Zusammenhang besteht zwischen den ausgewählten Aktionsparametern und ihrem Einfluß auf die soziale Struktur in der Schule und dem Abbau oder der Verstärkung von Strukturen in den Systemen außerhalb der Schule?

Wir müssen erkennen, daß es nicht möglich ist, jedes denkbare Element des Totalsystems abzugrenzen und seinen Wert bzw. seinen individuellen Beitrag zu den Bildungs-Outputs des Systems zu bestimmen. Dennoch ist es für jedes Evaluationsmodell unerläßlich, möglichst viele als signifikant erachtete Faktoren jeweils zu erkennen und ihren Einfluß zu ermitteln; denn je besser wir diese Faktoren isolieren, desto genauer wird unsere Analyse sein können.

In einem nächsten Schritt gilt es zu analysieren, wie unser Modell auf verschiedene Evaluationssituationen angewendet werden kann.

Anwendungsmöglichkeiten des Aufwands-Effektivitäts-Modells

Wie wir bereits dargelegt haben, liefern die herkömmlichen Kosten-Nutzen-Ansätze nicht die notwendigen Daten oder erfüllen nicht die bildungspolitischen Erfordernisse, um die wir uns hier bemühen. In diesem Abschnitt wird deshalb das von uns vorgeschlagene Aufwands-Effektivitäts-Analyse-Modell näher erläutert und seine Anwendung auf verschiedene Evaluationssituationen beschrieben. Für Zwecke dieses Beitrags wird unter »Programm« die Gesamtheit der Bemühungen einer Entscheidungseinheit zur Erreichung eines bestimmten Ziels oder eines Zielbündels verstanden. Auf das Bildungswesen übertragen, versteht man z. B. unter einem Programm die Sekundarschulbildung, die Hochschulbildung usw. Jedoch ist es schwierig, sämtliche Bemühungen zur Erreichung eines Teilziels, wie etwa Grundschülern das Lesen zu lehren, aufzulisten und zu beschreiben; d. h., es würde außerordentlich schwierig sein, die Kosten- und Programmelemente aller sich auf die Leseleistung der Kinder beziehenden Aspekte des gesamten Schulprogramms zu betrachten.

(1) Deswegen sind Programmalternativen verschiedene mögliche Wege, um dieselben oder ähnliche Ziele zu erreichen. Im Bildungswesen könnten etwa öffentliche und private Schulen Programmalternativen sein; wenn

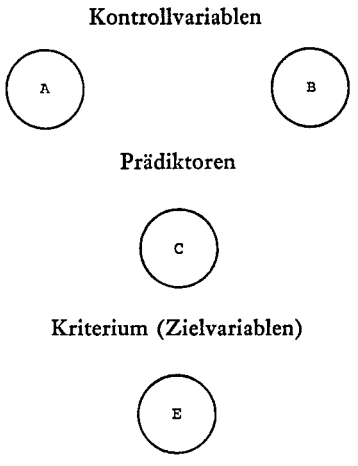
man davon ausgeht, daß unterschiedliche Schulen insgesamt oder zum Teil auf dieselben Ziele hinarbeiten, dann könnten die gesamten Programme dieser Schulen ebenfalls als Programmalternativen betrachtet werden. Unterschiedliche Schulen bieten unterschiedliche Programmalternativen. Folglich kann man den Erfolg verschiedener Programmalternativen zur Erreichung bestimmter Ziele der Programme evaluieren. Da das Niveau der Schüler bei den Programmen unterschiedlich ist, muß man davon ausgehen, daß die Outputs streuen; um die Programmalternativen zu evaluieren, muß man in der Lage sein, vorher die Unterschiede bei den Schüler-Inputs und bei den externen Systemen mit ihren Einflüssen festzustellen.

Dieser Begriff alternativer Programme kann noch erweitert werden. Falls Programme hinsichtlich ihrer unbeeinflußbaren Merkmale (Schüler-Inputs und externe Systeme) ähnlich sind, sich aber in der Höhe des finanziellen Inputs unterscheiden, kann man sie als Alternativprogramme zur Erreichung der gleichen oder ähnlicher Ziele ansehen. Man könnte die Aufwands-Effektivität alternativer Programme auch evaluieren – wobei sich alternative Programme durch die Höhe der finanziellen System-Inputs unterscheiden sollen – ohne daß man sich für die Art und Weise der Verwendung der finanziellen Mittel innerhalb des Systems interessiert (»Black box«-Ansatz).

Betrachten wir diese Art der Evaluation anhand des in Abb. 1 dargestellten Modells: Die Gruppe A von Variablen bezeichnet das externe System, Gruppe B die Schüler-Inputs, Gruppe C die finanziellen Inputs, Gruppe D die finanziell aufwendigen und beeinflußbaren Merkmale und Gruppe E die Ergebnisse. Bei Betrachtung dieses einfachen Diagramms und der darin aufgeführten Variablengruppen erkennt man, daß alternative Unterrichtsprogramme (bzw. die finanziellen Mittel der Schule) auf ihre Aufwands-Effektivität hin evaluiert werden können; dabei sind A und B die (unbeeinflußbaren) Kontrollvariablen, die finanziellen Inputs C die Prädiktoren und die Variablengruppe E das Kriterium (Zielvariable) (vgl. Abb. 2). Das Modell läßt sich für die Beantwortung folgender Frage anwenden: Zu welcher Veränderung des Ergebnisses (bei jeder einzelnen Ergebnisgröße) führt eine Erhöhung der finanziellen Inputs, gemessen in Dollar, wenn die Schüler-Inputs und die externen Systeme statistisch konstant gehalten werden?

(2) Eine zweite Form der Aufwands-Effektivitäts-Evaluation befaßt sich mit der Beurteilung bestimmter Unterrichtsprogramme. In diesem Fall würden wir bestimmte Unterrichtsgesamtprogramme von Schulen auf ihren jeweiligen Beitrag zu den Ergebnissen untersuchen, und zwar nachdem wir den Einfluß bestimmter unbeeinflußbarer Merkmale des jeweiligen Systems gebührend berücksichtigt haben. Wenn man also bloß Schulen als Insti-

Abbildung 2: Evaluation der Aufwands-Effektivität von einzelnen Unterrichtsprogrammen bzw. alternativem finanziellen Aufwand der Schule

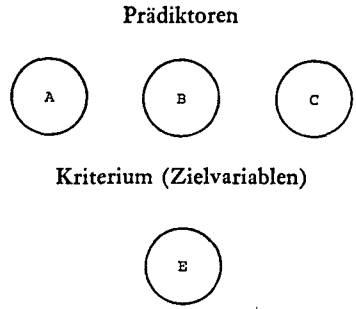


tutionen im Hinblick darauf evaluiert, was sie im Verhältnis zu den ihnen zur Verfügung stehenden menschlichen und finanziellen Ressourcen leisten, könnte das vorgestellte Modell für eine Aufwands-Effektivitäts-Analyse herangezogen werden. Mit anderen Worten, wenn man die finanziellen Inputs als vorgegebene Größen und folglich als Bestandteil des Systems betrachtet, so ist das Ausmaß, in dem es einer einzelnen Institution gelingt, ein von uns vorausgesagtes Ergebnisniveau zu erreichen, ein Maß für die Aufwands-Effektivität des Gesamtprogramms der Institution. Z. B. könnte man für eine Institution 1 mit den Schüler-Inputs S 1, den externen Systemmerkmalen E 1 und den finanziellen Inputs F 1 für die einzelnen Zielvariablen bestimmte zu erreichende Niveaus voraussagen: K 1,1, K 2,1, K 3,1 ... K i,1. Wenn die Institution diese Erwartungswerte für die Ergebnisse oder für die als vorteilhaft – zumindest nicht als nachteilig – angesehenen Wirkungen erreicht oder übertrifft, dann wird die Institution aufwandsgünstig (effizient) in bezug auf die einzelnen Ergebnisse geführt.

Die zweite Form von Aufwands-Effektivitäts-Untersuchung kann also für eine einzelne Schule unternommen werden. Die Evaluation eines einzelnen Schulprogramms würde aufgrund von statistisch abgeleiteten Erwartungswerten für dieses Programm unter Berücksichtigung seiner unbeeinflußbaren Merkmale erfolgen (vgl. Abb. 3). Die Aufwands-Effektivitäts-Evaluation einer Schule würde anhand von Werten erfolgen, die sich aus dem Verhältnis von erreichten zu erwarteten Ergebnissen bei den einzelnen

Zielvariablen errechnen. Eine Schule, deren erreichte Leistung den Erwartungswert einer Zielvariablen übertrifft, soll deshalb bezüglich dieser Zielvariablen als aufwandsgünstig gelten.

Abbildung 3: Evaluation der Aufwands-Effektivität von einzelnen Schulprogrammen



Aufwands-Effektivitäts-Wert (Bezirk i, Zielvariable l)

= F (Erreichter Wert für $E_{l,i}$, Erwartungswert für $E_{l,i}$)

d. h. $\dfrac{\text{Erreichter Wert für } E_{l,i}}{\text{Erwartungswert für } E_{l,i}}$

oder Erreichter Wert für $E_{l,i}$ – Erwartungswert für $E_{l,i}$

(3) Wir können den vom PPBS (Planning Programming Budgeting System) entlehnten Begriff der »Erfüllung einer gestellten Aufgabe auf verschiedenen Wegen« als eine brauchbare Grundlage für eine dritte Form von Aufwands-Effektivitäts-Evaluation ansehen. Die gestellte Aufgabe beinhaltet, daß das zu erreichende Ergebnis (Output) und das Programm vorher festgelegt wurden. Auf jeder Stufe des Programms stellt sich die Frage: Können wir durch eine mögliche Änderung der Produktions- oder Verteilungstechnik (a) den zeitlichen Ablauf der Produktion oder der Verteilung verbessern (d. h. die Programmziele in kürzerer Zeit erfüllen und dabei weniger Zeit der Schüler in Anspruch nehmen) oder (b) die Quantität und Qualität des Outputs anheben (d. h. innerhalb des Programms eine größere Anzahl von Schülern ausbilden, bzw. ein höheres Niveau bei den Lernzielen oder geringere unerwünschte Wirkungen erreichen) oder (c) die Einheitskosten oder Gesamtkosten der Produktion bzw. Verteilung senken (im Bildungswesen würde das bedeuten, die gleichen Ziele mit einem ge-

ringeren Aufwand an Dollar zu realisieren)? Bei der »Erfüllung einer gestellten Aufgabe auf verschiedenen Wegen« geht man von einem bestimmten oder vorgegebenen Programm aus und erweitert die Möglichkeiten für die Kombination von verschiedenen Inputverwendungen, wodurch das Programm abgewandelt wird. Für das hier anstehende Problem scheint dies die geeignete Methode zu sein. Die Frage bezüglich alternativer Bildungsprogramme führt zwar zu Antworten, die über die Aufwands-Effektivität von Gesamtbildungsprogrammen Aufschluß geben, lenkt den Blick aber nicht auf die Merkmale des Systems, die für die Erzeugung unterschiedlicher Bildungsergebnisse verantwortlich sind.

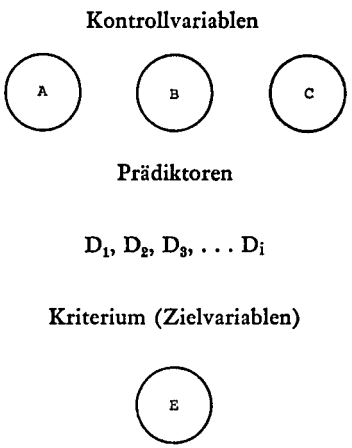
Es gibt natürlich von Ort zu Ort hinsichtlich der Qualität der verfügbaren bzw. zur Auswahl stehenden Ressourcen beträchtliche Unterschiede. Wenn der Ökonom z. B. Lehrer, Material usw. als Inputs des Systems betrachtet, rechnet er mit qualitativen Unterschieden der Inputs. In diesem Modell, das für Entscheidungen der Bildungsbehörden entwickelt worden ist, werden Kostenfaktoren wie Lehrer, Lehrbücher, Verwaltungs- und Hilfspersonal als finanzielle, beeinflußbare Merkmale des Systems angesehen. Jeder dieser Aktionsparameter stellt eine mögliche Verwendung des finanziellen Inputs dar.

Eine der wesentlichen Aufgaben des Staates liegt darin, den örtlichen Schulbezirken eine freie Entscheidung über die Verwendung qualitativ ausreichender Input-Faktoren zu ermöglichen, um ein effizientes Arbeiten des Schulbezirks sicherzustellen. In mancherlei Weise kommen die Bundesländer dieser Verpflichtung nach. Zum Teil hängen Wahlmöglichkeiten bei den zu verwendenden Input-Faktoren von der Wirtschaft des Bundeslandes, den unterschiedlichen Arbeitsmarktverhältnissen, dem Zugang zu höherer Bildung usw. ab. Die Landesregierung setzt bei den Wahlmöglichkeiten bezüglich der Qualität der zu verwendenden Inputs Grenzen durch landesrechtlich festgelegte Bildungserfordernisse und Landesbestimmungen für die Lehrbefähigung. Was man mit dem finanziellen Input in einem Schulbezirk erreichen kann (die Kaufkraft der finanziellen Mittel), wird also zum Teil von der Landesregierung, von der geographischen Region und u. U. sogar von den Gegebenheiten in der jeweiligen Gemeinde bestimmt.

Wir wiesen darauf hin, daß es nicht möglich ist, gleichzeitig den Ertrag zu maximieren und den Aufwand zu minimieren. Im Hinblick auf das hier aufgeworfene Problem bedeutet dies, daß es unmöglich ist, zur gleichen Zeit Programmziele in kürzerer Zeit zu erreichen, die Einheitskosten der »Produktion« der Bildungsergebnisse zu verändern und Bildungsziele auf einem höheren Niveau zu verwirklichen. Mehrere dieser Faktoren müssen als Nebenbedingungen des Programms vorgegeben werden, und jeweils nur ein Faktor kann als Gegenstand der Aufwands-Effektivitäts-Ana-

lyse bezeichnet werden. Eine Betrachtung der Aufwands-Effektivitäts-Evaluation bestimmter finanzieller Aktionsparameter des Systems (Lehrer, Lehrbücher, Verwaltungspersonal, Einrichtung) wurde schon angeregt. Zweck des Vorschlages ist es, durch die Ausübung der Wahlmöglichkeiten bezüglich der Verwendung der Ressourcen innerhalb des Systems den Output zu maximieren, während die Höhe des gesamten finanziellen Inputs und die Schüler-Inputs einschließlich der aufgewendeten Zeit als vorgegebene Nebenbedingungen in das Modell eingehen. In unserem Modell erfordert dieser Prozeß die Berücksichtigung der Variablengruppen A, B und C als Kontrollvariablen, jeweils einzelne Variablen D als Prädiktoren und die Variablengruppe E als Kriterien (vgl. Abb. 4).

Abbildung 4: Evaluation der Aufwands-Effektivität von Wahlmöglichkeiten bezüglich der Verwendung der Inputs



Eine andere Frage erhebt sich in diesem Zusammenhang zwangsläufig: Wenn die Eigenschaften der Schüler-Inputs und der finanziellen Inputs des externen Systems statistisch konstant gehalten werden, welches ist dann die Wirkung jedes einzelnen mit finanziellem Aufwand verbundenen Aktionsparameters des Systems auf erhöhte Bildungs-Outputs? Eine derartige Evaluation erfordert neben der Darstellung der Beziehung zwischen den finanziellen Aktionsparametern und den verschiedenen Ergebnisgrößen eine Untersuchung der Kostenfunktionen für die finanziellen Aktionsparameter.

Es gilt dann, die durch den Einsatz einer zusätzlichen Einheit bei jedem

finanziellen Aktionsparameter bewirkte Änderung des Outputs festzustellen. Auf dieser Stufe der Analyse sind wenigstens drei wesentliche Probleme zu erwarten:

(a) Es würde schwierig sein, für die Aktionsparameter exakte Kostendaten zu erhalten;

(b) es würden Schwierigkeiten bei der Behandlung der Aufwands-Effektivitäts-Schätzungen auftreten, wenn Wechselbeziehungen zwischen den Systemen bestehen;

(c) allgemeine Aussagen könnten schwerlich auf Einzelfälle übertragen werden (wenn eine solche Verallgemeinerung angestrebt würde).

Was das erste Problem betrifft, wären primärerhobene Daten aus der Schulpraxis natürlich wünschenswert. Jedoch liefert das Rechnungswesen gewöhnlich diese Information nicht. In den Fällen, in denen primäre Daten nicht erhältlich sind, müßte man eine Kostenfunktion aufstellen und daraus die Kosten ableiten; bei der Untersuchung einer Reihe von Fällen ließen sich Daten gewinnen, indem man aus dem Vorhandensein und dem Umfang verschiedener Aktionsparameter auf eine Kostenfunktion schließt, etwa die laufenden Bildungsausgaben. Auf diese Weise könnte man eine Kostenkurve erhalten, die die Produktionskosten jeweils den Aktionsparametern zuordnet. Eine solche Produktionsfunktion könnte man ableiten, indem man historische Daten oder Zeitreihendaten zugrunde legt, wie etwa eine Untersuchung von Adelson, Alkin, Carey und Helmer (1967); eine Produktionsfunktion läßt sich aber auch mit Hilfe von Querschnittsdaten konstruieren (Katzman 1967).

Für das zweite Problem, die zwischen den Systemen existierenden Wechselbeziehungen, läßt sich keine einfache Lösung finden. Man kann versuchen, die einzelne Variable von ihren Kovarianten durch geeignete statistische Verfahren zu isolieren. Aus den Ergebnissen über die Wechselbeziehungen zwischen den Kovarianten kann man dann jeweils den Erwartungswert für die Veränderung bestimmen, der sich auf den Einsatz einer zusätzlichen finanziellen Einheit in einem bestimmten Zwischenglied zurückführen läßt. Vielleicht könnten durch eine systematische Beurteilung Art und Umfang der gegenseitigen Abhängigkeiten aufgedeckt und isoliert werden. Ausgehend von den statistischen Daten, könnten dann den Elementen des Systems entsprechende Kostenvektoren zugeordnet werden. Darüber hinaus könnte evtl. mit Hilfe von Verfahren der Netzplantechnik ein tieferer Einblick in die Daten gewonnen werden.

Eine andere Lösungsmöglichkeit besteht in der Berücksichtigung systematisch gewonnener Expertenurteile, z. B. mit Hilfe der Delphi-Methode (Gordon/Helmer 1964; ferner Adelson/Alkin/Carey/Helmer 1967). Es könnte durchaus sinnvoll sein, eine Gruppe fachlich qualifizierter Entschei-

dungsträger aus dem Bildungsbereich mit verschiedenen Erfahrungen und Interessen zusammenzustellen. Sie könnten beauftragt werden, die Art und den Umfang der Wechselbeziehungen zwischen den Variablen zu prüfen und aus diesen Beziehungen ein Urteil über die Aufwands-Effektivität jedes einzelnen im System vorhandenen Aktionsparameters zu fällen. Dieser Delphi-Prozeß, der die verschiedenen Erkenntnisse zusammenfaßt, könnte – durch Diskussion und Darstellung abweichender Meinungen, Rückkoppelung bei den Teilnehmern und mehrere ergänzende Durchgänge für dasselbe Verfahren – zur Übereinstimmung oder mindestens doch zu einem Verständnis für die Minderheitenmeinungen führen.

Das dritte Problem bezieht sich auf die Schwierigkeit, Aussagen allgemeiner Art auf Einzelfälle zu übertragen. Eine mögliche Lösung dieses Problems hängt von der Entwicklung einer Typologie der Schule ab, deren Resultate sich als Moderator-Variable bei der Vorhersage der Ergebnisse in der Analyse verwenden läßt. Schwierigkeiten bestehen hinsichtlich des Einsatzes statistischer Verfahren (z. B. der aus einer Reihe von Daten abgeleiteten Regressionskoeffizienten) für die Schätzung der Erwartungswerte bei den Zielvariablen (Ergebnissen) im Einzelfall. Die Genauigkeit eines vorhergesagten Ergebnisses für eine einzelne Schule wird in hohem Maße von dem Typ der Schule abhängen, wie die Schule nämlich ihre finanziellen Aktionsparameter variiert. Um ein einfaches Beispiel zu nennen: Man würde von einer Veränderung der Zahl der Schüler pro Schulpsychologen an der Beverley Hills High School nicht die gleiche Wirkung erwarten wie an einer kleinen ländlichen Oberschule. Sicher spielt der Schultyp als Moderator-Variable bei der Vorhersage des Ergebnisses eine Rolle. Die Forschungsergebnisse von Klein, Rock und Evans (1967) beim Educational Testing Service über die Gruppierung von Variablen empfehlen sich vielleicht für die Lösung dieses Problems.

Zusammenfassung

In diesem Beitrag stellten wir den Unterschied zwischen Kosten-Nutzen-Analyse und Aufwands-Effektivitäts-Analyse dar. Wir zeigten, daß sich die Kosten-Nutzen-Analyse fast ausschließlich auf finanzielle Erträge bezieht und deshalb für die Beurteilung von Bildungsprozessen – da hier viele Ergebnisse nicht ökonomisch definiert werden können – von begrenztem Wert ist.

Weiterhin gaben wir einen Überblick über die verschiedenen Komponenten eines Modells, mit dessen Hilfe unserer Ansicht nach Entscheidungsträger Aufwands-Effektivitäts-Evaluationsuntersuchungen im Bildungs-

wesen durchführen können. In dem Modell wiesen wir auf die Notwendigkeit einer Betrachtung folgender Faktoren hin: Schüler-Inputs – d. h. Merkmale der in das System eintretenden Schüler; Bildungs-Outputs – d. h. kognitive und nicht-kognitive Veränderungen, die bei den Schülern eintreten, nachdem sie mit einem Unterrichtsprogramm konfrontiert worden sind; finanzielle Inputs – d. h. die für die Durchführung des Unterrichtsprogramms verfügbaren finanziellen Mittel; externe Systeme – d. h. die soziale, politische, rechtliche und ökonomische Struktur der Gesellschaft; und schließlich Aktionsparameter – d. h. jene Faktoren des Programms, die volkswirtschaftliche Werte (Ressourcen) verzehren und die durch die Verwaltung beeinflußt werden können.

Schließlich zeigten wir die Anwendungsmöglichkeiten des Aufwands-Effektivitäts-Modells in unterschiedlichen Evaluationssituationen und machten deutlich, wie man ein Modell für die Aufwands-Effektivitäts-Evaluation verschiedener finanzieller Inputs und einzelner Schulprogramme benutzen kann. Abschließend legten wir dar, daß das Aufwands-Effektivitäts-Evaluations-Modell die verschiedenen Möglichkeiten bei »Erfüllung einer gestellten Aufgabe auf verschiedenen Wegen« bewerten könnte.

Gene V. Glass

Die Entwicklung einer Methodologie der Evaluation

Das biologische Gesetz der Allometrie besagt, daß das Wachstum eines Organismus durch seine Form begrenzt wird. Organismen sind dadurch gekennzeichnet, daß ihr Wachstum, z. B. im Gegensatz zu Stalagmiten und Stalaktiten, an einem bestimmten Punkt zum Stillstand kommt. Man stelle sich vor, daß die Erbinformationen (genetic code) ein würfelförmiges Wachstum determinieren. Wenn die Umwelt eines solchen Organismus in bezug auf Erreichbarkeit von Nahrung, Stoffwechselumsatz usw. die Entwicklung von 8 Größeneinheiten für seine Erscheinungsform zuläßt, dann kann er sich nur bis zu zwei Größeneinheiten in jeder Dimension entwickeln. Muß ein Organismus kugelförmig wachsen, dann kann bei 8 Wachstumseinheiten sein Durchmesser maximal etwa 2,5 Einheiten betragen. Wenn jedoch die Erscheinungsform des Organismus quadratisch und nur eine Zelle stark ist, dann erlauben seine 8 Wachstumseinheiten es ihm (bei voller Reife), eine sehr große Fläche einzunehmen.

Ein Insekt atmet durch seine Haut; dadurch wird seine Größe von vornherein begrenzt. Wenn nämlich ein Insekt so groß wie ein Mensch wäre, würde seine sauerstoffaufnehmende Oberfläche nicht ausreichen, es am Leben zu erhalten. Denn beim Wachstum von 3 mm auf 1,80 m würde sein Volumen in soviel größerem Maße als seine Oberfläche zunehmen, daß es ersticken müßte. Die menschliche Lunge besteht aus einer so großen sauerstoffaufnehmenden Fläche, daß ein Wachstum von 1,80 m möglich ist. So begrenzt in der Biologie die Form das Wachstum.

Kenneth Boulding (1953, 21–32) übertrug das biologische Gesetz der Allometrie auf eine Vielzahl nicht-biologischer Phänomene. Dieses Gesetz kann bei der Untersuchung von Organisationen sinnvoll angewendet werden. Die Entwicklung einer sozialen Organisation wird durch die von ihr gewählte Form bestimmt. Das Entwicklungspotential einer Organisation bestimmt sich durch solche Dinge wie die für sie erreichbare Technologie und ihre Zukunftsperspektiven. Eine Organisation, die sich auf halbwöchentliche, direkte, persönliche Übermittlung von Informationen an alle

Mitglieder verlassen muß, kann wohl kaum größer werden als 100 Mitglieder. Durch die Verwendung von Telefonanlagen könnte die Organisation ihre Mitgliederzahl verdoppeln. Wenn jedoch die Organisation mit nur einer Kommunikation zwischen ihren Mitgliedern im Jahr auskommt, dann kann sie sehr viel größer werden. Vor 1860 hatte die Bundesregierung nie mehr als 5000 Beschäftigte. Bei den damals zur Verfügung stehenden technischen Hilfsmitteln (d. h. z. B. Büromaterial und Schreibkräfte) hätte die Zahl der Beschäftigten nicht erhöht werden können, ohne die Arbeitsfähigkeit der Organisation zu gefährden. Ziel und Stand der Entwicklung von Organisationen haben in der Gegenwart und für die Zukunft eine Konzeption von sich selbst. General Motors könnten schnell die in der Welt führenden Hersteller von Damenunterwäsche werden. Diese Rolle dürfte allerdings mit dem Selbstkonzept von General Motors nicht übereinstimmen; deshalb werden sie weiterhin Autos herstellen.

Allometrie steuert die Entwicklung der Organisation von Menschen, Dingen und Ideen. Die Entwicklung einer wissenschaftlichen Disziplin wird teilweise durch die von ihr gewählte Form bestimmt. Ihre Form ist in einem Entwicklungsgesetz enthalten, das von den Begründern der Disziplin teils zufällig gefunden, teils planmäßig erarbeitet wurde. Die Elemente dieses Entwicklungsgesetzes bestimmen z. B. die Gegenstände des Interesses, die zu ihrer Untersuchung benutzten Methoden und Verfahren, d. h. den Charakter der Disziplin.

Das Gesetz der Allometrie findet somit offensichtlich im sozialen Bereich eine Erweiterung: Form begrenzt Entwicklung (Wachstum), Entwicklung begrenzt Nützlichkeit. Einige ökonomische, soziale und wissenschaftliche Organisationen haben eine Organisationsform, die ihre Entwicklung hemmt und ihren gesellschaftlichen Nutzen einschränkt. Die Entwicklung anderer Organisationen schlägt fehl oder ist überflüssig.

Ziel meines Beitrags ist es, vier Modelle pädagogischer Evaluation darzustellen, ihre Konzeption zu bestimmen sowie ihre Entwicklungsmöglichkeiten und ihren gesellschaftlichen Nutzen zu beurteilen.

Ich werde Tylers Modell, das Akkreditationsmodell, das Management-System-Evaluationsmodell und das Zielkomplex-Modell (composite-goal model) untersuchen.

Pädagogische Forschung und Evaluation

Vor einer Analyse der vier Evaluationsmodelle soll zunächst zwischen pädagogischer Evaluation und pädagogischer Forschung eine Unterscheidung getroffen werden. Diesen Versuch, Forschung und Evaluation zu unter-

scheiden, sollte man weder als überflüssig noch als kleinlichen Aristotelismus ansehen. Denn abstrakte, verbale Definitionen beeinflussen das Verhalten. So wird manches Projekt der pädagogischen Forschung unzulänglich durchgeführt, weil man es Evaluation nennt; doch weit mehr Evaluationsuntersuchungen sind nutzlos, weil sie als pädagogische Grundlagenforschung behandelt werden.

Einfache verbale Definitionen von Forschung und Evaluation schließen sich somit nicht gegenseitig als wertlos aus. Es ist unzureichend, Forschung als Suche nach dem Verständnis von Phänomenen in Systemen von in Beziehung stehenden Phänomenen zu definieren, in denen Verständnis als die Fähigkeit, vorherzusagen und zu kontrollieren, bestimmt wird. Auch Evaluation versucht vorherzusagen und zu kontrollieren, versucht die Sachverhalte mit Methoden vorherzusagen und zu kontrollieren, die sich von den Inhalten und Methoden der Forschung unterscheiden.

Die Schwierigkeit, zwischen pädagogischer Forschung und pädagogischer Evaluation zu unterscheiden, ergibt sich aus dem Mangel an treffenden Beispielen für beide Bereiche. Die meisten empirischen Untersuchungen über pädagogische Probleme verbinden Evaluations- und reine Forschungsfragen in unterschiedlichem Ausmaß. Der Versuch, innerhalb der pädagogischen Untersuchungen zwei Gruppen zu bilden, wäre ähnlich verwirrend wie jeder vergleichbare Versuch einer Unterscheidung zweier Begriffe in den Sozialwissenschaften. Es würden sich zwei kleine Gruppen mit der Bezeichnung *Forschung* und *Evaluation* und eine große mit der Bezeichnung *Anderes* ergeben. Wissenschaftler, die Taxonomien in den Sozial- und Verhaltenswissenschaften aufstellen, erfahren die Schwierigkeiten besonders, denen sich Zoologen in geringerem Umfang gegenüber sehen, wenn sie Wale und Tümmler in ihre Kategoriensysteme einordnen.

Obwohl man den Unterschied zwischen Forschung und Evaluation durch die Analyse von Projekten oder Untersuchungen kaum feststellen kann, lassen einzelne Probleme oder Fragen sich durchaus als Forschung oder Evaluation einordnen. Doch sogar dabei wird die Unterscheidung dadurch erschwert, daß beide Bereiche sich lediglich in bezug auf zusammenhängende Charakteristika, wie z. B. die Motive des Forschers, die Beziehung bestimmter Ergebnisse zu anderen, die Verwendung der Ergebnisse, unterscheiden lassen, so daß die Bereiche unmerklich ineinander übergehen. In Forschung und Evaluation wird empirisch und theoretisch gearbeitet; in beiden Bereichen verwendet man zum großen Teil dieselben Techniken (inferenzstatistische Analysen, experimentelle Versuchsanordnungen, Psychometrie, Umfrageanalysen usw.); Forschung und Evaluation führen zu Ergebnissen, die nützlich und aussagekräftig sind. Und dennoch unterscheiden sich Forschung und Evaluation deutlich.

Die Autoren von »Research for Tomorrow's Schools: Disciplined Inquiry for Education« (Cronbach/Suppes 1969, 20–21) unterscheiden zwischen *entscheidungsorientierter* (decision-oriented) und *schlußfolgerungsorientierter* (conclusion-oriented) Forschung:

> Bei einer entscheidungsorientierten Untersuchung ist es Aufgabe des Forschers, die von den Entscheidungsträgern gewünschten Informationen zu liefern; zu Entscheidungsträgern zählen z. B. Beamte der Schulverwaltung, Regierungsvertreter, Projektleiter. Die entscheidungsorientierte Untersuchung ist eine Auftragsuntersuchung. Der Entscheidungsträger glaubt, daß er Informationen für die Planung seiner Handlungen braucht, und stellt dem Forscher entsprechende Fragen. Die schlußfolgerungsorientierte Untersuchung ist dagegen durch das Engagement und die Hypothesen des Forschers charakterisiert. Der Entscheidungsträger kann bestenfalls das Interesse des Forschers für ein Problem wecken. Der Forscher formuliert dann seine eigene Fragestellung, die meist eher eine allgemeine Frage als eine Frage über eine bestimmte Institution ist. Das Ziel besteht darin, das ausgewählte Problem begrifflich zu fassen und zu verstehen; ein einzelnes Ergebnis ist lediglich ein Mittel dazu. Deshalb konzentriert sich der Forscher auf Personen und Einrichtungen, von denen er aufschlußreiche Erkenntnisse erwartet.

Schlußfolgerungsorientierte Untersuchungen fallen zum großen Teil unter das, was hier als Forschung bezeichnet wird; der Begriff »entscheidungsorientierte Untersuchung« charakterisiert Evaluation.

Als eine erste noch nicht befriedigende Unterscheidung könnte man sagen, daß pädagogische Evaluation den *Wert*, pädagogische Forschung dagegen die wissenschaftliche *Wahrheit* einer Sache einzuschätzen versucht. Sieht man davon ab, daß Wahrheit ein hoher Wert ist und von daher alles, was wahr ist, wertvoll ist, leistet diese Unterscheidung recht gute Dienste, um Forschung und Evaluation gegeneinander abzugrenzen. Die Unterscheidung kann präziser gefaßt werden, wenn man Wert mit gesellschaftlichem Nutzen gleichsetzt und wissenschaftliche Wahrheit an Hand von zwei ihrer vielen Merkmale identifiziert:

1. empirische Überprüfbarkeit (verifiability) eines allgemeinen Phänomens [1] mit allgemein-verbindlichen Forschungsmethoden;
2. logische Konsistenz.

Die Unterscheidung zwischen dem Nachweis eines Wertes (Evaluation) und der wissenschaftlichen Wahrheit (Forschung) erhält nun mehr Gewicht.

Evaluation zielt direkt auf die unmittelbare Bewertung gesellschaftlichen Nutzens. Forschung mag den Nachweis von gesellschaftlichem Nutzen bringen, jedoch nur indirekt, weil empirische Überprüfbarkeit eines allgemeinen Phänomens und logische Konsistenz möglicherweise von grund-

legendem gesellschaftlichen Nutzen sein können. Um Evaluatoren und Forscher unterscheiden zu können, empfiehlt es sich zu fragen, ob man eine Untersuchung als Fehlschlag ansehen würde, wenn sie keine Informationen über den Nutzen des untersuchten Phänomens lieferte. Als Forscher wird man wahrscheinlich die Frage verneinen.

Forschung zielt auf die Abschätzung von drei unterschiedlichen Aspekten eines Gegenstands:

1. empirische Überprüfbarkeit von Forschungsgegenständen mit Hilfe allgemein-verbindlicher Methoden,
2. logische Konsistenz,
3. gesellschaftlicher Nutzen.

Exakte Forschung versucht abzuschätzen, bis zu welchem Grad jeder Aspekt Wirklichkeit ist. In Abbildung 1 sind einige Forschungsgebiete der Psychologie in bezug auf das Ausmaß klassifiziert, in dem sie jedes der obigen drei Phänomene zu beurteilen versuchen.

Die drei Winkel der Pyramide in der Abbildung repräsentieren drei un-

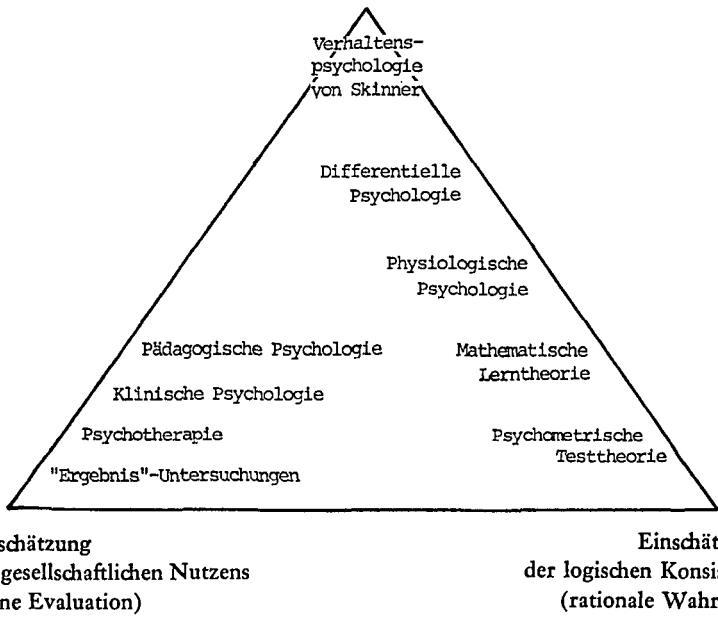


Abb. 1: Klassifikation psychologischer Forschungsansätze in bezug auf ihre Ziele.

terschiedliche Forschungsintentionen. Je näher ein Forschungsgebiet an einen der Winkel in dieser Pyramide heranreicht, desto stärker versucht es, die durch den Winkel repräsentierte Forschungsintention zu verwirklichen.

Das Tylersche Evaluationsmodell

Das erste Modell der Curriculumevaluation entstand im Verlauf der Eight-Year-Study. Dieses Modell wurde während der dreißiger Jahre von Ralph W. Tyler und dem Evaluations-Team der Eight-Year-Study erarbeitet. Die von Tyler und seinen Mitarbeitern entwickelten Evaluationsverfahren finden sich in Veröffentlichungen von Smith und Tyler (1942) und Tyler (1951). Folgende Aspekte charakterisieren das Tylersche Evaluationsmodell:

(1) *Formulierung der Ziele.* Bestimmung der allgemeinen Ziele des Curriculum.

(2) *Klassifikation der Ziele.* Entwicklung eines Zielkatalogs zur rationellen Abwicklung der theoretischen und praktischen Arbeit.

(3) *Definition der curricularen Ziele in Verhaltensbegriffen.* Dieses Merkmal wurde zum Kern des Tylerschen Modells. Einige moderne Methoden der Evaluation, die sich stark auf die Formulierung spezifischer Verhaltensziele stützen, sind nicht über Tylers Gedanken zur Evaluation hinausgekommen.

(4) *Entwurf von Situationen, in denen die Erreichung der Lernziele nachgewiesen werden kann.*

(5) *Entwicklung oder Wahl von Bewertungstechniken* (standardisierte Tests, informelle Tests, Fragebogen usw).

(6) *Sammlung und Interpretation von Verhaltensdaten.* Der letzte Schritt im Evaluationsprozeß besteht in der Messung des Schülerverhaltens und dem Vergleich zwischen den Verhaltensdaten mit den vorher formulierten Verhaltenszielen. Das Curriculum wird dann wegen seiner so nachgewiesenen Erfolge anerkannt und wegen seiner Fehlschläge kritisiert.

Curriculumevaluation nach Tyler berücksichtigt fast ausschließlich das *Verhalten* der Schüler. Die Ziele müssen in Verhaltensbegriffen formuliert werden; lediglich Verhaltensdaten in bezug auf das angezielte Verhalten sind vom Evaluator zu berücksichtigen. Die Curriculum-Evaluatoren bewerten nur die *Ergebnisse* des Unterrichts und nicht die *Mittel*, die zu diesen Ergebnissen führen.

Die Auffassung moderner Curriculum-Evaluatoren, lediglich die *Ergebnisse* der Erziehung und nicht die Mittel der Erziehung zu evaluieren, läßt sich nicht rechtfertigen. Mit Ausnahme des Elementarwissens (z. B. Schreiben und Rechnen) zielen die meisten Lernziele auf Verhaltensweisen, die sich wohl erst Jahre *nach* dem Ende des Unterrichts zeigen. Einige Ziele

sind der Sache nach unbeobachtbar, z. B. daß ein Schüler nach Erreichen seiner Volljährigkeit in geheimer Wahl intelligent und rational entscheiden kann.

Für einen großen Teil des Gesamtcurriculum – vielleicht seinen größten Teil – können die wirklichen von den Pädagogen angestrebten Verhaltensweisen nicht beobachtet werden. Deshalb muß der Unterricht durch die Beobachtung *stellvertretender* Ereignisse oder Verhaltensweisen evaluiert werden. *Stellvertretende* Verhaltensweisen stehen anstelle der letztlich angezielten Verhaltensweisen, die aus ökonomischen oder ethischen Gründen nicht beobachtbar sind. Ein Verhalten in einer stellvertretenden Situation läßt nur bedingt Schlüsse über das entsprechende Verhalten in der wirklichen oder letztlich gemeinten Situation zu. Ein großer Teil der Evaluation, der in der Einschätzung von Leistungsdaten in bezug auf Verhaltensziele besteht, schafft nur einen geringen Nachweis darüber, ob der Schüler das tatsächliche Unterrichtsziel, das im allgemeinen in der Übertragung oder Verallgemeinerung auf eine nicht-schulische Situation besteht, erreicht hat oder erreichen wird.

Wenn man im Rahmen der Evaluation eine solche Beweisführung mit stellvertretenden Verhaltensweisen akzeptiert, müssen auch andere Formen stellvertretender Verhaltensweisen akzeptiert werden. Zu diesen anderen Formen gehören nicht ausschließlich Schülerverhaltensweisen. Daß eine bestimmte Unterrichtseinheit logisch relevant ist, daß ein Lehrplan frei ist von unnötigen Unterbrechungen und daß Tests als Strafmittel benutzt werden, sind ebenso *stellvertretende Hinweise* darauf, ob Schüler das Unterrichtsziel erreichen oder nicht. Somit gibt es zwingende Gründe dafür, in der Curriculumevaluation Schülerverhalten nicht nur an in Verhaltensbegriffen formulierten Zielen zu messen. Man muß ein breiteres Spektrum von Daten in Betracht ziehen. Auch die Lehrer, die Curriculummaterialien, die Organisationspläne usw. müssen beobachtet und beurteilt werden. In vielen Fällen sollten die daraus gewonnenen Daten denen des Schülerverhaltens vorgezogen werden.

Im traditionellen Denken über pädagogische Evaluation war man der Überzeugung, daß Urteile subjektiv sind und daher sich nicht für eine Evaluationsuntersuchung eignen. Zweifellos sind Urteile subjektiv, aber sie können objektiv gesammelt und dargestellt werden. Darüber hinaus macht die Subjektivität von Werturteilen diese zu wichtigen Determinanten für den Erfolg eines Curriculum. Es ist sinnlos, festzustellen, daß das Urteil eines Schulleiters subjektiv ist, wenn sein Urteil, daß ein Curriculum wertlose Ziele hat, ihn veranlaßt, die Weiterentwicklung des Curriculum durch Entzug seiner Förderung zu verhindern. Urteile, Einstellungen und Gefühle der Befriedigung sind subjektiv. Jedoch können sie über

den Erfolg oder Mißerfolg eines Curriculum entscheiden und objektiv gemessen werden. Daher müssen sie vom Evaluator berücksichtigt werden.

Viele gegenwärtige Veröffentlichungen über Evaluationsmethoden sind von Tyler beeinflußt (vgl. Bruner 1966; Cronbach 1963; Carroll 1965). An Tylers Modell erinnert auch Cronbachs Beitrag von 1963, in dem er die detaillierte Analyse von curricularen Zielen, die Notwendigkeit, Schülerleistungen mit Verhaltenszielen zu vergleichen, und die Irrelevanz des Vergleichs von Curricula mit unterschiedlichen Zielen betont.

> Das Ziel, Curricula miteinander zu vergleichen, sollte nicht die Pläne für die Evaluation bestimmen ... Da die Ergebnisse von Gruppenvergleichen sehr fragwürdig sind, sollte man meiner Meinung nach eine gute Untersuchung in erster Linie so planen, daß sie es erlaubt, die Leistungen einer genau umschriebenen Gruppe am Ende eines Curriculum im Hinblick auf die wesentlichen Ziele und Nebenwirkungen zu bestimmen. (Cronbach 1963; 42–43, 47 *f.*)

Carrolls Ausführungen erinnern an Cronbach und damit indirekt auch an Tyler:

> Ich möchte Curriculumevaluation als den Prozeß bezeichnen, mit dem festgestellt wird, ob ein vorliegendes Curriculum seine Ziele erreicht, oder vielmehr, welche Ziele es unter welchen Bedingungen und für welche Schüler erreichen kann ... Aber in der Regel haben Curricula keine genau übereinstimmenden Ziele, und im allgemeinen wäre es unangemessen, sie zu vergleichen, weil das mehr oder weniger philosophische Fragen über die Vergleichbarkeit ihrer jeweiligen Ziele aufwerfen würde (Carroll, 1965).

Als direkte Erwiderung auf diese Einwände gegen den Vergleich von Curricula schrieb Scriven (1967):

> Die Schlußfolgerung scheint zwangsläufig zu sein, daß vergleichende Evaluation (ob nun sekundäre oder Ergebnisevaluation) die beste Methode für die Probleme der Evaluation darstellt.

Zwei ähnliche Gesichtspunkte wurden von Cronbach und Carroll zur Unterstützung ihrer Argumente vorgebracht: Carroll behauptet, der Vergleich zwischen Curriculum A und Curriculum B sei nutzlos, weil man von diesem Vergleich nicht auf Vergleiche von A mit anderen konkurrierenden Curricula generalisieren kann. Cronbach (1963) führte aus:

> Bestenfalls kann ein solcher Versuch zwei bereits bestehende Curricula miteinander vergleichen. Wenn man sich sehr bemüht, das schlechtere Curriculum zu optimieren, führt dies wahrscheinlich zur Umkehrung des Urteils über den Versuch.

Carroll und Cronbach sprechen sich gegen die vergleichende Versuchsmethode aus, weil das, was sie erreichen soll, besser von der Forschung ver-

wirklicht wird. Wenn der vergleichende Versuch in der Evaluation kritisiert wird, weil der Vergleich der Curricula A und B keine Informationen darüber liefert, wie der Vergleich von A mit einem unbekannten und nicht näher bezeichneten Curriculum C aussehen würde (wie Carroll behauptet), dann ist diese Methode auch abzulehnen, weil sie keine Informationen darüber liefert, ob später einmal ein Curriculum entwickelt werden wird, das besser als alle heute vorhandenen ist. Überdies vergleicht eine heute durchgeführte vergleichende Evaluation nur die gegenwärtigen Versionen von zwei oder mehreren Curricula.

Cronbachs Feststellung, daß eine größere Anstrengung, das schlechtere von zwei Curricula zu verbessern, dies wahrscheinlich besser als das konkurrierende Curriculum machen würde, ist wahrscheinlich richtig. Welche Auswirkung würde jedoch eine ähnliche größere Anstrengung auf das Curriculum haben, das zunächst besser war? Falls man nicht einen groben Fehler bei der Weiterentwicklung des zunächst überlegenen Curriculum macht, werden trotz größerer Anstrengungen an *beiden* Curricula beide bei späteren Evaluationsuntersuchungen ihre relative Qualität behalten.

Carroll wies darauf hin, daß Curricula gewöhnlich nicht die gleichen Ziele haben und daß ihr Vergleich philosophische Probleme über die Vergleichbarkeit von verschiedenen Lernzielen aufwirft. Die *Wahl* zwischen zwei konkurrierenden Curricula mit in hohem Maße unterschiedlichen Zielen zu treffen wirft philosophische oder ethische Fragen oder Fragen über den relativen Wert bestimmter von einer Gesellschaft anerkannter Wertvorstellungen nur auf, löst sie jedoch nicht. Diejenigen, die Entscheidungen über die Adaptation von Curricula und Innovationen treffen, stehen vor der Aufgabe, diese Fragen zu lösen. Ich bezweifle, daß sie sich adäquat lösen lassen und eine rationale Entscheidung getroffen werden kann, bevor nicht empirische Daten darüber vorliegen, wie gut ein Curriculum seine eigenen Ziele, die Ziele konkurrierender Curricula und allgemeine Ziele erreicht.

Viele Entscheidungen zwischen konkurrierenden Curricula werden unvermeidbar philosophische Fragen nach dem Wert aufwerfen. Es ist nicht Aufgabe des Evaluators, diese Fragen selbst zu beantworten; aber er spielt in der Zusammenarbeit mit dem Curriculumentwickler, den Schulpsychologen, Beamten der Schulverwaltung bei der Klärung der Fragen und der Sammlung der entsprechenden empirischen Daten eine äußerst wichtige Rolle.

Nach einer der wichtigsten kritischen Äußerungen Cronbachs trägt die vergleichende Methode der Evaluation nur wenig zum Verständnis des Curriculum bei:

»Bestenfalls kann ein solcher Versuch zwei bereits bestehende Curricula

miteinander vergleichen. Wenn man sich sehr bemüht, das schlechtere Curriculum zu optimieren, führt dies wahrscheinlich zur Umkehrung des Urteils über den Versuch« (Cronbach 1963, 42, *47 f.*).

Scriven (*1967, 65, 84*) antwortete Cronbach auf diesen Punkt:

> ... Verständnis ist nicht unser *einziges* Ziel in der Evaluation. Wir sind ebenso an Fragen der Unterstützung, Ermutigung, Annahme, Belohnung, Verbesserung usw. interessiert.
>
> In einigen Fällen können diese wichtigen Fragen zwar durchdacht werden, jedoch nicht dadurch vollständig beantwortet werden, daß man die Überlegenheit eines Curriculum nachweist.

Obwohl sich Cronbachs und Scrivens Auffassungen in diesem Punkt unterscheiden, haben sie doch ähnliche Zielsetzungen. Man wird Scriven zustimmen müssen: Probleme der Einführung eines Curriculum, Entscheidungen zwischen konkurrierenden Curricula usw. erfordern eine vergleichende Evaluation. Cronbachs Ausführungen dagegen scheinen sich eher an den Curriculumentwickler als an denjenigen zu wenden, der ein Curriculum auswählt. Der Curriculumentwickler will wahrscheinlich Daten finden, die die Vor- und Nachteile seiner Materialien weit genauer zeigen als die Daten, die er aus einem Vergleich seines Materials mit dem eines konkurrierenden Curriculum erhält. Auf die Mitteilung, daß sein Curriculum in einem vergleichenden Versuch mit seinem Hauptkonkurrenten unterlegen ist, würden die meisten Curriculumentwickler wahrscheinlich auf eine der zwei folgenden Arten reagieren:

(1) Sie würden behaupten, daß der Versuch ungültig, subjektiv und ungerecht war, oder

(2) sie würden behaupten, daß ihr Curriculum mit seinem Konkurrenten nicht hinsichtlich seiner zentralen Ziele verglichen wurde.

In beiden Fällen werden ihnen diese Daten für die weitere Entwicklungsarbeit nicht nützlich erscheinen. Sie können sogar insofern einen nachteiligen Effekt haben, als sie die Curriculumentwickler veranlassen, die Ziele ihrer Materialien zu ändern und von nun an Ziele nicht wegen ihres intrinsischen Wertes, sondern wegen ihrer leichteren Erreichbarkeit zu vertreten.

Wenn der Curriculumentwickler wissen will, *wie* und *warum* seine Materialien in einer bestimmten Weise wirken, werden ihm Vergleichsdaten wenig nützen. Dennoch ist vergleichende Evaluation auf einer bestimmten Ebene notwendig. Die Kritik, die sich gegen Vergleiche von Curricula richtet und statt dessen feststellt, welche Lernziele von welchen Schülern erreicht werden, setzt sich darüber hinweg, daß in der Aufstellung der Ziele für jedes Curriculum bereits ein Vergleich enthalten ist. Niemand wird z. B. so töricht sein, für ein Curriculum folgendes Ziel zu formulieren:

Schreiben sie zehn Wörter pro Minute mit nicht mehr als fünf Fehlern! Denn bestehende Curricula sind diesem Curriculum bereits überlegen. In einer Phase der Evaluation eines Curriculum müssen die impliziten Vergleiche aufgedeckt und untersucht werden.

Ob man ein vergleichendes oder nicht vergleichendes Vorgehen wählen soll, wurde im einzelnen analysiert, weil sich in diesem Punkt das Tylersche Modell und einige andere Modelle deutlich unterscheiden. Man kann zu Recht sagen, daß der Vergleich zwischen Schülerleistung und vorher formulierten Verhaltenszielen – anstelle des Vergleichs von Schülerleistung mit der Leistung unter anderen Bedingungen – für das Tylersche Modell charakteristisch ist.

Im Laufe fast eines halben Jahrhunderts wurde Tylers Evaluationsmodell immer weiter ausgearbeitet, bis es alle seine Möglichkeiten entwikkelt hatte. Die Beharrlichkeit seiner Verteidiger (vgl. z. B. Walbesser 1963 und 1966) und sein orthodoxer Charakter deuten darauf hin, daß sein Potential verwirklicht wurde und daß es aus der Sicht seiner Vertreter volle Verwendbarkeit erreicht hat. Das heißt, wir haben das Tylersche Modell in ausgereifter Form vor uns. Worin liegt der Nutzen dieses Modells? Ist es den gegenwärtigen Erfordernissen pädagogischer Evaluation angemessen?

Zu Beginn des zweiten Jahrzehnts des zwanzigsten Jahrhunderts wurde mit etwa 4 % nur ein kleiner Teil des in den Vereinigten Staaten für öffentliche Erziehung aufgewandten Geldes durch Steuern erhoben und von der Bundesregierung verteilt. Ermächtigt durch Gesetze, wie die Smith-Hughes und Smith-Lever-Gesetze, wurden diese Mittel in erster Linie für die Berufsausbildung und für die Landgemeinden ausgegeben. Die Art und der Umfang der für öffentliche Erziehung durch die Bundesregierung verteilten Mittel änderte sich zwischen 1920 und 1958 nur wenig. Konfrontiert mit neuen Problemen und zunehmendem öffentlichen Interesse für Erziehung, erließ der Kongreß den National Defense Education Act von 1958, den Elementary and Secondary Education Act von 1965 und den Education Professions Development Act von 1967. Damit verdoppelten sich beinahe die finanziellen Aufwendungen des Bundes für das öffentliche Erziehungswesen; sie stiegen in den Jahren zwischen 1958 und 1968 von druchschnittlich 4 % auf 7 %.

Der Hauptanteil dieser Ausgaben wird eher für Innovationen und Reformen im Erziehungswesen verwendet als für die bloße Ausstattung der Schulen oder das Herstellen neuer Schulbücher. Obwohl der aus Bundesmitteln stammende Betrag für innovative Programme, gemessen an den Gesamtausgaben für das Erziehungswesen, gering ist, hat er doch auf viele Schulen eine starke Auswirkung gehabt.

Für das große Interesse an der Entwicklung von Modellen der pädagogischen Evaluation gibt es drei Gründe:

Erstens steigt der Anteil der Finanzen an, die von seiten des Bundes für die öffentlichen Schulen aufgebracht werden. Nach einigen Voraussagen werden 1990 etwa 50 % der Kosten für den *tertiären Bildungsbereich* von der Bundesregierung aufgebracht werden. Durch diese Neuverteilung der Finanzen wird auch die Notwendigkeit, Curricula zu evaluieren, d. h. zu beschreiben und zu beurteilen, größer werden. Wenn alle Bildungsausgaben von der örtlichen Gemeinde aufgebracht werden, ist eine unmittelbare Rückmeldung über den Erfolg neuer Curricula gewährleistet, die von den Steuerzahlern in den Gemeinden bei ihrer Entscheidung berücksichtigt werden kann.

Wenn jedoch die Kosten eines neuen Curriculum auch mit den Steuergeldern aus anderen Bundesländern finanziert werden, dann können Fehlleistungen in der Entwicklung des Curriculum eher von der örtlichen Gemeinde verschleiert werden. Deshalb war die Forderung, formale Evaluation gesetzlich zu verankern, und die dann tatsächlich nachfolgende Gesetzgebung sinnvoll.

Der zweite und dritte Grund für die zunehmende Bedeutung der Evaluation sind die Bürgerrechtsbewegung und das bildungspolitische Engagement der Lehrer. Diese beiden Gründe sollen hier nicht weiter erörtert werden. Denn es wird fast täglich in den Massenmedien deutlich, daß Minderheitengruppen und eine aggressive Lehrerschaft sich gegen das pädagogische Establishment wenden. Jede Seite beruft sich mit zunehmender Häufigkeit auf empirische Ergebnisse über die Auswirkung von Erziehung, um ihre Ansichten zu erklären. Ein Soziologe, Dan Lortie an der Universität Chicago, sagte einem staatlich geprüften Evaluator voraus, daß er eine Funktion ausüben würde, die der des staatlich geprüften Wirtschaftsprüfers ähnlich wäre. Seine Voraussage wird eintreffen, wenn die folgenden Vorstellungen aus dem Bericht der National Advisory Commission on Civil Disorders (1968, 451) realisiert werden:

> Um die öffentlichen Schulen in verstärktem Maße dazu zu bringen, Rechenschaft abzulegen (accountability), sollten die Ergebnisse ihrer Leistung der Öffentlichkeit zugänglich gemacht werden. Solche Informationen sind in einigen, aber nicht in allen Städten zugänglich. Wir sehen keinen Grund, nützliche und relevante Unterlagen über die Leistung der Schulen (nicht der einzelnen Schüler) der Öffentlichkeit vorzuenthalten, und empfehlen daher, daß alle Schulsysteme ihre Aufmerksamkeit darauf richten, die Öffentlichkeit voll zu informieren.

Die Forderung von seiten der Öffentlichkeit und der Bürokratie nach Evaluation überraschte die Wissenschaftler. Innerhalb kürzester Zeit wur-

de Evaluation zu einem zentralen Problem, wobei man zunächst die Frage beantworten mußte, was denn Evaluation eigentlich sei.

Die Wissenschaftler, die sich als erste mit Veröffentlichungen an einen großen Kreis von Pädagogen wenden konnten, waren auch schon an der Curriculumbewegung der fünfziger Jahre beteiligt. Ihren Veröffentlichungen lag schon mehrere Jahre vor 1965 ein bestimmtes Verständnis von Evaluation zugrunde. Sie betrachteten Evaluation als einen untergeordneten Teil der Curriculumforschung und -entwicklung. Für die Bundesgesetzgebung entwarfen sie *Evaluationsrichtlinien,* die auf Tylers Evaluationsmodell beruhten. Modelle der Curriculumevaluation waren in der Pädagogik durchaus bekannt. Sie hatten ihren Ursprung in den Bereichen des pädagogischen Testens und der Curriculumentwicklung und zielten daher bis in die späten sechziger Jahre hinein vornehmlich auf objektive Leistungsmessung, Lernzieltaxonomien und in Verhaltensbegriffen formulierte Lernziele.

Bald wurde deutlich, daß die in der jüngsten Bundesgesetzgebung geforderte Art der Evaluation nicht Curriculumevaluation im traditionellen Sinn, sondern eine umfassendere Form der Evaluation war. Benötigt wurde nicht nur ein Verfahren zur Verbesserung des Curriculum, worunter man im allgemeinen gedrucktes Unterrichtsmaterial verstand. Man brauchte vielmehr ein Evaluationsmodell, mit dem man den Wert von Bildungseinrichtungen einschätzen konnte, die so verschieden waren, wie z. B. ein fahrbares Lernlaboratorium für Kinder von nicht ortsgebundenen Arbeitern, ein Computersystem zur Wiederauffindung von Forschungsergebnissen für Lehrer und ein Theater für sozial benachteiligte Kinder.

Das Tylersche Modell der formativen Curriculumevaluation eignet sich nicht für die Evaluation der Lehrerkompetenz, der Ausstattung von Bildungseinrichtungen, der Organisationspläne, der Begründung eines Curriculum oder des Kosten-Effektivitäts-Verhältnisses. Solche Probleme sind für den sich am Tylerschen Modell orientierenden Curriculum-Evaluator von geringem Interesse. Wenn jedoch Evaluatoren gegenüber ihren Auftraggebern und den Adressaten der Erziehung die volle Verantwortung tragen sollen, müssen sie sich solchen Problemen stellen. Daher wird sich das Tylersche Modell der Evaluation kaum so weiterentwickeln lassen, daß es die neuen Aufgaben der pädagogischen Evaluation erfüllen kann.

Das Akkreditations-Modell

Akkreditation ist die älteste Form von Evaluation. Organisationen wie die North Central Association of Colleges for Teacher Education und das National Council for the Accreditation of Teachers of Education bemüh-

hen sich, offensichtliche Unzulänglichkeiten in der Bildung von Schülern und Studenten zu identifizieren. Ausbildungsprogramme, bei denen Mängel gefunden werden, werden nicht zugelassen. Die Nichtanerkennung von Examina der als unzulänglich angesehenen Sekundar- oder Hochschulen führen im allgemeinen zu einer freiwilligen und raschen Verbesserung der Bedingungen, so daß sie den Normen entsprechen.

Die North Central Association (NCA) hat eine Entwicklungsgeschichte, die für Akkreditionsinstitutionen typisch ist[2]. Sie wurde 1895 von den Präsidenten der North Western University und den Universitäten von Michigan, Wisconsin, Chicago zusammen mit drei Sekundarschulleitern gegründet. Aufgabe der Gesellschaft war es, engere Beziehungen zwischen Hochschulen und Sekundarschulen zu schaffen. Deshalb kamen die Mitglieder der Gesellschaft aus der Verwaltung der öffentlichen und privaten Sekundarschulen und Hochschulen. Die NCA wurde während der neunziger Jahre des 19. Jahrhunderts zu einem Zentrum des Gedankenaustausches; damals stieg die Zahl ihrer Mitglieder auf 97 Institutionen (58 Sekundarschulen, 36 Hochschulen, 3 weitere Schulen) und 32 private Mitglieder. Zwischen 1901 und 1910 entwickelte die NCA die sie fortan kennzeichnende charakteristische Akkreditationspolitik. Vorher ließen kleinere Hochschulen und Universitäten in zunehmendem Maße Bewerber mit sehr ungleichen Sekundarschulvoraussetzungen aus sehr unterschiedlichen geographischen Regionen zum Studium zu. Auf der Jahrestagung der NCA von 1901 sprach Dekan Forbes von der Universität von Illinois über die Notwendigkeit der Zusammenarbeit der im Norden der zentralen Gebiete der USA gelegenen Hochschulen und Universitäten, um einheitliche oder mindestens gleichwertige Aufnahmeanforderungen zu erreichen. Daraufhin richtete die Gesellschaft drei Kommissionen zur Akkreditation von Schulen ein, das Committee on Unit Courses of Study, das Committee on *High School Inspection* und das *Committee on College Credit for High School Work*.

Das Committee on Unit Courses of Study und das Commitee on College Credit for High School Work lieferten auf der Jahrestagung von 1902 keine konstruktiven Arbeitsberichte und lösten sich langsam auf. So verpaßte die Gesellschaft die Gelegenheit, die Akkreditation auf die Schülerleistung zu gründen. Vielleicht war der Zeitpunkt ungünstig. Die Entwicklung des pädagogischen Testens sollte erst einige Jahre später in vollem Ausmaß erfolgen. Bis dahin gab es keine Technologie des Testens, auf die man sich beziehen konnte[3]. Diese Entwicklung veranschaulicht ein anderes Wachstumsgesetz: Wenn die nötigen Rohstoffe in der Umwelt nicht vorhanden sind, kann sich der Phänotyp trotz guter Entwicklungsmöglichkeiten des Genotyp nicht voll entwickeln.

Das Committee on High School Inspection erwies sich als das einflußreichste. Im Unterschied zu den beiden anderen Kommissionen konnte es sich auf die Erfahrungen seiner Vorgänger stützen. Bereits während der neunziger Jahre des 19. Jahrhunderts gab es in vielen Staaten eine staatliche Aufsicht über die Sekundarschule. Das High School Inspection Committee schlug vor, Sekundarschulen die Mitgliedschaft innerhalb der North Central Association zu gewähren, wenn sie folgende vier Bedingungen erfüllten:

(1) Alle Lehrer sollten ein Abschlußexamen eine NCA-Hochschule haben,
(2) die Lehrer sollten nicht mehr als vier Stunden täglich unterrichten,
(3) die Ausstattung der Arbeitsräume und der Bibliothek der Schule sollte angemessen sein,
(4) das »allgemeine intellektuelle und moralische Niveau« der Schule sollte sich im Verlauf einer sorgfältigen, verständnisvollen Inspektion als angemessen herausstellen.

Im Lauf der Jahre wurden die Richtlinien des Committee on High School Inspection in die Akkreditationskriterien aufgenommen. Bei den 1945 gebräuchlichen Kriterien für Sekundarschulen wurden folgende Schwerpunkte gesetzt:

(1) »Allgemeines intellektuelles und moralisches Niveau« der Schule
(2) Schulanlage
(3) Unterrichtsausstattung
(4) Bibliothek
(5) Finanzen und Personal
(6) Politik des Boards of Education
(7) Organisation und Verwaltung der Schule
(8) Lehrerqualifikation (Examina, Unterrichtsfächer)
(9) Pflichtstundenzahl der Lehrer
(10) Erfüllung der Bedürfnisse und Interessen der Schüler durch das Curriculum
(11) Schulpsychologische Beratung
(12) die Schule als Bildungs- und Freizeitzentrum für die ganze Gemeinde.

In den Akkreditations-Kriterien kommt das Anliegen der Schulverwaltung zum Ausdruck. Daher werden nicht nur die Auswirkungen der Erziehung auf die Schüler, sondern auch die Prozesse und Mittel der Erziehung berücksichtigt. Die prozeßorientierte Evaluation der frühen Jahre der NCA erfolgte in dem Glauben, daß die Änderung von Wahlfächern, Curriculumeinheiten, Anforderungen an die Lehrerausbildung und die Schulanlage bedeutsame Auswirkungen auf die Qualität des Lernens haben würden. Bei der Entwicklung dieser Kriterien während der ersten Hälfte

dieses Jahrhunderts zog die North Central Association keine Verhaltenswissenschaftler, Psychometriker und Statistiker zu Rate, die doch eine bedeutende Rolle bei der Entwicklung anderer Evaluationsmodelle spielten. Für eine produktive Zusammenarbeit zwischen der NCA und Wissenschaftlern aus den genannten Bereichen ergaben sich zwar des öfteren Möglichkeiten, die jedoch nicht aufgegriffen wurden.

Schon 1898 befaßte sich die NCA mit dem Englischunterricht. Das ging auf ein Interesse der stärker wissenschaftlich orientierten Mitglieder der Gesellschaft zurück. Auf die Frage, wie einheitliche Anforderungen in Englisch aufgestellt werden könnten, reagierten sie mit einer über zwanzigjährigen Auseinandersetzung und einer Reihe von umfangreichen Berichten. Mit Ausnahme der Akkreditation von Sekundarschulen – einem Ergebnis der Arbeit des Committee on High School Inspection – formulierte und diskutierte die North Central Association lediglich zahlreiche Probleme, ohne sie jedoch zu lösen.

Seit der Gründung der NCA wurden Unterrichtsergebnisse unter Bezugnahme auf die damals verbreitete Vermögenspsychologie (faculty psychology) verstanden. Auf der Jahrestagung von 1897 wurde beschlossen, »die Aufgaben, die am besten zur Entwicklung der Fähigkeiten eines Schülers geeignet sind, im Rahmen der verschiedenen Curricula vorrangig zu behandeln ...« Die Vermögenspsychologie wurde in den ersten Jahren des 20. Jahrhunderts von Thorndikes Assoziationstheorie und Watsons Behaviorismus abgelöst. Vielleicht erkannten die Verhaltenswissenschaftler und die Mitglieder der NCA, deren Aufgabe die Akkreditation war, daß sie in ihrem Verständnis der Schüler und der Lernprozesse so weit voneinander entfernt waren, daß eine Zusammenarbeit unmöglich war.

Zu Beginn der frühen zwanziger Jahre versuchte das Committee of Unit Courses of Study, Normen für die Evaluation der Unterrichtsergebnisse zu entwickeln. Das geschah abermals weitgehend unabhängig von den damals sich allmählich entwickelnden Bereichen der pädagogischen Psychologie und des pädagogischen Testens. Die Arbeit dieser Kommission endete mit der Formulierung einer Reihe allgemeiner Unterrichtsziele:

(1) Vermittlung wertvollen Wissens
(2) Entwicklung von Einstellungen, Interessen, Motiven, Idealen
(3) Entwicklung des Gedächtnisses, des Urteilsvermögens und der Phantasie
(4) Vermittlung wertvoller Persönlichkeitszüge und nützlicher Fertigkeiten.

Als der Exekutivausschuß 1940 empfahl, man solle sich bei der Akkreditation mehr auf die Qualität des Unterrichts konzentrieren, ließen die Bemühungen dieser Kommission allmählich nach.

Wenn man festzustellen versucht, warum die NCA bei der Evaluation nicht die Schülerleistung als Ergebnis des Unterrichts berücksichtigte, darf man den Einfluß der Persönlichkeitsmerkmale und der Arbeitsgebiete der Gesellschaftsmitglieder nicht unterschätzen. Sie scheinen sich für fähig gehalten zu haben, eher die Prozesse als die Ergebnisse der Erziehung zu evaluieren.

Die Methoden der Akkreditation sind immer noch wenig von den Methoden der Verhaltens- und Sozialwissenschaften beeinflußt. Normen für die Beurteilung von Schulen gewinnt man in der Regel durch Expertenbefragung. Der Wert eines Curriculum bzw. Schulprogramms wird im allgemeinen nach entsprechenden Schulbesuchen von Experten beurteilt. Zu einem solchen Urteil kommt man also gewöhnlich nicht durch die objektive Untersuchung der Schüler- und Lehrerleistung, durch Repräsentativbefragung über Einstellungen und Meinungen, durch Datenanalyse usw. Unter den Evaluationsmodellen zeichnet sich das Akkreditationsmodell durch die Berücksichtigung von Expertenurteilen sowie umfassende Beschreibung und Beurteilung der Schulverwaltung, Organisation und Finanzierung aus. Doch stagniert das Akkreditationsmodell seit einigen Jahren in seiner Entwicklung. Wie das Tylersche Modell hat es mit seiner vollen Entwicklung auch seine Grenzen erreicht. Das Akkreditationsmodell hat mit seiner Institutionalisierung das letzte Stadium einer Disziplin erreicht. Wenn eine Disziplin ihre Identität durch die Institutionalisierung mit Hilfe einer administrativen Hierarchie, von Fachkongressen und zahlreichen eigenen Publikationen wie dem *North Central Association Quarterly* erreicht, dann ist die Wahrscheinlichkeit künftiger revolutionärer Veränderungen gering. So kann die Institutionalisierung der Akkreditation in der North Central Association, der American Association of Colleges for Teacher Education, dem National Council for the Accreditation of Teachers Education (NCATE) als die volle Entwicklung des Akkreditationsmodells angesehen werden. Die Frage ist jedoch, ob die gegenwärtigen Erfordernisse pädagogischer Evaluation von diesem Modell erfüllt werden.

Evaluatoren, die sich mit der entsprechenden Methodenforschung befassen, können viel von den im Zusammenhang mit der Akkreditation gewonnenen Erfahrungen lernen. Beachtenswert ist die Komplexität der Akkreditation und die Berücksichtigung der nicht verhaltensbezogenen und schülerbezogenen Aspekte der Schule. Wertvoll sind ferner die für die Beobachter und den Lehrkörper ausgearbeiteten Evaluationsbogen. Hoffentlich wird man diese Verfahren in der Evaluation weiterhin verwenden. Obwohl das Akkreditationsmodell »den Vorteil schneller Ergebnisse und der Ausnutzung der Kompetenz des Evaluators bietet, läßt es offensichtlich viel hinsichtlich Objektivität und Validität zu wünschen übrig.« (Guba/Stuffle-

beam 1968, 11). Wenn das Akkreditationsmodell grundsätzliche Mängel hat – meiner Meinung nach hat es sie –, dann liegen sie darin, daß man die für die Beurteilung zugrunde gelegten Normen nicht empirisch zu rechtfertigen versucht und daß die Evaluation der Erziehungsprozesse nicht durch die Berücksichtigung ihrer Konsequenzen für die Lernenden ergänzt wird. Minimalforderungen an eine Schule werden durch Expertenurteile gewonnen, die selten durch empirische Forschungsergebnisse abgesichert werden können. Schulen erhalten manchmal nicht die Akkreditation, weil sie im Verhältnis zur Schülerzahl zu wenig Schulpsychologen beschäftigen oder weil ihre Lehrer bestimmte Qualifikationsnachweise nicht erbringen können; dabei geht aus keinem gültigen Forschungsergebnis hervor, daß ein ungünstiges Zahlenverhältnis zwischen Schulpsychologen und Schülern u. a. eine schlechtere Erziehung bewirkt. Die Auseinandersetzungen zwischen der Universität von Wisconsin und dem National Council for the Accreditation of Teachers of Education in den frühen sechziger Jahren ist ein Beispiel dafür, wie eine Akkreditationsinstitution versuchte, ungültige und ungerechtfertigte Normen auf ein gutes Lehrerausbildungsprogramm anzuwenden.

Die Formulierung der Normen für die schulischen Medienprogramme durch die American Library Association und die National Education Association (1969) ist für den Prozeß der Aufstellung von Evaluationsnormen charakteristisch. Sie wurden von einer aus 28 Personen bestehenden Kommission aus den beiden Gesellschaften in Zusammenarbeit mit Vertretern von fast 30 professionellen pädagogischen Gesellschaften entwickelt. Bezeichnenderweise hatte keine dieser Organisationen Erfahrungen mit empirisch-pädagogischer Forschung. Um die Normen für schulische Medien zu gewinnen, verwendete man daher folgende Verfahren:

Nach einer Tagung des Beratungsausschusses und nach den ersten zwei Tagungen der gemeinsamen Kommission wurden die vorläufigen Empfehlungen für die quantitativen Normen für Medienzentren in einzelnen Schulen und für das gemeinsame Programm in besonderen Sitzungen während der im Jahre 1967 stattfindenden Kongresse des Department of Audiovisual Instruction, der American Association of School Librarians und der National Education Association zur Diskussion vorgelegt. Man bat um Stellungnahmen und erhielt entsprechende Reaktionen. Diese Normen wurden außerdem auf zahlreichen anderen Konferenzen und Tagungen diskutiert. Mehrere tausend Teilnehmer hatten Gelegenheit, ihre Ansichten über die Normen darzulegen. Viele taten das und machten Verbesserungsvorschläge. Diese Meinungsäußerungen wurden aufgearbeitet und von den Mitgliedern der gemeinsamen Kommission bei der Zusammenstellung der Normen sorgfältig berücksichtigt.

Der verbesserte Entwurf der Normen wurde dann über zweihundert in Fragen der Schulbibliothek und der audiovisuellen Medien kompetenten Personen und

den leitenden Mitgliedern der Organisationen, die das Projekt finanziell förderten, den Präsidenten der Gesellschaften in den Einzelstaaten und anderen vorgelegt. Weitere Stellungnahmen aus der Praxis wurden von den Mitgliedern der gemeinsamen Kommission beim Fortgang ihrer Arbeit berücksichtigt. Dann trafen sich die Mitglieder des Beratungsausschusses, um den von der gemeinsamen Kommission genehmigten Entwurf durchzusehen; nach Berücksichtigung ihrer Empfehlungen wurden die Normen den leitenden Gremien der American Association of School Librarians und des Department of Audiovisual Instruction vorgelegt. (American Library Assoc. 1969, VIII, XV).

Zufrieden berichtete die gemeinsame Kommission, daß sehr viele Personen zu Rate gezogen worden waren und die Möglichkeit hatten, die Formulierung der Normen zu beeinflussen. Die Kommission versuchte ihre Arbeit zu rechtfertigen und ihre Kriterien durch den Konsens von Experten abzusichern, wobei sie noch durch die Stellungnahme mehrerer tausend Pädagogen unterstützt wurde.

Es ist jedoch zweifelhaft, ob die Befragung von Pädagogen mit dem Ziel, Meinungen über anerkannte Normen für Medienprogramme zu erhalten, wirklich die empirische Validierung der Normen ersetzen kann. Die Vergrößerung der die Normen aufstellenden Gruppe vermehrt lediglich die Möglichkeit zur Selbsttäuschung und zur bloßen Berücksichtigung der Eigeninteressen, es sei denn, die vorgeschlagenen Normen werden kompromißlosen Versuchen unterworfen, ihre Validität mit empirischen Daten zu beweisen.

Wie würden die Normen für Medienprogramme abschneiden, wenn sie einem objektiven empirischen Test ausgesetzt würden? Zweifellos nicht allzu gut. Denn unter den Normen für Medienprogramme finden sich unter anderem die folgenden:

(1) mindestens 20 Bibliotheksbücher pro Schüler,
(2) 3–6 Zeitungen in Elementarschulen, 6–10 Zeitungen in den Sekundarschulen,
(3) 6 Band- oder Schallplattenaufnahmen pro Schüler,
(4) Lese- und Aufenthaltsräume für jeweils höchstens 100 Schüler,
(5) 20–40 qm Raum für die Aufbewahrung von Zeitschriften.

Ohne Widerspruch fürchten zu müssen, kann man annehmen, daß bei einer Befragung, die die abhängigen Variablen wie »Wohlstand der Gemeinde« und »Fähigkeit der Schüler« statistisch kontrolliert, sich keine höhere Schülerleistung auf einer Skala für die Schulen zeigen würde, die im Unterschied zu anderen Schulen systematisch Zeitschriften sammeln. Eine solche Befragung würde wahrscheinlich ergeben, daß einige Schulen durch die Aufbewahrung von Zeitschriften Raum und Geld verschwenden.

Die Autoren der Normen für Medienprogramme wollten die Schulen

auch davon überzeugen, einen Medienfachmann für 250 und einen Medienassistenten für 2000 Schüler zu beschäftigen. Allerdings fehlt die Möglichkeit, diese Normen durchzusetzen. Eines der besten innovativen Medienprogramme wurde 1969 vom Ontario Institute for Studies in Education entwickelt. Viele Schulen können durch Telefon und Fernsehkabel an ein zentrales Medienzentrum angeschlossen werden. Innerhalb weniger Minuten nach der telefonischen Anfrage eines Lehrers kann das Zentrum einen Film oder eine Fernsehaufzeichnung aus seiner Sammlung in eine bestimmte Klasse übertragen. Ein solches Programm erfüllt die meisten Normen für Medienprogramme nicht.

Dennoch wird man anerkennen, daß im allgemeinen die der Akkreditation zugrunde gelegten Normen nicht ohne Wert sind. Sie sind beispielhaft in ihrer Komplexität und Detailliertheit. Es besteht jedoch die Gefahr, daß Normen unreflektiert durchgesetzt werden. Dies geschieht leicht dann, wenn nicht mit erprobten Methoden bewiesen werden kann, daß sie wertvolle pädagogische Ergebnisse bewirken.

Evaluation wird den *Wert* eines Programms nicht erhöhen, wenn sie die Berücksichtigung von Normen verlangt, von denen nicht bewiesen werden kann, daß sie zu wertvollen Zielen führen. Die Verfahren der pädagogischen Akkreditation werden gegenwärtig von erziehungswissenschaftlichen Forschern angegriffen, die empirisch nachweisen können, welche Normen gültig sind. Es besteht wenig Hoffnung auf eine produktive Zusammenarbeit zwischen diesen beiden Gruppen. Der von Anfang an im Akkreditationsmodell bestehende Fehler läßt sich wahrscheinlich nicht korrigieren; so wird sich aus ihm eine wirklich brauchbare und notwendige Methodologie der Evaluation nicht entwickeln lassen.

Das Management-System-Evaluationsmodell

Mehrere neuere Versuche, die Ansätze pädagogischer Evaluation zu systematisieren, haben zu einer Gruppe mit ähnlichen methodischen Verfahren geführt. Die Modelle von Alkin (1967; 1969), Guba und Stufflebeam (1968) und Stufflebeam (1969) sind für diese Gruppe charakteristisch und sollen hier diskutiert werden.

Guba und Stufflebeam (1968, 24) definieren Evaluation wie folgt:
Definition: *Pädagogische Evaluation ist (1) der Prozeß, durch den man (2) nützliche (3) Informationen (4) erhält und (5) für das Fällen von Entscheidungen (6) zur Verfügung stellt.*
Begriffsbestimmung:

(1) *Prozeß*: Eine bestimmte und fortlaufende Handlung, die viele Methoden und eine Reihe von Schritten oder Operationen umfaßt;

(2) *nützlich*: Angemessen in bezug auf vorherbestimmte Kriterien, die von Evaluator und Adressat gemeinsam entwickelt wurden;

(3) *Informationen*: Deskriptive oder interpretative Daten über (greifbare oder nicht greifbare) Einheiten und ihre Beziehungen;

(4) *erhält*: Bereitstellen von Daten durch Prozesse wie Sammeln, Ordnen, Analysieren und Berichten und durch formale Verfahren wie Messungen und statistische Methoden;

(5) *für das Fällen von Entscheidungen*: Wahl zwischen Handlungsalternativen als Antwort auf pädagogische Bedürfnisse oder pädagogische Probleme;

(6) *zur Verfügung stellt*: Das Ordnen in Systeme oder Sub-Systeme, die den Bedürfnissen oder Zielen der Evaluation am besten entsprechen.

Guba und Stufflebeam behaupten, Evaluation solle als die Informationssammlung für Entscheidungsträger angesehen werden. Nach ihrer Auffassung soll Evaluation den mit der Durchführung des Programms beauftragten Entscheidungsträgern behilflich sein, indem sie ihnen Daten zur Verfügung stellt. In ihren Veröffentlichungen über Evaluation konzentrieren sich diese Autoren auf die Vorbereitung von Entscheidungen, Entscheidungstypologien und die Wechselbeziehungen zwischen Entscheidungen in verschiedenen pädagogischen Kontexten.

Alkin (1969, 3–4) definiert Evaluation ähnlich:

Evaluation ist der Prozeß, in dem festgestellt wird, welche Entscheidungen getroffen, welche Informationen ausgewählt, gesammelt und analysiert werden müssen, um zusammenfassende Ergebnisse zu liefern, die den Entscheidungsträgern bei der Wahl zwischen Alternativen nützlich sind. ... Der Entscheidungsträger und nicht der Evaluator bestimmt, welche Fragen zu stellen sind bzw. welche Entscheidungen zu treffen sind. Seine Aufgabe ist es, vom Entscheidungsträger in Erfahrung zu bringen, für welche Entscheidungen Informationen nötig sind.

Alkin hebt hervor, daß Evaluatoren dem Entscheidungsträger lediglich Daten zur Verfügung stellen, nicht aber selbst Urteile abgeben sollen: »Die Information wird vom Evaluator zur Verfügung gestellt, aber der Entscheidungsträger muß den relativen Wert der Alternativen in einer Gesamtbeurteilung abschätzen.« (1969, 13). Obwohl Alkin seine Behauptung nicht zu rechtfertigen versuchte, hätte er es doch auch wenigstens wie die Autoren von »Disciplined Inquiry for Education« (1969, 26–27) tun können, die eine ähnliche Behauptung folgendermaßen begründeten:

Die Aufgabe jeder (entscheidungsorientierten) Untersuchung ist es, dem Entscheidungsträger Informationen an die Hand zu geben, nicht aber ihm zu sagen, was er zu tun hat. .... Die Entscheidung ist Aufgabe eines Beamten der Schulverwaltung und nicht eines Forschers; nur der Beamte der Schulverwaltung oder sein Beratungsgremium sind in der Lage, die politischen, ökonomischen und pädagogischen Aspekte der Entscheidung abzuwägen.

Die Logik dieser Empfehlung ist nicht einsichtig. Sie enthält z. B. die Annahme, daß die Evaluation eines Curriculum nicht die politischen und ökonomischen Aspekte der Entscheidungen berühren soll. Ohne Zweifel ist jedoch jede Evaluation, die diese Gesichtspunkte nicht berücksichtigt, unvollständig. Ferner ist es sehr fragwürdig, ob die subjektiven Eindrücke der Beamten der Schulverwaltung und ihrer Beratungsgremien neue relevante Informationen zu den objektiven Daten über politische, ökonomische und soziologische Fragen beitragen können, um die Ungewißheit im Hinblick auf die Folgen von Entscheidungen zu vermindern. Darüber hinaus ist der Standpunkt völlig unhaltbar, daß die Gewichtung, die die Entscheidungsträger den Informationsquellen beimessen, die private Angelegenheit der Beamten der Schulverwaltung und ihrer Beratungsgremien ist. Evaluationsdaten sind wertlos, gleichgültig, wie sorgfältig sie auch gesammelt wurden, wenn sie willkürlich oder unverständig zu Werturteilen zusammengezogen werden, die Einfluß auf Entscheidungen haben. Die Gewichtung von mehreren Skalen mit dem Ziel, den Gesamtwert von Alternativen zu bestimmen, muß transparent gemacht und vom Evaluator genau untersucht werden.

Die Versuche, Evaluationsmodelle zu entwickeln, die auf die Sammlung von Daten für den Entscheidungsprozeß abzielen, sind in mancher Hinsicht unzulänglich. Sie vernachlässigen zwei wesentliche Bestandteile der Scrivenschen Definition der Evaluation, nämlich daß die Evaluation darin besteht, ... »Verhaltensdaten mit einem gewichteten Satz von Skalen zu kombinieren, mit denen entweder vergleichende oder numerische Beurteilungen erlangt werden sollen, und in der Rechtfertigung
(a) der Datensammlungsinstrumente, (b) der Gewichtungen und (c) der Kriterienauswahl« (Scriven 1969, 40, *61*).

Evaluatoren, wie Guba und Stufflebeam, die sich mit entscheidungsorientierten Methoden der Evaluation befassen, behaupten, daß in ihrem Denken und in ihren Modellen Werte eine Rolle spielen, weil eine Entscheidung immer der Ausdruck eines Wertes ist: Wenn der Entscheidungsträger A gegenüber B vorzieht, so wertet er offensichtlich A höher als B. Deshalb liegen nach Meinung der Autoren den Entscheidungen Wertvorstellungen auf jeden Fall zugrunde.

Guba und Stufflebeam (1968, 28) behaupten, daß »das Verfahren, das hier als Evaluation beschrieben wird, der ursprünglichen Bedeutung des Begriffs *evaluieren* eher entspricht als das Verfahren, das gegenwärtig so bezeichnet wird. Wir würden dafür eintreten, daß, wenn man einen Begriff ändern wollte, es der Begriff für die gegenwärtige Praxis sein müßte. Werte sind besonders wichtig, wenn eine Auswahl getroffen werden muß. Dieses Auswählen ist der wesentliche Teil im Entscheidungsprozeß. Wir

schlagen daher vor, daß die Evaluation sich auf die Erarbeitung von Kriterien konzentrieren sollte, auf die man sich bei Entscheidungen beziehen kann. Durch das Formulieren solcher Kriterien erhalten wir eine Orientierungshilfe für die Art der Informationen, die gesammelt werden sollten, und darüber, wie sie analysiert und berichtet werden sollten. Der Begriff *Evaluation* scheint für das hier beschriebene Verfahren besonders geeignet zu sein, da dieses Verfahren einen ausgeprägten Gebrauch von Wertkonzepten macht«.

Für einen »wertorientierten Evaluator« sind jedoch im Verfahren der Messungen an Wertskalen, der Zusammenfassung von Meßwerten zu Wertaussagen und der Rechtfertigung der Messung und der Mittel, von den Meßwerten zu Wertaussagen zu kommen, Entscheidungen enthalten. Die Alternative, die auf einer gewichteten Kombination von Wertskalen den höchsten Punktwert erzielt, wäre die bessere Alternative. Ein entscheidungsorientiertes Evaluationsmodell kann jedoch angewandt werden, ohne die Aufmerksamkeit auf den Prozeß zu lenken, in dem ein Entscheidungsträger von Informationen zu einem Gesamturteil kommt.

Werte mit *Präferenzen* gleichzusetzen ist in den Wirtschaftswissenschaften seit langem üblich. Für den Wirtschaftswissenschaftler, mindestens in der Vergangenheit, drückt sich der Wert eines Produkts in den Präferenzen für dieses Produkt aus: Wenn der Verbraucher 5 Dollar für A bezahlt, dann ist der Wert von A 5 Dollar. Eine derartig vereinfachende Definition von Wert beurteilt eine gute und eine schlechte Evaluation gleich; ein 5-Dollar-Produkt ist so wertvoll wie jedes andere 5-Dollar-Produkt. Frauen bezahlen 5 Dollar für ca. 30 g Schönheitscreme (*Marktwert*), obwohl die Bestandteile der Creme, d. h. Material und Arbeit, nur 25 Cent kosten (der *tatsächliche Wert* des Produkts). Daß die Creme für 5 Dollar auf dem Markt gehandelt werden kann, ist Beweis für den irrationalen Glauben des Verbrauchers, daß teure Produkte auch gleichzeitig Produkte von hoher Qualität sein müssen. (Eine Kosmetikfirma setzte vor einiger Zeit den Preis einer teuren Schönheitscreme, die mit mehr als 1000 % Gewinn verkauft worden war, erheblich herab, mußte jedoch feststellen, daß der Absatz sehr zurückging!) Der Unterschied zwischen entscheidungsorientierten und wertorientierten Evaluationstheoretikern ist derselbe Unterschied, der in der Preisfestsetzung der Schönheitscreme besteht, deren Wert die einen mit 5 Dollar ansetzten, weil Frauen diesen Preis dafür bezahlen, und die anderen mit 25 Cent, weil die Gesamtinvestition eben nur soviel beträgt. Ein ähnlicher Mangel an Logik findet sich häufig in der pharmazeutischen Industrie: einige der renommierten Arzneimittel verkaufen sich weit besser als weniger bekannte identische Arzneimittel, obwohl erstere dreißigmal mehr kosten als letztere. Die Analogie zur pädagogischen Evaluation

ist leider nur zu deutlich. Beamte der Schulverwaltung haben sich oft für die Unterrichtsmethode A anstelle der Methode B entschieden, nur weil A teurer war, obwohl Evaluationsdaten eine andere Entscheidung nahelegten. Die für solche Verwaltungsbeamte typischen Überlegungen sind: Sicherlich wären all diese teuren Erfindungen nicht gemacht und die wertvollen Materialien nicht produziert worden, wenn sie nicht eine Verbesserung gegenüber alten Methoden darstellten; die neuen Methoden müssen einfach besser sein.

Man könnte die direkte Einschätzung von Werten gänzlich außer Acht lassen, wenn die Präferenzen der Entscheidungsträger immer ein logischer, rationaler, intelligenter Ausdruck ihrer Wertvorstellungen wären. In Wirklichkeit sind die meisten Entscheidungsträger durch den Entscheidungsprozeß überfordert; viele von ihnen fühlen sich wegen ihrer Unfähigkeit, ihre Entscheidungen zu rechtfertigen, unsicher. Deshalb empfiehlt es sich nicht, Evaluation als die Darbietung von Daten für Entscheidungsträger anzusehen, mit denen diese dann machen können, was sie wollen.

Evaluation kann in einem Curriculum viele *Rollen* übernehmen; sie kann den Herstellern durch die Ergebnisse in entsprechenden Leistungstests helfen; sie kann durch die Bereitstellung von Daten die schulische Durchführung des Curriculum erleichtern usw. Gleichwohl muß es immer das *Ziel* der Evaluation sein, eine Antwort auf die entscheidende Frage zu liefern: Ist das untersuchte Curriculum wertvoller als seine Konkurrenten, oder ist es an sich wertvoll genug, beibehalten zu werden?

Guba und Stufflebeam schließen sich der Auffassung früherer Kritiker an, die sich gegen die Verwendung vergleichender Versuchspläne (experimental design) für die Curriculumevaluation gewandt haben. Sie kommen zu dem Schluß, daß »die Anwendung von Versuchsplänen auf Probleme der Evaluation bei oberflächlicher Betrachtung sinnvoll zu sein scheint, da in der Vergangenheit experimentelle Forschung und Evaluation dazu dienten, Hypothesen über die Auswirkungen verschiedener Versuchsbedingungen (treatments) zu überprüfen. Bei diesen Überlegungen gibt es jedoch einige schwierige Probleme« (Guba/Stufflebeam 1968, 14).

Die meisten der angeblichen Probleme ergeben sich jedoch aus Gubas und Stufflebeams eigenwilliger Auffassung von vergleichenden Versuchen in den Sozialwissenschaften. Nach ihrer Meinung müssen z. B., damit Versuchsanordnungen mit Vergleichsgruppen gültige Resultate ergeben, »... die Bedingungen in den Versuchs- und Kontrollgruppen während des gesamten Versuchs konstant gehalten werden, d. h. sie müssen während des ganzen Versuchs den ursprünglich festgelegten Bedingungen entsprechen. Die Bedingungen in der Versuchsgruppe bzw. Kontrollgruppe dürfen während des Prozesses der Curriculumentwicklung nicht modifiziert werden,

da man sonst keine Aussagen darüber machen kann, was evaluiert wird.« (Guba/Stufflebeam 1968, 13). Offensichtlich beunruhigen sie Versuchsbedingungen, die so eng und streng definiert werden, daß sie den Entscheidungsträgern nicht die Möglichkeit geben, während des Versuchs modifizierend einzugreifen. Jedoch sind derart einschränkende Bedingungen für gültige Vergleichsuntersuchungen nicht erforderlich. Man kann ohne weiteres Bedingungen für pädagogische Untersuchungen so formulieren, daß Entscheidungsträger die Möglichkeit haben, das Bildungsprogramm den jeweiligen Erfordernissen anzupassen. Ein Forscher in der Medizin, der ein Arzneimittel mit Hilfe eines Placebo evaluiert, kann auch andere Medikamente einnehmen lassen, um Nebenwirkungen zu kontrollieren oder die Dosierung entsprechend seinen Beobachtungen über den Rückgang der Krankheit zu verändern. Eine solche Entscheidung stellt nicht die Gültigkeit des Vergleichs zwischen Medikament und Placebo in Frage, da sie ein notwendiger Teil des *Kontextes* ist, der evaluiert wird, nämlich die Behandlung der Krankheit X durch das Medikament A. Natürlich kann der Entscheidungsträger den Kontext einer Behandlung so ändern, daß die ursprünglich definierte Behandlung nicht länger evaluiert wird, so z. B., wenn der Forscher aufhört, das Medikament einzugeben. Dies bedeutet jedoch nicht, daß er nicht innerhalb des Kontextes eines gut geplanten Versuchs variierend eingreifen kann, ohne die Gültigkeit des Vergleichs zu beeinträchtigen.

Nach Auffassung von Guba und Stufflebeam erfordern Versuche mit Vergleichsgruppen, daß »... alle Schüler, die am Versuch teilnehmen, den gleichen Bedingungen ausgesetzt werden, für die sie ursprünglich vorgesehen wurden ...« (1968, 13). Versuchsanordnungen mit Vergleichsgruppen erfordern jedoch nichts dergleichen. Offensichtlich stellen sich die Autoren unter »Versuchsbedingung« eine sich nicht ändernde, in sich abgeschlossene Bedingung vor. Eine Versuchsbedingung in einem Versuch mit Vergleichsgruppen innerhalb der Sozialwissenschaften ist oft eine Abstraktion, ein Konstrukt mit definierenden Merkmalen, aus denen ein Kontext entsteht. Man kann nur den durch das Konstrukt gebildeten Kontext evaluieren. Der Kontext braucht sich nicht aus der Notwendigkeit zu ergeben, daß alle Versuchspersonen die gleiche *Menge* von etwas erhalten. Wirtschaftswissenschaftler führten in New Jersey gegen Ende der sechziger Jahre Versuche über die negative Einkommensteuer durch. Personen im negativen Einkommensteuerplan wurden mit Personen im herkömmlichen Steuerplan hinsichtlich solcher Variablen, wie Zahl der Arbeitslosen, Konsum- und Spargewohnheiten usw. verglichen. Für die negative Einkommensteuer ist kennzeichnend, daß sich ihr Betrag von Person zu Person unterscheidet; daraus wird jedoch keiner den Schluß ziehen, daß der Vergleich ungültig wäre. Tatsächlich

brauchen nicht alle Versuchspersonen derselben Bedingung ausgesetzt zu werden, wie das für die Evaluation von individuellem Unterricht erforderlich wäre.

Guba und Stufflebeam (1968, 14–15) behaupten, daß die Anwendung eines vergleichenden Versuchsplans auf Probleme der Evaluation »... mit dem Grundsatz in Konflikt gerät, daß Evaluation zur kontinuierlichen Verbesserung eines Curriculum dienen soll«, und daß sie zwar »... für Entscheidungen nach Beendigung eines Projekts nützlich, aber als Hilfsmittel für Entscheidungen während der Planung und Implementation eines Projekts fast nutzlos sei.« Die Brauchbarkeit eines vergleichenden Versuchsplans für Entscheidungen nach Abschluß eines Projekts wird von zwei weiteren Autoren hervorgehoben. Die von Guba und Stufflebeam aufgezeigten Schwierigkeiten wurden, nachdem Cronbach (1963) dieselben Probleme erörtert hatte, bereits durch Scrivens Unterscheidung zwischen formativer und summativer Evaluation geklärt.

Guba und Stufflebeam kritisieren den vergleichenden Versuchsplan, weil es fast unmöglich ist, Störvariablen (confounding variables) durch Zufallsstichproben oder mit anderen Verfahren zu kontrollieren oder zu eliminieren. Doch auch Cronbach hatte bereits auf das gleiche Problem aufmerksam gemacht: »Man gefährdet die Interpretation eines Versuchs, wenn man die Klassen nicht parallelisiert, die zu vergleichende Curricula benutzen. Leider sind solche Fehler fast unvermeidbar.« (1963, 42, 48). Man versucht nicht, Vergleichsgruppen zu parallelisieren; eine solche Parallelisierung von Gruppen ist schon frühzeitig in der Geschichte der Versuchsplanung als unmöglich erkannt worden. Im vergleichenden Versuchsplan werden Gruppen nach dem Zufallsprinzip gleichwertig gemacht, wodurch in Wirklichkeit jedoch noch keine Gleichwertigkeit geschaffen wird. Die nach dem Versuch sich herausstellenden Unterschiede werden dann daraufhin untersucht, ob sie so klein sind, daß sie der ursprünglichen Zuordnung nach dem Zufallsprinzip zugeschrieben werden können, oder ob sie so groß sind, daß die Versuchsbedingungen für den Unterschied verantwortlich zu machen sind. Gültige Versuche mit Vergleichsgruppen sind nicht möglich, weil Gruppen nicht vollständig parallelisiert werden können. Gültige, auf Wahrscheinlichkeitsaussagen beruhende Vergleiche sind jedoch möglich; das geht schon aus der zunehmenden Zahl gut geplanter Versuche mit Vergleichsgruppen in der Pädagogik hervor. Gewiß sind gültige Versuchspläne schwierig und nur unter erheblichem Kostenaufwand durchzuführen; aber die pädagogischen Forscher und Evaluatoren müssen davon überzeugt werden, daß solche Versuchspläne im allgemeinen die finanziellen Aufwendungen wert sind.

Schließlich legen Guba und Stufflebeam dar (1968, 16), daß »ein viertes

Problem bei der Anwendung herkömmlicher Versuchspläne darin liegt, *daß innere Validität durch die Kontrolle äußerer Variablen nur auf Kosten äußerer Validität erreicht werden kann.«* Diese Behauptung klingt so überzeugend, daß sie den mit den Methoden empirischer Forschung wenig vertrauten Leser überzeugt: Innere und äußere Validität sind *nicht* diametral entgegengesetzt. Das Planen von Versuchen, die in hohem Maße beide Arten von Validität aufweisen, schafft lediglich eine Reihe technischer Probleme für die Untersuchungsverfahren, die Datensammlung und die statistische Analyse (vgl. Bracht/Glass, 1968).

Das Tylersche und das Management-System-Modell betonen eher bestimmte Rollen der Evaluation, als daß sie sich bemühen, das Ziel der Evaluation zu erreichen. Herkömmliche Modelle der Curriculumevaluation haben sich vor allem darauf konzentriert, verschiedene Rollen bei der Entwicklung oder Durchführung eines Curriculum zu übernehmen. In einigen Fällen haben sich die Verfechter dieser Modelle sogar dagegen ausgesprochen, überhaupt den Versuch zu unternehmen, das Ziel der Evaluation zu erreichen. Das Ziel der Evaluatoren, die sich am Management-System-Modell orientieren, ist eher die Unterstützung der Beamten der Schulverwaltung als die Beurteilung von Wertfragen. Den Curriculumentwicklern bei der Durchführung des Curriculum behilflich zu sein, so daß sie ihre Aufgaben besser erfüllen können, *ist ein naheliegendes Ziel der Evaluation; das letzte Ziel der Evaluation besteht jedoch darin, Fragen nach dem Wert zu beantworten.* Ein Evaluator, der den Gesamtwert eines Curriculum beurteilt, stellt für die Lehrer und Beamten der Schulverwaltung eine Bedrohung dar, mit denen er in besserem Verhältnis stehen könnte, wenn er seine Aufgabe lediglich darin sähe, ihnen zu helfen. Trotzdem ist er verpflichtet, Urteile zu fällen und darf sich nicht dieser Verpflichtung entziehen.

Das Zielkomplex-Modell

Das Evaluationsmodell, das ich Zielkomplex-Modell (composite-goal model) nennen möchte, geht auf Scriven (1967) zurück.

Scriven (1967, 40, *61*) definiert Evaluation wie folgt:

> Evaluation an sich ist ein methodisches Vorgehen, das im Grunde genommen *gleich ist*, unabhängig davon, ob man Kaffeemaschinen, Lehrmaschinen, Pläne für ein Haus oder ein Curriculum zu evaluieren versucht. Es besteht einfach im Sammeln und Kombinieren von Verhaltensdaten mit einem gewichteten Satz von Skalen, mit denen entweder vergleichende oder numerische Beurteilungen erlangt werden sollen, und in der Rechtfertigung (a) der Datensammlungsinstrumente, (b) der Gewichtungen, (c) der Kriterienauswahl.

Scrivens Definition der Evaluation (in der die *komplexen* Wertkriterien hervorgehoben werden) liefert das beachtenswerte Evaluationsmodell, das wir als Zielkomplex-Modell bezeichnen. Meiner Meinung nach ist das Zielkomplex-Modell der Evaluation das einzige der hier diskutierten Modelle, das zu einer brauchbaren Methodologie der Evaluation führen kann.

Folgende Faktoren begründen den Wert des Zielkomplex-Modells: Das Ziel der direkten Werteinschätzung (worin es sich vom Management-System-Evaluationsmodell unterscheidet), das Anliegen, die ausgewählten Kriterien und Ziele zu rechtfertigen (worin es sich vom Akkreditationsmodell unterscheidet), und schließlich die Möglichkeit, in verschiedenen Kontexten anwendbar zu sein, die heute nach pädagogischer Evaluation verlangen (worin es sich vom Tylerschen Modell unterscheidet). Das Zielkomplex-Modell ist das einzige der hier diskutierten Modelle, nach dem wirklich Evaluation stattfinden kann. Der Prozeß, durch den man auf rationale Weise zu einer vertretbaren Einschätzung des Wertes eines Verfahrens oder eines Gegenstandes kommt, wird durch Scrivens dreiteilige Definition der Evaluation gut beschrieben. Das Akkreditationsmodell eignet sich nicht dazu, zu umfassenden und vertretbaren Werturteilen zu gelangen. Das Tylersche und das Management-System-Modell sind ohne Zweifel brauchbare Modelle. Sie sind jedoch keine Modelle für den Prozeß der Evaluation; sie sind vielmehr Modelle der Entwicklung bzw. der Implementation von Curricula. Zu einem großen Teil steht die Entwicklung des Zielkomplex-Modells noch bevor. Wenn das Modell seine volle Ausprägung und Brauchbarkeit erreichen soll, müssen für bestimmte in seiner Definition enthaltene Merkmale entsprechende technische Verfahren entwickelt werden.

Die Weiterentwicklung des Zielkomplex-Modells der Evaluation

Um zu einer Verbesserung des Zielkomplex-Modells zu gelangen, sollte man die Scrivensche Definition der Evaluation in den Mittelpunkt stellen. Die Evaluatoren haben bis heute nur wenige der Techniken entwickelt, die für die Anwendung des Zielkomplex-Modells erforderlich sind. Deshalb bedarf jedes Element der Scrivenschen Definition noch näherer Ausführung:

(a) Welche Daten sollen auf welchem Allgemeinheits- bzw. Spezifitätsgrad gesammelt werden?
(b) Wie soll man Daten gewichten und in Gruppen zusammenfassen, um zu Einschätzungen des Wertes des untersuchten Gegenstandes zu kommen?
(c) Wie können die Verfahren der Datensammlung, die Gewichtung und

Zusammenfassung der Daten in Gruppen und die Auswahl der Ziele gerechtfertigt werden?

Jede dieser Fragen erfordert bisher noch nicht bekannte Techniken der Evaluation. Im folgenden werde ich daher die angeschnittenen Fragen erläutern und einige Hinweise geben, wie die notwendigen Techniken gefunden werden können.

A. Sammlung von Daten

Zwei ungelöste Probleme bei der Sammlung der Evaluationsdaten bestehen in der Bestimmung der richtigen Ebene des Allgemeinheitsgrads, auf der die am meisten aussagekräftigen Daten liegen, und in der Festsetzung von Prioritäten für die Sammlung dieser Daten.

Allgemeinheitsgrad und Spezifitätsgrad von Daten

Ein Gegenstand, der so komplex wie ein Curriculum ist, kann auf zahlreichen Ebenen der Spezifität untersucht werden (Krathwohl 1965). Evaluatoren sollten darauf achten, eine große Sammlung zur Auswahl von Daten anzulegen. Sie sollten sich vergegenwärtigen, daß alles, was für das Curriculum vorausgesetzt wird, was während seiner Durchführung geschieht und aus ihm als Ergebnis resultiert, für den Erfolg des Curriculum sehr wichtig sein kann. Sie werden auch darauf hingewiesen, daß sie nicht nur den tatsächlichen Ablauf beobachten, sondern auch die dem Ablauf zugrunde liegenden Intentionen berücksichtigen müssen. Aber niemand hilft den Evaluatoren festzusetzen, welcher Allgemeinheits- bzw. Spezifitätsgrad sich für die Intentionen und Beobachtungen empfiehlt. Da aber Hinweise und Richtlinien dafür fehlen, kann es den Evaluatoren leicht mißlingen, die wesentlichen Merkmale des Curriculum aufzuzeigen. Tyler (1966) bezeichnete das Problem der Festsetzung des richtigen Spezifitätsgrads für die Formulierung von Lernzielen als die gegenwärtig schwierigste Aufgabe der Unterrichtsforscher. Er stellte fest, daß Verhaltensziele manchmal so spezifisch formuliert werden, daß selten bewußt gelehrt und daher auch nur schwer gelernt werden kann, spezifische Fakten zu generalisieren. Aus den Beobachtungen eines Bildungsprogramms kann man zu grundsätzlichen Aussagen gelangen, wenn die berücksichtigten Daten auf einem höheren Allgemeinheitsgrad liegen.

Die folgende Episode ist ein Beispiel dafür, wie der Beobachtung eine Methode zugrunde liegen muß, damit irrelevante Daten vermieden werden. Ein Bewohner des Mars wurde zur Erde geschickt, um ihre Bewohner zu beobachten. Nach seiner Rückkehr zum Mars schrieb er folgenden Be-

richt: »Den Planeten Erde bewohnen viele Milliarden geflügelter sechs- und achtbeiniger Kreaturen. Ihr kurzes Dasein ist frei von äußeren Gefahren, abgesehen davon, daß ab und zu große zweibeinige Kreaturen, von denen es auf dem ganzen Planeten nicht mehr als dreieinhalb Milliarden gibt, in ihre Lebenswelt eindringen.« Der Marsbewohner machte wirklich ein paar zutreffende Beobachtungen. Wir jedoch – in unserem Egozentrismus – denken, daß er das Charakteristische des Planeten Erde verfehlte, weil er die falschen Dinge beobachtete.

Auf welcher Ebene sollte der Evaluator nach den wichtigen Phänomenen in einem Curriculum suchen? Sollten »intendierte Prozesse« in Form eines genau Minute für Minute spezifizierten Stundenplans oder in einer groben wöchentlichen Aufzeichnung allgemeiner Themen und Aktivitäten angegeben werden? Sollte er das kognitive Ergebnis »Kenntnis der Tiergattungen« oder das Ergebnis »Beurteilung der Species, des Geschlechts und der Gattung des tasmanischen Teufels« messen? (Versuche, diesen Fragen auszuweichen durch den Hinweis, diese müßten vom Curriculumentwickler und nicht vom Evaluator beantwortet werden, widersprechen einer soliden, produktiven Konzeption der Evaluation).

Die Evaluatoren, die sich vor allem mit den Methoden der Evaluation befassen, müssen sich noch sehr darum bemühen, festzulegen, ob man generelle oder spezifische Phänomene beobachten sollte; ohne eine ausgearbeitete Methodologie werden zu viele Bemühungen in der Evaluation entweder zu irrelevanten Vereinfachungen oder wertlosen Verallgemeinerungen führen.

Prioritäten für Evaluationsdaten

Einige Evaluatoren sind der Ansicht, daß praktisch alle erreichbaren Daten gesammelt und analysiert werden sollten. In neueren Veröffentlichungen zur Methodologie der Evaluation überrascht und beeindruckt die Vielzahl und Vielfältigkeit der Variablen, die der Beobachtung für wert gehalten werden. Nach Stake (1967a) ergeben sich die Daten der Evaluation aus Beschreibungen und Beurteilungen von *Voraussetzungen*, *Prozessen* und *Ergebnissen* sowie aus den Kontingenzen zwischen ihnen. Stake sieht in einem außerordentlich breiten Spektrum von Erscheinungen die Elemente für die Datenmatrix der Evaluation.

Neuere Veröffentlichungen zur Evaluation haben zu einer erfreulichen Erweiterung der Konzeption und einer verstärkten Aufmerksamkeit gegenüber einer großen Anzahl von potentiell wertvollen Daten angeregt, die vorher übersehen worden waren oder für nebensächlich gehalten wurden. Im Grunde war die Erweiterung der Datenmatrix der Evaluation

teilweise eine Reaktion auf die enge und unreflektierte Bevorzugung bestimmter Daten durch einseitige Behavioristen. Diese Behavioristen lassen für die Evaluation des Unterrichts lediglich beobachtbare Daten gelten, die sich auf Verhaltensziele beziehen. Einige Evaluatoren zögern, Prioritäten für Evaluationsdaten zu setzen. Denn sie befürchten, jene kurzsichtigen und für die vergangenen Jahrzehnte charakteristischen Versuche, Probleme der Evaluation in Angriff zu nehmen, könnten sich bei einem neuen System von Prioritäten schnell wiederholen. Es besteht aber kein Anlaß, enge und unnötig begrenzte Evaluationsversuche zu befürchten, wenn es eher darum geht, eine *Methodologie* für die Aufstellung von Prioritäten für Daten zu entwickeln, als darum, ein neues System von Prioritäten zu schaffen.

Einer Entscheidung liegen zwei oder mehrere alternative Handlungsmöglichkeiten zugrunde. Die Entscheidung treffen bedeutet lediglich, eine dieser Alternativen zu wählen. Die Vergegenwärtigung der bevorstehenden Entscheidungen wird zum großen Teil bestimmen, welche Daten gesammelt und wie sie analysiert werden. Für jede Entscheidung bedarf es relevanter Daten. Setzt man unter den anstehenden Entscheidungen Prioritäten, bedeutet das zugleich auch, daß man Prioritäten für die zu sammelnden Daten aufstellen muß. Prioritäten können auch danach aufgestellt werden, inwieweit man empirische Daten für eine Entscheidung braucht. Ein System von Prioritäten für die Sammlung von Evaluationsdaten kann bestimmt werden durch die bevorstehenden zu fällenden Entscheidungen sowie durch die notwendige Berücksichtigung von unvorhergesehenen Entscheidungen, die mit Sicherheit im Verlaufe der Untersuchung zu treffen sein werden.

Eine vorläufig brauchbare Methodologie zur Festsetzung von Prioritäten bei der Sammlung von Evaluationsdaten kann folgende Aspekte beinhalten:

(1) Finanzieller Aufwand der Sammlung verschiedener Daten;

(2) Abschätzung der Wahrscheinlichkeit, mit der die einer Entscheidung zugrunde liegenden Alternativen durch Daten gestützt werden, falls diese gesammelt werden sollten;

(3) der finanzielle Aufwand der Implementation jeder Entscheidungsalternative.

Die drei Komponenten dieser sich noch im Anfangsstadium befindlichen Methodologie sollen im folgenden ausgeführt werden; ich habe verdeutlicht, wie jede für sich die Prioritäten bei der Datensammlung festlegen würde:

(1) Der finanzielle Aufwand für die Sammlung verschiedener Daten.

Nehmen wir an, daß alle Faktoren mit Ausnahme der unterschiedlichen

Aufwendung für die Sammlung der Evaluationsdaten gleich sind. Dann werden die Mittel für die Evaluation dadurch am besten ausgegeben, daß man möglichst viele Entscheidungen trifft. Denn nach unserer Annahme sind die verschiedenen Entscheidungen gleich kostspielig, gleich wertvoll, und nach unseren vorläufigen Erwartungen unterstützen die für jede Entscheidung gesammelten Daten mit gleicher Wahrscheinlichkeit jede Alternative der Entscheidung.

(2) Die der Entscheidung vorausgehende Annahme, daß jede der Entscheidung zugrunde liegende Alternative durch die gesammelten Daten gestützt wird.

Angenommen, alle Faktoren außer den folgenden sind gleich: Für Entscheidung 1 gibt es zwei Alternativen: A und B. Die Wahrscheinlichkeit – vielleicht aufgrund einer persönlichen Schätzung des Evaluators –, daß die Daten, falls sie gesammelt werden, A stützen, beträgt für (A) = .90, für (B) = .10.

Für Entscheidung 2 gibt es zwei Alternativen: C und D.

Die Wahrscheinlichkeit, daß die relevanten Daten C stützen, wird auf (C) = .50 geschätzt: Also beträgt die Wahrscheinlichkeit für D ebenfalls (D) = .50. Daher kann man mit ziemlicher Sicherheit annehmen, daß die Ergebnisse der Datensammlung für Entscheidung 1, aber nicht für Entscheidung 2 sprechen. Offensichtlich ist daher die Priorität für die Sammlung der Daten für Entscheidung 2 höher als die Priorität der Datensammlung für Entscheidung 1. Wenn unsere Schätzungen der Wahrscheinlichkeit einen hohen Gültigkeitsgrad haben, kann Entscheidung 1 ohne die Sammlung empirischer Daten getroffen werden.

(3) Der finanzielle Aufwand der Implementation der Alternativen einer Entscheidung.

Jeder Entscheidung liegen zwei oder mehr Alternativen zugrunde, für deren Implementation der finanzielle Aufwand abgeschätzt werden kann. Die Alternativen A und B der Entscheidung können bei ihrer Verwirklichung 10 000 Dollar bzw. 11 000 Dollar kosten. Die finanzielle Aufwendung für die Verwirklichung der Alternativen C und D der Entscheidung 2 können 1000 Dollar bzw. 5000 Dollar betragen. Gesetzt den Fall, daß nur eine einzige Entscheidung auf Grund von Daten getroffen werden kann, die andere aber durch das Werfen einer Münze entschieden werden muß: Welche der beiden Entscheidungen soll dann aufgrund empirischer Daten getroffen werden? Die Antwort hängt nicht nur von den Kosten der Alternativen ab, sondern auch vom Gewinn, den die Implementation jeder der beiden Alternativen, und vom Verlust, den die Implementation der schlechteren der beiden Alternativen mit sich bringt.

Trotz des offenbar vielversprechenden Ansatzes solcher rudimentären

Strategien der Entscheidung und trotz der Leichtigkeit, mit der sie formuliert werden können, setzen aber wahrscheinlich alle ein zu großes *apriorisches* Wissen voraus, um eine unmittelbare Anwendung in der pädagogischen Evaluation finden zu können. Schon die Annahme, daß alle Alternativen einer Entscheidung schon vor der Datensammlung bekannt sind, ist bereits dem heutigen Stand der pädagogischen Technologie nicht mehr angemessen. Dennoch können couragierte Forscher mit unzulänglichen Methoden eher zu Ergebnissen kommen als risikoscheue Forscher, die auf erprobte Techniken warten. Boulding (1969, 7–8) tritt dafür ein, die ersten relativ gut entwickelten Verfahren der Kosten-Nutzen-Analyse zu verwenden:

> Der ganze Bereich der Kosten-Nutzen-Analyse, z. B. im Hinblick auf monitäre Einheiten, also »reale« Dollar bei konstanter Kaufkraft, ist von äußerster Bedeutung für die Evaluation gesellschaftlicher Entscheidungen und selbst gesellschaftlicher Institutionen. Wir können ohne weiteres zugestehen, daß der »reale« Dollar, der sonderbarerweise bloß in der Einbildung existiert, ein gefährlich unvollkommenes Maß für die Qualität des menschlichen Lebens und der menschlichen Werte ist. Trotzdem stellt er eine brauchbare erste Annäherung dar, und im Hinblick auf die Evaluation von schwierigen Entscheidungen ist es äußerst nützlich, erste Annäherungswerte zu besitzen, die sich modifizieren lassen. Ohne diese wird alle Evaluation zu einer zufälligen Auswahl, basierend auf bloßen Vermutungen.

Trotz des weitverbreiteten Interesses an der Kosten-Nutzen-Analyse und dem Planning Programming and Budgeting System haben solche Methoden das Bildungswesen bisher nur auf makroökonomischer Ebene beeinflußt. Evaluatoren haben sich bisher wenig mit der Abschätzung von Kosten und dem Verhältnis zwischen Kosten und Nutzen befaßt. Das Problem der Aufstellung von Prioritäten bei der Sammlung von Evaluationsdaten könnte zu einer größeren Berücksichtigung der Kosten- und Ressourcen-Allokation führen.

B. Die Gewichtung der Daten

Fast jede summative Evaluation ist vergleichend. Normalerweise beinhaltet summative Evaluation die Messung konkurrierender Curricula in bezug auf Leistung oder Ziele und die Zusammenfassung der Daten zu einem Urteil über die Überlegenheit eines Curriculum. Die Evaluatoren haben der Verarbeitung von Informationen zu einem summativen Urteil bisher kaum Bedeutung zugemessen. Scriven machte darauf aufmerksam, daß der Prozeß der Kombination von Verhaltensdaten ein Prozeß der *Summierung gewichteter Ziel- oder Leistungsskalen* ist; jenes Programm, das den höchsten Gesamt-

punktwert erreicht, wird wahrscheinlich bevorzugt. Die Gewichtung für die Einschätzung leitet sich vom menschlichen Urteil und den statistischen Eigenschaften der Skalen ab. Die Evaluatoren können auf eine hochentwickelte psychometrische Theorie des Messens von Urteilen und der Zusammenfassung von Informationen zu gewichteten Gesamtwerten zurückgreifen. In dem Modell, das mit der Summierung von gewichteten Ziel- oder Leistungsskalen arbeitet, wird eine durchschnittliche Leistung, die die Leistung auf verschiedenen Skalen berücksichtigt, erarbeitet. Wenn Programm A auf Skala 1 schlechter ist als B, kann man es dennoch B vorziehen, da Programm A in bezug auf Skala 2 bessere Leistungen erbringt und somit seine Unterlegenheit auf Skala 1 ausgleicht.

Das Modell, das mit der Summierung von gewichteten Ziel- oder Leistungsskalen arbeitet, ist dennoch nur eins von mehreren denkbaren Modellen zur Integration von Daten in summative Schlußfolgerungen. Es gibt auch nicht-kompensatorische Modelle, in denen geringe Punktwerte auf einer Skala nicht durch hohe Punktwerte auf anderen Skalen ausgeglichen werden können. Mit solchen nicht-kompensatorischen Modellen ist die Integration von Daten in eine summative Entscheidung lediglich eine Frage der Wahl des Programms, das durch die größere Zahl von ungewichteten Skalen überlegen ist; dabei wird der Grad der Überlegenheit jedoch nicht berücksichtigt. Viele Entscheidungsträger benutzen ein auf dem *Mini-Max-Prinzip* basierendes Entscheidungsmodell. Das Mini-Max-Prinzip geht davon aus, daß es sich empfiehlt, auf jeden Fall Fehlschläge zu vermeiden, auch wenn die Möglichkeit zu größeren Erfolgen besteht. Anstatt seine Erfolge zu maximieren, will der nach dem Mini-Max-Prinzip handelnde Entscheidungsträger vor allem die Möglichkeiten eines maximalen Mißerfolgs minimieren. Obwohl Curriculum A auf fast allen Skalen Curriculum B weit überlegen ist, kann der Entscheidungsträger, der nach dem Mini-Max-Prinzip handelt, sich für B entscheiden, weil die Unzufriedenheit der Lehrer mit dem Arbeitsaufwand für die Vorbereitung für A die Gefahr eines Widerstands heraufbeschwört, den er auf alle Fälle vermieden wissen will.

Die Wissenschaften vom Management hatten in letzter Zeit Bayessche Entscheidungsmodelle in der Wirtschaft angewandt. Diese Modelle verbinden Informationen und menschliches Urteil zu Entscheidungsstrategien (vgl. Schlaifer 1959). Evaluatoren können durch die Berücksichtigung der Modelle der Integration von Informationen und Urteilen und ihre Zusammenfassung in summative Entscheidungen erheblich zur Weiterentwicklung ihrer Disziplin beitragen.

Wenn die Methoden der Kombination von Informationen zu summativen Wertaussagen nicht angewandt werden, wird dieser Prozeß von Vorurtei-

len, vorschnellen Schlüssen und Irrationalität beherrscht sein. Wenn man dies einsieht, könnte das der erste Schritt auf dem Wege zur Verbesserung dieses wichtigen Verfahrens sein.

C. Die Rechtfertigung der Instrumente zur Datensammlung, Gewichtung der Einzelwerte und ihrer Zusammenfassung zu einem Gesamtwert und Auswahl der Ziele

(1) Rechtfertigung der Instrumente zur Datensammlung

Jahrzehntelanges Forschen mit quantitativen Methoden auf den Gebieten der Pädagogik, Soziologie und Psychologie haben zu gut ausgearbeiteten Theorien des Messens und vielen brauchbaren Instrumenten der Datensammlung geführt. Psychometrische Theorien der Reliabilität der Kriterien- und der Konstruktvalidität haben viel für die Praxis der Evaluation geleistet. Jedoch gibt es noch ungelöste Probleme im Zusammenhang mit der Verwendung und Rechtfertigung menschlicher Urteile als Daten der Evaluation. Scriven (1967) und Stake (1967a) treten für die Berücksichtigung von Urteilen bei der Evaluation ein. In zunehmendem Maße erkennen die Evaluatoren, daß – im Gegensatz zu der wissenschaftlichen Forderung nach Objektivität – Menschen Informationen äußerst effizient und effektiv verarbeiten können. In diesem Jahrzehnt hat die Evaluation durch die Berücksichtigung der Möglichkeit, Informationen zu sammeln, zu speichern, zu integrieren und Urteile abzugeben, gewonnen.

Leider haben die Evaluatoren sich darauf beschränkt, zu behaupten, daß Urteile wertvolle Daten sind, die mit Hilfe der Psychometrie ausgewertet werden können. Die Psychometrie jedoch trägt zum Prozeß der Urteilsfindung nur Methoden bei, die zur Messung der Übereinstimmung von Urteilen und zur Beschreibung einzelner Aspekte der Urteile dienen können. Zur Zeit haben die Evaluatoren noch keine Methoden, um die Validität von Urteilen abzuschätzen. Vielleicht kann die Validität eines Urteils am besten dadurch erhöht werden, daß man die wenigen Personen heranzieht, die durch ihre genaue Kenntnis der Umstände besonders gut geeignet sind, gültige Urteile abzugeben. Ein erfahrener Beamter der Schulverwaltung strebt genauso nach fundierten Urteilen wie der Evaluator. Er interessiert sich weniger für die Messung der »Homogenität« der Urteile. Es ist sogar so, daß er widersprüchliche Urteile erwartet. Aufgabe der Beamten der Schulverwaltung ist es nicht, Meinungsverschiedenheiten zu beseitigen oder Urteile einander anzugleichen, sondern zu entscheiden, wessen Urteil in einer bestimmten Frage angemessen ist. In den einfachsten sozialen Organisationen lernen die beteiligten Personen schnell, die Gültigkeit der In-

formation, die eine Person liefert, zu bestimmen. In Organisationen von der Familie bis zur Körperschaft findet unter den Mitgliedern eine Interaktion statt, um die Kenntnisse jedes einzelnen festzustellen. In einer Familie wird man dem Urteil des Kleinkindes, welches die beste Farbe für das Wohnzimmer ist oder ob der Keller von Gespenstern bevölkert ist, kaum Bedeutung beimessen; man wird ihm jedoch ein Urteil darüber zutrauen, ob es Hunger oder Durst hat. Aufgabe eines Beamten der Schulverwaltung ist es, festzustellen, wer die besten Kenntnisse als Basis für seine Entscheidung liefern kann. Dabei ist es eins der größten Probleme, daß die Beamten beim Aufstieg in die Verwaltungshierarchie den Kontakt mit den Praktikern verlieren, deren Information sie benötigen. Ohne Interaktion mit den Lehrern verliert der Verwaltungsbeamte bald das Gefühl dafür, wen er zu einem bestimmten Problem befragen muß. Auf die Evaluation bezogen, heißt dies: Wessen Urteil ist der Beachtung wert und wessen nicht? Diese Frage ist viel schwieriger zu beantworten als die Frage, ob die Beurteiler A und B die gleichen Meinungen vertreten. Auf jeden Fall nehmen diejenigen, die sich die Frage nach der Gültigkeit von Urteilen nicht stellen, dem Prozeß der Urteilsbildung in der Evaluation seine Bedeutung.

Es gibt jedoch wichtige Fälle, in denen die Gültigkeit der Urteilsdaten, d. h. ihr Wahrheitsgehalt oder ihre Zuverlässigkeit, irrelevant ist, wenn Urteile als Begleitfaktoren oder Prädiktoren zukünftiger Handlungen untersucht werden. In einem solchen Fall ist es unvernünftig, die Sammlung der Urteile eines potentiellen Entscheidungsträgers mit dem Argument abzulehnen, sie seien subjektiv. Wenn z. B. die positive oder negative Einstellung eines Schulleiters gegenüber dem innovativen Charakter eines neuen Curriculum mit 90 Prozent Wahrscheinlichkeit seine Annahme oder Ablehnung nahelegt, lohnt es sich kaum, danach zu fragen, ob der Schulleiter ein kompetenter Beurteiler von innovativen Curricula ist. Ungeachtet der Fähigkeit, über solche Phänomene zu urteilen, kann eine wichtige und funktionelle Beziehung zwischen Einstellung und Handlung beobachtet werden. Die Übereinstimmung in einer Gruppe von Beurteilern ist für den Evaluator nicht immer wichtig; noch ist die Gültigkeit des Urteils immer von Interesse. Die Verläßlichkeit der Urteilsdaten kann unabhängig von ihrer Gültigkeit erwogen werden. Gegenwärtig haben die Evaluatoren nur wenige Methoden aus der Psychometrie zur Untersuchung der Übereinstimmung von Urteilen von Personen übernommen, sie haben aber keine Methoden für die Untersuchung der Gültigkeit der Urteile dieser Personen zur Verfügung.

2. *Die Rechtfertigung der Gewichtung der Einzelwerte und ihrer Zusammenfassung zu einem Gesamtwert*

Das zentrale Problem des Zielkomplex-Modells der Evaluation besteht darin, die Daten auf verschiedenen Skalen zu einer einzigen Wertbeurteilung zusammenzufassen. Ungeachtet der verschiedenen möglichen Methoden, mit denen man Leistungsdaten zusammenfassen kann, wird ein Evaluator vielleicht eine Schwierigkeit darin sehen, nach verschiedenen Kriterien erbrachte Leistungen gleichzusetzen. Soll zum Beispiel, wenn für ein Mathematikcurriculum der Sekundarschule ein zusammengesetzter Meßwert zu bestimmen ist, der Erwerb von Fertigkeiten, Probleme zu lösen, doppelt oder halb soviel wie die Fähigkeit, sich an Fakten zu erinnern, gewichtet werden? Daß Evaluatoren diese berechtigten Fragen selten ernst nehmen, deutet auch auf eine fehlende Technik für den Umgang mit diesen wichtigen Problemen hin.

Im Zusammenhang mit der Verbesserung der Technik der Curriculumentwicklung gewinnt das Problem an Bedeutung, wie man Kriterien gewichten soll, um eine zusammengesetzte Wertskala zu entwickeln. Eine verbesserte Technik der Curriculumentwicklung sollte den Curriculumautoren helfen, die von ihnen erstrebten Ziele zu erreichen. Die typische empirische Evaluation der Zukunft wird sich vielleicht mit der Bestätigung begnügen, daß jedes Curriculum seine Ziele erreicht; einige der Ziele wären allein seine speziellen Ziele, andere hätte es mit allen verglichenen Curricula gemeinsam. Die tatsächliche Bestimmung seines Wertes wird dann in der Gewichtung der Verhaltensdaten zu einer gewichteten Leistungsskala bestehen.

Die Antwort auf das Gewichtungsproblem liegt wahrscheinlich in der Entdeckung einer grundlegenden Maßeinheit für Nutzen, Gewinn oder Wert, die für alle Lernziele gültig ist. Das Fehlen dieser Maßeinheit für das Messen pädagogischer Werte erinnert an die Entwicklung der deskriptiven Linguistik. Die Linguistik machte jahrelang geringe Fortschritte, weil die Mannigfaltigkeit sprachlicher Äußerungen die Kodifizierung erschwerte. Die Definition des Phonems als kleinste Einheit, die wenigstens zwei gesprochene Worte unterschied, bedeutete eine revolutionäre Entdeckung für linguistische Untersuchungen. Seitdem machte die Linguistik große Fortschritte. Ebenso wurde die psychologische Schlafforschung durch die Entdeckung der raschen Augenbewegungen (REM) neu belebt. Wir nähern uns vielleicht einer ähnlichen Situation in der Entwicklung der Evaluation, in der die Entdeckung einer für alle Curricula gültigen Maßeinheit die echte Einschätzung des Wertes von Curricula erlaubt und der stagnierenden Methodologie der Evaluation neue Impulse vermitteln wird.

3. Rechtfertigung der Auswahl von Zielen

Im Unterschied zum Tylerschen Modell, in dem Ziele ohne Fragen akzeptiert werden, oder auch zum Akkreditationsmodell, in dem Ziele zwar beurteilt, manchmal jedoch unzulänglich beurteilt werden, stellt das Zielkomplex-Modell auch die Frage, ob die Ziele eines Curriculum überhaupt erstrebenswert sind.

»So muß richtig verstandene Evaluation gleichermaßen Leistungsmessung in bezug auf die Ziele und die Verfahrensweisen für die Evaluation der Ziele einschließen.« (Scriven 1967, 52, 72)

Dagegen betonte Tyler (1951, 48) noch nicht die Notwendigkeit, die Ziele selbst zu evaluieren: »Evaluation bezeichnet einen Bewertungsprozeß, der die Billigung spezifischer Werte und die Verwendung zahlreicher Beobachtungsverfahren enthält einschließlich quantitativer Verfahren als Grundlagen für Werturteile.«

Angenommen, der Entwickler eines Curriculum in der Politischen Bildung für die 9. Klasse in Iowa beschließt, eine ein halbes Jahr dauernde Einheit über moderne Weltprobleme um die Hälfte zu kürzen und statt dessen eine Einheit über die Geschichte Iowas einzuführen, dann würde man vom Evaluator, der nach dem Tylerschen Modell arbeitet, erwarten, daß er dem Curriculumentwickler behilflich ist, die Lernziele der neuen Einheit besser zu formulieren, und daß er ihm Beweise für den Erfolg seines Materials liefert. Der Evaluator, der nach dem Akkreditationsmodell arbeitet, wird wahrscheinlich Bedenken anmelden, diese Einheit in das Curriculum einzugliedern, weil das dazu führen könnte, die Geschichte von Iowa zu einem Prüfungsgegenstand für Lehrer zu machen. Der Evaluator, der nach dem Management-System-Modell arbeitet, würde zu bestimmen versuchen, welche Daten der Curriculumentwickler benötigt, um seine Materialien in den Schulen einzuführen.

Von dem Evaluator, der nach dem Zielkomplex-Modell vorgeht, könnte man erwarten, daß er feststellt, ob Schüler der 9. Klasse in Iowa ein halbes Jahr lang die Geschichte Iowas durchnehmen *sollten*. Er kann wahrscheinlich herausfinden, daß 85 % der betroffenen Schüler der 9. Klasse den Staat mit 23 Jahren verlassen und niemals zurückkehren. Er kann zu dem Schluß kommen, daß in einer derartig mobilen Gesellschaft die Verwendung eines vollen Semesters für die Geschichte Iowas nicht gerechtfertigt werden kann. Scriven weist darauf hin, daß Evaluation sich der Frage der Rechtfertigung von Zielen stellen muß, und führt aus:

Natürlich, wenn wir *nicht* wissen, daß (und im allgemeinen auch nicht, wie) ... Leistung Gewinn bringt, ist es ein Widerspruch, Leistungsmessung als Evaluation anzusehen, und gerade dieser Widerspruch findet sich in einem großen Teil der

Curriculumevalution, wo dann von derartigen gesammelten Daten keine haltbaren Schlüsse über den Nutzen gezogen werden können. Eine gute Konzeptanalyse (des relevanten Konzepts des Nutzens im Hinblick auf die in ihm enthaltenen qualitativen Bestimmungen) und eine gute Versuchsplanung sind notwendige Voraussetzungen *für jegliche Leistungsmessung im Evaluationsprozeß* (Scriven *1966*, 6,7).

Man ist überrascht, wie viele Wissenschaftler noch immer darauf bestehen, daß Wissenschaft keine Wertfragen zu stellen habe. Wenn ein bekannter Psychometriker zur Feder greift, wird der Leser wie nach einer modernen Fassung des *de gustibus non disputandum* behandelt, das unbegründet auf wissenschaftliche Forschung und ihre Anwendung generalisiert wird:

In Diskussionen über Methoden und Ziele der Wissenschaft wird oft darauf hingewiesen, *daß sie sich lediglich mit der Aufdeckung funktionaler Beziehungen zwischen* Variablen befaßt, ohne sich dafür zu interessieren, ob die Variablen oder die funktionalen Beziehungen wertvoll sind. Sie kann sich nicht mit moralischen, ethischen oder gesellschaftlichen Werten beschäftigen, außer wenn sie versuchen würde, Variablen in diesen Gebieten zu definieren und Beziehungen zwischen ihnen aufzudecken... Das bedeutet nicht, daß Wissenschaftler als Personen sich nicht um Werturteile und moralische und ethische Fragen bemühen sollten. Es bedeutet lediglich, daß diese Überlegungen kein angemessener Forschungsgegenstand für wissenschaftliche Methoden oder Verfahren sind. Leider wurde diese Unterscheidung nicht nachdrücklich und klar genug getroffen. Viele Leute haben Schwierigkeiten, Wertvorstellungen und wissenschaftliche Vorstellungen auseinanderzuhalten. Wenn Werturteile gefällt werden und Lernziele oder allgemeine Ziele im Hinblick auf diese Werturteile formuliert werden, dann ist es die legitime Rolle der Wissenschaft, Methoden zur Erreichung dieser Ziele zu entwickeln, zu formulieren oder zu untersuchen; jedoch kann Wissenschaft keine Aussage darüber machen, ob diese Ziele angestrebt werden sollen. Wissenschaftliche Methoden können bestimmen, ob das Erreichen bestimmter Lernziele die Verwirklichung anderer Lernziele erleichtern wird, aber sie können keine Aussage darüber machen, ob die Lernziele gut oder schlecht sind, außer wenn sie das Erreichen anderer Lernziele fördern (Horst *1966*, *335*).

Nur wenige Wissenschaftstheoretiker würden mit Horst übereinstimmen. Die moderne Auffassung über die Beziehung der Wissenschaft zu Werten kommt in der Formulierung der Aufgabe zum Ausdruck, die Kaplan sich selbst im zehnten Kapitel seines Buches »The Conduct of Inquiry« (*1964*, 373) stellt:

Die These, die ich vertreten möchte, besagt, daß nicht alle Wertfragen unwissenschaftlich sind, sondern daß in der Tat einige von ihnen von der wissenschaftlichen Forschung aufgeworfen werden und daß *diejenigen*, die den wissenschaftlichen Idealen zuwiderlaufen, unter Kontrolle gebracht werden können, sogar von den Wissenschaften, in denen die Wertfragen die größte Rolle spielen.

Der noch immer skeptische Leser wird verwiesen auf Glanville Williams Buch »The Sanctity of Life and the Criminal Law«, das eine logisch und wissenschaftlich meisterhafte empirische Analyse der moralischen und gesellschaftlichen Aspekte der Geburtenkontrolle, Sterilisation, künstlichen Befruchtung, Abtreibung, des Selbstmordes und der Euthanasie darstellt. Wenn Philosophen und Sozialwissenschaftler einer Lösung dieser schwierigen Fragen näherkommen können, dann brauchen Pädagogen sich nicht von der Schwierigkeit der Einschätzung des relativen gesellschaftlichen Wertes einiger Curricula entmutigen zu lassen.

Pädagogische Veröffentlichungen enthalten wertende Äußerungen über die einen oder anderen Curricula oder Unterrichtsmethoden. Die Bestimmung des relativen Wertes von »heuristischem Lehren« und »darstellendem Lehrervortrag« (discovery and dispository teaching) muß auf einer Analyse der Definitionen der beiden Begriffe und empirischen Längsschnittuntersuchungen der Wirkungen jeder Methode auf das Behalten von Wissen und auf die Entwicklung von Interesse, Motivation, Berufsplänen, Persönlichkeit usw. beruhen. Die gegenwärtigen Erörterungen über die Überlegenheit des heuristischen Lehrens über das darstellende Lehren verlangen nach ernsthaften Versuchen, die Begriffe logisch zu analysieren und aussagekräftige empirische Daten zu sammeln.

Zur Rechtfertigung von Bildungszielen bedarf es ohne Zweifel logischer und empirischer Analysen. Philosophen können wesentlich dazu beitragen, das Problem der Rechtfertigung der Auswahl von Zielen zu lösen, indem sie die logische Konsistenz zwischen curricularen Zielen und der Philosophie des Curriculum bzw. der Begründung des Curriculum und der Übereinstimmung mit den philosophischen Grundgedanken der Erziehung untersuchen. Man kann Wissenschaftler fragen, ob die für ihre Disziplin relevanten Ziele sich rechtfertigen lassen. So ist z. B. ein Biologe besonders kompetent, um zu beurteilen, ob Lysenkoismus wegen seines Wertes als Gegenstand wissenschaftlicher Forschung in einem Biologiekurs der Sekundarschule gelehrt werden sollte. Sozialwissenschaftler können von allen Wissenschaftlern wahrscheinlich am meisten zur Lösung der Probleme der Auswahl von Zielen beitragen.

Die Psychologie wird für die Rechtfertigung eines curricularen Ziels oft sehr relevant sein. Man betrachte als Beispiel das von der American Association for the Advancement of Science (AAAS) entwickelte naturwissenschaftliche Curriculum für die Primarstufe. Die Autoren dieses Curriculum betrachten Naturwissenschaft als Sammlung einer kleinen Zahl transferierbarer Prozesse und wissenschaftlicher Methoden.

Die AAAS-Materialien setzen sich zum Ziel, diese heuristischen Fertigkeiten dem Schüler zu vermitteln; der Kontext ihrer Anwendung, d. h. die

Inhalte des naturwissenschaftlichen Curriculum, wird für wesentlich weniger wichtig gehalten. Einige Kritiker haben das AAAS-Curriculum angegriffen; sie vertreten die Auffassung, daß es auf der Vermögenspsychologie des 19. Jahrhunderts beruht. Sie behaupten, psychologische Forschung habe gezeigt, daß das Gedächtnis nicht als eine Sammlung von Anlagen oder Fähigkeiten angesehen werden kann, die durch Gebrauch verbessert und sodann in einer Vielzahl von Situationen angewendet werden können. Die Frage, ob die AAAS-Materialien auf einer solchen Vorstellung vom Lernenden basieren und ob ein solches Konzept als eine Theorie des Verhaltens nutzlos ist, können nur Psychologen qualifiziert beantworten. Die Antworten könnten sicherlich Einfluß auf die Rechtfertigung des prozeßorientierten Charakters der AAAS-Curriculummaterialien haben.

Pädagogische Forschung, die die Auswahl von Bildungszielen rechtfertigen kann, ist dringend erforderlich. Es fehlen uns die elementarsten Daten – etwa aus Längsschnittuntersuchungen – darüber, wie Wissen behalten wird und Interessen entstehen. Wie sollen wir wissen, ob ein Curriculumentwickler gut beraten ist, wenn er sich mit der Förderung des Interesses an Mathematik beschäftigt, anstatt vielmehr mathematische Inhalte zu lehren? Wenn Längsschnittbefragungen zeigen, daß mathematische Inhalte innerhalb von 5 Jahren nach Beendigung des formalen Unterrichts vergessen werden, daß aber das Interesse an Mathematik fortbesteht und zu weiterer Beschäftigung und positiver Einstellung gegenüber den Wissenschaften führt, dann ist die Auswahl der Ziele der Curriculumentwickler wahrscheinlich gerechtfertigt. Offensichtlich wird pädagogische Evaluation auch von anderen Wissensgebieten abhängig sein, um mit ihrer Hilfe Fragen nach der Rechtfertigung der Auswahl von Zielen zu beantworten.

Schlußfolgerung

Wie jedes komplexe Werk des Menschen hat die Methodologie der Evaluation kein wirkliches Entwicklungspotential; das einzige Entwicklungspotential ist ein Plan für ihre zukünftige Entwicklung im Geist ihrer Schöpfer.

III Unterrichtsbeobachtung und Evaluation

Einführung

Ob und inwieweit Unterrichtsbeobachtung zur Evaluation eines Bildungsprogramms (Curriculum, Schulversuch, usw.) gehört, hängt davon ab, welches Erkenntnisinteresse zugrunde liegt. Das Erkenntnisinteresse ist wiederum abhängig von den Adressaten der Evaluation (Lehrern, Schülern, Eltern, Schulverwaltungen, Bildungspolitikern, Erziehungswissenschaftlern) und den Evaluatoren. Bei der Entwicklung eines Evaluationsplans wird man sich daher zunächst folgende drei Fragen stellen, von deren Beantwortung es abhängt, ob Unterrichtsbeobachtung stattfinden soll und wenn, in welcher Form und in welchem Ausmaß sie erfolgen soll:

(1) Welche Informationen benötigen die Adressaten der Evaluation?
(2) Welche Informationen hat man in ähnlichen Untersuchungen zu gewinnen versucht?
(3) Welche personellen und materiellen Möglichkeiten bestehen für die Durchführung der Evaluation?

Zweifellos ist die Frage nach dem Erkennisinteresse und der Art der benötigten Informationen die entscheidende Frage zu Beginn einer Evaluation. So dürften einige Adressaten mehr an einer ideologiekritischen, d. h. intrinsischen Evaluation der Ziele, Inhalte und Materialien eines Curriculum interessiert sein, andere an einer reinen Ergebnisevaluation und wie der andere an einer Aufwands-Effektivitäts-Analyse (vgl. Alkin 146 ff.).

Für viele Adressaten einer Evaluation ist es wichtig, Informationen darüber zu erhalten, wie die Wirklichkeit eines Schulversuchs aussieht und ob die Intentionen einer Innovation im Unterricht realisiert werden, d. h. ob und inwieweit es zu einer Verbesserung der Unterrichtswirklichkeit kommt. Im amerikanischen Bereich ist es hier in den letzten Jahren zu einer intensiven Forschungsarbeit gekommen. Man begnügte sich nicht länger mit der bloßen Erfolgskontrolle in einer Ergebnisevaluation, sondern wollte verstärkt Informationen darüber erhalten, wie bestimmte Ergebnisse erreicht worden sind. Wurde dies bereits zu einem Anliegen derer, die mit Hilfe von Verhaltenszielen das Schwergewicht der Evaluation auf die Er-

gebnisevaluation legten, so mußte es für die Pädagogen zu einer zentralen Frage werden, die die besondere Wichtigkeit der Unterrichtsprozesse auch im Hinblick auf die Zieldefinition des Unterrichts betonten (Eisner 1969; Stenhouse 1971; Brügelmann 1972; Wulf 1972 b).

Im Rahmen einer Evaluationsuntersuchung kommt der Unterrichtsbeobachtung auch insofern erhebliche Bedeutung zu, als sie dazu beitragen kann, die Gründe für Unzulänglichkeiten bei der Realisierung eines Schulversuchs aufzudecken. Es lassen sich mit ihrer Hilfe »Pattern« im Unterrichtsverhalten von Lehrern und Schülern entdecken, die z. B. die Realisierung bestimmter Innovationen behindern. Darüber hinaus kann Unterrichtsbeobachtung unmittelbare Informationen über die Angemessenheit bestimmter, an einem Bildungsprogramm orientierter Lehr- und Lernstrategien liefern und bei etwaiger Unzulänglichkeit zu entsprechenden Modifikationen der im Rahmen der Realisierung der Innovationen erfolgenden Lehrerfortbildung führen. In den letzten Jahren dienten Unterrichtsbeobachtungssysteme in den USA außer zur Erforschung von Lehrer- und Schülerverhalten auch zur Curriculumevaluation (vgl. z. B. Oliver/Shaver 1966) und vor allem zur Lehrerausbildung. In der Lehrerausbildung kommen ihnen wenigstens drei Funktionen zu:

(1) Sie können dazu dienen, Lehrerverhalten und Lehrerausbildungsprogramme zu evaluieren;

(2) sie sensibilisieren den Lehrer, der sie zur Analyse von Unterrichtsverhalten im Rahmen seiner Ausbildung anwendet, für Unterrichtsinteraktionen und helfen ihm bei der Selbstevaluation seines Unterrichts;

(3) sie geben dem Lehrer in einem ihn unmittelbar berührenden Bereich eine Einführung in die Anwendung von Verfahren empirischer Forschung.

Daher haben Unterrichtsbeobachtungssysteme an vielen amerikanischen Universitäten einen festen Platz im Curriculum der Lehrerausbildung.

Unsere oben formulierten Fragen nach den erkenntnisleitenden Interessen der Adressaten der Evaluation und den dementsprechend in den verschiedenen Untersuchungen gewählten Schwerpunkten führen unmittelbar zu der Notwendigkeit, eine Bestandsaufnahme der bestehenden Unterrichtsbeobachtungssysteme zu machen. Dabei muß man das den Systemen zugrunde liegende Erkenntnisinteresse aufdecken und die dazu entwickelten Beobachtungsverfahren analysieren. Daraufhin kann man entscheiden, ob das System zur intendierten Evaluation der Unterrichtsprozesse beitragen kann oder ob man ein anderes verwenden oder gar ein eigenes Beobachtungssystem entwickeln muß. Eine systematische Analyse der vorliegenden Beobachtungssysteme kann außerdem Auskunft darüber geben, wo die bisherigen Schwerpunkte der Unterrichtsbeobachtung bzw. -evaluation lagen. Einen solchen Versuch machen die 15 umfangreichen Bände der

Mirrors for Behavior (Simon/Boyer, 1967; 1970), in denen eine Sammlung von 79 Unterrichtsbeobachtungsinstrumenten dargestellt wird, die nach 7 Kategorien klassifiziert worden sind. Ohne hier auf eine nähere Analyse dieser Anthologie eingehen zu können, sollen doch wenigstens kurz die Kategorien genannt und die Zuordnung der Instrumente wiedergegeben werden, aus der bereits die Schwerpunkte der bisherigen Unterrichtsbeobachtung deutlich werden und die zugleich eine Antwort darauf geben, inwieweit diese Systeme im Rahmen der Evaluation verwendet werden können und inwieweit neue Systeme entwickelt werden müssen, die neuen und anderen Erkenntnisinteressen gerecht werden.

Die sieben Dimensionen der Unterrichtsbeobachtung sind:

1. affektive – der emotionale Inhalt der Kommunikation (62 Systeme)
2. kognitive – der intellektuelle Inhalt der Kommunikation (48 Systeme)
3. psychomotorische – nicht sprachliche Kommunikation (17 Systeme) (mit Körperhaltung, Gesichtsausdruck, Gesten usw.)
4. Aktivität – ein Geschehen, das sich auf eine Person oder einen Gegenstand (z. B. Lesen und Schreiben) bezieht (37 Systeme)
5. Inhalt – worüber gesprochen wird (27 Systeme)
6. soziologische Struktur – Soziologie der Interaktion einschließlich der Frage, wer spricht zu wem in welcher Rolle (26 Systeme)
7. äußere Bedingungen der Schulumwelt – Beschreibungen des Schulraumes, in dem die Beobachtung stattfindet, einschließlich der benutzten Materialien und Einrichtungen (10 Systeme).

Deutlich liegt der Schwerpunkt der Unterrichtsbeobachtung in der affektiven Dimension. Ihr folgen die kognitive, die aktivitäts- und inhaltsbezogene und die soziologische Struktur-Dimension. Entsprechend den Bedürfnissen der Adressaten wird man also im Hinblick auf die Dimensionen der Untersuchung eine Auswahl treffen müssen. Darüber hinaus bietet sich eine weitere Spezifikation des Beobachtungssystems an. Es muß entschieden werden, ob die Unterrichtsbeobachtung sich auf den Lehrer (14 Systeme), die Schüler (7 Systeme) oder auf die Interaktion zwischen ihnen (67 Systeme) beziehen soll (12 Systeme für den außerschulischen Bereich).

So nützlich die Anthologie »Mirrors for Behavior« für den ist, der Unterrichtsprozesse evaluieren will, so wird doch ihr Wert durch die weitgehend deskriptive Zusammenstellung der Systeme eingeschränkt. Es erfolgt keine Bewertung und kritische Sichtung der vorliegenden Instrumente, so daß eine effektive Nutzung auf erhebliche Schwierigkeiten stößt. Bei einer sorgfältigen Analyse der Beobachtungsinstrumente ergibt sich sodann, daß viele Systeme m. E. nicht genügend evaluativ sind, d. h. für die Zwecke der Evaluation nicht in ausreichendem Maße die Dimension der Bewertung der Unterrichtsaktionen berücksichtigen (vgl. auch Scriven 1967, 49, 69).

Nach Aufarbeitung der wohl umfangreichsten Bibliographie zur Unterrichtsbeobachtung im »Summary«-Band der Mirrors for Behavior erschien der Abdruck dieser beiden Beiträge von Bellack und Nuthall sinnvoll, da sie eine gute Einführung in die Grundproblematik der Unterrichtsbeobachtung liefern und zugleich den Forschungsstand in diesem Bereich kritisch zusammenfassen und auf neue Aufgabenfelder hinweisen. Die beiden Beiträge machen nicht den Versuch, alle 79 Systeme zu beschreiben. Sie wählen einige der Systeme aus, die für die Entwicklung des ganzen Untersuchungsbereichs von zentraler Bedeutung sind. Die Auswahl erfolgt so, daß die Beobachtungssysteme für die verschiedenen Dimensionen der Unterrichtsbeobachtung exemplarisch sind.

Der Beitrag Bellacks, der selbst eines der am besten ausgearbeiteten Beobachtungssysteme entwickelt hat, versteht sich als eine Einführung in die Methoden der Beobachtung des Unterrichtsverhaltens von Lehrern und Schülern, der zugleich in didaktisch gelungener Weise einige konzeptuelle und methodische Probleme der Entwicklung von Beobachtungssystemen darstellt. Sodann bietet er die exemplarische Beschreibung des auf die affektive Dimension zielenden Flanderschen Systems der Interaktionsanalyse, des Bellackschen Systems der Analyse der kognitiven Dimensionen des Unterrichts und des Oliver/Shaverschen Systems zur Analyse der affektiven, der kognitiven und der Verfahrens-Dimension, das im Zusammenhang mit der Evaluation des Public Issues Harvard Social Studies Project entwickelt wurde. Schließlich werden noch einige Probleme der Entwicklung und des Gebrauchs dieser Analyse-Systeme dargelegt.

Die Ausführungen Nuthalls, der früher an der Entwicklung des bekannten Smith/Meuxschen Systems mitgearbeitet hatte, sind in vieler Hinsicht eine Ergänzung und Erweiterung des Bellackschen Beitrags. Der Autor beschreibt, vergleicht und kritisiert einige weitere Beobachtungssysteme. Ferner werden zahlreiche Ergebnisse der Unterrichtsbeobachtung im Hinblick auf ihren Beitrag zum Verständnis der Ursachen für bestimmte Lernvorgänge dargestellt und analysiert, die von erheblichem Interesse sein dürften. Abschließend wird versucht, die Forschungsergebnisse aufzuarbeiten, die die Schülerleistung in Beziehung zu den Unterrichtsinteraktionen und dem Lehrerverhalten setzen – eine für die Evaluation von Unterrichtsprozessen besonders wichtige Frage.

Die beiden Beiträge bieten zusammen eine Einführung in den Stand der Entwicklung von Systemen zur Unterrichtsbeobachtung und -evaluation, die bislang dem deutschen Leser nicht zugänglich waren, und dürften daher einen wesentlichen Beitrag zur Erweiterung unseres Verständnisses der Unterrichtsinteraktion und der Verfahren zu ihrer Erforschung und Evaluation liefern.

ARNO A. BELLACK

Methoden zur Beobachtung des Unterrichtsverhaltens von Lehrern und Schülern

Während der letzten zehn Jahre wurde durch das aufkommende Interesse an der Erforschung des Unterrichtsverhaltens eine bedeutsame Entwicklung in der pädagogischen Forschung in Gang gesetzt. Es handelt sich hierbei um das Wiederaufgreifen eines seit langem existierenden Anliegens der Erziehungswissenschaftler. Bei den gegenwärtigen Unterrichtsforschungen, die sich von den Arbeiten vorangegangener Perioden unterscheiden, wird auf systematische Verhaltensbeobachtung von Schülern und Lehrern im Unterricht großer Wert gelegt.

Dieser Beitrag stellt eine Auseinandersetzung mit der neueren Entwicklung von Methoden der Verhaltensbeobachtung im Unterricht dar. Er gliedert sich in drei Abschnitte:

I. Entwicklung von Systemen zur Verhaltensbeobachtung im Unterricht
II. Beispiele für Beobachtungssysteme
III. Probleme und Konsequenzen in der Entwicklung und im Gebrauch von Systemen zur Analyse des Lehrer- und Schülerverhaltens.

I. Entwicklung von Systemen zur Verhaltensbeobachtung im Unterricht

Bei der Konstruktion eines Systems zur Beobachtung des Unterrichtsverhaltens muß der Forscher zwei fundamentale Fragen beantworten: (1) Welche Dimensionen des Unterrichtsverhaltens sollen beobachtet werden? (2) Auf welche Art und Weise sollen die Beobachtungen durchgeführt werden? Diese beiden Fragen sind natürlich eng miteinander verknüpft, denn Beobachtungsgegenstand und Beobachtungsmethode sind voneinander abhängig. Der Prozeß der Unterrichtsbeobachtung wird von den Konzeptionen des Forschers über das Verhalten, das er untersucht, und von den Verfahrensweisen bestimmt, mit denen er die ihm zur Verfügung stehenden Daten sammelt.

Welches Unterrichtsverhalten soll beobachtet werden?

Da der Forscher keinesfalls alles in der Klasse beobachten kann, muß er eine Entscheidung darüber fällen, welche Aspekte des Lehrer- und Schülerverhaltens er zum Gegenstand seiner Untersuchungen nimmt. »Jede Beobachtung wird gemacht«, schreibt Kaplan, »sie ist das Produkt einer aktiven Wahl, nicht eines passiven Geschehens. Keine einzige Interpretation ist notwendigerweise die Folge von dem, was beobachtet wurde; es gibt immer viele Möglichkeiten, Verhalten zu klassifizieren« (Kaplan 1964, 133).

Wie der Forscher bei der Klassifikation des Unterrichtsverhaltens verfährt, wird vom Ziel seiner Untersuchung bestimmt, ebenso vom konzeptuellen Rahmen, der den Beobachtungsgegenstand, die Beobachtungsmethode und den Umfang der Beobachtung angibt. Flanders (1965) z. B. entwickelte auf Grund seines theoretischen Interesses am Klassenklima eine Reihe von Kategorien, um die verschiedenartigen Einflüsse des verbalen Lehrerverhaltens zu beschreiben. Im Gegensatz dazu konstruierten B. O. Smith und Meux (1962) auf Grund ihres Interesses an den kognitiven Dimensionen des Unterrichtsgeschehens einen Katalog von Beobachtungskriterien, um logische Operationen im Unterricht zu kategorisieren. Kounin (1967), dessen Interesse sich auf Klassenführung und Disziplin konzentrierte, entwickelte ein Instrument zur Klassifikation von Dimensionen des Lehrerstils, wie er sich bei der Kontrolle des Schülerverhaltens zeigt.

Zur systematischen Beobachtung des Unterrichtsverhaltens, das für die Zielsetzung und den theoretischen Anspruch relevant ist, muß der Wissenschaftler ein zuverlässiges und gültiges Instrument in Form von Schätzskalen (rating scales) bzw. ein Beobachtungssystem entwickeln, das in operationalisierter Form die zu kategorisierenden oder zu messenden Verhaltensdimensionen spezifiziert.

Schätzskalen: Beim Gebrauch von Schätzskalen besteht die Aufgabe des Beobachters darin, das beobachtete Verhalten an einem Ort auf einem Kontinuum oder in einem in einer bestimmten Reihenfolge vorgegebenen Kategoriensystem zu lokalisieren (vgl. Remmers 1963). Ein Beispiel für eine experimentell entwickelte Schätzskala ist die von Ryan (1960) für seine Erforschung von Lehrercharakteristika konstruierte Skala. In dieser Untersuchung bewertete der Beobachter, nachdem er einer Unterrichtsstunde beigewohnt hatte, jede der sechsundzwanzig Dimensionen des während des Unterrichts beobachteten Verhaltens und übertrug anschließend seine Bewertung in Form von Schätzungen auf 7-Punkte-Skalen. Die 22 Dimensionen des Lehrerverhaltens bezogen sich u. a. auf folgende Kategorien: parteiisch – fair; barsch – freundlich; zurück-

haltend – temperamentvoll; autokratisch – demokratisch und unsicher – sicher. Zu den Dimensionen des Schülerverhaltens gehörten: teilnahmslos – lebhaft; sich auf andere verlassend – die Initiative selbst ergreifend; Widerstand leistend – sich verantwortlich fühlend. Jede Dimension war in einem speziellen Wörterverzeichnis definiert, wobei die relevanten Verhaltensaspekte, auf denen die Schätzung basierte, gesondert angegeben wurden.

Die Verfahren, in denen die Schätzung *nach* der Beobachtung erfolgte – wie bei den von Ryan (1960) entwickelten – unterliegen allerdings bedeutsamen Einschränkungen, wenn mit ihrer Hilfe die Aktivitäten im Unterricht erfaßt werden sollen. Die Einschätzungen durch die Beurteiler liefern lediglich allgemeine Eindrücke und Erinnerungen an das tatsächliche Unterrichtsgeschehen, nicht jedoch exakte Aufzeichnungen des Lehrer- und Schülerverhaltens. Deshalb und wegen gewisser Unzulänglichkeiten der Schätzskalen als Meßinstrumente werden sie nur in wenigen gegenwärtigen Untersuchungen verwendet, die die direkte Beobachtung des Unterrichtsgeschehens zum Gegenstand haben, obwohl sie in früheren Untersuchungen häufig Anwendung fanden.

Beobachtungssysteme: Die Systeme zur Beobachtung des Unterrichtsverhaltens liefern dem Beobachter eine Reihe von Kategorien, denen das jeweilige Verhalten zugeordnet wird. Meistens sind Verhaltensstimuli oder operationale Definitionen der Kategorien und Kodierungsanweisungen vorhanden, um dem Beobachter die Entscheidung zu erleichtern, welcher der Kategorien das beobachtete Verhalten zuzuordnen ist. Während der letzten Jahre wurde eine große Anzahl von Beobachtungssystemen entwickelt. Simon und Boyer (1967) katalogisierten in ihrer Anthologie von Beobachtungsinstrumenten, »Mirrors for Behavior«, 26 Verfahren; doch enthält sie lediglich die Hälfte aller in der letzten Zeit entwikkelten Systeme [1].

Zwischen den vorhandenen Beobachtungssystemen zeigen sich große Unterschiede. Es bestehen Differenzen

1. in den Dimensionen des zu klassifizierenden Unterrichtsverhaltens,
2. in der Art des zur Beobachtung entwickelten Kriterienkatalogs,
3. in dem Bezugsrahmen des Beobachters für das Kodieren,
4. in der beim Kodieren zu benutzenden Verhaltenseinheit,
5. im Anwendungsbereich.

Diese Unterschiede verdeutlichen die Grundproblematik bei der Konstruktion von Beobachtungsinstrumenten.

1. Dimensionen des Unterrichtsverhaltens

Wie bereits dargelegt, werden bei den vorliegenden Systemen zur Verhaltensanalyse bestimmte Aspekte besonders hervorgehoben, andere vernachlässigt. Die vorhandenen Beobachtungskataloge enthalten ein weites Spektrum von Verhaltensweisen. Biddle (1967) fand bei einer Aufarbeitung der Forschungen über Unterrichtsverhalten, daß kürzlich entwikkelte Instrumente sich befassen

a) mit dem Lehrerverhalten im Sinne von Handlungsweisen, Methoden und charakteristischen Rollen,
b) mit dem Zuhörer- und Zielverhalten von Lehrern und Schülern,
c) mit der Lehrer-Schüler-Interaktion,
d) mit äußeren Strukturen – wie Lehrinhalten und administrativen Maßnahmen,
e) mit inneren Strukturen – wie Kommunikationsstruktur, Aktivitätsstruktur, charakteristischen Rollen, sozialen Funktionen.

Im Gegensatz dazu haben Simon und Boyer (1967) verschiedene Beobachtungsverfahren in drei Gruppen nach folgender Systematik kategorisiert:

a) affektive Systeme, mit deren Hilfe man das emotionale Klassenklima erfassen kann und die Art und Weise, wie es durch die Reaktion der Lehrer auf die Emotionen, Ideen oder Handlungsweisen der Schüler bedingt wird,
b) kognitive Systeme, die sich mit Denkprozessen befassen und den diese zum Ausdruck bringenden verbalen Verhaltensmustern,
c) multidimensionale Systeme, mit denen man sowohl die kognitiven als auch die affektiven Dimensionen des Verhaltens erfassen kann.

Das Flanderssche System der Interaktionsanalyse (1965) und das Hughessche System zur Klassifizierung der Funktionen des Lehrerverhaltens (1959) sind bedeutsame Beispiele für das erste System. Das System von Bellack und anderen zur Beschreibung des Sprachverhaltens im Unterricht (1966) und B. O. Smiths' und Meux' Verfahren zur Analyse logischer Operationen des Unterrichtens (1962) sind typisch für die zweite Gruppe. Olivers und Shavers Beobachtungssystem zur Beschreibung kontrastierender Unterrichtsstile in dem Fach Sozialkunde (social studies) (1966) sowie das System von Joyce zur Erfassung affektiver und kognitiver Aspekte der verbalen Lehrerkommunikation (vgl. Joyce/Harootunian 1967) sind Repräsentanten der dritten Kategorie.

Die vorhandenen Systeme zur Beobachtung des Unterrichtsverhaltens können ebenso nach der Art der Kommunikation klassifiziert werden, d. h. danach, ob sie die Aufmerksamkeit auf (a) verbale Aspekte, (b) nicht-ver-

bale Aspekte oder (c) verbale und nicht-verbale Aspekte der Kommunikation konzentrieren. Die meisten Systeme richten sich ausschließlich auf verbales Verhalten, während einige Systeme verbale und nicht-verbale Dimensionen beinhalten. Zwanzig der Systeme, die in »Mirrors for Behavior« (Simon und Boyer 1967) beschrieben sind, befassen sich mit verbaler Kommunikation, nur sechs beziehen sich sowohl auf verbales als auch auf nicht-verbales Verhalten. Soweit der Autor informiert ist, ist Galloway (1962) der einzige Forscher, der für seine Untersuchungen über nicht-verbale Kommunikation, wie die Gestik des Lehrers, den Ausdruck der Stimme und die Mimik, ein Instrument nur zur Beschreibung des nicht-verbalen Verhaltens entwickelt hat.

Die Entscheidung, welche Aspekte des Lehrer-Schüler-Verhaltens zu untersuchen sind, ist prinzipiell ein theoretisches Problem. Es sind jedoch auch wichtige methodische Erwägungen miteinbezogen, z. B. ist dafür zu sorgen, daß die Kategorien in einem vorgegebenen System sich gegenseitig ausschließen und einen definierten Bereich erschöpfend repräsentieren. Das Kriterium der Vollständigkeit bedeutet natürlich nicht, daß jedes Verhalten während einer bestimmten Beobachtungsperiode klassifiziert wird. Vielmehr ist es erforderlich, daß das zu beobachtende Verhaltensuniversum (z. B. soziales Klima, logische Operationen, Kommunikationsprozesse) klar definiert und eine repräsentative Stichprobe von Kategorien aus diesem Verhaltensuniversum entnommen wird. Zum Beispiel schließen sich die folgenden fünf von Aschner und Gallagher (1965) entwickelten allgemeinen Kategorien zur Klassifikation von Gedankenprozessen, die sich im verbalen Unterrichtsverhalten widerspiegeln, gegenseitig aus und sind auch repräsentativ für ein Universum kognitiver Funktionen: mechanisches Denken, Gedächtnistätigkeit, konvergentes Denken, evaluatives Denken, divergentes Denken.

In dem Stadium der Forschungsarbeit, in dem die relevanten theoretischen Konzepte noch nicht vollständig entwickelt sind, ist es häufig schwer, Kategorien aufzustellen, die sich gegenseitig ausschließen und die repräsentativ sind. Jedoch muß der Wissenschaftler im Verlauf seiner Forschungsarbeit diese Kriterien ständig beachten und sich nicht nur auf die Analyse der empirischen Beobachtungsdaten, sondern ebenso auf die Verbesserung der zur Analyse der Daten verwendeten Begriffe konzentrieren.

2. Verschiedene Arten von Beobachtungssystemen

In der Erforschung des Unterrichtsverhaltens werden heute weitgehend zwei Arten von Beobachtungssystemen angewandt; Medley und Mitzel (1963) bezeichnen diese als Kategoriensysteme und Zeichensysteme.

Ein Kategoriensystem beschränkt die Beobachtung auf spezifische Verhaltensdimensionen im Unterricht, wobei es eine Reihe von Kategorien liefert, in die jede beobachtete Verhaltenseinheit eingestuft wird. Das daraus resultierende Ergebnis zeigt für jeden Beobachtungszeitraum die vollständige Anzahl der aufgetretenen Verhaltenseinheiten an sowie die Anzahl innerhalb jeder einzelnen Kategorie. Die von Hughes (1959) in ihrer Untersuchung über die Interaktion in der Primarstufe entwickelte Reihe von Kategorien ist ein typisches Beispiel dafür. Der dieses System benutzende Beobachter klassifiziert das verbale und nicht-verbale Lehrerverhalten nach sechs Hauptfunktionen: Kontrollfunktionen, unterstützende Funktionen, Funktionen, die die Vermittlung von Unterrichtsinhalten zum Gegenstand haben, Funktionen, die als persönliche Antwort dienen, Funktionen von positivem und von negativem Gefühlswert. Im Gegensatz dazu liefert ein Zeichensystem dem Beobachter eine Reihe spezifischer Verhaltensweisen, die während einer Beobachtungsperiode auftreten oder nicht auftreten können. Die Beobachter werden angewiesen, auf diese spezifischen Verhaltensweisen zu achten, und das Ergebnis zeigt, welche dieser Verhaltensweisen während des Beobachtungszeitraums aufgetreten sind. Ein Beispiel für ein solches Beobachtungssystem ist die von Medley und Mitzel (1958) entwickelte Beobachtungsskala (OSCAR). Ein Beobachter, der dieses Instrument verwendet, registriert lediglich die Verhaltensweisen, die von den 71 Items erfaßt werden. Zu diesen in mehrere Abschnitte gruppierten Items gehören etwa solche: »der Lehrer trägt vor«, »der Lehrer beantwortet die Fragen der Schüler«, »der Lehrer veranschaulicht etwas an der Tafel«, »der Lehrer reagiert sarkastisch«, »der Schüler spricht zur Gruppe«, »der Schüler flüstert«, »der Schüler liest oder arbeitet an seinem Platz«.

Während Zeichensysteme gewöhnlich aus einer großen Itemanzahl bestehen, die sich auf konkrete, spezifische Verhaltensweisen beziehen und daher ein geringes Maß an Schlußfolgerungen seitens des Beobachters erfordern, sind Kategoriensysteme gewöhnlich aus einer geringeren Anzahl von Items zusammengestellt und besitzen ein höheres Abstraktionsniveau, das einen höheren Grad an schlußfolgerndem Denken beim Beobachter voraussetzt. Nicht alle Kategoriensysteme sind auf dem gleichen Abstraktionsniveau, aber in der Regel auf einem höheren als die Zeichensysteme konstruiert. Die in den beiden vorhergehenden Abschnitten erwähnten Beispiele veranschaulichen den Unterschied zwischen mehr oder weniger schlußfolgerndes Denken erfordernden Kategorien. Der Beobachter ist weniger auf Schlußfolgerungen angewiesen, wenn er die Handlung eines Lehrers zu interpretieren hat, wie z. B. »Beantwortung der Schülerfragen« oder »Veranschaulichung an der Tafel«. Im Gegensatz dazu er-

fordert es einen höheren Grad an schlußfolgerndem Denken, eine Aussage des Lehrers als Kontroll- oder Unterstützungsfunktion zu interpretieren. Es sollte ebenso festgehalten werden, daß Kategorien, die in einem hohen Maße schlußfolgerndes Denken erfordern und die die Verhaltensbeobachtung als Basis für Schlußfolgerungen über Motive oder Auswirkungen des Verhaltens beinhalten, den Forscher bei der Validitätsbestimmung mit schwierigeren Problemen konfrontieren, als es bei Kategorien der Fall ist, die sich mit deskriptiven Verhaltensbeobachtungen befassen und oft als solche bezeichnet werden, die Augenscheinvalidität besitzen.

Medley und Mitzel (1963) stellen fest, daß Kategoriensysteme häufiger in empirischen Untersuchungen angewandt werden, die auf hochdifferenzierten Theorien beruhen, während Zeichensysteme dann verwendet werden, wenn die Theorie nicht genügend aussagekräftig ist. Es ist selbstverständlich, daß die theoretische Orientierung des Forschers und der Grad der Entwicklung seiner Theorie das Niveau der Konzeptualisierung und folglich auch den Grad des erforderlichen Maßes an schlußfolgerndem Denken bestimmen.

3. Bezugsrahmen des Beobachters

Die Wissenschaftler versuchen, die Vorgänge im Schulunterricht wenigstens unter drei Aspekten zu kategorisieren:

a) die Absicht oder das Motiv des Handelnden
b) die Auswirkungen des Verhaltens auf den Adressaten
c) die objektiven Verhaltenscharakteristika.

Beispiele dieser drei Aspekte finden sich in neueren Untersuchungen über Unterrichtsprozesse.

Das analytische System von Withall (1949), das er in Verbindung mit seinen Forschungen über sozial-emotionales Klassenklima entwickelt hat, erfordert vom Beobachter eine Interpretation der Aussagen des Lehrers dahingehend, ob ihre Intention eine schülerunterstützende, problemstrukturierende oder direktive ist. Hughes (1959) andererseits klassifizierte in ihrer Untersuchung über das Unterrichtsverhalten von Primarschullehrern die Funktionen des verbalen und nicht-verbalen Lehrerverhaltens nach ihren erwarteten Wirkungen und deren Bedeutung für die Schüler. Im Gegensatz dazu beschreiben Smith und Meux (1962) in ihrer Untersuchung über die logischen Aspekte des Unterrichtsgeschehens die logischen Bestandteile der Lehrer- und Schüleraussagen, wobei sie über Motive des Sprechers oder seine Wirkungen auf die Zuhörer keine Aussagen machen.

Die Erfahrungen dieser und anderer Forscher verdeutlichen, daß das Unterrichtsverhalten auf der Basis aller drei Aspekte zuverlässig kodiert werden kann. Welcher der drei Aspekte für eine Untersuchung angemes-

sen ist, kann nur entschieden werden, wenn man das Ziel der Untersuchung berücksichtigt. Biddle (1967) hat mehrere mögliche Zielsetzungen genannt, die für die Unterrichtsforschung in Frage kommen, sowie Konzeptionen entwickelt, die diesen Zielen angemessen sind. Zum Beispiel ist er der Auffassung, daß Urteile über die Intention des Lehrers angemessen sind, wenn man primär an den Determinanten des Lehrerverhaltens interessiert ist. Wenn man sich allerdings mit der Qualifikation des Lehrers befaßt, sind Urteile über die Wirkung des Lehrerverhaltens auf die Lernprozesse der Schüler angebrachter. Wenn andererseits die individuellen und sozialen Determinanten des Verhaltens erforscht werden sollen oder kontrastierende Modelle der Interaktion im Klassenraum zu testen sind, wäre es sinnvoll, objektive Verhaltenscharakteristika als Grundlage zu nehmen.

4. Verhaltenseinheiten

Eine entscheidende Aufgabe bei der Erstellung eines Beobachtungssystems besteht darin, die Verhaltenseinheit zu spezifizieren, die als Basis für das Kodieren benutzt wird. Für die vielen Möglichkeiten, die Verhaltenseinheit zu definieren, gibt es zwei grundsätzlich verschiedene Methoden: (a) Bestimmung einer willkürlichen Zeiteinheit, (b) genaue Beschreibung einer analytischen Einheit, die häufig durch das Kategoriensystem selbst vorgegeben wird.

Bei den Systemen, bei denen eine willkürliche Zeiteinheit benutzt wird, muß der Beobachter ein Protokoll über die in dieser festgesetzten Zeit auftretenden Verhaltensweisen erstellen. Flanders (1965) z. B. verlangt vom Beobachter alle drei Sekunden ein Urteil darüber, ob der Lehrer innerhalb der kurzen Periode direkten oder indirekten Einfluß auf die Schüler ausübt. In gleicher Weise schreibt Spaulding (vgl. Simon/Boyer 1967) Zeitperioden von 3–10 Sekunden vor, für die der Beobachter ein Urteil über das Verhalten des Lehrers hinsichtlich seiner Kontrollfunktion in der Klasse abgibt, wobei er Kategorien wie »Setzen von Verhaltenszielen« und »Anordnen bestimmter Aktivitäten« verwendet. Medley und Mitzel (1958) andererseits erwarten, daß der Beobachter Verhaltensweisen in 5-Minuten-Perioden protokolliert.

Der grundlegende Vorteil der Setzung beliebiger Zeiteinheiten liegt in ihrem mechanischen Charakter, der eine Hilfe für die Regulierung des Beobachtungsprozesses darstellt. Weiterhin kann diese Methode bei der Analyse des verbalen und des nicht-verbalen Verhaltens angewandt werden. Jedoch liegt die Problematik der Zerlegung des Unterrichtsgeschehens in willkürliche Zeiteinheiten darin, daß die Ergebnisse nicht die na-

türlich eintretenden Verhaltensmuster oder den Ablauf des Verhaltens widerspiegeln, wie es sich allmählich entwickelt.

Anstatt eine beliebige Zeiteinheit anzugeben, entscheiden sich viele Forscher für Analyse-Einheiten verschiedener Art. Diese Einheiten repräsentieren einzelne Elemente des verbalen und/oder nicht-verbalen Verhaltens, die beim Kodieren verwendet werden, um Unterrichtsgeschehnisse in Einzelkomponenten zu zerlegen. Analyse-Einheiten werden von Wissenschaftlern unterschiedlich definiert, und zwar im Sinne von:

a) Aktivität im Unterricht,
b) verbaler Kommunikation zwischen zwei oder mehreren Sprechern,
c) Kommunikation oder Mitteilung eines einzelnen Sprechers,
d) Ausdruck einer Gedankeneinheit durch einen Sprecher.

Diese Analyse-Einheiten können durch folgende Beispiele veranschaulicht werden:

a) In seiner Untersuchung über Unterrichtsverhalten und mangelndes Leistungsvermögen der Schüler führte Perkins (1964, 1965) als grundlegende Analyse-Einheiten sechs Arten von Klassenaktivitäten auf: Großgruppendiskussion, Klassenvortrag, Einzel- oder Projektarbeit, die nicht allen Schülern zugewiesen ist, Bearbeitung einer allen zugewiesenen Aufgabe, Kleingruppenarbeit, mündliche Berichterstattung. Aspekte des Schüler- und Lehrerverhaltens wurden im Kontext mit diesen Aktivitätsarten kodiert.

b) In ihrer Untersuchung der logischen Dimensionen des Unterrichtsverlaufs bezeichneten B. O. Smith und Meux (1962) als fundamentale Verhaltenseinheit den verbalen Austausch zwischen zwei oder mehreren Sprechern, wobei drei Phasen zu unterscheiden sind: Anfangs- oder Eröffnungsphase, Durchführungsphase und Endphase.

c) Jackson (1965) definierte drei Arten von verbalen Äußerungen als fundamentale Einheiten zur Beschreibung der Kommunikation in der Primarstufe: Äußerungen, die sich auf (a) inhaltliche Ziele, (b) Verfahrensweisen und Regeln bei der Gruppenarbeit, (c) Mitteilungen zur Aufrechterhaltung der Disziplin und Ordnung beziehen.

d) In ihrer Untersuchung über die Lehrstrategien zur Schulung kognitiver Fähigkeiten bei Primarschülern bestimmten Taba u. a. (1964, 1966) die fundamentale Analyse-Einheit als »Gedanken-Einheit«, definiert als »Äußerung oder Serie von Äußerungen, die einen mehr oder weniger vollständigen Gedanken zum Ausdruck bringen, eine spezifische Funktion besitzen und entsprechend dem jeweiligen Niveau des Gedankens klassifiziert werden können« (1966, 134).

5. Anwendungsbereich

Die Beobachtungssysteme unterscheiden sich auch dadurch voneinander, daß sie in verschiedenem Ausmaß auf andere Forschungsvorhaben und Populationen anwendbar sind.

Einige Systeme wurden in der Absicht entwickelt, sie für verschiedene Klassen, Altersstufen, Fächer und Schülergruppen verwenden zu können. Zum Beispiel ist das System der Interaktionsanalyse von Flanders in Primar- und Sekundarstufen anwendbar, in denen Lehrer mit Schülern verbal in Interaktion treten. Gleichermaßen wurde das Verfahren zur Analyse der strukturellen und funktionellen Aspekte des Kommunikationssystems im Unterricht von Biddle und Adam bewußt für die Anwendung in Primar- und Sekundarstufen entwickelt, in denen zahlreiche Fächer unterrichtet werden.

Im Gegensatz dazu sind einige Beobachtungssysteme zur Analyse von Unterrichtsprozessen nur für bestimmte Altersstufen geeignet. Das System von Hughes (1959) zur Beschreibung der Lehrerfunktionen ist in Primarstufen anwendbar, während das System zur Analyse logischer Operationen des Unterrichtens von B. O. Smith und Meux (1962) lediglich für Klassen der Sekundarstufe in Frage kommt, in denen Fächer mit stärkerem Wissenschaftscharakter unterrichtet werden.

Andere Systeme sind auf bestimmte Fächer oder Schülergruppen beschränkt. Das System von Wright (1959) wurde zur Untersuchung des Mathematikunterrichts entwickelt, während das Verfahren von Oliver und Shaver (1966) vornehmlich zur Erforschung des Sozialkundeunterrichts (socialstudies) eingesetzt werden kann.

Einige Forscher befassen sich nur mit bestimmten Schülergruppen: Kounin (1967) mit emotional gestörten Schülern, Perkins (1964, 1965) mit leistungsschwachen Schülern und Smith und Geoffrey (1968) mit Schülern von sozioökonomisch niedrigerem Status.

Ob die Forscher versuchen sollten, umfassende multidimensionale Systeme zu konstruieren, die auf einen großen Bereich von Unterrichtssituationen anwendbar sind, oder sich darauf beschränken sollten, Instrumente für eine begrenzte Anzahl von Unterrichtssituationen zu entwickeln, ist eine umstrittene Frage, auf die später näher eingegangen wird.

Wie werden Unterrichtsbeobachtungen durchgeführt?

Registrieren und Kodieren von Verhaltensweisen

Eine Vielzahl verschiedener Registrier- und Kodiertechniken zur Fixierung von Verhaltensweisen wurde im Zusammenhang mit den oben erörterten

Beobachtungssystemen angewandt. Einige Forscher (z. B. Flanders 1965; Jackson 1965; Medley/Mitzel 1958) fordern vom Beobachter, das Verhalten so zu kodieren, wie es sich tatsächlich ereignet hat; d. h. der Beobachter, dem die Aufgabe des Kodierens zukommt, transformiert die beobachteten Ereignisse in Symbole, die gezählt und tabuliert werden können. Der Hauptvorteil beim unmittelbaren Kodieren des Verhaltens liegt darin, daß man direkten Zugang hat zu visuellen Reizen wie Mimik und Gestik der Lehrer und Schüler und zu situationsgebundenen Faktoren während des Unterrichts, die für die genaue Interpretation des beobachteten Verhaltens relevant sein können. Ein schwerwiegender Nachteil liegt darin, daß es äußerst schwierig ist, die komplexen Verhaltensweisen unmittelbar in dem Augenblick zuverlässig zu kodieren, in dem sie während des Unterrichts auftreten. Daher protokollieren viele Forscher zunächst das Verhalten während des Unterrichts und kodieren es anschließend mit Hilfe des Protokolls.

Bei der Erforschung des Unterrichtsverhaltens wurden verschiedene Arten der Berichterstattung verwendet, u. a. schriftliche Protokolle, Tonbandaufnahmen und audiovisuelle Aufzeichnungen. Hughes (1959) sammelte in ihrer Untersuchung über den Unterricht in der Primarstufe Daten in Form von schriftlichen Protokollen, die in der Hauptsache aus einem chronologischen Bericht über das verbale und nicht-verbale Lehrerverhalten im Unterricht bestanden. Zwei geübte Beobachter protokollierten gleichzeitig das Unterrichtsgeschehen; das endgültige Protokoll über eine Beobachtungsperiode bestand lediglich aus Beschreibungen, denen beide Beobachter zustimmten. Es wurde kein Versuch unternommen, das Verhalten zu kategorisieren; die Beobachter berichteten lediglich in Kurzfassung über die Aussagen und Handlungen der Lehrer. Beim Kodieren klassifizierte man später die protokollierten Daten in sieben Hauptkategorien hinsichtlich der Lehrerfunktionen.

Hughes vertritt die Auffassung, daß das Sammeln von Unterrichtsdaten in Form von schriftlichen Protokollen verschiedene Vorteile hat, z. B. ermöglicht dieses Vorgehen den Forschern, den Unterrichtsverlauf so zu fixieren, daß er hinsichtlich seiner spezifischen Qualität sowie seiner charakteristischen Merkmale untersucht werden kann; die Kontinuität des Lehrerverhaltens wird festgehalten; die Protokolle sind insofern neutral, als während der Beobachtung des Unterrichtsgeschehens keine Bewertung abgegeben wird. Diese Neutralität der schriftlichen Berichte muß jedoch ernsthaft in Frage gestellt werden, da die Protokolle über das Unterrichtsverhalten aus zweiter Hand durch den Beobachter abgefaßt werden, dessen Subjektivität der Wahrnehmung sich unvermeidbar in seinem Bericht über das Geschehen widerspiegelt.

Diese Unzulänglichkeit von schriftlichen Berichten kann weitgehend überwunden werden durch den Gebrauch von Tonbändern oder audiovisuellen Aufzeichnungen, die fortlaufende, objektive Verhaltensprotokolle liefern. In mehreren Untersuchungen zur Erforschung des verbalen Unterrichtsverhaltens (z. B. Bellack u. a. 1966; Taba u. a. 1964; Taba 1966; und B. O. Smith/Meux 1962) sind die Ereignisse mit Hilfe von Fernsehkameras aufgezeichnet worden. Häufig werden von Tonbandaufzeichnungen maschinenschriftliche Protokolle angefertigt; derjenige, dem die Aufgabe des Kodierens zufällt, hat bei der Kategorisierung des Verhaltens Zugang zu den Tonbandaufzeichnungen und den maschinenschriftlichen Protokollen. In Anbetracht der Schwierigkeit, das komplexe verbale Verhalten im Unterricht der Realität entsprechend zu beobachten und zu analysieren, haben diese Aufzeichnungen über den Unterrichtsverlauf offensichtlich Vorteile, da sie wiederholt von verschiedenen Perspektiven her analysiert werden können. Die prinzipielle Unzulänglichkeit von Tonbandaufnahmen liegt darin, daß sie keine Informationen über nicht-verbales Verhalten liefern, das als solches sehr bedeutsam sein oder als Hinweis auf die adäquate Interpretation des verbalen Verhaltens dienen kann.

Dieser Schwierigkeit kann durch die Verwendung von audiovisuellen Aufzeichnungen begegnet werden, die ein umfassendes Protokoll der Unterrichtsabläufe liefern, da sie verbales und nicht-verbales Verhalten erfassen. Kounin (1967) und Biddle und Adams (1967) haben Untersuchungen abgeschlossen, in denen Beobachtungsdaten während des regulären Unterrichts mit transportablen Videorekordern aufgezeichnet wurden. Die Hauptprobleme liegen in den damit verbundenen Kosten und den technischen Schwierigkeiten beim Gebrauch der komplizierten Apparatur. Der Leser, der am Einsatz von Videorekordern in der Unterrichtsforschung interessiert ist, wird auf den Bericht von Biddle und Adams verwiesen (1967).

Beziehungen zwischen Beobachter und Beobachteten

Ein häufig erhobener Einwand gegen Beobachtungsverfahren liegt darin, daß die Gegenwart eines Beobachters oder eines Aufnahmeinstruments derart ablenkend ist, daß das beobachtete Verhalten nicht als typisches Verhalten angesehen werden kann. Dies scheint jedoch nach den Urteilen erfahrener Forscher keine stichhaltige Kritik zu sein. Heyns und Lippitt stellen in ihrem umfangreichen Überblick über systematische Beobachtungstechniken in der sozialpsychologischen Forschung fest, daß die Wissenschaftler, die Beobachter zur Erforschung des Unterrichtsgeschehens einsetzen, die allgemeine Auffassung teilen, daß »die Beobach-

ter, wenn überhaupt, nur sehr geringe Wirkung ausüben. Dieser Überzeugung sind auch experimentell arbeitende Forscher, die in einer Vielfalt von Situationen und auf verschiedenartigen Fachgebieten arbeiten« (Heyns/Lippitt 1954, 399).

Die meisten Forscher im Bereich der Erziehungswissenschaft scheinen diese Auffassung zu teilen. Biddle und Adams (1967), deren komplizierte audiovisuelle Apparatur aus zwei Kameras und zwei Mikrophonen bestand, die in den Klassenräumen aufgestellt wurden, in denen sie ihre Beobachtungen vornahmen, kommentieren die Auswirkungen dieser Apparatur auf Schüler und Lehrer folgendermaßen:

»Alle beteiligten Lehrer wurden informell über den Effekt der Aufnahmeapparatur interviewt. Einige berichteten über gewisse Spannungsgefühle zu Beginn der ersten protokollierten Unterrichtseinheit, stellten jedoch fest, daß diese Spannung dann aufhörte, wenn sie sich im Unterricht engagierten. Die Forscher bemerkten, daß einige Lehrer sich an den Tagen besonders sorgfältig zu kleiden schienen, an denen die Aufnahmen stattfanden, obwohl ein derartiger Effekt bei den Schülern nicht festgestellt wurde, die – es soll noch einmal daran erinnert werden – nicht wußten, an welchen Tagen die Aufnahmen gemacht wurden. Gelegentlich zeigten sich deutliche Anzeichen dafür, daß die Schüler sich der Aufnahmekameras bewußt waren, da sie auf die vermuteten Kameras starrten oder in der Pause vor ihnen Theater spielten. Jedoch wurde überall bestätigt, daß die Kameras die Schüler offensichtlich kaum ablenkten« (Biddle/Adams 1967, 217).

Ähnliche Auffassungen, daß die von Beobachter und Aufnahmegerät ausgehenden Wirkungen auf das Unterrichtsverhalten als unbedeutend angesehen werden können, vertraten u. a. auch: B. O. Smith und Meux (1962), Hughes (1959), Flanders (1965) und Bellack u. a. (1966).

Einige Wissenschaftler trafen besondere Vorkehrungen, um die Wirkungen zu verringern, die von Beobachtern und von der Aufnahmeapparatur ausgehen. Zum Beispiel weisen B. O. Smith und Meux (1962) darauf hin, daß die Informationen für die mit ihnen zusammenarbeitenden Lehrer die Versicherung enthielten, daß die Untersuchung keinerlei Bewertung ihrer Lehrfähigkeit bezwecke, daß bei der Veröffentlichung der Forschungsergebnisse völlige Anonymität gewährleistet sei und daß nur Mitglieder des Forscherteams Einblick in die Tonbänder und Manuskripte hätten. Außerdem wurden die Aufnahmegeräte einige Tage vor der Unterrichtsstunde, in der die Beobachtungen stattfanden, im Klassenraum installiert, damit sich Schüler und Lehrer daran gewöhnen konnten.

Ohne die Probleme zu unterschätzen, die durch Beobachter und Aufnahmegerät entstehen, scheint es, daß diese Einflüsse zum großen Teil

ausgeschaltet werden können, wenn entsprechende Vorsichtsmaßnahmen getroffen werden. Da jedoch nicht angenommen werden kann, daß Lehrer und Schüler durch die Anwesenheit eines Beobachters und durch das Vorhandensein eines Aufnahmegerätes trotz derartiger Maßnahmen völlig unbeeinflußt bleiben, sollte man sich an die einfache, offensichtliche Tatsache erinnern, daß es besser ist, »etwas über das Lehrer-Schüler-Verhalten in Erfahrung zu bringen, während sie unter Beobachtung stehen, als überhaupt keine Kenntnisse über das Lehrer-Schüler-Verhalten gewinnen zu können« (Medley/Mitzel 1963, 248).

Die Zuverlässigkeit beim Kodieren

Die Zuverlässigkeit eines Beobachtungsinstruments für das Unterrichtsgeschehen kann definiert werden als »das Ausmaß, in dem die Messung, unter konstanten Bedingungen wiederholt, konstante Ergebnisse erbringt« (Kaplan 1964, 200). Unter den relevanten Bedingungen bei der Verhaltensmessung während des Unterrichts sind von besonderer Bedeutung die Beobachter, die die Messungen vornehmen, sowie die Stabilität des zu messenden Verhaltens. Daher kann die Zuverlässigkeit von Instrumenten zur Beobachtung von Unterrichtsabläufen geschätzt werden durch den Grad der Übereinstimmung zwischen voneinander unabhängigen Beobachtern (Übereinstimmungskoeffizient) und auf Grund der Stabilität der Verhaltensdimensionen unter Beobachtung (Stabilitätskoeffizient). Obwohl beide Arten der Zuverlässigkeit offensichtlich von Bedeutung sind, wird der Stabilitätskoeffizient bei der Erforschung des Unterrichtsverhaltens gegenwärtig kaum berechnet; größere Aufmerksamkeit wird der Schätzung des Ausmaßes an Übereinstimmung zwischen den Beobachtern geschenkt. Die gebräuchlicheren statistischen Indizes für diese Schätzungen sind der Prozentsatz der Übereinstimmung zwischen den kodierenden Personen und der Korrelationskoeffizient; von einigen Wissenschaftlern wird auch die Varianzanalyse angewandt.

Zur Sicherung einer adäquaten Zuverlässigkeit beim Gebrauch von Beobachtungsskalen haben Wissenschaftler bestimmte Verfahrensweisen entwickelt. Von zentraler Bedeutung ist die sorgfältige Entwicklung der Beobachtungsskala selbst, indem man besonders auf die präzise Definition der Kategorien und der Analyse-Einheit und auf die Formulierung von Kodier-Regeln zur Anleitung der kodierenden Personen achtet. Wahrscheinlich resultieren die größten Schwierigkeiten beim zuverlässigen Kodieren aus der unklaren Definition der zu kodierenden Verhaltenseinheit; ohne eine derartige präzise Definition kann kein hoher Grad an Übereinstimmung zwischen den Beobachtern erzielt werden. Die sorgfältige Schu-

lung der Beobachter ist zweifellos ein wesentliches Mittel zur Sicherung der Zuverlässigkeit; die Erfahrung der Forscher bestätigt die Wichtigkeit intensiver Schulungsprogramme für die beobachtenden und kodierenden Personen und für die Notwendigkeit eines systematischen Vorgehens zur Kontrolle der Beobachtungen und Kodierungen.

Statistische Analyse der Daten

Eine detaillierte Diskussion über die Methoden einer statistischen Analyse der Beobachtungsdaten aus dem Unterrichtsgeschehen geht über den Rahmen dieses Beitrags hinaus. Dennoch ist darauf hinzuweisen, daß neuere Theorien und Techniken es ermöglichen, bestimmte Dimensionen von Unterrichtsprozessen zu erforschen, die zuvor nur schwer, wenn überhaupt, statistisch zu erfassen waren. Biddle und Adams (1967) und Bellack u. a. (1966) haben z. B. die zeitliche Verlaufscharakteristik von Ereignissen im Unterrichtsgeschehen statistisch mit Hilfe einer Markoff-Kette beschrieben. In der letzteren Untersuchung versuchten die Wissenschaftler festzustellen, ob bestimmte pädagogische Verhaltensmuster dahin tendieren, andere pädagogische Verhaltensmuster nach sich zu ziehen. Dies wurde statistisch durch die Markoff-Kette beschrieben, mittels derer es möglich ist, die bedingte Wahrscheinlichkeit des Übergangs von einem Verhaltensmuster auf ein anderes zu bestimmen (Kemeny/Snell 1962). Nimmt man die verschiedenen pädagogischen Verhaltensmuster als Einheiten an, dann wurde die Wahrscheinlichkeit des Übergangs von einem Verhaltensmuster zu einem anderen erforscht, um festzustellen, ob ein pädagogisches Verhaltensmuster die Tendenz hat, das jeweils unmittelbar nachfolgende Verhalten zu beeinflussen.

Die statistische Auswertung von Beobachtungsdaten erfuhr eine wesentliche Erleichterung durch den Einsatz von Computern. Gage (1969) berichtete über wichtige Ergebnisse aus der Arbeit von Allen und Snow am Stanford Center for Research and Development in Teaching, die durch den Einsatz von Computern ermöglicht wurden. Allen und Snow entwickelten eine Taxonomie von Verhaltensweisen im Unterricht, indem sie Computer zum Speichern und zum Reproduzieren von Items verwendeten, die sich auf Lehrer- und Schülerverhalten beziehen. Gage beschreibt die Ziele und das Verfahren dieses Projektes:

»Es wurden mehr als 1000 Items gespeichert; diese können vom Computer in einer endgültigen Fassung reproduziert und gedruckt werden, so daß sie den Beobachtern, Bewertern und Wissenschaftlern für die Inhaltsanalyse zur Verfügung stehen. Die Items, die reproduziert werden sollen, können entsprechend vielen differierenden Dimensionen exakt angegeben werden. Die von den Beobachtern oder

den für die Inhaltsanalyse verantwortlichen Wissenschaftlern gewonnenen Daten, die die Häufigkeit, Intensität oder die Verhaltenskorrelate betreffen, die durch diese Items gekennzeichnet werden, können zusammen mit den Items im Computer gespeichert werden. Auf diese Weise werden die Erfahrungswerte gesammelt, die für die Gültigkeit und Zuverlässigkeit der Items bei verschiedenen Lehrertypen, Beobachtern, Lehrinhalten, Effektivitätskriterien und dergleichen relevant sind. Kurz, das Ideal einer universalen Taxonomie des Unterrichtsverhaltens, die für viele Zwecke nützlich ist, kann durch ein auf dem Computer basierendes System – wie in Stanford – erreicht werden« (Gage, 1969, 1452).

II. Beispiele für Beobachtungssysteme

In diesem Abschnitt werden kurz drei Analyse-Systeme beschrieben, die drei Haupttypen von vorhandenen Beobachtungssystemen charakterisieren: (1) Das auf das affektive Geschehen zielende System der Flandersschen Interaktionsanalyse, das sich mit dem sozialen Klima im Klassenraum befaßt (1965); (2) das System von Bellack u. a. zur Analyse des Sprachverhaltens; es enthält vor allem kognitive Dimensionen des Unterrichts (1966); (3) das System von Oliver und Shaver zur Beschreibung von Erziehungsstilen im Sozialkunde-Unterricht, ein multidimensionales System, das sich auf affektive und kognitive Aspekte der Aktivitäten im Unterricht konzentriert (1966).

Ein System zur Analyse affektiver Dimensionen des Unterrichts

Konzeptueller Rahmen: Mit Flanders' System zur Interaktionsanalyse können vornehmlich affektive und interpersonelle Komponenten der Unterrichtsprozesse erfaßt werden. Flanders behauptet, daß die Art und Weise, in der der Lehrer durch seine verbale Kommunikation das Verhalten von Schülern zu beeinflussen sucht, das wichtigste Kennzeichen der Lehrer-Schüler-Beziehung im Unterricht ist. Seine Bemühungen sind darauf gerichtet, zwei vom Lehrer angewandte kontrastierende Beeinflussungsmethoden zu analysieren:

(a) Direkte Beeinflussung, die »aus den verbalen Äußerungen des Lehrers besteht und die die Handlungsfreiheit einschränken, indem entweder die Aufmerksamkeit auf ein Problem gerichtet wird oder der Lehrer sich auf seine Autorität beruft oder beides zusammen erfolgt.

(b) Indirekte Beeinflussung, die sich in solchen verbalen Äußerungen des Lehrers zeigt, die die Handlungsfreiheit des Schülers durch Ermutigung zur verbalen Teilnahme und Initiative am Unterricht vergrößern« (Flanders 1965, 9).

Kategorien und Methode der Analyse: Das Kategoriensystem von Flanders, das für Situationen im Unterricht entwickelt wurde, in denen Lehrer und Schüler aktiv in verbaler Interaktion stehen, beschreibt verbale Handlungen des Lehrers, die direkten und indirekten Einfluß ausüben. Das Beobachtungsschema besteht aus den folgenden zehn Items:

Flanders Kategorien der Interaktionsanalyse

Lehrer-äußerungen	Indirekte Beeinflussung	1. akzeptiert Gefühle 2. lobt oder ermutigt 3. akzeptiert oder bewertet Ideen der Schüler 4. stellt Fragen
	Direkte Beeinflussung	5. trägt etwas vor 6. gibt Anweisungen 7. kritisiert oder rechtfertigt seine Autorität
Schüleräußerungen		8. Schüleräußerungen – Antwort 9. Schüleräußerungen – Eigeninitiative
		10. Ruhe oder Unruhe

Sieben Kategorien sind zur Klassifikation der Lehreräußerungen, zwei zur Klassifikation der Schüleräußerungen vorgesehen. In die zehnte Kategorie fallen Pausen, kurze Ruheperioden oder Phasen der Unruhe. Von den sieben auf die Lehreräußerungen bezogenen Kategorien repräsentieren die Items 1–4 indirekte, die Items 5–7 direkte Einflußnahme.

Die Daten werden von einem Beobachter während des Unterrichts gewonnen. Am Ende jeder 3-Sekunden-Periode entscheidet der Beobachter, welche der zehn Kategorien die Kommunikationsereignisse dieser Periode am besten beschreibt. Er notiert die Numerierungen der jeweiligen Kategorien, während er gleichzeitig die Kommunikation in der nächsten 3-Sekunden-Periode feststellt. Insgesamt erhebt er also ca. 20 Beobachtungen pro Minute. Seine Notierungen bestehen folglich aus einer von oben nach unten in einer Kolonne geschriebenen Sequenz von Zahlen, so daß die ursprüngliche Sequenz von Ereignissen beibehalten bleibt. Wenn sich eine Veränderung im Unterrichtsablauf ergibt, zieht der Beobachter einen Doppelstrich und vermerkt die Zeit. Handlungsperioden auf diese Weise festzustellen ist somit eine zweite Art der Kategorisierung, die zur Ergänzung des Systems zur Klassifizierung verbaler Aussagen

dient. Flanders nennt fünf Handlungsarten: Einführung von neuem Material, Bewertung von Hausaufgaben oder Testen, Klassendiskussion, Beaufsichtigung und Förderung individueller Unterrichtsarbeit, sonstige übliche Schularbeit.

Nach einer bestimmten Beobachtungsperiode überträgt der Beobachter die Numerierungen, die er in Zeitreihen registriert hat, auf eine 10 x 10-Matrix. Die Numerierungen sind in der Matrix jeweils paarweise registriert. Wenn z. B. die Reihenfolge der Numerierungen auf Grund der Beobachtungen lautet: 10, 6, 7, 5, 4, 8, 1, 4 8, besteht das erste Paar aus den Numerierungen 10–6, das zweite Paar lautet 6–7, das dritte 7–5 etc. Für das erste Paar, 10–6, wird die Registrierung in die Zelle der Reihe 10, Spalte 6 eingetragen, für das zweite Paar, 6–7, in die Zelle der Reihe 6, Spalte 7. Diese Art der Analyse hilft bis zu einem gewissen Grad, die Sequenz der Ereignisse im Unterrichtsablauf festzuhalten. Die Analyse der Matrixdaten liefert verschiedene Informationen über den Interaktionsprozeß im Unterricht: Der Umfang der Lehrer- und Schüleräußerungen, die Zeitspanne, die in Ruhe oder Unruhe verbracht wurde, das Verhältnis von indirektem und direktem Einfluß, den der Lehrer ausübt (I/D-Quotient), die Sequenzen der verschiedenen Arten von Lehrer- und Schüleräußerungen u. ä.

Zuverlässigkeit: Flanders und seine Mitarbeiter entwickelten systematische Verfahrensweisen zur Schulung der Beobachter, um die Zuverlässigkeit des Kodierens zu sichern. Die Zuverlässigkeit der Beobachtungen, die sich durch die Übereinstimmung der Ergebnisse der einzelnen Beobachter ergibt, wird mit Hilfe des Scott-Koeffizienten geschätzt. In einer von Flanders (1965) berichteten neueren Untersuchung lagen die Reliabilitätskoeffizienten von Scott für geschulte Beobachter ausnahmslos über .86.

Anwendungsbereich: Das System der Interaktionsanalyse wurde häufig bei einer Vielfalt von deskriptiven und experimentellen Untersuchungen angewandt, kürzlich auch bei der Lehrerausbildung und -fortbildung, um den Lehrern eine Rückmeldung über ihr eigenes Verhalten im Unterricht zu ermöglichen.

Ein System zur Analyse kognitiver Dimensionen des Unterrichts

Konzeptueller Rahmen: In ihrer Untersuchung »The Language of the Classroom« (1966) konzipierten Bellack und seine Mitarbeiter den Unterrichtsablauf als eine Art Sprachinteraktion (language game), wobei sie Wittgensteins Auffassung von der Sprache als einem durch Regeln gelenkten verbalen Verhalten übernahmen. Die Grundeinheit zur Beschreibung dieser Interaktionen im Unterricht stellt der pädagogische Impuls

(paedagocial move) dar. Diese Impulse werden entsprechend den pädagogischen Funktionen, die sie in der Unterrichtsdiskussion erfüllen, in vier Hauptkategorien eingeteilt:

Strukturieren: Strukturierende Impulse dienen den pädagogischen Funktionen der Setzung des Kontexts für nachfolgendes Verhalten bei entweder erfolgreich beginnender oder schleppend anlaufender Interaktion zwischen Schülern und Lehrern. Die Lehrer beginnen z. B. häufig eine Unterrichtsstunde mit einem strukturierenden Impuls, mit dem sie die Aufmerksamkeit auf das während dieser Stunde zu diskutierende Thema oder Problem lenken wollen.

Auffordern: Pädagogische Impulse in dieser Kategorie sollen verbale Antworten hervorlocken, um die Schüler, von denen Aufmerksamkeit oder eine physische Reaktion erwartet wird, zu ermutigen. Alle Fragen, Befehle, Anweisungen und Bitten sind zu den Aufforderungen zu zählen.

Antworten: Diese pädagogischen Impulse stehen in einer reziproken Beziehung zu den Aufforderungen und treten nur im Zusammenhang mit ihnen auf. Ihre pädagogische Funktion liegt darin, die durch die Aufforderungen ausgedrückten Erwartungen zu erfüllen; so werden Schülerantworten auf Lehrerfragen als Antworten auf die Aufforderungen klassifiziert.

Reagieren: Diese pädagogischen Impulse werden durch ein strukturierendes, aufforderndes, antwortendes oder der Reaktion vorausgehendes Verhalten veranlaßt, sind aber nicht direkt durch sie verursacht. Aus pädagogischer Sicht dienen diese Verhaltensweisen der Modifikation (durch Klärung, Synthetisierung oder Erweiterung) und/oder der Bewertung (positiv oder negativ) dessen, was vorher gesagt wurde. Reaktionen unterscheiden sich von Antworten folgendermaßen: Während eine Antwort immer direkt durch eine Aufforderung hervorgerufen wird, werden die Reaktionen durch vorausgehende pädagogische Impulse lediglich verursacht; z. B. ist die Bewertung einer Schülerantwort durch den Lehrer als eine Reaktion gekennzeichnet.

Die das verbale Lehrer- und Schülerverhalten beschreibenden Impulse treten im Unterrichtsablauf in bestimmten periodischen Verhaltensmustern oder Kombinationen auf, die als Unterrichtszyklen (teaching cycles) bezeichnet werden. Ein typischer Unterrichtszyklus besteht z. B. aus der Aufforderung eines Lehrers, auf die der Schüler mit einer Antwort reagiert, auf die wiederum eine Lehrerreaktion folgt. Der sequentielle Ablauf der Unterrichtszyklen wird durch eine Markoff-Kette beschrieben.

Die Analyse der sprachlichen Interaktionen im Unterricht wäre ohne Beschreibung des Inhalts der pädagogischen Impulse unvollständig. Zwei grundlegende Arten des Inhalts können unterschieden werden:

(a) Inhalte mit fachspezifischer Bedeutung, die sich auf den zur Diskussion stehenden Lehrstoff beziehen (in dieser Untersuchung: Internationaler Handel);

(b) Inhalte mit unterrichtsspezifischer Bedeutung, die sich auf die zugewiesene Aufgabe, auf die Materialien und Verfahrensweisen im Unterrichtsgeschehen beziehen.

Die Inhalte mit fach- und unterrichtsspezifischer Bedeutung werden zusammen mit ihren assoziierten logischen Bedeutungen beobachtet und registriert, die sich auf die mit den fachspezifischen und unterrichtsspezifischen Inhalten – wie definieren, interpretieren, Fakten darlegen, erklären, Meinungen äußern und begründen – in Verbindung stehenden kognitiven Prozesse beziehen. Wenn ein Schüler also die Fragen eines Lehrers beantwortet, indem er eine Definition von »Preis« gibt, ist der Impuls des Schülers als eine *Antwort* zu kodieren, die fachspezifische Bedeutung als *Preis*, und der logische Prozeß als *definieren*. Zusätzlich wird der Schüler als der Sprecher gekennzeichnet und die Länge seiner Antwort nach der Anzahl der Schreibmaschinenzeilen gemessen.

Kodierungsverfahren: Das Kodieren erfolgt nach Gesichtspunkten des Beobachters, der den pädagogischen Bedeutungsgehalt dem verbalen Verhalten des Redners entnimmt. Die mit der Kodierung beauftragten Personen hören die Bandaufnahmen ab, die vom Unterrichtsgeschehen aufgenommen wurden, und verwerten ebenso die Niederschriften der Tonbandaufnahmen. Jeder pädagogische Impuls wird folgenden Analysekategorien entsprechend kodiert:

(1) Sprecher
(2) Art der pädagogischen Impulse
(3) fachspezifische Bedeutung
(4) fachspezifisch-logische Bedeutung
(5) Antwortenlänge bei (3) und (4)
(6) unterrichtsspezifische Bedeutung
(7) unterrichtsspezifisch-logische Bedeutung
(8) Antwortenlänge bei (6) und (7)

Das folgende Beispiel stellt einen kodierten pädagogischen Impuls dar:

L	/	STR	/	IE	/	ERK	/	4	/	V	/	F	/	2
(1)		(2)		(3)		(4)		(5)		(6)		(7)		(8)

Diese Angaben werden wie folgt interpretiert: Ein *Lehrer* (1) gibt einen *strukturierenden* (2) Hinweis, indem er etwas über *Import* und *Export* (3) in einer Länge von *vier* (5) Schreibmaschinenzeilen *erklärt* (4); ferner informiert er die Klasse *faktisch* (7) darüber, wie sie im weiteren Verlauf des Unterrichtsgeschehens *verfahren* (6) soll, und zwar in einer Länge von *zwei* (8) Zeilen. Im folgenden wird eine kurze Zusammenfas-

sung der Analysekategorien gegeben, wobei die acht Numerierungen mit den oben genannten übereinstimmen:

1. Sprecher
   - Lehrer (L)
   - Schüler (S)
   - audiovisuelles Gerät (A)
2. Art des pädagogischen Impulses
   - strukturieren (STR)
   - auffordern (AUF)
   - antworten (ANT)
   - reagieren (REA)
3. Fachspezifische Bedeutungen (Lehrinhalt – in dieser Untersuchung: Internationaler Handel; die Kategorien basieren auf einer Inhaltsanalyse eines Textes über den internationalen Handel, der in den an der Untersuchung teilnehmenden Klassen zugrunde gelegt wurde)
   - Handel (HA)
   - Produktions- und/oder Spezialisierungsfaktoren (PSF)
   - Importe und/oder Exporte (IE)
   - ausländische Investitionen (AI)
   - Handelsbarrieren (HB)
   - Förderung des freien Handels (FFH)
   - Kriterien, die für den Handel von Bedeutung sind (KHB)
   - Gründe, die gegen den Handel sprechen (GH)
4. Fachspezifisch-logische Bedeutungen (kognitive Prozesse, die bei der Behandlung von Lehrinhalten miteingeschlossen sind)
   - Analytische Prozesse
     - definieren (DEF)
     - interpretieren (INT)
   - Empirische Prozesse
     - formulieren von Faktoren (F)
     - erklären (ERK)
   - Evaluative Prozesse
     - Meinungen äußern (MÄ)
     - begründen (BEG)
5. Antwortenlänge in Schreibmaschinenzeilen bei 3. und 4.
6. Unterrichtsspezifische Bedeutungen (Faktoren, die auf die Klassenführung bezogen sind)
   - Zuweisung einer Aufgabe (ZA)
   - Material (M)
   - Person (P)
   - Verfahren (V)

Aussage (A)
logischer Prozeß (LP)
Sprachtechnik (ST)
allgemeine Handlungen (AH)
verbale Handlungen (VH)
physische Handlungen (PH)
kognitive Handlungen (KH)
emotionale Handlungen (EH)

7. Unterrichtsspezifisch-logische Bedeutungen (kognitive Prozesse, die mit unterrichtsspezifischen Bedeutungen assoziiert sind)
   analytische, empirische, evaluative Prozesse (wie bei 4.)
   Einschätzen (Bezug zur Metakommunikation, gewöhnlich eine evaluative Reaktion)
      positiv (POS)
      zustimmend (ZU)
      wiederholend (WDH)
      genau beschreibend (GB)
      nicht zustimmend (NZU)
      negativ (NEG)
   außer-logische Prozesse
      ausführen (AUS)
      anleiten (ANL)
8. Antwortenlänge in Schreibmaschinenzeilen bei 6. und 7.

Zuverlässigkeit: Um die Zuverlässigkeit des Kodierens zu schätzen, kodierten zwei Gruppen des Forscherteams Stichproben von Protokollen, die nach dem Zufallsprinzip ausgewählt worden waren, und verglichen die Ergebnisse miteinander. Der Prozentsatz der Übereinstimmung wurde für jede der fundamentalen Kategorien des Analyse-Systems berechnet, und zwar hinsichtlich der Anzahl und Dauer der pädagogischen Impulse. Das Ergebnis erbrachte einen übereinstimmend hohen Zuverlässigkeitsgrad für alle Kategorien. Die Übereinstimmung lag zwischen .84 und .96.

Anwendungsbereich: Mit geeigneten Modifikationen in den Kategorien, die als »fachspezifische Bedeutung« klassifiziert sind, liefert dieses Verfahren eine nützliche Technik zur Erforschung des Unterrichtsverhaltens in einer Vielzahl von Fächern der Primar- und Sekundarstufe. Vor kurzem wurden am »Teachers College« Dissertationen beendet, in denen diese Methode der Analyse zur Erforschung des Mathematikunterrichts in der Sekundarstufe und des Leseunterrichts in der Primarstufe verwendet wurde.

Ein System zur Analyse affektiver und kognitiver Dimensionen des Unterrichts

Konzeptueller Rahmen: Im Zusammenhang mit ihrer Untersuchung »Teaching Public Issues in the High School« (1966) entwickelten Oliver und Shaver ein Beobachtungssystem zur Analyse von zwei Unterrichtsstilen: Lehrervortrag und sokratisches Gespräch. Der theoretische Rahmen, in dem diese beiden Unterrichtsweisen definiert sind, beinhaltet drei Dimensionen der Interaktion:

a) kognitive Dimension,
b) affektive oder sozial-emotionale Dimension,
c) Verfahrensdimension.

Was die kognitiven Aspekte betrifft, »tendiert der Lehrervortrag zur Darstellung. Hierbei wird angenommen, daß der Tatbestand eindeutig dargestellt werden kann und daß der Lehrer nur Informationen vorzutragen und zu erklären hat oder eine analytische Struktur vorgibt, durch die die Informationen gegliedert werden können« (Oliver/Shaver 1966, 178). Im Gegensatz dazu »ist der sokratische Stil zweifellos dialektisch. Er setzt voraus, daß das Problem nur in einem entgegengesetzte Aspekte beinhaltenden Kontext geklärt werden kann, in dem verschiedene Standpunkte dargestellt und verteidigt werden« (a. a. O., 178). Hinsichtlich der affektiven Aspekte der Interaktion ist es naheliegend, daß sowohl der sokratisch arbeitende als auch der vortragende Lehrer den Schüler unterstützen muß. Da jedoch »die offene Kontroverse auf dem kognitiven Niveau in den affektiven Bereich übergeht, tendiert die sokratische Methode dahin, daß eine Diskussion stark mit negativen Affekten aufgeladen wird« (a. a. O., 179). Da der Lehrer jedoch im Zentrum des Dialogs mit den Schülern steht, werden keine systematischen Unterschiede zwischen vortragenden und sokratisch unterrichtenden Lehrern in dieser Dimension erwartet.

Kategorien für die Analyse: Oliver und Shaver entwickelten eine Reihe von Kategorien, die für die oben beschriebenen Unterrichtsstile des Lehrervortrags und des sokratischen Gesprächs bedeutsam sind. Ihr Untersuchungsinstrument besteht aus den folgenden drei Gruppen von Kategorien:

Affektive und sozial-emotionale Kategorien

1. Solidarität
2. geringer positiver Affekt
3. Entspannung
4. Spannung
5. geringer negativer Affekt

6. Widerspruch
   neutral (keine affektive Mitteilung)

Kognitive Kategorien

7. Hinweis auf Inkonsistenz
8. Beschreibung
9. Evaluation
10. Wiederholung, Zusammenfassung, Herausstellung des Wesentlichen
11. Klärung
12. Hinweis auf Analogien
    nicht-kognitiv

Verfahrenskategorien

13. Anordnung aufgabenorientierten Verhaltens
14. Kontrolle abweichenden Verhaltens

Der Beobachter, der dieses Instrument benutzt, muß auf die kognitive oder sonstige für den Unterrichtsverlauf relevante Bedeutung in jeder Aussage oder in jeder Gedankeneinheit schließen, die vom Lehrer zum Ausdruck gebracht wird. Jede derartige einstufbare Handlung wird in einer kognitiven oder Verfahrenskategorie und in einer affektiven Kategorie eingetragen. Die Funktionen der kognitiven Kategorien werden folgendermaßen beschrieben: »Die primäre Funktion der kognitiven Kategorien (7–12) liegt darin, Fragen über Unterschiede in bezug auf den intellektuellen oder logischen Inhalt in den nach den beiden verschiedenen Stilen stattfindenden Unterrichtsabläufen zu beantworten. Von zentraler Bedeutung für die Unterscheidung zwischen den beiden Stilen ist das Ausmaß, in dem sich die Lehrer mit *deskriptiver Information* bei kontroversen Fragen im Verhältnis zu *Werturteilen* befassen, die über diese Fragen abgegeben werden. Die Kategorien 8 und 9 sind speziell dafür vorgesehen, Unterschiede dieser Art aufzudecken. Die Kategorie 7 soll die Versuche des Lehrers zum Ausdruck bringen, bei den Schülern persönliche Wertkonflikte dadurch hervorzurufen, daß er den Eindruck erweckt, gegensätzliche Urteile in gleichen Situationen abzugeben (das kontroverse Problem und eine Analogie). Die Kategorie 12 (Analogie) hat so eine offensichtliche Bedeutung. Die Kategorien 10 und 11 wurden miteinbezogen, um das kognitive Subsystem zu vervollständigen« (Oliver/Shaver 1966, 291–292).

Zuverlässigkeit: Die Beobachter wurden anhand von Tonbandaufnahmen der Klassendiskussionen sorgfältig in der Anwendung des Verfahrens geschult. Eine Bearbeitung des Chi-Quadrat-Verfahrens wurde zur Schätzung der Übereinstimmung zwischen den Beobachtern verwendet. Alle Chi-Quadrat-Werte lagen deutlich unter dem 5 %-Niveau, das die Wissenschaftler als Kriterium für die Untersuchung angenommen hatten.

Anwendungsbereich: Obwohl dieses analytische Verfahren zur Verwendung im Fach Sozialkunde entwickelt wurde, in dem öffentliche Probleme im Zentrum der Diskussion stehen [2], nehmen die Autoren an, daß es ebenso zur Analyse des Unterrichtsgeschehens in anderen Fächern, z. B. im naturwissenschaftlichen und Literaturunterricht, verwendet werden kann.

III. Probleme und Konsequenzen in der Entwicklung und im Gebrauch von Systemen zur Analyse des Lehrer- und Schülerverhaltens

In diesem Abschnitt wendet sich die Diskussion zwei Problemen und Konsequenzen zu, die den Erziehern bei der Entwicklung und Anwendung von Beobachtungssystemen im Unterricht begegnen.

Beziehungen zwischen den Beobachtungssystemen

Das erste Problem erwächst aus dem Bestreben der Forscher, ein neues Analyse-System für nahezu jede neue Untersuchung zu entwickeln, wohl in der Annahme, die von anderen entwickelten Verfahren seien ungeeignet oder inadäquat. Es mag in der Tat zutreffen, daß Aspekte untersucht werden sollen, für die eine Beobachtungsskala noch nicht entwickelt wurde, oder daß die vorhandenen Systeme nur eine geringe Gültigkeit und Zuverlässigkeit besitzen. Die Entwicklung zahlreicher verschiedener Beobachtungssysteme hat zur Folge, daß den Wissenschaftlern nur sehr wenige Daten zur Verfügung stehen, die einen Vergleich verschiedener Untersuchungen zulassen.

Die umfangreiche Konstruktion von Beobachtungssystemen ist zweifellos das Ergebnis unterschiedlicher Konzeptionen der Forscher. In diesem Zusammenhang sei daran erinnert, daß »in den frühen Stadien der Entwicklung jeder Wissenschaft verschiedene Menschen, die mit dem gleichen Bereich von Phänomenen, nicht aber mit denselben Einzelphänomenen konfrontiert sind, diese unterschiedlich beschreiben und interpretieren« (Kuhn 1962, 17). So kann von den Wissenschaftlern auf dem Gebiet der Unterrichtsforschung – einem Sektor, der sich gerade als wissenschaftlicher Forschungsbereich abzuzeichnen beginnt – nicht erwartet werden, daß sie ihren Forschungen ein gemeinsames theoretisches Konzept zugrunde legen. In diesem Stadium der Entwicklung sind in Widerspruch stehende und konkurrierende Ansätze zur Beschreibung von Unterrichtsabläufen zu erwarten und sogar zu unterstützen.

Jedoch schließt dieser Stand der Dinge nicht die Möglichkeit von Un-

tersuchungen zur Bestimmung von Faktoren oder Dimensionen aus, die den verschiedenen Beobachtungssystemen zugrunde liegen, so daß Differenzen und Ähnlichkeiten zwischen ihnen klarer unterschieden werden können. Ein faktorenanalytisches Verfahren, das dem Vergleich von Beobachtungssystemen zum Zweck der Feststellung von Unterschieden zwischen den ihnen zugrunde liegenden Dimensionen dient, wurde von Gage vorgeschlagen (1969, 1450):

»Die Verhaltensweisen von Lehrern in einer großen Stichprobe können aufgrund vieler Variablen solcher Art, wie sie von Flanders (1965), Smith u. a. (1962), Bellack u. a. (1966), Medley und Mitzel (1958) u. a. entwickelt wurden, erfaßt werden. Dann können die Interkorrelationen der aus den Variablen gewonnenen Punktwerte der Faktorenanalyse unterzogen werden. Die ermittelten Faktoren können die Dimensionen in relativ knapper oder einfach strukturierter Form definieren.«

Wenn vergleichende empirische Untersuchungen von vorhandenen Systemen verfügbar werden, wird es möglich sein, viele Aspekte berücksichtigende Beobachtungsskalen zu konstruieren, die Konzepte einer Vielfalt verschiedener Systeme in sich vereinen. Meux (1967) hat z. B. die Entwicklung eines neuen multidimensionalen Systems vorgeschlagen, um Einheiten mit lernzielorientierten Inhalten und innerhalb dieser Einheiten Perioden, in denen initiiert und reagiert wird, sowie fachspezifische und unterrichtsspezifische Aspekte, Gruppenprozesse und Denkprozesse einzubeziehen. Er ist der Auffassung, daß Unterrichtsstrategien konstruiert werden können, die weit effektiver als die besten der zur Zeit von unseren Lehrern verwendeten sind, wenn Kategorien wie die oben genannten von verschiedenen Beobachtungssystemen in dem vorgeschlagenen multidimensionalen System kombiniert werden. »Diese Möglichkeit neuer Kombinationen von gegenwärtigen Praktiken ist von einigen Autoren übersehen worden, die der Auffassung sind, daß Beobachtungsverfahren uns prinzipiell niemals über die gegenwärtige Praxis hinausbringen werden« (Meux 1967, 549–550).

Mit ähnlichen Zielsetzungen hat Biddle (1967) eine andere Variante eines Idealsystems entworfen, das drei umfassende Konzepte miteinbeziehen soll:

(a) Übergreifende Konzepte für das Verhalten im Unterricht, wobei strukturelle und funktionelle Aspekte des sozialen Systems – dargestellt durch den Klassenverband – mitberücksichtigt werden sollen;

(b) Konzepte für die beobachtbaren, sich gegenseitig beeinflussenden Handlungen der Unterrichtsteilnehmer;

(c) Konzepte für den Sprachgebrauch in der Klasse einschließlich Syntax und Phonologie.

Ob in diesem Entwicklungsstadium der Unterrichtsforschung die Wissenschaftler ihr Augenmerk mehr auf die Konstruktion von multidimensionalen Verfahren wie die von Meux und Biddle vorgeschlagenen richten sollen, oder ob das primäre Bemühen auf die Konstruktion von Beobachtungssystemen von begrenztem Rahmen konzentriert werden soll, ist eine umstrittene Frage. Es ist jedoch wenig sinnvoll, wenn Forscher sich auf das eine oder andere Vorgehen beschränken. In Anbetracht unseres begrenzten Wissens über den Unterrichtsprozeß könnte man vernünftigerweise erwarten, daß beide Ansätze zu einem besseren Verständnis der Unterrichtsprozesse beitragen.

Mögliche Anwendungsbereiche für Beobachtungsverfahren

Die Beobachtung von Lehrer- und Schülerverhalten ist zu einer anerkannten Methode für die Sammlung von Daten in der Unterrichtsforschung geworden. Die mit Hilfe der Beobachtungstechniken gewonnenen Daten geben Anlaß zu der Hoffnung, daß sich unser Verständnis für das komplexe Leben im Klassenraum vertieft. Der zunehmende Gebrauch von Beobachtungstechniken in der Unterrichtsforschung hat Pädagogen, die in der Lehrerbildung und Schulaufsicht tätig sind, zu dem Vorschlag angeregt, die gleichen Beobachtungsverfahren für die Lehrerausbildung und -fortbildung einzusetzen. Dieser Vorschlag wirft mehrere Probleme auf, die wenigstens kurz erörtert werden sollen.

Für den Einsatz von Beobachtungssystemen in der Lehrerausbildung und -fortbildung wird von Simon und Boyer (1967) folgende Begründung gegeben:

»Zunächst stellen die Systeme einen Spiegel für den Lehrer dar, um eine Rückmeldung über sein eigenes Lehrverhalten in den im angewandten System berücksichtigten Dimensionen zu erhalten. Dieses Feedback stellt für die Lehrer eine günstige Gelegenheit dar, ihr Verhalten aufgrund der Informationen darüber, wie sie sich im Unterricht verhalten, selbst zu modifizieren. Ferner – und dies ist eventuell ein noch wichtigerer Aspekt – sind viele dieser Systeme aufgrund einer theoretischen Dimension konstruiert worden, die Verhaltensweisen beinhaltet, die bei der Verwendung im Unterricht eine Hilfe zur Förderung der Entwicklung der Schüler darstellen sollen, wie es zur Zeit in den Schulen in der Weise noch nicht der Fall ist. Wenn ein Lehrer eines dieser Systeme anwendet, erhält er ein Feedback über die Verhaltensweisen, die er *nicht* zeigt, wie auch über die, die er zeigt. Dem Lehrer bietet sich somit eine Möglichkeit, neue Verhaltensweisen zu lernen und auf diese Weise sein Verhaltensrepertoire mit Hilfe von Methoden zu erweitern, die den Lehrern im allgemeinen nicht zur Verfügung stehen« (Simon/Boyer 1967, 19–20).

Kaum einer würde die Wichtigkeit konzeptueller Instrumente für die

Ausbildung und Fortbildung von Lehrern leugnen, die sie befähigen, den Unterrichtsprozeß zu analysieren und auf diese Weise das Verständnis für ihre Aufgaben als Lehrer zu vertiefen. Wenn die verschiedenen Analyseverfahren, die im Zusammenhang mit Forschungsprojekten entwickelt wurden, den Lehrern als Instrumente vorgestellt werden, mit denen sie den Unterricht von verschiedenen Perspektiven her betrachten können, sind die Lehrer in der Lage, aus den Analysen große Vorteile zu ziehen. In einem kürzlich erschienenen Bericht der American Association of Colleges for Teacher Education (1968) wird der Vorschlag gemacht, eine Phase in der Lehrerausbildung einer solchen Analyse des Unterrichts zu widmen.

Jedoch ist ein Wort der Vorsicht angebracht. Die Versuchung, Beschreibungen des Lehrerverhaltens in Vorschriften umzuwandeln, die von Lehrern zu befolgen sind, muß vermieden werden. Denn bisher konnten nur wenige Korrelationen zwischen den Leistungen der Schüler und vielen Unterrichtsvariablen in den vorhandenen Beobachtungssystemen festgestellt werden. Obwohl die Beobachter, die die verschiedenen Analyseverfahren anwenden, dem Lehrer zuverlässige Beschreibungen vieler schwieriger Dimensionen der Unterrichtsabläufe liefern können, dürfen sie selbstverständlich nicht vorschreiben, wie Lehrer auf der Basis dieser Beschreibungen unterrichten sollen. Denn Entscheidungen hinsichtlich des sich daraus ergebenden Unterrichtsverhaltens sind vom Wissen um Konsequenzen verschiedener Verfahren und vom Urteil hinsichtlich des Wertes dieser Konsequenzen abhängig.

GRAHAM A. NUTHALL[1]

Ausgewählte neuere Untersuchungen zur Unterrichtsinteraktion und zum Lehrverhalten

Ein kritischer Bericht

Den meisten Lesern wird aufgefallen sein, daß in letzter Zeit die Zahl der Untersuchungen über das Lehrerverhalten und die Lehrer-Schüler-Interaktion im Unterricht zugenommen hat. Einige neuere Untersuchungen machen das Ausmaß dieser Zunahme deutlich (Lawrence 1966; Amidon/Simon 1965; Gage/Unruh 1967; Biddle 1967). Obwohl derartige Untersuchungen nicht neu sind, haben viele Studien noch immer Versuchscharakter. Daher wurden bisher nur wenige von ihnen in größerer Auflage veröffentlicht; und es wurde selten versucht, die bisherigen oder noch erreichbaren Ergebnisse dieser neuen Forschungsrichtung kritisch einzuschätzen. Dieser Beitrag soll einen kritischen Überblick über eine Auswahl neuerer Untersuchungen geben, die für diesen Forschungszweig sehr wichtig zu sein scheinen. Besondere Aufmerksamkeit soll dabei jenen Befunden zukommen, die unsere Kenntnis über die Beziehungen zwischen den Verhaltensweisen des Lehrers und den Lernprozessen bzw. Leistungen des Schülers verbessern.

Von ein paar Ausnahmen abgesehen, wurde das Lehrerverhalten und die Unterrichtsinteraktion nach dem klassischen Modell aller wissenschaftlichen Untersuchungen erforscht. Nach einer Reihe erkundender und abtastender Felduntersuchungen führten gezielte Versuche, Taxonomien des Unterrichtsverhaltens zu erarbeiten, zu mehreren hochentwickelten deskriptiven Untersuchungen über Lehrer und Schüler *in situ*. Im Anschluß an diese ersten »ökologischen« Überblicke wurde dann versucht, die beobachteten Lehrvariablen mit der gemessenen Schülerleistung zu korrelieren. Und neuerdings gibt es auch sorgfältig kontrollierte Untersuchungen halbexperimenteller Art. Da die Wirklichkeit jedoch stets verworrener ist als das Modell, finden immer noch interessante, wichtige Entwicklungen auf jeder der verschiedenen Entwicklungsstufen statt.

Daher werden die Untersuchungen, die analysiert werden sollen, unter drei Rubriken betrachtet: Zuerst soll die Entwicklung der Methoden behandelt werden, die zur Beobachtung und Analyse des gesamten Unter-

richtsgeschehens dienen. Dabei wird nicht versucht, alle der mehr als dreißig verschiedenen Systeme zu behandeln; es sollen lediglich einige interessante neue Verfahren sowie einige Modifikationen alter Systeme beschrieben werden. Danach sollen einige Ergebnisse der Unterrichtsbeobachtung dargestellt werden; dabei soll besonders berücksichtigt werden, wie weit sie zu unserem Verständnis der Ursachen des Lernens beitragen. Im letzten Teil werden dann Untersuchungen behandelt, in denen versucht wird, Schülerleistung in Beziehung zu Lehrerverhalten oder Unterrichtsinteraktion zu bringen. Sie reichen von einfachen Korrelationsuntersuchungen der Beziehungen in normalen Klassen, die ihren üblichen Unterricht erhielten, bis hin zu Versuchen, die Lehrvariablen experimentell zu manipulieren.

Verfahren zur Analyse von Unterrichtsverhalten

Revision eines alten Systems

Das bei weitem am häufigsten gebrauchte Verfahren für die Analyse des Unterrichtsverhaltens ist das Flanderssche Verfahren der Interaktionsanalyse (Amidon/Flanders 1963). Nach diesem Verfahren erstellt ein Beobachter ein fortlaufendes Protokoll der Unterrichts-Aktivitäten von Lehrer und Schülern. Alle drei Sekunden (bei wichtigen Veränderungen noch öfter) notiert der Beobachter die Kategorie, die das augenblickliche Geschehen am besten beschreibt. Das Ergebnis ist ein kontinuierliches Protokoll, das in bezug auf die relative Häufigkeit verschiedener Verhaltensarten und hinsichtlich der paarweisen Verhaltensfolgen anhand einer zwei-dimensionalen Matrix analysiert werden kann. Der ursprüngliche Zweck der Interaktionsanalyse bestand darin, Schätzwerte zu erhalten, bis zu welchem Grad die Lehrer ihre Schüler »indirekt« oder »direkt« beeinflussen. Das ursprüngliche, von dieser Analyse abgeleitete Maß (I/D Verhältnis) sollte diesen Aspekt des »affektiven Klimas« (affective climate) im Unterricht ausdrücken.

Während dieses Verfahren in seiner ursprünglichen Form noch immer sehr häufig benutzt wird, haben auch mehrere Verbesserungsversuche stattgefunden. In seinem Bericht über neue Entwicklungen bei der Interaktionsanalyse wies Amidon (1966) auf das Fehlen »kognitiver« Kategorien hin. Er beschrieb eine Reihe von Kategorien zur Identifikation verschiedener kognitiver Ebenen bei Fragen und Antworten. Teilweise basierten sie auf der Arbeit Aschners und Gallaghers (Aschner/Gallagher u. a. 1965), die auf Guilfords fünf kognitive Prozesse zurückgreift (Gedächtnis, Er-

kenntnis [cognition], konvergierende und divergierende Produktion und Bewertung). Bisher gibt es offenbar noch keine Untersuchungen, die eine Modifikation des Flandersschen Verfahrens in bezug auf kognitive Variablen benutzen; einige empirische Untersuchungen deuten darauf hin (Furst 1967), daß das Konzept des »kognitiven Klimas« im Unterricht weiter ausgearbeitet werden sollte. Ein solcher Meßwert des »kognitiven Klimas« im Unterricht könnte zeigen, welches »kognitive Niveau« (cognitive level) für einen Lehrer charakteristisch ist. L. Siegel und L. C. Siegel (1967) entwickelten ein ähnliches Konzept für das »intellektuelle Klima« (intellectual climate).

Eine weitere Modifikation des Flandersschen Verfahrens, typisches Lehrerverhalten zu messen, versuchte Honigman (1968). Er berichtete über die Entwicklung eines dreidimensionalen Kategorien-Systems und fügte der ursprünglichen affektiven Dimension eine kognitiv-inhaltliche (cognitive-substantive) sowie eine Verfahrensdimension (class management) hinzu. Dieses System wird als »Multidimensionale Analyse der Unterrichtsinteraktion« bezeichnet (Multidimensional Analysis of Classroom Interaction; MACI). Dieser Versuch, die Skala der Verhaltensweisen, auf die der Beobachter zu achten hat, auszuweiten, hat Honigman offenbar dazu gebracht, einige einfallsreiche Aufzeichnungstechniken zu entwickeln, die das Protokollieren erleichtern sollen. Honigmans Ausführungen sind auch bemerkenswert, weil sie einen Versuch beschreiben, die Validität (concurrent validity) der »Multidimensionalen Analyse der Unterrichtsinteraktion« zu prüfen. Bei dieser Untersuchung wurden 75 Lehrer, die an einem Sommerkurs teilnahmen, nach einem Zufallsstichprobenverfahren in drei gleich große Gruppen geteilt. Die eine Gruppe sah die Fernsehaufzeichnung einer Unterrichtsstunde, die zweite las nur das nach der »Multidimensionalen Analyse der Unterrichtsinteraktion« von einem Beobachter geführte Protokoll derselben Stunde, und die dritte Gruppe las ein Protokoll, das nach dem Flandersschen Verfahren geführt worden war. Der Kriteriumswert (criterion measure) wurde auf Grund der Angaben aller drei Gruppen in einem Fragebogen ermittelt, der aus 32 Fragen über verschiedene (affektive, kognitive und Verfahrens-)Aspekte der Stunde bestand. Folgende Hypothese wurde aufgestellt: Die Antworten der Gruppe, die das nach Honigmans »Multidimensionaler Analyse der Unterrichtsinteraktion« angefertigte Protokoll gelesen hatte, müßten mehr Ähnlichkeiten mit den Antworten der Gruppe haben, die die Life-Übertragung der Stunde gesehen hatte, als mit denen der Gruppe, die das nach dem Flandersschen Kategorien-System angefertigte Protokoll gelesen hatte. Ein solches Ergebnis war bei dem Teil des Fragebogens bestimmt zu erwarten, der sich auf die kognitive Dimension der Unterrichtsstunde bezog. Die Er-

gebnisse zeigten jedoch, daß die »Multidimensionale Analyse der Unterrichtsinteraktion« zwar besser die affektiven Aspekte, nicht aber die kognitiven und die Verfahrens-Aspekte der Unterrichtsstunde herausarbeitete. Es konnte noch nicht nachgewiesen werden, daß Honigmans Beobachtungssystem wirklich multidimensional ist.

Die Entwicklung neuer Systeme

Während nach dem Flandersschen Verfahren der Interaktionsanalyse und Honigmans »Multidimensionaler Analyse der Unterrichtsinteraktion« der Beobachter seine Aufzeichnungen in gleichmäßigen Zeitabständen machen muß, gibt es andere Analyse-Systeme, nach denen Beobachtungen in natürlichen Zeiteinheiten oder zu natürlichen Übergängen im Unterrichtsverhalten gemacht werden. Weil die nötige Identifikation und Analyse dieser Zeiteinheiten oder Übergangspunkte oft recht schwierig ist, werden in der Regel Tonband- oder Videorecorder-Aufzeichnungen des untersuchten Verhaltens gemacht. Bellack und andere (1966) berichteten z. B. von einer genauen Analyse der von ihnen als »Unterrichtszyklus (teaching cycles) bezeichneten Einheit. Die Analyse beruht auf Transkriptionen von Tonbandaufnahmen mehrerer Unterrichtsstunden, die über das gleiche ökonomische Thema von verschiedenen Lehrern gehalten wurden. Ein »Unterrichtszyklus« ist von allen natürlichen Einheiten im Ablauf der Diskussion im Unterricht am leichtesten als Einheit zu identifizieren; er besteht in der »Frage-Antwort-Reaktions-Sequenz«, oder in einer ihrer vielen Varianten. Dies war auch die Grundeinheit einer Untersuchung von Nuthall und Lawrence (1965).

In zwei sehr verschiedenen Untersuchungen wird von der Entwicklung neuer analytischer Systeme berichtet, die auf natürlichen Einheiten beruhen. Die erste Untersuchung beschreibt den Versuch, mit Hilfe eines brauchbaren psychologischen Konzepts das Unterrichtsverhalten zu erforschen. In der anderen Untersuchung wird der weit komplexere Versuch gemacht, Methoden zur Aufzeichnung aller denkbaren wichtigen Verhaltensweisen im Unterricht zu finden. MacDonald und Zaret (1967) berichteten von einem Versuch, den Begriff der »Offenheit« in den menschlichen Beziehungen (wie er von Rogers und anderen in der Psychotherapie entwickelt wurde) auf die Interaktion im Unterricht zu beziehen. Sie entwikkelten ein Verfahren, den Grad an Offenheit in den protokollierten verbalen Reaktionen der Lehrer gerade zu den Zeitpunkten im Unterrichtsgespräch einzuschätzen, zu denen die Lehrer Antworten von Schülern kommentieren und die weitere Diskussionsrichtung bestimmen. Sie stellten die Hypothese auf, daß die Lehrer, die sich während dieser wichtigen Situati-

on in der Unterrichtsdiskussion offener verhielten, mehr produktive Antworten ihrer Schüler provozieren würden. Die Lehrer dagegen, die verschlossener sind oder sich stärker an den Rollenerwartungen orientieren, erhalten dadurch wahrscheinlich weniger produktive Antworten von ihren Schülern und erhöhen gleichzeitig auch die Zahl der nur reproduktiven Antworten. Durch Tonbandaufzeichnungen von Unterrichtsstunden in der Sozialkunde (social studies) in neun Klassen einer Elementarschule fanden sie in acht der neun Klassen statistisch signifikante Beziehungen zwischen dem Grad der Offenheit des Lehrers und der Zahl der produktiven Antworten der Schüler. Obwohl sicherlich noch mehr sorgfältig kontrollierte Daten nötig sind, scheint doch der Wert des Begriffs der Offenheit für die Unterrichtsinteraktion nachgewiesen worden zu sein. Es wäre höchst interessant zu wissen, ob es zwischen Flanders' Begriff der »Indirektheit« und MacDonalds und Zarets Begriff der »Offenheit« eine enge Beziehung gibt.

In ihrem Bericht über ihre umfangreiche Untersuchung über Unterrichtsaktivitäten beschrieben Biddle und Adams (1967) die Entwicklung eines Beobachtungssystems, das wohl das umfangreichste aller vorliegenden Systeme ist. Im Gegensatz zu dem System von MacDonald und Zaret beruht das System von Biddle und Adams nicht auf einer spezifischen Theorie oder einem theoretischen Konstrukt, obwohl es etwas soziologisch orientiert ist. Es scheint der wohlüberlegte Versuch zu sein, die Voraussetzungen dafür zu schaffen, daß möglichst viele verschiedene, leicht identifizierbare Aspekte aller Aktivitäten im Unterricht in einem Protokoll aufgenommen werden können. Der Umfang des Systems hängt offenbar davon ab, ob eine Videorecorder-Anlage vorhanden ist, die eine detaillierte Analyse verbaler und verhaltensbezogener Aspekte des Unterrichtsgeschehens ermöglicht. Das System von Biddle und Adams basiert auf der Unterscheidung zwischen strukturellen und funktionellen Aspekten der Unterrichtsaktivitäten. Der strukturelle Aspekt umfaßt Positionen und Rollen aller Unterrichtsteilnehmer (was sie tun, mit wem, wo, usw.). Der funktionelle Aspekt umfaßt die Art des Inhalts (oder der Bedeutung) der Interaktion und die Art der Behandlung dieses Inhalts. Um also Strukturen beschreiben zu können, wird identifiziert: Wer spricht (Sender), zu wem (Empfänger), und wer hört zu (Zuhörer), also die Kommunikationsstruktur und Gruppierung aller Individuen und das Verhältnis der Individuen zu den physikalischen Dimensionen der Klasse. Um Funktionen zu beschreiben, identifizieren sie die drei in Honigmans »Multidimensionaler Analyse der Unterrichtsinteraktion« enthaltenen Dimensionen, und differenzieren weiter zwischen relevanter und irrelevanter Thematik. Eine Sonderkategorie erfaßt nicht-sprachliche Vorgänge (z. B. Drill, praktische Übungen, usw.).

Im allgemeinen hat man bisher dazu geneigt, möglichst viele verschiedene Aspekte des Unterrichtsverhaltens zu erfassen, ohne die feineren Unterscheidungen und Unterordnungen zu berücksichtigen, die unter jedem dieser Aspekte vielleicht zu treffen wären. Das Ergebnis ist ein System, das in gewissem Sinne unbegrenzt erweiterbar ist, vorausgesetzt, daß die Videorecorder-Aufnahmen der Unterrichtsvorgänge ausreichend präzise sind.

Die Zukunft der Systeme zur Analyse der Unterrichtsinteraktion

Während der frühen Stadien der Entwicklung der ersten wichtigen Systeme zur Analyse des Unterrichtsverhaltens äußerten mehrere Autoren die Befürchtung, die Vielzahl verschiedener Systeme könne zu einer beträchtlichen Verwirrung führen. Wenn jeder sein eigenes Analyse-System aufbaue, so wurde befürchtet, werde eine gewaltige Informationssammlung über Unterrichtsverhalten zustande kommen, die nicht systematisiert werden könne und aus der noch nicht einmal die einfachsten Thesen über Unterrichtsgeschehen entwickelt werden können, die die meisten Forscher benötigen. Schon jetzt könne man die Situation verwirrend finden und vielleicht den Wert so vieler scheinbarer Wiederholungen in Frage stellen. Dazu führt Komisar aus:

> Die Produktion »neuer« Kategoriensysteme bringt uns rasch dem Chaos näher. Scheinbar ist keiner der Forscher willens oder in der Lage, uns zu erklären, *warum* gewisse Kategorien gewählt werden oder in welchem Verhältnis die Kategorien eines Forschers zu denen eines anderen stehen (Komisar 1968, 22).

Andererseits wies Biddle darauf hin, daß jeder Forscher unbedingt sein eigenes System entwickeln müsse, um den Bedingungen seines eigenen Forschungsprojekts gerecht zu werden, zumal so vielen der neuen Systeme jegliche zuverlässige oder systematisch-theoretische Grundlage fehle (Biddle 1968, 31).

Mit dem Erscheinen einer Anthologie der 26 »bekanntesten und gebräuchlichsten Beobachtungssysteme für den Unterricht« (Simon/Boyer 1967) ist hoffentlich ein Wendepunkt erreicht worden [2]. In Zukunft wird der Konstrukteur eines »neuen« Kategoriensystems sein Werk theoretisch rechtfertigen müssen. So wird vielleicht die nötige Aufmerksamkeit auf die Erklärung und das Verständnis des Unterrichtsverhaltens gelenkt. Wie in den folgenden Abschnitten dieses Beitrags gezeigt werden soll, ist trotz der vielen protokollierten Details über die Geschehnisse in zahlreichen Unterrichtssituationen immer noch sehr wenig über die Ursachen oder Wirkungen der Geschehnisse bekannt.

Beziehungen zwischen Beobachtungssystemen

Es wurde bereits mehrmals versucht, die Beziehungen zwischen verschiedenen Systemen zu erforschen. In einigen Untersuchungen wandte man dazu beispielsweise zwei verschiedene Systeme parallel an. Medley und Hill (1968) berichteten über die Ergebnisse eines Vergleichs des Flandersschen Verfahrens der Interaktionsanalyse mit ihrer eigenen, unlängst abgeänderten Fassung des »Beobachtungsplans und -protokolls« (Observation Schedule and Record; OScAR 4V). Bei dieser Untersuchung wurden 70 Lehrer und Lehrerinnen der Sekundarschule während ihres ersten Berufsjahres von je zwei Beobachtern mindestens viermal systematisch beobachtet. Jedesmal protokollierte ein Beobachter nach dem Flandersschen System und der andere nach OScAR 4V. Entsprechend diesen parallelen Protokollen wurden für jeden Lehrer insgesamt 75 Meßwerte interkorreliert (38 nach Flanders und 37 nach OScAR). Da es offenbar schwierig ist, in eine so große Korrelationstabelle (75 × 75) Sinn hineinzubringen, wurde eine Faktorenanalyse durchgeführt. Es gelang, zehn Faktoren zu identifizieren, die die Unterschiede zwischen den Lehrern beschrieben. Von diesen zehn Faktoren wurden fünf mit beiden Systemen gemessen, drei nur mit OScAR, und zwei nur mit dem Flandersschen System. Die beiden Systeme scheinen sich bis auf einige unabhängige Merkmale zu überschneiden. Es ist allerdings sehr schwierig, die Besonderheit jedes Systems aus den untersuchten Faktoren abzuleiten.

In einer von Pearson [3] geleiteten Untersuchung wurde das Flanderssche Verfahren als zusätzliches Verfahren zur Analyse der »evaluativen Abschnitte« (evaluative ventures) bei Gesprächen in Klassen der Sekundarstufe I (intermediate school) benutzt. Ein »evaluativer Abschnitt« ist einer von neun Abschnitten oder themen-zentrierten Gesprächseinheiten (topic-centered units), die Smith und Meux (1967) identifiziert haben, unter dem die Diskussion über eine Wertfrage oder die Bedeutung eines Sachverhalts verstanden wird. Die vorläufige Analyse der Daten weist darauf hin, daß die Meßwerte des Flandersschen Systems mit der Zahl der evaluativen Gespräche, mit ihrer Dauer und ihrer relativ logischen Komplexheit in Beziehung stehen. Mit anderen Worten: Das »affektive Klima« ist erwartungsgemäß nicht unabhängig von den logischen und semantischen Aspekten der Diskussion in einer Schulklasse.

Durch Beobachtungssysteme gewonnene Ergebnisse

Einige neuere Untersuchungen brachten Daten über die Art des Lehrverhaltens und der Unterrichtsinteraktion; zu ihnen gehören auch die Untersuchungen von Bellack u. a. (1966) sowie von Smith und Meux (1967), deren Ergebnisse sich auf umfassende Forschungen stützen, die hauptsächlich verbale Interaktionen, semantische und thematisch-inhaltliche Aspekte betreffen.

Die Sequenz der Unterrichtsinhalte

Smith und Meux (1967) berichteten über die Ergebnisse eines Versuchs, die Organisation und sequenzielle Anordnung der Unterrichtsinhalte mit Hilfe von Aufzeichnungen von Diskussionen in Sekundarschulen zu analysieren. Sie definierten zunächst den Begriff »Unterrichtsstrategie« (teaching strategy) als eine besondere Sequenz von Teilthemen innerhalb des vorgegebenen Unterrichtsthemas. Davon ausgehend analysierten und beschrieben sie die »Strategien« in acht verschiedenen, zuvor identifizierten Kategorien. Diese acht Kategorien waren folgende: Ursachen, Begriffe, Bewertungen, Interpretationen, Verfahren, Regeln, Gründe und besondere Informationen (causes, concepts, evaluations, interpretations, procedures, rules, reasons and particular information). Dadurch fanden Smith und Meux z. B. heraus, daß Diskussionen über Begriffe zu der häufigsten Art von Diskussionen gehören. Nachdem sie diejenigen Teile des Gesprächs isoliert hatten, in denen ein genannter Begriff Mittelpunkt der Diskussion war, isolierten sie auch die Informationen über Begriffe, in denen die Begriffe ganz allgemein beschrieben oder besprochen wurden. Die kurzen Abschnitte der Diskussion, in denen nur eine einzige Information über einen Begriff beschrieben oder besprochen wurde, bezeichneten sie als »Begriffs-Impulse« (conceptual moves). So betrachtet, besteht der Begriff Unterrichtsstrategie also aus einer bestimmten Reihenfolge verschiedener »Begriffs-Impulse«.

Drei Hauptarten der »Begriffs-Impulse« ließen sich auf Grund umfangreicher Tonbandaufzeichnungen von Unterrichtsstunden in Sekundarschulen identifizieren: (1) Gesprächsteile, in denen Informationen über *Beispiele* des Begriffs besprochen wurden (»Beispiels-Impulse«) (instantial moves); (2) Gesprächsteile, in denen der Begriff mit anderen *verglichen* oder kontrastiert wurde (»Vergleichs-Impulse«) (comparative moves); und (3) die Gesprächsteile, in denen die Kriterien für den Begriff festgesetzt oder direkt *beschrieben* wurden (»Beschreibungs-Impulse«) (descriptive moves). Tabelle 1 enthält eine Liste dieser verschiedenen Arten der Impulse.

Tabelle 1
Impulsarten in »Begriffs-Abschnitten«*
(Types of Moves in Concept Ventures)

I *Beschreibungs-Impulse*
1. Beschreibung eines Merkmals
2. Definition ausreichender Bedingungen
3. Klassifikation des Begriffs
4. Klassifikatorische Beschreibung
5. Definition der Beziehungen zwischen Merkmalen
6. Zerlegung in Einzelteile

II *Vergleichs-Impulse*
7. Vergleich durch Analogie
8. Vergleich durch Differenzierung
9. Vergleich von Beispielen eines Begriffs

III *Beispiel-Impulse*
10. Beschreibung positiver Beispiele
11. Aufzählung aller Beispiele
12. Beschreibung eines Nicht-Beispiels
13. Beschreibung der Produktion von Beispielen
14. Bestätigung einer Sache als Beispiel

IV *Gebrauchs-Impulse* (usage moves)
15. Metaunterscheidungen bei Verwendung eines Begriffs

* aus: B. O. Smith, The Strategies of Teaching (1967)

Verschiedene Lehrer wurden hinsichtlich der in ihren Klassen üblichen Sequenzen der Impulse beschrieben und verglichen. Dabei stellte sich z. B. heraus, daß eine wechselnde Sequenz der Erörterung von Kriterien und die positiven Beispiele eines Begriffs eine der verbreitetsten Unterrichtsstrategien ist.

Unterschiede im Lehrerverhalten beim gleichen Curriculum

Die Praxis vieler Forscher, irgendwelche leicht zugänglichen Stichproben von Lehrern oder Klassen zu untersuchen, hat einige Unzufriedenheit geschaffen. Neuere Untersuchungen berichten dagegen von sorgfältig erhobenen Daten. Danach kann nun einiges darüber ausgesagt werden, wie das Verhalten im Unterricht durch gewisse, als unabhängig bekannte Variablen verändert werden kann. Gallagher (1966) berichtete über einen interessanten Versuch, die Unterschiede im Unterrichtsverhalten von Lehrern zu beschreiben, die dasselbe Thema desselben Curriculum mit vergleichbaren Klassen behandeln. Gallagher und andere (1966) entwickelten ein System der »Themenanalyse« (topic analysis), das die unterschiedlichen Gesprächsgänge über das gleiche Thema in verschiedenen Klassen identifizieren sollte. Als »Thema« wurde der Teil der Diskussion bezeichnet, in dem »sich die Klassendiskussion auf eine gegebene Handlung, einen Begriff

oder ein Prinzip konzentriert«. Jedes Thema wurde nach der Art des Inhalts klassifiziert (Informationswissen, Verfahrensfertigkeiten), nach dem Grad der Abstraktheit in der Diskussion (Daten, Begriff, Generalisierung) und nach dem Denkstil, der dem Gespräch zu entnehmen ist (Beschreibung, Expansion, Erklärung, Rechtfertigung, Bewertung usw.). Mit diesem System wurden die Tonbandprotokolle der Diskussionen in den Klassen von sechs Biologielehrern analysiert, die das Thema »Photosynthese« mit Hilfe der Materialien der neuen Biological Science Curriculum Study (»blue book version«) behandelt hatten. Gallagher fand signifikante Unterschiede zwischen den Lehrern bezüglich des Inhalts der Themen und im Grad der Abstraktheit, nicht aber im Denkstil. Auch der Anteil der Zeit, in der der Lehrer sprach, variierte signifikant von Klasse zu Klasse. Signifikante Unterschiede gab es auch für drei der vier Denkarten: Beschreibung, Erklärung und Expansion. Gallaghers Analyse machte deutlich, daß auch bei konstant gehaltenem Curriculum eine große Vielfalt in den behandelten Themen und in der Art ihrer Behandlung auftritt.

Unterschiede im Unterrichtsverhalten in bezug auf das Schuljahr, das Alter des Lehrers und die Unterrichtsinhalte

Weitere Informationen über die Art der Unterschiede zwischen Lehrern liefert der Bericht einer Untersuchung von Biddle und Adams (1967). Diese Wissenschaftler machten Fernsehaufzeichnungen von 32 Unterrichtsstunden, die von 16 Lehrern in drei verschiedenen Schuljahren (im ersten, sechsten und elften) gehalten wurden. Diese Unterrichtsstunden wurden gleichmäßig auf Mathematik und Sozialkunde verteilt; ebenso bildete man aus den Lehrern zwei gleich große Gruppen (älter oder jünger als 30 Jahre). Die Zusammenfassung einiger ausgewählter Ergebnisse soll hier wiedergegeben werden:

(1) *Schuljahr.* Viele Unterschiede ergaben sich als eine Funktion des Schuljahrs. Die ersten Klassen verbrachten viel Zeit mit Handlungen (wie Singen, Vorlesen) und mit Anweisungen zur Organisation des Unterrichts. Die sechsten Klassen wirkten am wenigsten traditionell, da es in ihnen mehr Gruppenarbeit und viele Interaktionen gab. Erwartungsgemäß waren die elften Klassen stärker thematisch orientiert und verbrachten mehr Zeit für die intellektuelle Auseinandersetzung mit relevanten Unterrichtsinhalten.

(2) *Alter des Lehrers.* Die Klassen der älteren Lehrer schienen insofern traditioneller zu sein, als ihre Handlungen öfter vom Lehrer gesteuert wurden. Sie beschäftigten sich auch mehr damit, Informationen weiterzugeben, als zu intellektuellen Auseinandersetzungen (z. B. zu klärenden und bewer-

tenden Diskussionen u. ä.) hinzuführen. Die Klassen mit jüngeren Lehrern zeigten weniger Lehrer-Kontrolle; hier waren die Lehrer öfter mit kleinen Arbeitsgruppen beschäftigt oder überhaupt nicht mit einbezogen.

(3) *Fachspezifische Inhalte.* Die Diskussion in den Mathematikstunden neigte dazu, in den unteren Klassen einen geringeren und in der elften Klasse einen höheren »intellektuellen« Gehalt zu haben als in den entsprechenden Unterrichtsstunden der Sozialkunde. Bei den letzteren gab es dagegen mehr Diskussionen über irrelevante Themen.

Interessante Zusatzbefunde:

(4) Wenn der Lehrer als Sender (Redner) vor der ganzen Klasse spricht, besteht der Inhalt seiner Rede aller Wahrscheinlichkeit nach in der Vermittlung einfacher Informationen über das Thema oder aus Anweisungen zur Organisation des Unterrichts.

(5) Wenn der Lehrer etwas zum Thema vorträgt, steht er wahrscheinlich vor der Klasse; wenn er dagegen etwas zur Organisation des Unterrichts äußert, steht er oft seitlich zur Klasse.

(6) Die Schüler, die in der Klasse entlang der Mittellinie sitzen, nehmen eher an der Diskussion teil und werden häufiger vom Lehrer angesprochen. Die Schüler, die weiter von der Mittellinie weg sitzen, sprechen entschieden seltener und werden auch seltener angesprochen.

Diese kurze Zusammenfassung kann den vielen Daten, die Biddle und Adams sammelten, zwar kaum gerecht werden; sie soll jedoch auf die allgemeine Bedeutung dieser umfassenden Dokumentation struktureller und funktioneller Aspekte des Unterrichtsverhaltens hinweisen.

Verbale Lehrerreaktion

Deutlich unterscheidet sich von der Studie von Biddle und Adams eine Untersuchung, über die Zahorik (1968) berichtete. Während Biddle und Adams möglichst viele Aspekte des Unterrichtsverhaltens berücksichtigten, ohne ihrer Analyse dabei eine bestimmte Richtung zu geben, richtete Zahorik seine Aufmerksamkeit auf einen einzigen Aspekt, der zweifellos von beträchtlicher Bedeutung ist. Zahorik entwickelte ein System zur Analyse und Bewertung der Reaktionen des Lehrers auf Schülerantworten. Zahoriks Daten wurden aus Tonbandprotokollen gewonnen, die in acht Klassen des dritten und in sieben Klassen des sechsten Schuljahrs aufgezeichnet worden waren. Man hatte die Lehrer gebeten, den Inhalt eines aktuellen Nachrichtenmagazins zu behandeln, wodurch man eine gewisse Kontrolle über den Unterrichtsstoff in den beiden protokollierten Stunden erhielt. Die erste Stunde sollte eine Einführung in die Lektüre des Nachrichtenmagazins geben, und die zweite Stunde sollte dann daran anknüpfen.

Aus den Reaktionen der Lehrer auf die Antworten ihrer Schüler bildete Zahorik 14 verschiedene Kategorien. Sie werden in Tabelle 2 aufgeführt:

Tabelle 2
Kategorien der verbalen Lehrer-Reaktionen *

1. Lobende Bestätigung
2. Tadelnde Verneinung
3. Lobende Bestätigung + Tadelnde Verneinung
4. Positive Antwort
5. Negative Antwort
6. Positive Antwort + Negative Antwort
7. Positive Erklärung
8. Negative Erklärung
9. Erweiterung der Reaktion: Entwicklung
10. Erweiterung der Reaktion: Verbesserung
11. Wiederholung der Aufforderung: mehrere Antworten
12. Wiederholung der Aufforderung: eine Antwort
13. Fortsetzung des Unterrichts: neues Thema
14. Verschiedenes

* aus: J. A. Zahorik, Classroom Feedback Behavior of Teachers, *Journal of Educational Research* 62, 1968, 147–150.

Die protokollierten Lehrerreaktionen bestanden aus einer oder einer Sequenz solcher Äußerungen. Außer den Angaben über die Häufigkeit dieser verschiedenen Reaktionen verschaffte sich Zahorik auch noch Lehrer-Einschätzungen (ratings) über die Richtigkeit der Schülerantworten und Schüler-Einschätzungen (ratings) über den Wert einer Auswahl von Lehrerreaktionen. Seine Ergebnisse wiesen darauf hin, daß von den 175 verschiedenen Arten von Lehrerreaktionen nur sechzehn häufiger wiederkehrten. Am häufigsten kamen folgende vor:

(1) Positive Antwort mit anschließender Überleitung zu neuem Thema (8,5 Prozent)
(2) Bitte um Erweiterung der Antwort, ohne irgendwelche Hilfen zu geben (8,3 Prozent)
(3) Einfache lobende Bestätigung mit anschließender Überleitung zu neuem Thema (7,8 Prozent)
(4) Einfache lobende Bestätigung, dann positive Antwort (Wiederholung), dann Überleitung zu neuem Thema (5,8 Prozent)
(5) Keine Antwort, Überleitung zu neuem Thema (5,1 Prozent)

Die Häufigkeit einer Reaktionsart variierte mit der Schulklasse, mit der

Phase des Unterrichts und mit den Urteilen der Lehrer über die Richtigkeit der Antwort. Durch diese Untersuchung konnte deutlich gemacht werden, daß die Rückmeldung durch den Lehrer keine Verstärkung (reinforcement) in dem Sinne ist, in dem Psychologen wie Skinner diesen Begriff verwenden. Auch betrachten die Lehrer Schülerantworten nicht als ein Verhalten, das Verstärkung braucht. Diese Untersuchung hat ein überraschendes Ergebnis: »Sie zeigt, daß es dem Lehrer nicht in erster Linie auf das sofortige Lernen ankommt und daß es ihm nicht das Wichtigste ist, was das Kind während der Interaktion sagt und tut« (Zahorik 1968).

Diese Ergebnisse sollten keinen Lehrer verwundern, es sei denn, er hätte sich von den Schriften einiger behavioristischer Theoretiker davon überzeugen lassen, daß Schüler nur das lernen, was sie tun, und daß ihr Tun aktiv verstärkt werden muß. Zur Verstärkung des Unterrichtsverhaltens sind viele vage Generalisierungen vorgeschlagen worden, besonders in Hinsicht auf das verbale Verhalten von Schülern. Zahoriks Untersuchung sollte dazu beitragen, entstandene Mißverständnisse zu beseitigen.

Damit soll freilich nicht gesagt werden, die Reaktionen des Lehrers auf die Antworten der Schüler seien unwichtig. Eine Untersuchung von Emmer (1968) weist darauf hin, daß einige Kategorien der Lehrerreaktionen (z. B. Flanders' Kategorie 3: »akzeptiert oder verwendet die Ideen des Schülers«) damit zusammenhängen, wie oft die Schüler durch Fragen die Initiative ergreifen. Emmers Ergebnisse, die aus der Untersuchung von sechzehn Lehrern des zweiten Schuljahres stammen, sind nicht unwidersprochen geblieben; sie machen jedoch deutlich, daß die Schüler um so freier mitmachen, je mehr der Lehrer ihre Ideen akzeptiert und darauf aufbaut. Wenn man davon ausgeht, daß es eine Beziehung zwischen der Beteiligung der Schüler an der Unterrichtsdiskussion und ihrer Leistung gibt – und einiges weist darauf hin (z. B. Gallagher 1966) –, dann sollte man auch nach der Beziehung zwischen der Art der verbalen Lehrerreaktionen und der Schülerleistung fragen. Das bedeutet jedoch nicht einfach, daß Schüler sich besser daran erinnern, was sie im Unterricht getan oder gesagt haben.

Unterschiede im Lehrerverhalten während des ersten Berufsjahrs

Medley berichtete über Beobachtungen an 70 Lehrern der Sekundarschule während des ersten halben Jahres ihrer Berufspraxis. Das verwendete Beobachtungssystem war das System »Observation Schedule and Record« (OScAR 4V), das Medley und seine Mitarbeiter im Laufe mehrerer Jahre entwickelt hatten. Als signifikante Veränderung im Verhalten dieser Lehrer ergab sich: (1) sie neigten dazu, weniger ergänzende Fragen (d. h., anders formulierte oder zusätzliche Fragen) zu stellen, (2) weniger Reaktionen auf

Schülerantworten zu geben, und (3) mehr Antworten als falsch abzulehnen, gleichzeitig jedoch die Schüler zum Antworten aufzumuntern. Diese Ergebnisse geben vielleicht die zunehmende Geschicklichkeit der Lehrer bei der Formulierung und bei der Auswahl des rechten Zeitpunkts ihrer Fragen wieder. Sie sehen immer seltener einen Anlaß, sich selbst zu verbessern, weil sie immer eindeutigere und angemessenere Fragen stellen.

Die Beziehung zwischen Lehrerverhalten und Schülerleistung

Affektive Dimensionen der Unterrichtsinteraktion

Das Unbehagen nach früheren Mißerfolgen und der Glaube daran, daß alle wichtigen Variablen im Unterricht noch entdeckt werden müssen, haben in neuerer Zeit Forscher dazu geführt, an das Problem der Ursachen von Schülerleistung nur mit beträchtlicher Vorsicht heranzugehen. Ohne Widerspruch zu fürchten, behauptete Bloom in einer Rede über den Stand der pädagogischen Forschung (1966), die Erforschung von Lehrmethoden habe gezeigt, daß die meisten etwa gleich wirkungsvoll sind. Dabei waren einige bedeutende Unterschiede bereits damals bekannt. Flanders (1965) hatte schon die Ergebnisse seiner umfassenden Untersuchung berichtet, in der er eine eindeutige Beziehung zwischen seinem Maß »Indirektheit« in der Unterrichtsinteraktion und der Schülerleistung in der Sozialkunde und in der Mathematik hatte zeigen können. Seinen Untersuchungen zufolge schien es so, als ob die »indirekteren« Lehrer ein Unterrichtsklima schufen, in dem die Schüler Sympathien für ihre Lehrer empfanden und in standardisierten Leistungstests besser abschnitten.

In einem kleinen Kontrollversuch bestätigte La Shier (1967) diese Ergebnisse. La Shier gebrauchte das Flanderssche Verfahren zur Bewertung der Leistung von zehn Schulpraktikanten (student teachers), die Schüler der achten Klasse in einer sechswöchigen Arbeitseinheit über »tierisches Verhalten« aus dem Biological Science Curriculum Study (BSCS) unterrichteten. Den Schülern wurde in den entsprechenden Inhalten der Biologie ein Vor- und ein Nachtest gegeben, und sie wurden gebeten, einen Fragebogen über ihre Einstellungen zu dem Unterricht auszufüllen. Als Kriteriumswert diente die durchschnittliche Verbesserung der mittleren Leistung jeder Klasse im Vergleich zur Ausgangsleistung (California Test of Mental Maturity). Die Ergebnisse wiesen darauf hin, daß die »Indirektheit« (I/D Verhältnis) der Praktikanten in signifikanter Beziehung zum Leistungsgewinn und den positiven Einstellungen bei den Schülern stand.

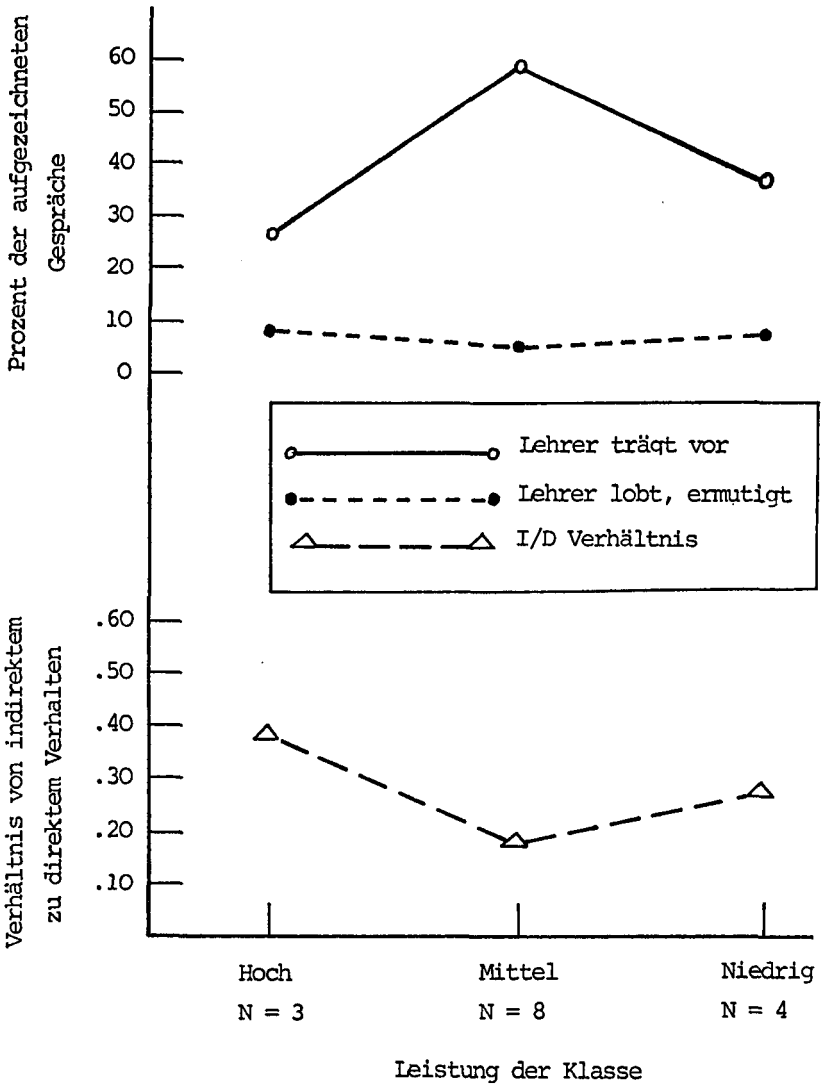
Das Verhältnis zwischen der Einstellung der Schüler und ihrem Leistungszuwachs war ebenfalls signifikant.

In einer noch umfassenderen Untersuchung benutzte Furst (1967) Daten von Bellack und anderen (Bellack u. a. 1966) in dem Bemühen, die Schülerleistung mit Maßen des Flandersschen Verfahrens der Interaktionsanalyse und mit Maßen des Bellackschen Analyse-Systems in Beziehung zu bringen. Bellacks Daten bestanden aus schriftlichen Protokollen von je vier Unterrichtsstunden, die fünfzehn Lehrer in ihren New Yorker Sekundarschulen gehalten hatten. Der Unterrichtsstoff war ein standardisierter vierstündiger Kurs in Wirtschaftswissenschaften. Auf Grund der Ergebnisse der Leistungstests (korrigiert nach Schülerintelligenz und Klassengröße) wurden die Lehrer aufgeteilt. Drei Lehrer wurden den Klassen mit den besten Leistungen, vier denen mit den schwächsten Leistungen und acht denen mit den mittleren Leistungen zugeordnet. Diese drei Lehrergruppen wurden nach drei Maßen des Flandersschen Verfahrens und nach drei des Bellackschen Analyse-Systems miteinander verglichen. Die Ergebnisse zeigten, daß die drei Klassen mit der höchsten Leistung sich von den anderen in folgendem unterschieden: mehr ausführliche, indirekte Gesprächsbeiträge der Lehrer, mehr positive als negative unmittelbare Reaktionen auf Schülerantworten, intensivere Unterrichtsbeteiligung seitens der Schüler. Bei näherer Betrachtung zeigen Fursts Ergebnisse freilich keine einfache Beziehung zwischen diesen Maßen und der Schülerleistung. Eine Auswahl der Ergebnisse wird in Abbildung 1 dargestellt. Hieraus ist zu ersehen, daß eine Kurvilinearität zwischen der Leistung und dem Prozentsatz der verschiedenen Arten von protokolliertem Lehrer- und Schülerverhalten besteht. Ob diese deutliche Kurvilinearität in Fursts Daten allerdings einmalig ist oder ob dieser Sachverhalt von anderen Forschern nur übersehen worden ist, ist nicht bekannt. Aber das weist darauf hin, daß bei einer anderen Lehrer-Stichprobe möglicherweise ganz andere Ergebnisse gefunden worden wären.

Noch einige Ergebnisse von Furst müssen erwähnt werden. Sie legten die Hypothese nahe, Lehrer mit hohen Leistungsergebnissen gäben in ihren Unterrichtsdiskussionen eine mittlere Anzahl pädagogischer Impulse zur verbalen Strukturierung des Unterrichts (verbal structuring moves), zeigten ein mittleres Verhältnis im Frage-Antwort-Austausch und einen hohen Grad an Verschiedenheit in den logischen Prozessen. Ein zusammengesetzter Meßwert, der aus den Einzelwerten dieser »kognitiven« Variablen bestand, bestätigte diese Hypothese. Ohne weitere Details über die Art dieser Meßwerte und ihre Beziehung zur Leistung ist es jedoch unmöglich, die Bedeutung dieser Ergebnisse richtig einzuschätzen.

Powell (1968) versuchte zu zeigen, daß man auf Grund der Beobachtun-

Abbildung 1. Meßwerte der Klassen mit hohen, mittleren und niedrigen Leistungen, gewonnen mit dem Flandersschen Verfahren der Interaktionsanalyse (aus: Furst, 1967).



gen von Lehrern in ihrer normalen schulischen Umgebung auf signifikante Beziehungen zwischen »Indirektheit« und besseren Schülerleistungen bei standardisierten Leistungsprüfungen schließen kann. Er verwendete in verschiedenen Klassen des dritten Schuljahres das Flanderssche Verfahren und wiederholte das Ganze im folgenden Jahr in denselben Klassen (viertes Schuljahr). Er wählte die Schüler aus, die während ihrer ersten drei Schuljahre dieselben »indirekten« bzw. »direkten« Lehrer hatten, und im vierten Jahr dann umgekehrt bei »direkten« bzw. »indirekten« Lehrern waren. Die Ergebnisse zeigen, daß die Leistung im Rechnen, nicht aber im Lesen, mit der »Indirektheit« des Lehrers während der ersten drei Schuljahre eng zusammenhängt. Und im vierten Schuljahr hatte es offenbar überhaupt keine Wirkung, ob die Schüler bei einem »direkten« oder »indirekten« Lehrer waren. Der Autor schloß daraus:

> Von den Ergebnissen dieser Untersuchung her scheint es eindeutig, daß der indirekte Unterricht keinen deutlichen Gesamtvorteil bringt (Powell 1968, 4).

Das Problem bei solchen Untersuchungen liegt in der Neigung des Forschers, eine Eigenschaft wie »Indirektheit« als feststehendes Attribut zu betrachten. Wenn der eine Forscher die beobachteten Lehrer in die »indirektesten« und die »direktesten« einteilt, kann seine Trennungslinie an einer anderen Stelle liegen als die eines anderen Forschers, der eine andere Lehrerstichprobe untersucht. Die Lehrer, die in der einen Untersuchung als »indirekt« bezeichnet werden, können in einer anderen Untersuchung als »direkt« bezeichnet werden. Dies macht vielleicht nicht sehr viel aus, sofern das Verhältnis zwischen den Faktoren des Unterrichtsklimas und der Leistung einfach linear ist. Aber wenn dieses Verhältnis kurvilinear ist, wie die Daten in Fursts Untersuchung andeuten, dann wird eine sorgfältigere Analyse der Daten erforderlich.

Der vielleicht sorgfältigste und umfangreichste Versuch, Aspekte des Unterrichtsklimas und des Lehrereinflusses mit der Schülerleistung in Beziehung zu bringen, wurde von Soar (1967; 1968) berichtet. Soar begann seine Untersuchung mit der Sammlung umfangreicher Daten über die Schüler in 55 Klassen (drittes bis sechstes Schuljahr) in vier verschiedenen Elementarschulen. Er ergänzte die standardisierten Leistungstests für Wortschatz, Lesen und Rechnen mit Kreativitätstests (Minnesota), Angsttests (Children's Manifest Anxiety Scale), Tests über Neigung zur Abhängigkeit (Flanders 1960) und Fragebogen, mit deren Hilfe die Einstellungen der Schüler gegenüber Lehrern und Mitschülern abgeschätzt werden sollten. Die gleichen Tests wurden nach einem Jahr und am Ende des zweiten Jahres wiederholt. Die Werte des Lernzuwachses der Schüler über den Zeitraum von einem bzw. zwei Jahren wurden als Kriteriumswerte verwendet.

In jeder der 55 Klassen wurden Beobachtungen angestellt, und zwar mit dem Flandersschen Verfahren der Interaktionsanalyse und mit einem anderen Verfahren, das Teile von Medley und Mitzels OScAR und Fowlers Ablehnungs-Zustimmungs-Skala (Hostility Affection Schedule) einschloß. Alle Meßwerte aus diesen Beobachtungen wurden faktorenanalysiert; es wurde eine Neun-Faktoren-Lösung erarbeitet, die »das Bild vom Unterricht am deutlichsten wiedergab«.

Für jede Klasse wurde ein Faktorenwert für jeden dieser neun Faktoren errechnet. Diese Faktorenwerte wurden dann mit dem durchschnittlichen Lernzuwachs korreliert, den die Schüler in den Tests erreicht hatten. Die Ergebnisse können so zusammengefaßt werden:

(1) Es gab keinen signifikanten Zusammenhang zwischen den beiden Faktoren, die die Ursache für die höchsten Prozentwerte der Varianz waren, und einem der Maße des Leistungszuwachses. Diese Faktoren schienen vielmehr Maße für die Gesamtdauer der sprachlichen Äußerungen der Lehrer und für die Möglichkeit der Schüler zu sein, eine Diskussion anzufangen.

(2) Der Faktor mit der engsten Beziehung zur Leistungsverbesserung der Schüler wurde als »ausgedehnter Lehrervortrag im Gegensatz zu schneller Lehrer-Schüler-Interaktion« bezeichnet (extended discourse vs. rapid teacher-pupil interchange). Soar bemerkt dazu:

> »Hier spiegelt sich offenbar ein Zyklus wider, in dem zuerst der Lehrer 15 bis 20 Sekunden lang redet, eine Frage stellt oder eine Anweisung gibt, und dann die Schüler eine Weile lang reden, ... kein längerer Vortrag im üblichen Sinn also, sondern eine Reihe von Lehreräußerungen, die der Problem-Strukturierung dienen können, oder die Darbietung einer begrenzten Zahl von Informationen, die die Schüler weiterentwickeln und verwenden sollen.« (Soar 1967, 7).

Die Parallele zwischen dieser Beschreibung und Fursts »mittlerem Grad der Steuerung der Lernaktivitäten« ist evident.

(3) Die Häufigkeit verbaler Schärfe oder Kritik durch den Lehrer stand in negativer Beziehung zum Lernzuwachs der Schüler im Rechnen und in einiger Beziehung zu ihrer Angst und ihrer Neigung zur Abhängigkeit.

(4) Ein Faktor zeigte viel Ähnlichkeit mit Flanders Maß der »Indirektheit«. Er stand jedoch weniger zur »Direktheit« des Lehrers in negativer Beziehung als vielmehr zum Ausmaß an »Ruhe und Unruhe« im Unterricht. Als Faktor korrelierte er zwar nicht mit der Verbesserung der Schülerleistung, stand aber in positiver Beziehung zur Steigerung der Kreativitätswerte und der Interessen der Schüler am Unterricht. Soar fiel es schwer zu erklären, warum der Faktor, der am engsten mit der »Indirektheit« des Lehrers zusammenhing, keine Korrelation mit der Leistung zeigte. Er

korrelierte jedoch mit dem jeweiligen Schuljahr, wobei Soar davon ausging, daß bei einigen früheren Untersuchungen man vielleicht das Ausmaß der Leistung mit der Höhe des Schuljahrs vermischt habe.

Im zweiten Teil seiner Untersuchung erforschte Soar den Einfluß des Lehrerverhaltens auf die Entwicklung der Schüler im Laufe eines zweiten Jahres. Er interessierte sich für das Ausmaß, in dem sich die Wirkungen des Lehrerverhaltens eines bestimmten Jahres vielleicht auf nachfolgende Jahre übertragen. Die Ergebnisse zeigten, daß die Kritik der Lehrer nur in geringem Maße von einem Jahr auf das nächste übertragen wird:

> Es scheint, als ob Ausdrücke negativer Affekte im Unterricht ihren stärksten Einfluß in der ersten Zeit danach hätten. Diesjährige Kritik ist von Bedeutung, die letztjährige nicht mehr.

Die Ergebnisse für die anderen Faktoren, die die Schülerleistung beeinflussen, zeigen kontinuierliche Auswirkungen und eine gewisse Interaktion zwischen den beiden Jahren. Der wohl auffallendste Befund im Hinblick auf die oben berichteten Daten von Furst besteht darin, daß ein mittleres Ausmaß von Lehrerkontrolle und »Indirektheit« die größten Vorteile zu bringen schien:

> ... in diesen Klassen erzeugte ein mittleres Maß an Kontrolle, entweder in Form von Kritik oder indirektem Unterricht, mehr wünschenswerte Veränderungen im Schülerverhalten als der übertriebene Mangel an Lehrerkontrolle. Vielleicht läßt sich daraus die Notwendigkeit ersehen, daß der Lehrer ein Minimum an Strukturierung schaffen muß, innerhalb derer die Schülerleistung maximiert werden kann (Soar 1967, 10).

Zweifellos wird es weitere Untersuchungen geben, die den Versuch machen werden, Schülerleistung und Wert für das affektive Klima im Unterricht miteinander in Beziehung zu bringen. Aber schon jetzt scheint genügend Beweismaterial vorzuliegen, um einige allgemeine Schlüsse ziehen zu können.

Erstens: Beobachtungsverfahren mit nur einem einzigen Kriterium wie dem Ausmaß der »Indirektheit« des Lehrers sind wahrscheinlich in ihrem Ansatz zu ungenau, um jemals klare Beziehungen mit Kriterien der Schülerleistung aufzeigen zu können. In einer Reihe von Beobachtungen werden sie wahrscheinlich zufällig mit mehreren anderen Aspekten des Unterrichtsverhaltens in Verbindung gesetzt.

Zweitens: Selbst wenn diese Beobachtungsverfahren eine gültige Dimension des Unterrichtsverhaltens erfassen, ist es unwahrscheinlich, daß diese Dimension eine einfache lineare Beziehung zur Schülerleistung aufweist. Wahrscheinlich ist es naiv zu erwarten, daß eine größere Freiheit der Schüler und ein freundlicheres Verhalten der Lehrer direkte und kontinuierliche Fortschritte in der intellektuellen Entwicklung des Schülers bringen.

Wie aus Soars Untersuchung deutlich hervorgeht, ist Kritik und verbale Schärfe der Lehrer gegenüber den Schülern ein wichtiger Faktor, aber das Ausbleiben dieser Kritik führt nicht automatisch zur Verbesserung des Lernens. Untersuchungen über das Unterrichtsklima sind schließlich Untersuchungen der Bedingungen, *unter denen Lernen stattfindet.* Daß es eine beträchtliche Vielfalt an Bedingungen gibt, die das Lernen nicht verhindern, sollte dabei nicht überraschen.

Die Suche nach signifikanten kognitiven Variablen

In einer neueren Abhandlung über die kognitiven Aspekte im Unterricht, bezweifelte Gage (1966 b) die Nützlichkeit, solche deskriptiven Verfahren wie die Systeme von Bellack und Smith zu entwickeln, bevor etwas über die mögliche Gültigkeit der kognitiven Aspekte des Unterrichtens bekannt ist, die analysiert werden sollen. Er zitierte eine Untersuchung, in der offensichtlich signifikante Unterschiede zwischen den Fähigkeiten der Lehrer, den Schülern ein Prinzip zu erklären, auftreten. Gage schlug einen systematischen Vergleich des Verhaltens solcher Lehrer vor, die bekanntermaßen gut bzw. schlecht erklären können; dieser Vergleich sollte wichtige Hinweise auf die wesentlichen Elemente solchen Unterrichts liefern. Im Anschluß daran berichteten Fortune, Gage und Shutes (1966) über die Ergebnisse eines Vergleichs der »Erklärungsfähigkeit« von 40 Praktikanten. Diese Praktikanten mußten in viertelstündigen Lektionen kleinen Gruppen von Schülern der Sekundarschule eine kontrollierte Themenreihe unterrichten. Als Kriteriumswert für jedes Thema wurde ein Test mit zehn Auswahl-Antwort-Aufgaben benutzt (multiple-choice item test), von denen die Praktikanten die Hälfte schon gesehen hatten, die andere Hälfte ihnen aber unbekannt war. Die Untersuchung war so aufgebaut, daß die Fähigkeit der Praktikanten zu erklären bzw. die Themen zu unterrichten über die verschiedenen Themen und die verschiedenen Schülergruppen hinweg verglichen werden konnte. Die Ergebnisse wiesen darauf hin, daß sich das Erklärungsvermögen eines Lehrers wahrscheinlich mit verschiedenen Themen verändert, bei verschiedenen Schülergruppen jedoch relativ konstant bleibt.

Vor einiger Zeit machten Gage und seine Mitarbeiter Fernsehaufzeichnungen von 43 Lehrern der Sozialkunde, die Schülern der zwölften Klasse Unterricht in zwei standardisierten Themen gaben. Die Themen bezogen sich auf aktuelle ökonomische, politische und soziale Entwicklungen in Jugoslawien und Thailand. Die Lehrer unterrichteten beide Themen in Lektionen von 15 Minuten, während ein drittes Thema (über Israel) allen Klassen in Form einer standardisierten Tonbandlektion vorgeführt wurde. Die Fähigkeit der Schüler, den Inhalt der standardisierten Tonbandlektion zu

lernen, diente dazu, die Kriteriumstestwerte an die anderen zwei Themen anzupassen. Die Schüler in jeder Klasse wurden außerdem darum gebeten, anhand standardisierter Fragebogen jede Lektion zu bewerten und den Grad ihrer Aufmerksamkeit während jeder Lektion anzugeben.

Ist die Erklärungsfähigkeit der Lehrer für verschiedene Themen relativ konstant? Podlogar, Rosenshine und Gage entdeckten (1968) in ihrer Datenanalyse, daß die Korrelation zwischen den durchschnittlichen Testwerten der Schüler für beide Themen zwischen .41 und .47 lag. Diese Korrelation ist signifikant und weist darauf hin, daß zwischen 16 und 20 Prozent der Varianz in der Klassenleistung auf einen Faktor der Fähigkeit des Lehrers zurückzuführen sein könnte. Aus den Bewertungen der Schüler wurde außerdem klar, daß sie die Erklärungsfähigkeit eines Lehrers ziemlich genau einschätzen konnten und daß die Einstufung ihrer eigenen Aufmerksamkeit signifikant mit dem Ausmaß ihres Lernens korrelierte.

Zwei etwas verschiedene Versuche, die wichtigen Komponenten der »Erklärungsfähigkeit« zu isolieren, sind mit diesen Daten gemacht worden. Hiller, Fisher und Kaess (1968) sowie Dell und Hiller (1968) berichteten von einem Versuch, mit Hilfe eines Computers die kritischen Elemente im verbalen Verhalten von Lehrern zu isolieren. Rosenshine (1968) berichtete die Ergebnisse eines Vergleichs innerhalb einer Teilstichprobe, die aus den erfolgreichsten und erfolglosesten Lehrern der Gesamtstichprobe gebildet worden war; dabei wurde sowohl das Verbal- als auch das Handlungsverhalten der Lehrer untersucht. Die Computer-Analyse erbrachte zwei Dimensionen des verbalen Lehrerverhaltens, die verläßlich mit dem Lehrererfolg bei den beiden Lektionen zusammenzuhängen schienen. Die erste dieser Dimensionen wurde als »verbale Flüssigkeit« (verbal fluency) bezeichnet. Sie bestand in einem zusammengesetzten Maß, das auf der Durchschnittslänge des gesprochenen Satzes und auf anderen Kennzeichen der Sprachflüssigkeit, wie etwa dem Anteil der »ähs« in der Rede des Lehrers basiert. Der zweite Faktor wurde »Unbestimmtheit« (vagueness) genannt: gleichfalls ein zusammengesetztes Maß, das sich aus der relativen Häufigkeit solcher Worte wie »fast, im allgemeinen, vielleicht usw.« errechnet. Die Forscher folgerten daraus nicht, daß diese verbalen Faktoren in direktem Zusammenhang mit dem Lernen des Schülers standen:

> Wir sind der Meinung, daß das Verhältnis zwischen unserem Maß und den Werten des Kriteriumstests vor allem die Korrelation und nicht die Ursache widerspiegelt. ... unser Maß der Unbestimmtheit dient wohl eher als Hinweis auf andere Verhaltensweisen, die ihrerseits das Verstehen und Behalten der Lektion kausal beeinflussen (Hiller/Fisher/Kaess 1968, 7).

In Rosenshines Analyse wurden viele Variablen untersucht; auf einige

ist bereits in früheren Untersuchungen hingewiesen worden, andere, z. B. die linguistischen Indikatoren, wurden erstmals hier untersucht. Die folgenden drei Variablen zeigten eine signifikante Beziehung zum Lehrererfolg:

(1) *Gesten und Bewegung.* Die fähigeren Lehrer zeigten eine größere Tendenz, sich frei in der Klasse zu bewegen und dabei Gesten zu machen.

(2) *Regel und Beispiel.* Die fähigeren Lehrer neigten dazu, Regeln vor und nach der Diskussion der Beispiele zu erklären, während die weniger fähigen Lehrer dazu neigten, die Regel nur einmal, entweder vor oder nach den Beispielen zu erklären.

(3) *Erklärende Bindeglieder* (explaining links). Die fähigeren Lehrer benutzten in der Regel häufiger Bindewörter wie »weil, weshalb, um ... zu, folglich, mittels, usw.«.

Rosenshine beendete seinen Bericht mit der Behauptung, die wichtigeren Variablen seien offensichtlich jene, die mit der Struktur der Kommunikation des Lehrers zu tun hätten. Die sequenzielle Anordnung der Gedanken und ihrer tragenden Elemente bedarf weiterer Erforschung. Man sollte jedoch nicht zu viel in die Ergebnisse dieser von Gage an der Stanford Universität angefangenen Untersuchungsreihe hineinlesen. Es darf nicht vergessen werden, daß, obwohl die Daten auf der Untersuchung einer größeren Zahl von Lehrern beruhen, nur zwei fünfzehnminütige Stichproben von dem Unterricht eines jeden Lehrers gemacht wurden [4]. Die Ergebnisse dieser Analyse lassen sich jedoch ohne weiteres mit jenen anderer Forscher wie beispielsweise Furst (1967) und Soar (1966; 1967) vereinbaren.

Ein Versuch mit Variationen in der sequenziellen Anordnung von Unterrichtsinhalten

Eine Untersuchung, die aus der Arbeit von Smith und Meux (1967) entstand, wirft etwas Licht auf die Bedeutung der Gedankensequenz des Lehrers. In dieser Untersuchung (Nuthall 1968) wurden vier alternative Strategien zum Unterrichten von Begriffen (concept teaching strategies) bzw. Sequenzen von »Begriffs-Impulsen« (sequence of conceptual moves) miteinander verglichen (vgl. die Beschreibung von Smiths Analyse von Begriffs-Impulsen). Es waren vier Unterrichtsstrategien, die Smith und seine Mitarbeiter in ihren Tonbandaufzeichnungen von Unterrichtsstunden in Sekundarschulen identifiziert hatten. Der verbale Inhalt dieser Strategien wurde in programmierter Textform einer Stichprobe von 412 Sekundarschülern dargeboten. Eine Varianzanalyse der Ergebnisse nachträglicher Kriteriumstests (delayed criterion tests) deutete an, daß die alterna-

tiven Strategien signifikant andere Wirkungen hatten. Die weitere Analyse ließ vermuten, daß die Wirksamkeit einer Unterrichtsstrategie davon abhängig ist, in welchem Maße sie auf dem schon vorhandenen Wissen der Schüler aufbaut und wie sehr sich Unterrichtsstrategie und vorhandenes Wissen gegenseitig beeinflussen. Bei den beiden Begriffen, die während dieses Versuchs unterrichtet wurden, war der Gebrauch von Beispielen und von Vergleichen mit anderen Begriffen je nach dem Vorwissen der Schüler verschieden.

Einige Schlußfolgerungen und Hinweise für künftige Forschung

Wie zu Beginn dieses Beitrags erwähnt wurde, haben viele Untersuchungen der letzten Jahre Versuchscharakter gehabt. Man sollte nicht versuchen, endgültige Schlußfolgerungen zu ziehen, auf denen Lehrer ihre Praxis aufbauen oder woraus Forscher experimentelle Hypothesen ableiten könnten. Einige Ergebnisse scheinen jedoch in den Schlußfolgerungen vieler Forscher wiederzukehren. Alle beziehen sich auf verbales Lehrerverhalten als einer signifikanten Ursache für Schülerleistung. Solange der Lehrer den Unterricht von Feindseligkeit und übermäßiger Kritik freihalten kann, wird wahrscheinlich die Wirksamkeit seines Einflusses auf die Schülerleistung von solchen Dingen abhängig sein wie:

(1) seiner Fähigkeit, den verbalen Kontext, in dem die Interaktion zwischen Lehrer und Schülern stattfindet, vorzubereiten bzw. zu »strukturieren«;

(2) seiner Fähigkeit, bei der Darbietung der Unterrichtsinhalte Gedanken mit einem Maximum an logischem Zusammenhang und einem Minimum an Unbestimmtheit und Ziellosigkeit zu organisieren;

(3) seiner Fähigkeit, die Schüler zur Teilnahme an Diskussionen anzuregen und sie für die Entwicklung und Erweiterung von Ideen zu interessieren.

Diese Themen sollten als Anregungen zu weiterer Forschung betrachtet werden, insbesondere für die Entwicklung experimentell brauchbarer Theorien des Unterrichtsverhaltens.

Dem kritischen Leser muß es aufgefallen sein, daß vielen der erwähnten Untersuchungen eine feste Richtung und eine kontrollierte Anordnung fehlen; dieser Mangel kann nur durch eine angemessene Theorie beseitigt

werden. Es drängt sich folgende Frage auf: Wenn diese Untersuchungen das Lernen der Schüler im Unterricht erfassen sollen, warum werden dann so wenige von ihnen durch die bekannten Lerntheorien beeinflußt? Wann wird eine Verbindung zwischen den etablierten psychologischen Theorien und dieser neuen Richtung der Erforschung des Unterrichtsverhaltens hergestellt?

Nach Meinung dieses Autors gibt es nur eine Antwort auf diese Fragen, daß nämlich die traditionelle psychologische Theorie keinen großen Wert haben kann, bevor die Unterrichtsforscher ihrerseits nicht signifikante theoretische Erklärungen des Unterrichtsgeschehens gefunden haben. Man braucht keine weitere Anpassung und Ausweitung bekannter Theorien, sondern die Schaffung einer neuen Theorie, die direkt auf das wirkliche Verhalten, das sie erklären soll, bezogen ist. Die in diesem Beitrag behandelten Untersuchungen weisen darauf hin, daß wir jetzt schon genug über einige wahrscheinlich wichtige Variablen wissen, um ihre Funktion wenigstens teilweise zu erklären.

Im Rahmen einer Unterrichtstheorie verdienen folgende Elemente des Unterrichtsverhaltens besondere Aufmerksamkeit:

(1) Nicht alle Schüler beteiligen sich immer am Unterricht; dennoch wird erwartet, daß alle lernen, und im allgemeinen lernen alle etwas. Eine Theorie des stellvertretenden Lernens (vicarious learning) beziehungsweise des Lernens bei nachlassender Teilnahme ist daher erforderlich.

(2) Die Bedeutung der Schülerteilnahme sollte unabhängig von Begriffen, Reaktion (response) und Verstärkung (reinforcement) eingeschätzt werden. Viele Autoren (z. B. Smith 1961) haben angedeutet, daß Lehrer die Beteiligung ihrer Schüler an Diskussionen als Informationsquelle über den Wissensstand und die intellektuellen Prozesse der Schüler benutzen. Das heißt, sie diagnostizieren die Schülerreaktionen und treffen spontane Entscheidungen aufgrund dieser Diagnosen. Jacksons neuere Veröffentlichung (1968), die auf Interviews mit erfolgreichen Lehrern beruht, weist darauf hin, daß Lehrer ihre sichersten Erfolgszeichen der subtilen »Rückmeldung« entnehmen, die sie von den Schülern ihrer Klasse erhalten. Als Praktiker sind sie sehr mißtrauisch und voller Zweifel gegenüber den Ergebnissen standardisierter Tests. Soweit diesem Autor bekannt ist, hat bisher noch keiner die Vorstellungen vom Lernen zu erklären versucht, aufgrund derer ein fähiger Lehrer die Diskussion beeinflußt. Wie interpretiert er die Reaktionen seiner Schüler? Welche Zeichen benutzt er, um den Verlauf seiner Handlungen zu bestimmen? In den Antworten auf diese Fragen muß die Erklärung für Ursache und Wirkung der Diskussion im Unterricht liegen.

(3) Die Bedeutung des Schülers als einer unabhängigen Größe für Verän-

derungen im Unterricht muß berücksichtigt werden. In einem neueren Artikel hat Turner (1968) Material aufgearbeitet, welches die Hypothese unterstützt, daß Schüler zumindest einige Aspekte des Lehrerverhaltens selbst bestimmen. Schüler können ihrerseits ziemlich genau den Erfolg ihrer Lehrer bewerten (vgl. Podlogar/Rosenshine/Gage 1968); Beobachter des Unterrichtsverhaltens werden bald darauf aufmerksam, daß Schüler die subtilen Hinweise ihrer Lehrer begreifen können und es auch tatsächlich tun. Jede Theorie, die darauf zielt, das Lernen im Unterricht zu erklären, sollte die Schüler auch als aktiv auswählende Teilnehmer berücksichtigen.
(4) Die Doppelrolle des Lehrers, die darin besteht, z. T. intellektuell und psychisch auf die Schüler Einfluß zu nehmen (Gage, 1966a) und z. T. ihre Motivation zu erhöhen, bedarf weiterer Untersuchungen. Die aktive Teilnahme des Schülers zu erreichen ist eine oft genau so kunstvolle und schwierige Aufgabe wie die Führung der Gedankengänge von Schülern. Das Problem der Motivation im Unterricht muß bei jeder Untersuchung der gedanklichen Interaktionen mitberücksichtigt werden. Hier ist die Erzeugung von verdeckten geistigen Reaktionen (covert mental responses) ähnlich schwierig wie die Stimulierung offener verbaler Reaktionen.

Schließlich ist kein kritischer Bericht über die Unterrichtsforschung ohne die Warnung vollständig, die auch verschiedene andere Autoren geäußert haben: Die Frage nach den besten Unterrichtsmethoden und den wirksamsten Mitteln, das Lernen zu fördern, darf nicht mit den Fragen darüber gekoppelt werden, wie sich Unterrichten und Lernen im Unterricht in Wirklichkeit abspielen. Erst wenn wir über die Beziehungen zwischen Unterrichtsverhalten, Lernen und Leistung der Schüler genauere Kenntnisse haben, wird es vielleicht möglich sein, wohlbegründete Unterrichtsprinzipien aus diesem Verständnis abzuleiten. Dieser kritische Bericht sollte deutlich machen, daß ein solches Verständnis noch nicht erreicht worden ist. Wahrscheinlich werden die Bemühungen um ein ausreichendes Verständnis des Unterrichts zusätzlich erschwert, wenn dieser Wunsch nach einem besseren Verständnis mit dem nach einem verbesserten Unterricht verwechselt wird.

IV Ausgewählte Beispiele zur Evaluation

Einführung

Häufig ist in der Evaluationsliteratur der letzten Jahre auf das Fehlen guter und leicht zugänglicher Evaluationsuntersuchungen hingewiesen worden (z. B. Westbury 1970; Cooley 1971). Im Unterschied zu den in den letzten Jahren zahlreich und sorgfältig entwickelten Evaluationsmodellen hat es wenige Berichte von Evaluationsuntersuchungen gegeben, die Eingang in Fachzeitschriften oder Veröffentlichungen über Evaluation gefunden hätten (vgl. Grobman 1968). So enthält eine von Baker (1969) zusammengestellte, 80 Titel umfassende Liste von Veröffentlichungen zur Curriculumevaluation nur sechs empirische Untersuchungen. Manchen Autoren ist das zum Anlaß geworden, in Anlehnung an Veröffentlichungen Schwabs (1970, 1971a, 1971b), die den Wert von Theorien für die Curriculumentwicklung radikal in Frage stellen, auch den Wert von Modellen und Theorien der Evaluation für die Praxis der Evaluation anzuzweifeln (z. B. Lewy 1972). Aus diesen Ausführungen geht deutlich hervor, daß es zwischen Modellen der Evaluation und der wirklichen Durchführung von Evaluationsuntersuchungen zahlreiche konzeptuelle, methodische und technologische Schwierigkeiten zu überwinden gibt, wozu die Modelle selbst wegen ihres hohen Generalisierungsgrades nur begrenzt beitragen können. Deshalb sollten in dieser Veröffentlichung dem Leser neben den anspruchsvollen Modellen aus Teil II auch konkrete Evaluationsuntersuchungen vor Augen geführt werden. Sie thematisieren als Ergänzung zu den Beiträgen der vorherigen Abschnitte neue Fragen und Probleme, die sich bei der Anlage einer Evaluationsuntersuchung und ihrer Durchführung unausweichlich ergeben. Auf Grund der Unterschiedlichkeit der Evaluationsberichte wird dabei das weite Spektrum der Möglichkeiten sichtbar, die es für eine sinnvolle Durchführung von Evaluationsuntersuchungen gibt. Sie werfen die Frage nach dem wissenschaftstheoretischen und bildungspolitischen Standort des Evaluators auf, ohne sie anders als für ihren eigenen Zusammenhang lösen zu können. Die hier ausgewählten Untersuchungen nehmen nicht in Anspruch, vorbildlich zu sein. Sie bieten durchaus Anlaß zur Kritik, die sie

z. T. in der offenen Darstellung der eigenen Fehler und Unzulänglichkeiten herausfordern. Jedoch gehören sie zu den wenigen interessanten, im Rahmen dieses Bandes reproduzierbaren Evaluationsuntersuchungen, die sich im angelsächsischen Bereich finden ließen.

Eine der bekanntesten Untersuchungen haben Ball und Bogatz mit ihrer Evaluation des ersten Jahres von Sesame Street vorgelegt, die als eine summative Evaluation die Leistungsfähigkeit dieser Sendungen und des Fernsehens als Unterrichtsmedium nachweisen konnte und die wertvolle Anregungen zur Modifikation der Sendereihe für das zweite Jahr brachte. Die Untersuchung ergab, (1) daß die Kinder, die am häufigsten fernsehen, auch am meisten lernen, (2) daß das, was in den Sendungen am stärksten behandelt wird, auch am besten gelernt wird und (3) daß das Programm keine besondere Beaufsichtigung der Kinder durch Erwachsene erforderlich macht. Diese Evaluationsstudie ist zugleich ein Beispiel für die Evaluation eines Bildungsprogramms, das nicht zum schulischen Bereich im engeren Sinne gehört.

Andersons Ausführungen zielen auf die Evaluation eines curricularen Programms. Dazu beginnen sie mit der Bestimmung des Standorts der Untersuchung im Kontext der Diskussion über die verschiedenen Ansätze zur Evaluation. Es folgt die exakte Einarbeitung dieser Ansätze in die Konzeptualisierung und Planung der Untersuchung. Sodann wird die Durchführung der Evaluation detailliert beschrieben und werden ihre Ergebnisse dargestellt. Die Untersuchung dient als Beispiel für die Evaluation eines curricularen Programms mit Hilfe einer Kontrollgruppe und soll die Brauchbarkeit dieser Form der Evaluation belegen.

In Cooleys Beitrag werden verschiedene in diesem Zusammenhang neue methodische und statistische Verfahren der Evaluation einer schulischen Innovation beschrieben, die für die Planung von Evaluationsuntersuchungen wichtig sein dürften. Sie wurden vom Autor und seinen Mitarbeitern für die Evaluation des umfassenden Schulversuchs des Learning Research und Development Center in Pittsburgh entwickelt, der mit dem Projekt Individually Prescribed Instruction durchgeführt wird.

Wesentlich neue Gesichtspunkte bringt auch der aus Großbritannien stammende Bericht über die Evaluation des Humanities Curriculum Project, die aufgrund der speziellen curricularen Vorstellungen des Projekts (vgl. Stenhouse 1971) vor besonderen Schwierigkeiten steht. Hier gilt es, eine Evaluation durchzuführen, ohne daß Lernziele im herkömmlichen Sinn als Kriterien der Evaluation verwendet werden. Das heißt, die Evaluation muß in Entsprechung zu der »Offenheit« des Projekts konzeptualisiert und durchgeführt werden. MacDonald gibt einen Erfahrungsbericht über die ersten zwei Jahre der Evaluation, in dem die zahlreichen Unzuläng-

lichkeiten offen dargelegt werden, und entwickelt einen Evaluationsplan für die nächsten zwei Jahre der Evaluation des Humanities Curriculum Project.

Das hier von MacDonald gewählte Verfahren der Evaluation erinnert in einigen Aspekten durchaus an Handlungsforschung (action research), die einen weithin neuen Bereich der pädagogischen Forschung darstellt. Handlungsforschung zielt auf die sofortige Lösung der untersuchten Probleme. Dafür ist sie bereit, den klassischen Forschungsplan aufzugeben und die entsprechenden Nachteile in Kauf zu nehmen. Sie beruht auf der Hypothese, daß Lehrer ihr Verhalten besonders dann verändern, wenn sie ihre Einstellungen ändern (Corey 1953). Dazu können sie im Rahmen der Lehrerfortbildung am besten gebracht werden, wenn sie – unter Beratung von Wissenschaftlern – sich der Erforschung ihrer eigenen Probleme zuwenden. Handlungsforschung ist u. a. durch zwei Aspekte gekennzeichnet:

(1) Ziel und Methode können nicht wie bei der klassischen empirischen Forschung in einer einfachen Zweck-Mittel-Relation gesehen werden. Die Ziele entstehen und verändern sich unter dem Einfluß der »Objekte« der Forschung im Laufe der Untersuchung, wodurch ebenfalls eine Veränderung der Forschungsverfahren bewirkt wird.

(2) Das Verhältnis von Subjekt und Objekt ist nicht durch die übliche Rollenverteilung gekennzeichnet. Die Distanz zwischen den Handelnden und den ihre Handlungen Erforschenden wird weitgehend aufgehoben. Das impliziert, daß die Validität und Reliabilität der Forschungsergebnisse nicht mehr gewährleistet ist und die Generalisierbarkeit der Ergebnisse in Frage gestellt werden muß.

Beim augenblicklichen Stand der Diskussion sollte man sich gegenüber der Handlungsforschung als einer Ausprägung der Evaluation offen zeigen und sich fragen, ob und in welchen Bereichen ihre Anwendung möglich und sinnvoll ist. Das neuerliche Interesse an dieser Form der Forschung in der BRD wird vielleicht bald zu ihrem besseren Verständnis führen.

Samuel Ball / Gerry Ann Bogatz

Das erste Jahr von Sesame Street

Eine Evaluation

Die Vorgeschichte der Untersuchung

Im Sommer 1968 begann das Children's Television Workshop (CTW), sein Programm Sesame Street zu planen. Alle Beteiligten stimmten überein, daß die Pläne eine unabhängige Evaluation der Auswirkungen des Programms einschließen sollten. Children's Television Workshop beauftragte das Educational Testing Service (ETS), eine gemeinnützige pädagogische Test- und Forschungsinstitution in Princeton, New Jersey, eine Evaluation durchzuführen, um festzustellen, in welchem Ausmaß die Fernsehsendung Sesame Street ihre gesetzten Ziele während des ersten Jahres erreicht hatte. Die Untersuchung versuchte u. a. folgende Fragen zu beantworten:

Was sind, im ganzen gesehen, die Auswirkungen von Sesame Street?

Was sind die modifizierenden Einflüsse von Alter, Geschlecht, vorausgehendem Leistungsstand und sozio-ökonomischem Status auf die Auswirkungen von Sesame Street?

Haben Kinder, die zu Hause Sesame Street sehen, im Vergleich zu Kindern, die es zu Hause nicht sehen, einen Gewinn davon?

Haben Kinder in Vorschulklassen, die Sesame Street als Teil ihres Schulcurriculum ansehen, einen Gewinn davon?

Haben Kinder aus spanisch-sprechenden Elternhäusern einen Gewinn von Sesame Street?

Wie beeinflussen die häuslichen Verhältnisse die Auswirkungen von Sesame Street?

Das innovative pädagogische Programm des Children's Television Workshop erhielt wesentliche Unterstützung von öffentlichen und privaten Stellen. Von Anfang an waren es die Carnegie Corporation New York, die Ford Foundation, das National Center for Educational Research and Development im U. S. Office of Education, das U. S. Office of Economic Opportunity und das National Institute of Child Health and Human Development. Unter den anderen Institutionen, die später für Unterstützung sorgten, wa-

ren die Corporation for Public Broadcasting, die National Foundation of Arts and Humanities und die John & Mary R. Markle Foundation.

Die Hauptergebnisse

In der ersten Sendeperiode von 26 Wochen zeigte Sesame Street, daß das Fernsehen ein wirkungsvolles Medium sein kann, um 3- bis 5jährigen Kindern wichtige einfache Sachverhalte und Fertigkeiten, wie z. B. das Erkennen und Benennen von Buchstaben und Zahlen, und komplexere höhere kognitive Fertigkeiten, wie das Klassifizieren und Sortieren nach einer Vielzahl von Kriterien, zu lehren. Die Forschungsergebnisse des Educational Testing Service erbrachten, daß Sesame Street Kinder aus sozial benachteiligten innerstädtischen Bezirken, aus mittelständischen Vororten und aus abgelegenen ländlichen Gebieten fördert. All diese Gruppen wurden in dieser Evaluation untersucht.

Die Leistungsfähigkeit des Bildungsfernsehens als Unterrichtsmedium wird durch drei Hauptergebnisse der Untersuchung deutlich:

1. Kinder, die am meisten zusahen, lernten auch am meisten. Das Ausmaß des Lernens – d. h. der Punktzuwachs, den ein Kind zwischen den Testergebnissen für bestimmte Fertigkeiten vor und nach dem Betrachten von Sesame Street zeigte – vergrößerte sich im Verhältnis zu dem Ausmaß der Zeit, die das Kind dem Programm zusah.

2. Die Fertigkeiten, die die meiste Zeit und Aufmerksamkeit durch das Programm erhielten, waren mit wenigen Ausnahmen auch die Fertigkeiten, die am besten gelernt wurden. Eine Analyse des Inhalts der Sendung zeigte z. B., daß mehr Zeit (13,9 %) als jedem anderen Gegenstand den Fertigkeiten, die mit Buchstaben in Beziehung stehen, gewidmet wurde. Auf diesem Gebiet der Buchstaben und Zahlen waren die Gewinne der Kinder am auffälligsten. Außer dem Erwerb von Fertigkeiten, die direkt und ausdrücklich gelehrt wurden, fand auch offensichtlich ein gewisser Lerntransfer statt, indem einige Kinder Dinge lernten, die in dem Programm nicht gelehrt wurden, z. B. ganze Wörter zu erkennen oder den eigenen Namen zu schreiben.

3. Das Programm erforderte keine ausdrückliche Beaufsichtigung durch Erwachsene, damit die Kinder auf den vom Programm umfaßten Gebieten lernten. Kinder, die Sesame Street zu Hause sahen, zeigten ebenso große Gewinne und in einigen Fällen sogar größere als Kinder, die in der Schule unter Aufsicht eines Lehrers das Programm sahen. Dieses Ergebnis hat besondere Bedeutung angesichts der Tatsache, daß mehr als vier Fünftel aller Drei- bis Vierjährigen und ebenso mehr als ein Viertel aller Fünfjährigen keinerlei Bildungseinrichtungen besuchen.

Das Hauptergebnis, daß Kinder, je länger sie die Sendung sehen, desto mehr auch lernen, gilt unabhängig von Alter, Geschlecht, geographischer Wohnlage, sozio-ökonomischem Status, Intelligenzalter und unabhängig davon, ob die Kinder zu Hause oder in der Schule das Programm sahen. In allen acht Bereichen, in denen die Kinder getestet wurden, vergrößerten sich die Lerngewinne mit der Häufigkeit des Zuschauens. Der Punktzuwachs war bei einigen Tests und Untertests jedoch höher, und einige Gruppen von Kindern zeigten einen höheren Punktzuwachs als andere. Die Dreijährigen erzielten die höchsten, die Fünfjährigen die geringsten Gewinne. D. h. dreijährige Kinder, die die Sendung sahen, hatten höhere Punktzahlen im Nachtest als diejenigen Vier- und Fünfjährigen, die die Sendung seltener sahen, selbst dann, wenn im Vortest die jüngeren Kinder niedrigere Punktzahlen als die älteren hatten. Dieses Ergebnis hat bedeutsame Folgen für die gesamte Erziehung, denn es legt nahe, daß dreijährige Kinder fähig sind, viele Fertigkeiten zu lernen, die gewöhnlich erst in späteren Jahren unterrichtet werden.

Ein ähnliches Ergebnis zeigte sich bei sozial privilegierten und sozial benachteiligten Kindern. Obwohl die sozial benachteiligten Kinder mit beträchtlich niedrigeren Leistungen in den Fertigkeiten, die gelehrt wurden, begannen, übertrafen diejenigen, die sehr oft und lange die Sendung sahen, die Kinder der Mittelschicht, die nur selten das Programm sahen. Diese Fernsehsendungen können offensichtlich die beträchtliche Kluft in der Bildung, die gewöhnlich sozial privilegierte und sozial benachteiligte Kinder trennt, schon bis zum Zeitpunkt des Eintritts in die erste Klasse verringern.

Ein auffallendes, obwohl sehr vorläufiges Ergebnis legt nahe, daß Sesame Street besonders effektiv sein könnte, den Kindern einige Fertigkeiten zu lehren, deren Muttersprache nicht Englisch ist und die in der Schule keine guten Leistungen erzielen. Eine sehr kleine Stichprobe von Kindern aus spanisch-sprechenden Elternhäusern im Südwesten erzielte bessere Gewinne als jede andere Untergruppe von Kindern.

Sesame Street hat einige seiner Ziele erfolgreicher verwirklicht als andere. Die Untersuchung liefert die Gründe und gibt Anhaltspunkte für die Verbesserung der Programmentwicklung. Es zeigte sich, daß in einigen Fällen der relative Mangel an Erfolg von einer anfänglichen Unterschätzung, in anderen Fällen von einer anfänglichen Überschätzung der Vorkenntnisse und Fertigkeiten der Kinder herrührte. Ein weiteres Ergebnis bestand darin, daß das Lernen erfolgreicher war, wenn die Fertigkeiten wie bei den Buchstaben direkt und nicht wie bei den Anfangslauten indirekt angesprochen wurden.

Stichprobe und Tests

Zu Beginn der Untersuchung wurden annähernd 1200 Kinder aus den fünf verschiedenen Regionen Boston (Massachusetts), Durham (North Carolina), Philadelphia (Pennsylvania), Phoenix (Arizona) und aus einem ländlichen Gebiet im Nordosten Kaliforniens ausgewählt. Die Stichprobe, die schließlich 943 Kinder zählte, bestand aus sozial benachteiligten Kindern aus innerstädtischen Bezirken, sozial privilegierten Kindern aus Vorortgebieten, Kindern aus ländlichen Gebieten und sozial benachteiligten spanischsprechenden Kindern. Im ganzen umfaßte die Stichprobe mehr Jungen als Mädchen und mehr Unterschicht- als Mittelschichtkinder. Unter den sozial benachteiligten waren mehr schwarze als weiße Kinder. Die meisten Kinder waren vier Jahre, einige waren drei und einige fünf Jahre alt. Die Mehrzahl der Kinder der Stichprobe sahen Sesame Street zu Hause und nicht im Vorschulunterricht.

Die Hersteller von Sesame Street hatten spezifische Lernziele für das Programm festgesetzt. Um den Lerngewinn im Hinblick auf diese Ziele und die Transferwirkungen zu bestimmen, wurden Meßinstrumente benutzt, die vom Educational Testing Service eigens für diese Evaluation entwickelt worden waren. Die acht Haupttests und ihre Untertests waren[1]:

Körperteiletest
- Auf die Körperteile zeigen
- Benennen der Körperteile
- Funktion von Körperteilen (zeigen)
- Funktion von Körperteilen (nennen)

Buchstabentest
- Erkennen von Buchstaben
- Benennen von Großbuchstaben
- Benennen von Kleinbuchstaben
- Vorgegebene Buchstaben in Wörtern finden
- Erkennen von Buchstaben in Wörtern
- Anfangslaute
- Wörter lesen

Formentest
- Erkennen von Formen
- Benennen von Formen

Zahlentest
- Erkennen von Zahlen
- Benennen von Zahlen
- Zahlverständnis (vgl. Beispielaufgabe 1)
- Zählen
- Addition und Subtraktion
- (Parallelisierter Untertest für Buchstaben, Zahlen und Formen)

Beziehungstest
- Umfangsbeziehungen
- Größenbeziehungen
- Positionsbeziehungen (vgl. Beispielaufgabe 2)

Sortiertest

Klassifikationstest (vgl. Beispielaufgabe 3)
- Klassifikation nach Größe
- Klassifikation nach Form
- Klassifikation nach Zahl
- Klassifikation nach Funktion

Puzzletest

Alle Tests waren nach dem gleichen Grundschema aufgebaut. Die Testmaterialien waren einfach, und die Tests wurden den Kindern einzeln von einem geschulten Erwachsenen aus ihrer Nachbarschaft gegeben. Ferner wurden Informationen über die familiären Verhältnisse eines jeden Kindes gesammelt und darüber, wie oft das Kind Sesame Street gesehen hatte. Die 943 Kinder der Stichprobe wurden in Quartilen aufgeteilt entsprechend der Länge der Zeit, die sie während der Dauer der Untersuchung Sesame Street gesehen hatten. Alle nachfolgenden Analysen sind auf diese Quartilen bezogen worden. Sie reichen von Quartil 1 (Q 1), in dem die Kinder Sesame Street selten oder nie sahen, bis Quartil 4 (Q 4), in dem die Kinder das Programm im Durchschnitt mehr als fünfmal in der Woche sahen. (Sesame Street war so populär, daß es nur wenige Kinder gab, die die Sendung wirklich nicht sahen; viele Kinder in Q 1 sahen das Programm gelegentlich.).

Gesamtergebnisse

Für die Stichprobe insgesamt gilt: Kinder aus den Quartilen, in denen die Sendung am meisten gesehen wurde, verhielten sich in allen Tests besser als Kinder aus den Quartilen, in denen sie weniger gesehen wurde. Kinder, die das Programm am meisten sahen (Q 4), hatten die höchsten Punktzahlen im Vortest (das bedeutet, daß sie schon mit einem Vorsprung anfingen), sie hatten die höchsten Punktzahlen im Nachtest, und sie erzielten den höchsten Punktzuwachs in der Zeit zwischen Vor- und Nachtest. Die allgemeine Tendenz, bei längerem und häufigerem Ansehen der Sendung einen höheren Punktzuwachs zu erzielen, war bei einigen Tests ausgeprägter als bei anderen. Besonders ausgeprägt war diese Tendenz bei den Buchstaben-, Zahlen- und Klassifikationstests; am wenigsten zeigte sie sich beim Körperteiletest.

Sozial benachteiligte Kinder

In der Gesamtstichprobe von 943 Kindern wurden 731 als sozial benachteiligt angesehen. Auch bei ihnen erhöhte sich der Punktzuwachs im Verhältnis zu der Häufigkeit, in der sie Sesame Street sahen. Im Hinblick auf die Gesamtpunktzahl für die 203 Testaufgaben, die im Vor- und Nachtest gleich war, gewannen die Kinder aus Q 1 19 Punkte, die Kinder aus Q 2 29 Punkte, die Kinder aus Q 3 38 Punkte und die Kinder aus Q 4 47 Punkte (siehe Tabelle 1)[2]. Ein Teil des Punktzuwachses, der von Kindern aus Q 1 erzielt wurde, muß weitgehend als eine Folge der Reifung angesehen werden, da viele von ihnen die Sendung niemals gesehen hatten. Die größeren Gewinne der Kinder in anderen Quartilen sind jedoch weitgehend eine Folge der Häufigkeit ihres Ansehens der Sendung. Dieselbe Beziehung ließ sich zwischen den verschiedenen Gesamtbeträgen für alle acht Haupttests beobachten. Den höchsten Punktzuwachs gab es bei den Buchstaben-, Zahlen- und Klassifikationstests (vgl. Tabelle 1).

Komplizierte statistische Analysen wurden durchgeführt, um zu bestimmen, ob die beobachteten Unterschiede sich zufällig eingestellt hatten, ob sie signifikant durch andere Faktoren herbeigeführt worden waren oder ob sie – wie es schien – weitgehend eine Folge der Häufigkeit des Ansehens der Sendung waren[3]. Die Häufigkeit des Zuschauens erwies sich bei weitem als die bedeutendste Variable; das bedeutet, ihr Einfluß schien gleichermaßen davon unabhängig zu sein, welches Geschlecht die Kinder hatten und ob sie das Programm zu Hause oder in der Schule sahen. Um die Auswirkungen der Häufigkeit des Zuschauens genau zu isolieren, wurde eine Spezialuntersuchung mit zwei Parallelgruppen von Kindern durch-

geführt (die »Age Cohorts Study« – die Altersgruppen-Untersuchung). Gruppe 1 war 53 bis 58 Monate zur Zeit des Vortests alt; Gruppe 2 war 53 Monate bis 58 Monate zur Zeit des Nachtests alt. Außer demselben Lebensalter zum Zeitpunkt des Vergleichs waren die beiden Gruppen auch von vergleichbarem Intelligenzalter und lebten in denselben Gemeinden. Es gab, kurz gesagt, keine beobachtbaren Unterschiede zwischen den beiden Gruppen in bedeutsamen Punkten wie Vorkenntnissen, IQ und häuslichen Verhältnissen. In jeder Gruppe waren mehr als 100 sozial benachteiligte Kinder, die keine Bildungseinrichtungen besuchten. Die Vortest-Punktzahlen von Gruppe 1 (bevor die Kinder Sesame Street gesehen haben konnten) wurden mit den Nachtest-Punktzahlen von Gruppe 2 verglichen, nachdem diese Kinder das Programm gesehen hatten. Die das Programm häufig sehenden Kinder in Gruppe 2 – Kinder aus Q 3 und Q 4 – erreichten über 40 Punkte mehr bei den 203 gemeinsamen Testaufgaben als die vergleichbaren Kinder in Gruppe 1, die die Sendung niemals gesehen hatten (vgl. Tabelle 2). In gleicher Weise signifikant ist die Tatsache, daß gelegentliche Zuschauer (Q 1) in Gruppe 2 sich nur um 12 Punkte von vergleichbaren Kindern in Gruppe 1, die Sesame Street nicht gesehen hatten, unterschieden. Zusammenfassend kann gesagt werden: Hielt man Auswirkungen der Reifung, IQ, Vorkenntnisse und häusliche Verhältnisse konstant, erzielten die häufigen Zuschauer große und bedeutsame Gewinne.

Obwohl die Häufigkeit des Zuschauens sich mit dem Alter der Kinder nicht auffallend veränderte, ergaben sich dennoch Änderungen in den Testergebnissen. Zum Zeitpunkt des Vortests schnitten Dreijährige, wie vorauszusehen war, weniger gut als Vierjährige und Vierjährige weniger gut als Fünfjährige ab. In bezug auf den Punktzuwachs im Nachtest waren die Ergebnisse jedoch gerade umgekehrt. Obwohl die Gruppe unter den Dreijährigen, die das Programm am häufigsten sahen, im Vortest mit einem niedrigeren Punktwert als irgendeine Gruppe der Fünfjährigen begann, erreichten die Dreijährigen, die die Sendung am häufigsten sahen, zum Zeitpunkt des Nachtests im Durchschnitt höhere Punktwerte als die Vierjährigen in Q 1, Q 2 und Q 3 und höhere als die Fünfjährigen in Q 1 und Q 2. Selbst Dreijährige, die das Programm nur zwei- oder dreimal in der Woche sahen, erzielten im Vergleich zu anderen Altersgruppen einen beachtlichen Punktzuwachs (vgl. Tabellen 3, 4, 5 und Abbildung 1).

Einige Testergebnisse hingen deutlich vom Alter ab. Unter den Kindern, die die Sendung häufig sahen, wurde beim Körperteiletest der höchste Punktzuwachs von den Dreijährigen erzielt; Drei- und Vierjährige erzielten bei den Zahlen einen höheren Punktzuwachs als Fünfjährige; und Fünfjährige erreichten beim Lesen von Wörtern (was einen Lerntransfer anzeigt) und bei den Anfangslauten (was in Sesame Street indirekt gelehrt wurde)

einen höheren Punktzuwachs als die anderen. Um es kurz zu sagen: Ziele, die indirekt gelehrt wurden, wurden von den älteren Zuschauern besser gelernt, und ein Lerntransfer zeigte sich bei ihnen deutlicher als erwartet werden konnte. Im allgemeinen galt: Wo spezifische Kenntnisse und Fertigkeiten direkt gelehrt wurden, erzielten die jüngeren Kinder einen höheren Punktzuwachs als die älteren.

Sozial privilegierte Kinder

169 Kinder in der Untersuchung wurden als sozial privilegiert angesehen. Sie erreichten im Vortest höhere Punktwerte als die anderen Gruppen und sahen im Durchschnitt einen größeren Teil der Sendungen als alle Gruppen der sozial benachteiligten Kinder. Eine relativ geringe Häufigkeit des Zuschauens brachte bei diesen Kindern relativ hohen Punktzuwachs (vgl. Tabelle 6 und Abbildung 2).

Spanisch-sprechende Kinder

Es wurden nur 43 spanisch-sprechende Kinder von der Untersuchung erfaßt. Sie unterschieden sich in dem Ausmaß, in dem sie vor dem Ansehen von Sesame Street mit der englischen Sprache in Berührung gekommen waren. Infolge dieser Unterschiede und dem geringen Umfang der Stichprobe können Schlußfolgerungen nur mit großer Vorsicht gezogen werden. Die größte Zahl der spanisch-sprechenden Kinder war in Q 1; lediglich ein Rest von 18 befand sich in der das Programm häufig sehenden Gruppe. Diese Kinder erzielten einen fast unglaublich hohen Punktzuwachs. Der Punktzuwachs der spanisch-sprechenden Kinder aus Q 3 war in der Tat so hoch wie der der anderen Kinder in Q 4. Beim Buchstabentest begannen die spanisch-sprechenden Kinder aus Q 4 mit den niedrigsten Punktwerten im Vortest und erreichten die höchsten Punktwerte im Nachtest. Andere Buchstaben-Untertests und die Tests über Zahlen, Formen, über das Sortieren, die Beziehungsverhältnisse und das Klassifizieren zeigten das gleiche Ergebnis: ein niedriger Start mit nachfolgendem sehr hohem Punktzuwachs für Kinder, die das Programm sehr häufig sahen.

Kinder aus ländlichen Gebieten

In der Untersuchung hatten die Kinder aus ländlichen Regionen in den Vortests relativ niedrige Punktwerte, erreichten aber in den Nachtests als Folge des Zuschauens einen hohen Punktzuwachs. Ihre Eltern hatten oft eine bessere Bildung als die Eltern der sozial benachteiligten Stadtkinder.

Ihre großen Gewinne legen nahe, daß Sesame Street als pädagogisches Medium für Kinder, die in abgelegenen Gegenden oder in kleinen Dörfern wohnen, sehr geeignet ist.

Sesame Street in den Schulen

Die Lehrer, deren Klassen Sesame Street im Rahmen der Untersuchung sahen, wurden gebeten, ihre Reaktionen auf das Programm anzugeben. Obwohl sie Sesame Street als Unterrichtsmittel für kleine Kinder anerkannten, waren sie über die Eignung des Programms für den Unterricht selbst geteilter Meinung. Einige vertraten nachdrücklich die Auffassung, daß die Sendung wertvolle Zeit verbrauche, die besser für andere Vorhaben verwendet werden könnte, andere meinten, daß sie eine wertvolle Bereicherung des Schultags sei.

Kinder, Eltern und Sesame Street

Die Kinder, die Sesame Street am meisten sahen und deshalb am meisten lernten, hatten Mütter, die oft mit ihnen zusammen die Sendung ansahen und oft zu ihnen über die Sendung sprachen. In diesen Familien hatten die Eltern in der Regel etwas höhere Erwartungen für ihre Kinder.

Schlußfolgerung

Hinsichtlich seiner selbst gesteckten Ziele war Sesame Street im allgemeinen sehr erfolgreich. Die Untersuchung des Educational Testing Service zeigt, daß drei- bis fünfjährige Kinder aus verschiedenen häuslichen Verhältnissen wichtige einfache und komplexe kognitive Fertigkeiten durch das Ansehen von Sesame Street erwarben. Die am meisten die Sendung sahen, erzielten auch die höchsten Gewinne. Die zusammenfassende Schlußfolgerung lautet: die Leistungsfähigkeit des Bildungsfernsehens als eines wirkungsvollen Mediums, um bestimmte Fertigkeiten sehr kleinen Kindern zu lehren, ließ sich durch Sesame Street nachweisen [4].

Abbildung 1

Vortestergebnis und Punktzuwachs im Gesamttest für alle sozial benachteiligten 3-, 4- und 5jährigen Kinder

(nach der Häufigkeit des Zuschauens in Quartilen unterteilt)

N = 127 3jährige
N = 433 4jährige
N = 159 5jährige

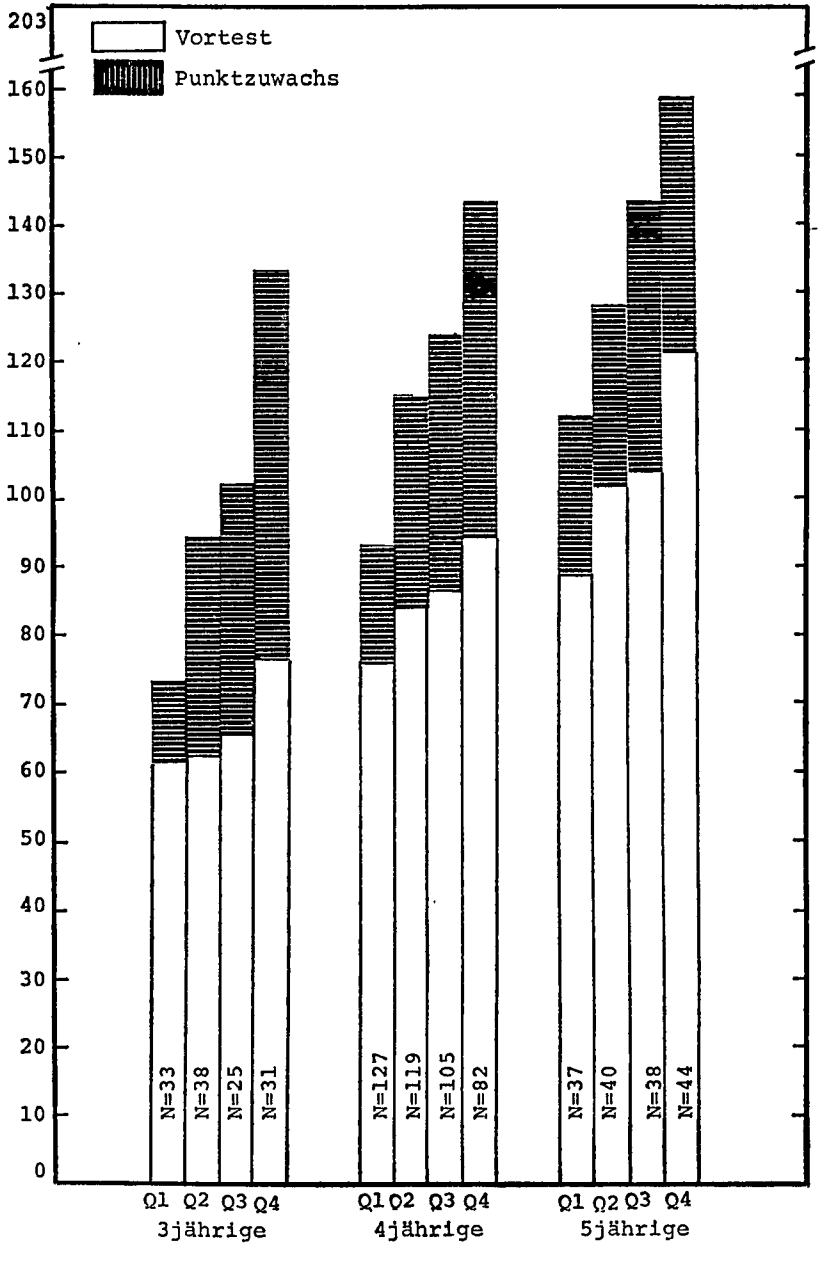


Abbildung 2

Vortestergebnis und Punktzuwachs im Gesamttest
für alle sozial privilegierten Kinder
(nach der Häufigkeit des Zuschauens in Quartilen unterteilt)
N = 169

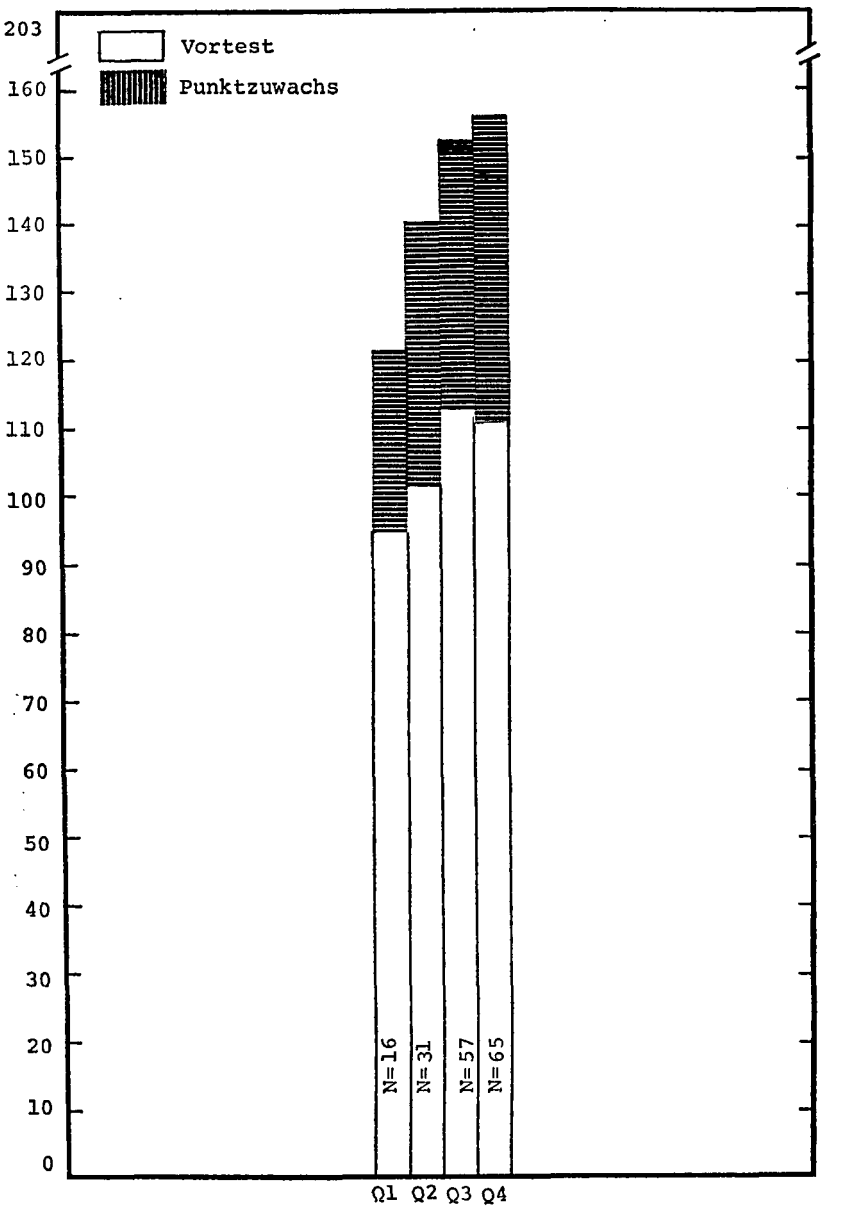
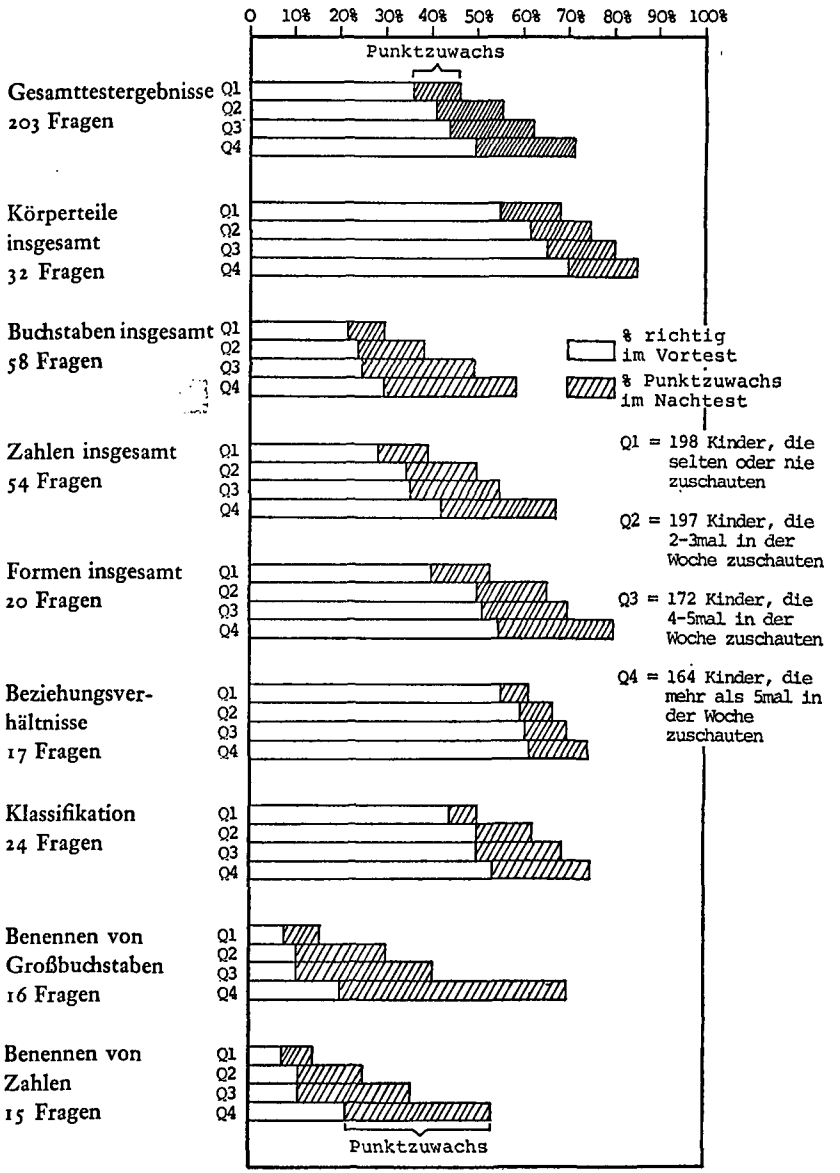


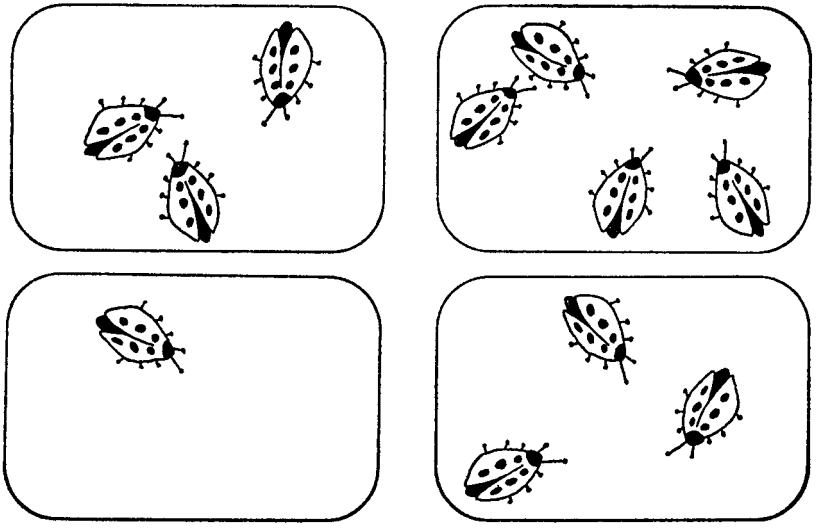
Abbildung 3

Zusammenstellung der Ergebnisse, die sozial benachteiligte Kinder in den verschiedenen Tests erzielten

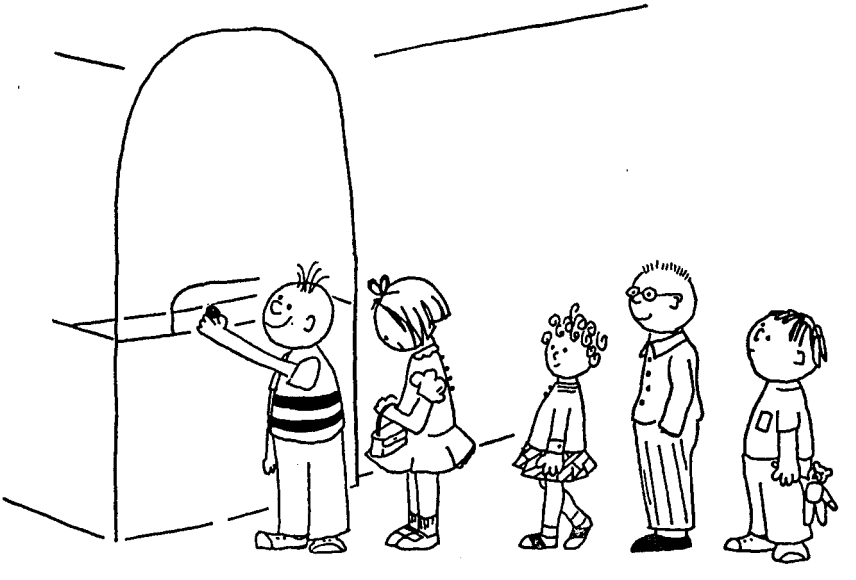
(Die von allen sozial benachteiligten Kindern in Vor- und Nachtest richtig beantworteten Testaufgaben sind in Prozenten angegeben.)



Testaufgabe (Beispiel 1)
Schau auf die Marienkäfer hier, hier, hier und hier. In welchem Kästchen sind fünf Marienkäfer?

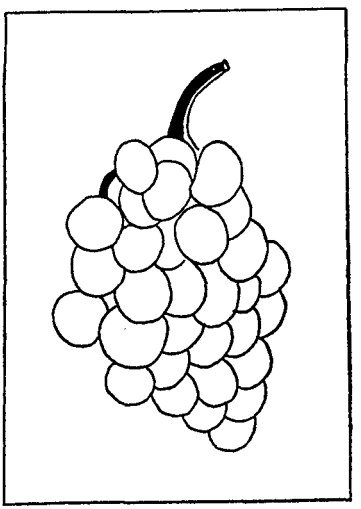


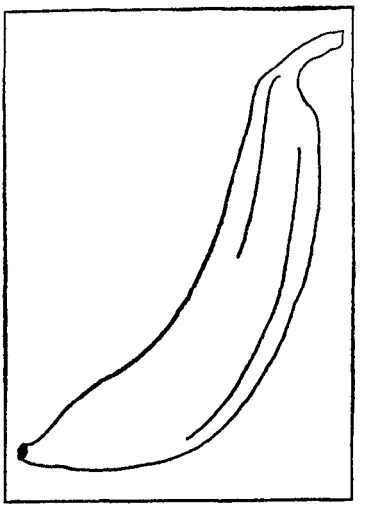
Testaufgabe (Beispiel 2)
Hier stehen Kinder in einer Reihe an. Sie warten, um in ein Kino gehen zu können. Welches Kind steht zuletzt in der Reihe?

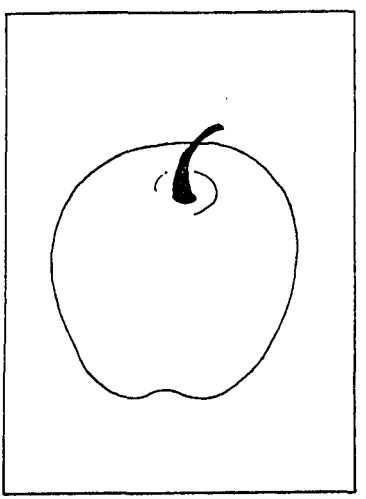


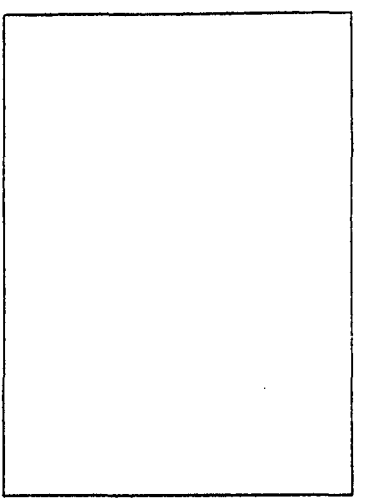
Testaufgabe (Beispiel 3)

Hier ist ein Bild von Weintrauben, von einer Banane und einem Apfel. Ein Bild fehlt. Wir wollen das Bild, das hierher paßt, herausfinden.



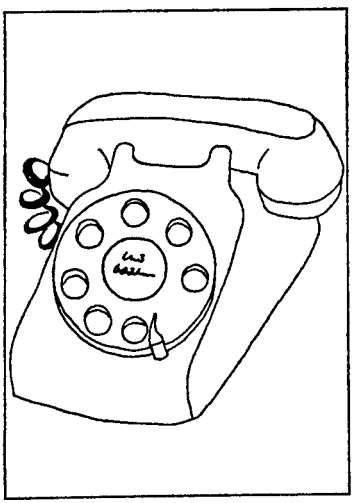


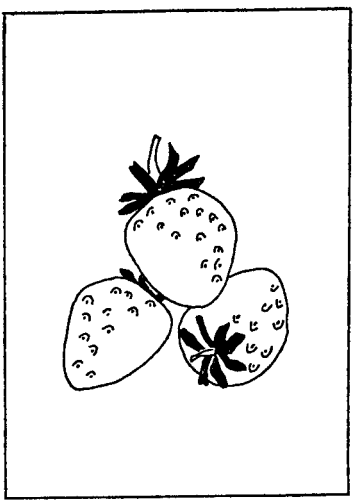


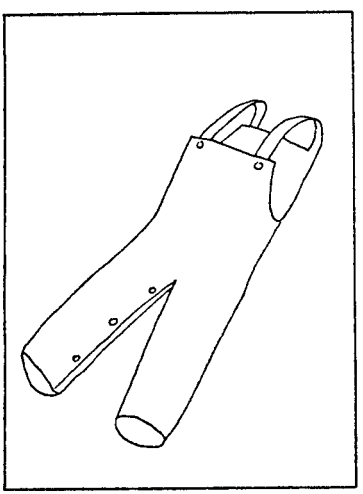


Testaufgabe (Beispiel 4)

. Hier siehst du ein Telephon, Erdbeeren, eine Hose und ein Buch.
Was davon paßt zu den Weintrauben, der Banane und dem Apfel?







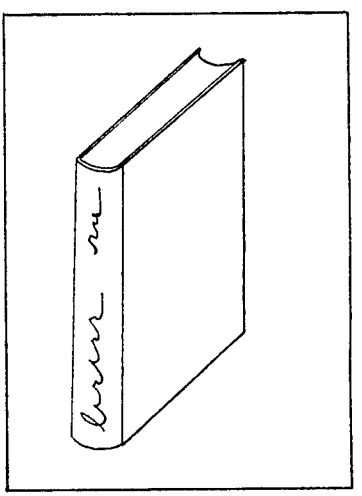


Tabelle 1

Vortestergebnis und Punktzuwachs für alle sozial benachteiligten Kinder

(in Quartilen unterteilt)

N = 731

Haupttests (ohne Untertests)	Maximal möglicher Punktwert *	Q 1 N = 198				Q 2 N = 197				Q 3 N = 172				Q 4 N = 164			
		Vortest		Punktzuwachs		Vortest		Punktzuwachs		Vortest		Punktzuwachs		Vortest		Punktzuwachs	
		$\overline{X}$	s	$\overline{X}$	s	$\overline{X}$	s	$\overline{X}$	s	$\overline{X}$	s	$\overline{X}$	s	$\overline{X}$	s	$\overline{X}$	s
Gesamttestergebnisse	203	75.62	24.73	18.63	20.04	84.42	27.60	29.11	22.51	87.74	27.63	37.97	25.29	97.54	32.16	47.36	26.15
Körperteile insgesamt	32	18.11	6.51	3.88	5.71	20.00	6.35	4.38	5.50	21.09	6.04	4.74	5.31	22.47	6.05	5.24	4.88
Buchstaben insgesamt	58	13.07	5.95	4.30	7.43	14.42	7.37	8.22	9.26	14.95	7.00	11.89	11.00	17.98	10.12	15.97	11.19
Formen insgesamt	20	8.43	3.50	2.29	3.77	9.89	4.01	3.15	4.05	10.04	3.64	4.29	4.07	10.64	3.50	5.49	3.52
Zahlen insgesamt	54	16.18	8.20	5.43	7.05	18.56	9.38	8.52	8.23	19.64	10.10	10.88	9.51	23.69	11.15	13.01	9.52
Parallelisierter Untertest	11	7.83	2.76	1.26	2.87	8.38	2.55	1.50	2.50	8.90	2.19	1.12	2.09	9.32	1.77	1.02	1.82
Beziehungsverhältnisse insges.	17	9.07	2.98	1.11	3.18	9.88	3.06	1.52	3.34	10.08	2.77	1.80	2.93	10.15	3.13	2.47	3.34
Sortieren insgesamt	6	2.30	1.33	0.47	1.85	2.54	1.44	0.81	1.82	2.52	1.50	1.38	1.76	2.73	1.39	1.64	1.71
Klassifikation insgesamt	24	10.57	4.15	1.67	4.41	11.98	4.63	2.96	4.78	12.06	4.68	4.56	4.97	12.88	4.60	5.32	4.67
Puzzles insgesamt	5	1.88	1.40	0.43	1.86	2.04	1.37	0.80	1.64	2.15	1.28	0.83	1.58	2.41	1.45	0.98	1.57

$\overline{X}$ = arithmetisches Mittel

s = Standard-Abweichung

* Die Differenz zwischen dem maximal möglichen Punktwert für den Gesamttest und der Summe der maximal möglichen Punktwerte der einzelnen Haupttests ergibt sich dadurch, daß einige Testaufgaben in mehreren Tests verwendet wurden.

Tabelle 2

Vortest- und Nachtestpunktwerte für sozial benachteiligte Kinder, die zu Hause der Sendung zuschauten

(nach der Häufigkeit des Zuschauens in Quartilen unterteilt)

Gruppe 1 = Kinder, die zum Zeitpunkt des *Vortests* 53–58 Monate alt waren

Gruppe 2 = Kinder, die zum Zeitpunkt das *Nachtests* 53–58 Monate alt waren

(Alterskohorten)

Haupttests (ohne Untertests)	Maximal möglicher Punktwert *	Q 1				Q 2				Q 3				Q 4			
		Gruppe 1 N = 31 Vortest		Gruppe 2 N = 26 Nachtest		Gruppe 1 N = 33 Vortest		Gruppe 2 N = 33 Nachtest		Gruppe 1 N = 27 Vortest		Gruppe 2 N = 18 Nachtest		Gruppe 1 N = 23 Vortest		Gruppe 2 N = 24 Nachtest	
		$\overline{X}$	s	$\overline{X}$	s	$\overline{X}$	s	$\overline{X}$	s	$\overline{X}$	s	$\overline{X}$	s	$\overline{X}$	s	$\overline{X}$	s
Gesamttestergebnisse	203	76.77	22.27	88.42	21.83	81.97	18.90	101.70	24.78	90.37	25.21	130.33	29.59	99.04	36.42	139.33	35.99
Körperteile insgesamt	32	17.87	6.49	21.04	6.01	20.24	5.74	22.91	5.84	21.93	5.57	26.83	3.73	22.87	5.51	26.75	4.58
Buchstaben insgesamt	58	14.06	6.45	14.65	3.91	13.09	3.65	18.24	6.82	14.81	5.90	26.83	11.89	18.52	11.33	31.92	14.18
Formen insgesamt	20	7.45	3.36	11.04	3.43	9.09	3.21	11.21	3.27	9.93	4.08	14.22	3.61	10.35	4.21	15.46	3.91
Zahlen insgesamt	54	16.77	7.06	19.00	7.64	17.97	7.10	23.76	9.63	20.37	9.42	32.67	10.67	23.96	12.42	35.54	11.77
Parallelisierter Untertest	11	7.97	2.93	9.31	1.85	8.45	1.99	9.97	1.16	8.78	2.28	10.33	0.59	9.17	1.67	10.00	1.50
Beziehungsverhältnisse insges.	17	9.61	2.35	10.65	2.78	10.33	2.98	11.30	2.27	10.81	2.32	12.39	2.48	10.26	3.77	18.00	2.52
Sortieren insgesamt	6	2.13	1.38	2.69	1.41	1.67	1.29	3.33	1.49	2.81	1.55	4.28	1.32	2.30	1.22	4.54	1.25
Klassifikation insgesamt	24	10.71	3.84	11.96	4.25	11.03	2.91	13.79	4.25	12.89	4.50	17.78	4.10	13.04	5.06	17.75	5.14
Puzzles insgesamt	5	2.03	1.56	2.31	0.93	2.55	1.37	2.55	1.39	2.26	1.02	3.44	1.38	2.52	1.44	2.92	1.35

$\overline{X}$ = arithmetisches Mittel

s = Standard-Abweichung

* Die Differenz zwischen dem maximal möglichen Punktwert für den Gesamttest und der Summe der maximal möglichen Punktwerte der einzelnen Haupttests ergibt sich dadurch, daß einige Testaufgaben in mehreren Tests verwendet wurden.

Tabelle 3

Vortestergebnis und Punktzuwachs für alle sozial benachteiligten 3jährigen Kinder

(nach der Häufigkeit des Zuschauens in Quartilen unterteilt)

N = 127

Haupttests (ohne Untertests)	Maximal möglicher Punktwert *	Q 1 N = 33				Q 2 N = 38				Q 3 N = 25				Q 4 N = 31			
		Vortest		Punktzuwachs		Vortest		Punktzuwachs		Vortest		Punktzuwachs		Vortest		Punktzuwachs	
		$\overline{X}$	s	$\overline{X}$	s	$\overline{X}$	s	$\overline{X}$	s	$\overline{X}$	s	$\overline{X}$	s	$\overline{X}$	s	$\overline{X}$	s
Gesamttestergebnisse	203	60.76	20.34	12.42	25.67	62.42	20.82	30.71	21.14	65.48	15.76	37.20	28.28	75.81	25.14	57.23	25.66
Körperteile insgesamt	32	13.88	5.21	3.03	6.26	15.76	5.77	4.79	5.91	16.72	5.44	6.64	6.94	18.84	6.26	8.00	5.52
Buchstaben insgesamt	58	10.73	5.99	3.79	9.20	10.18	4.95	7.35	8.99	11.32	3.99	10.52	9.71	11.91	6.65	20.13	12.14
Formen insgesamt	20	7.70	3.16	1.03	3.83	7.84	3.90	3.39	3.96	7.36	2.81	5.00	4.25	9.13	3.50	6.29	3.59
Zahlen insgesamt	54	11.21	6.40	2.94	9.34	11.37	6.08	9.34	7.53	13.00	5.39	8.08	10.02	16.38	8.39	14.13	9.79
Parallelisierter Untertest	11	6.94	2.70	0.94	3.43	6.53	3.33	3.05	3.04	7.00	2.68	2.40	2.72	8.25	2.53	2.03	2.74
Beziehungsverhältnisse insges.	17	7.42	2.46	1.39	3.55	8.45	3.13	1.79	3.46	8.24	2.62	1.76	3.44	8.72	2.39	3.23	2.70
Sortieren insgesamt	6	2.33	1.29	−0.12	1.73	2.21	1.36	0.42	1.73	2.44	1.26	0.92	1.85	2.41	1.10	1.52	1.59
Klassifikation insgesamt	24	8.67	3.53	1.27	3.59	8.50	4.43	4.53	4.69	9.12	3.48	4.44	4.81	10.56	4.66	5.71	3.68
Puzzles insgesamt	5	1.76	1.28	0.21	1.85	1.63	1.10	0.45	1.43	1.28	1.02	1.24	1.48	2.03	1.49	1.19	1.60

$\overline{X}$ = arithmetisches Mittel

s = Standard-Abweichung

* Die Differenz zwischen dem maximal möglichen Punktwert für den Gesamttest und der Summe der maximal möglichen Punktwerte der einzelnen Haupttests ergibt sich dadurch, daß einige Testaufgaben in mehreren Tests verwendet wurden.

Tabelle 4

Vortestergebnis und Punktzuwachs für alle sozial benachteiligten 4jährigen Kinder

(nach der Häufigkeit des Zuschauens in Quartilen unterteilt)

N = 433

Haupttests (ohne Untertests)	Maximal möglicher Punktwert *	Q 1 N = 127				Q 2 N = 119				Q 3 N = 105				Q 4 N = 82			
		Vortest		Punktzuwachs		Vortest		Punktzuwachs		Vortest		Punktzuwachs		Vortest		Punktzuwachs	
		$\overline{X}$	s	$\overline{X}$	s	$\overline{X}$	s	$\overline{X}$	s	$\overline{X}$	s	$\overline{X}$	s	$\overline{X}$	s	$\overline{X}$	s
Gesamttestergebnisse	203	75.13	22.21	18.24	18.40	84.09	23.25	30.60	24.35	86.63	23.64	38.50	25.44	93.79	29.50	49.01	24.62
Körperteile insgesamt	32	18.35	6.22	4.09	5.31	20.08	6.22	4.92	5.47	21.14	5.85	4.64	5.19	22.27	5.75	5.10	4.50
Buchstaben insgesamt	58	13.20	5.92	3.45	6.37	13.94	6.08	8.46	9.17	14.56	5.67	12.02	11.17	17.33	8.84	15.37	10.45
Formen insgesamt	20	8.21	3.42	2.55	3.82	9.87	3.67	3.31	4.40	9.94	3.59	4.32	4.13	10.38	3.39	5.63	3.72
Zahlen insgesamt	54	15.82	6.89	5.69	6.32	18.72	7.96	8.84	8.83	19.08	8.85	11.37	9.81	21.95	10.43	14.65	8.65
Parallelisierter Untertest	11	7.81	2.77	1.35	2.76	8.41	2.28	1.49	2.40	8.98	2.05	1.05	1.94	9.46	1.31	0.77	1.33
Beziehungsverhältnisse insges.	17	8.99	2.79	1.02	3.11	9.78	2.70	1.65	3.33	9.99	2.62	1.95	3.11	9.70	3.22	2.80	3.50
Sortieren insgesamt	6	2.05	1.28	0.62	1.91	2.48	1.40	0.95	1.84	2.47	1.46	1.36	1.81	2.52	1.28	1.87	1.59
Klassifikation insgesamt	24	10.52	3.84	1.17	4.60	12.01	3.98	2.91	4.66	11.83	4.33	4.86	5.02	12.33	4.38	5.77	5.07
Puzzles insgesamt	5	1.86	1.44	0.32	1.84	2.10	1.37	0.78	1.69	2.17	1.24	0.79	1.58	2.19	1.33	1.01	1.58

$\overline{X}$ = arithmetisches Mittel

s = Standard-Abweichung

* Die Differenz zwischen dem maximal möglichen Punktwert für den Gesamttest und der Summe der maximal möglichen Punktwerte der einzelnen Haupttests ergibt sich dadurch, daß einige Testaufgaben in mehreren Tests verwendet wurden.

Tabelle 5

Vortestergebnis und Punktzuwachs für alle sozial benachteiligten 5jährigen Kinder

(nach der Häufigkeit des Zuschauens in Quartilen unterteilt)

N = 159

Haupttests (ohne Untertests)	Maximal möglicher Punktwert *	Q 1 N = 37				Q 2 N = 40				Q 3 N = 38				Q 4 N = 44			
		Vortest		Punktzuwachs		Vortest		Punktzuwachs		Vortest		Punktzuwachs		Vortest		Punktzuwachs	
		$\overline{X}$	s	$\overline{X}$	s	$\overline{X}$	s	$\overline{X}$	s	$\overline{X}$	s	$\overline{X}$	s	$\overline{X}$	s	$\overline{X}$	s
Gesamttestergebnisse	203	88.68	29.20	23.08	19.14	101.23	30.69	26.75	17.30	104.13	30.82	38.97	25.73	120.91	29.78	37.32	26.37
Körperteile insgesamt	32	20.38	7.15	3.92	6.68	23.35	4.34	2.93	4.98	23.18	6.02	4.08	5.34	25.73	4.40	3.41	3.55
Buchstaben insgesamt	58	14.97	5.59	6.35	8.45	18.40	10.05	8.70	9.70	18.79	8.98	13.66	11.64	24.16	12.71	14.32	11.71
Formen insgesamt	20	9.35	3.74	2.81	3.06	11.08	4.15	3.30	3.04	11.97	3.15	3.39	3.58	12.20	3.15	4.64	3.25
Zahlen insgesamt	54	21.00	10.71	5.95	6.87	23.53	11.37	7.58	6.54	25.89	11.87	11.18	9.41	31.89	10.12	9.66	9.93
Parallelisierter Untertest	11	8.84	2.61	1.05	2.84	9.48	1.72	0.70	1.64	9.97	1.05	0.32	1.49	9.96	1.19	0.66	1.27
Beziehungsverhältnisse insges.	17	10.81	3.28	0.97	2.85	11.28	3.44	1.18	3.56	11.11	2.66	1.58	2.34	12.02	2.62	1.25	3.05
Sortieren insgesamt	6	2.89	1.33	0.62	1.74	2.83	1.50	0.95	1.85	2.74	1.67	1.71	1.63	3.27	1.57	1.36	1.87
Klassifikation insgesamt	24	12.05	5.07	3.19	4.08	14.28	4.74	2.45	5.08	14.05	4.98	4.13	4.64	15.49	4.24	4.18	4.66
Puzzles insgesamt	5	2.05	1.39	1.00	1.83	2.33	1.46	1.08	1.65	2.45	1.37	0.92	1.62	3.02	1.45	0.73	1.60

$\overline{X}$ = arithmetisches Mittel

s = Standard-Abweichung

* Die Differenz zwischen dem maximal möglichen Punktwert für den Gesamttest und der Summe der maximal möglichen Punktwerte der einzelnen Haupttests ergibt sich dadurch, daß einige Testaufgaben in mehreren Tests verwendet wurden.

Tabelle 6

Vortestergebnis und Punktzuwachs für alle sozial privilegierten Kinder

(nach der Häufigkeit des Zuschauens in Quartilen unterteilt)

N = 169

Haupttests (ohne Untertests)	Maximal möglicher Punktwert *	Q 1 N = 16				Q 2 N = 31				Q 3 N = 57				Q 4 N = 65			
		Vortest		Punktzuwachs		Vortest		Punktzuwachs		Vortest		Punktzuwachs		Vortest		Punktzuwachs	
		$\overline{X}$	s	$\overline{X}$	s	$\overline{X}$	s	$\overline{X}$	s	$\overline{X}$	s	$\overline{X}$	s	$\overline{X}$	s	$\overline{X}$	s
Gesamttestergebnisse	203	95.44	23.90	26.69	16.04	102.13	21.65	38.65	17.02	112.77	24.36	40.46	18.83	110.83	25.63	45.25	22.87
Körperteile insgesamt	32	24.13	5.77	3.19	4.97	25.74	4.90	2.52	4.31	26.37	5.64	2.35	4.28	25.71	4.79	3.14	4.50
Buchstaben insgesamt	58	15.19	8.79	8.06	9.26	16.81	7.03	12.48	10.10	19.25	10.21	17.09	9.99	18.62	8.86	19.63	11.46
Formen insgesamt	20	10.63	3.48	3.00	4.23	11.35	3,20	4.32	2.74	12.37	3.05	3.88	3.59	12.31	3.15	4.62	3.39
Zahlen insgesamt	54	22.13	10.37	8.69	5.38	24.13	8.65	12.06	6.79	28.07	9.80	12.16	8.17	27.50	10.83	12.40	7.68
Parallelisierter Untertest	11	9.31	1.45	0.81	1.17	9.90	1.01	0.39	1.20	9.67	1.09	0.65	1.11	9.32	1.60	1.05	1.74
Beziehungsverhältnisse insges.	17	10.63	2.58	1.56	2.85	10.48	2.34	2.10	2.69	11.58	1.96	1.19	2.19	11.71	2.57	1.38	2.64
Sortieren insgesamt	6	2.75	1.34	0.50	1.41	2.81	1.22	1.52	1.29	2.98	1.41	1.65	1.83	2.86	1.41	1.75	1.54
Klassifikation insgesamt	24	11.50	3.12	3.69	5.33	14.03	3.56	4.97	4.01	15.19	4.21	4.58	4.95	15.11	4.23	4.55	4.27
Puzzles insgesamt	5	2.75	1.18	0.13	0.96	2.23	1.15	1.23	1.41	2.93	1.42	0.79	1.59	3.15	1.21	0.48	1.60

$\overline{X}$ = arithmetisches Mittel

s = Standard-Abweichung

* Die Differenz zwischen dem maximal möglichen Punktwert für den Gesamttest und der Summe der maximal möglichen Punktwerte der einzelnen Haupttests ergibt sich dadurch, daß einige Testaufgaben in mehreren Tests verwendet wurden.

Richard C. Anderson

Eine vergleichende Felduntersuchung: Ein Beispiel vom Biologieunterricht in der Sekundarstufe[1]

Eine gebräuchliche, aber wenig sinnvolle Form der pädagogischen Forschung ist der Versuch, verschiedene Unterrichtsmethoden miteinander zu vergleichen. In den letzten Jahren gab es zahlreiche Vergleiche zwischen Vorträgen, die über das Fernsehen ausgestrahlt wurden, und Vorträgen, die direkt vor den Adressaten gehalten wurden, zwischen forschendem Lernen und darstellendem Lehrervortrag, zwischen schüler- und lehrerzentriertem Unterricht, zwischen programmiertem und Lehrbuch-Unterricht usw. Der Unterricht selbst war bei diesen Untersuchungen nur von geringer Bedeutung. Er war lediglich das Vehikel zur Evaluation einer Unterrichtsmethode; man nahm an, man könne die dabei erzielten Ergebnisse auf beliebige Unterrichtsinhalte übertragen. Gegenwärtig gibt es meiner Meinung nach eine allgemeine Übereinstimmung darüber, daß diese Annahme ungerechtfertigt war (Cronbach 1963; Lumsdaine 1965). Nichtsdestoweniger will ich darlegen, daß die vergleichende Untersuchung, wenn man sie anders einsetzt, einen Teil des Aufwands an Zeit und Mühe in der pädagogischen Forschung verdient.

Die Begründung lautet etwa so: Unsere Fähigkeit, die für die Schüler beste Unterrichtsart vorherzusagen, ist gering. Es gibt keine Unterrichtsmethoden, die sich gegenüber anderen Methoden stets als besser erwiesen haben. Es gibt keine Unterrichtsmerkmale, die notwendigerweise mit einer besseren Schülerleistung verknüpft sind. Weder kleine Lernschritte noch aktives Antworten, noch sofortige Leistungskontrolle und Erfolgsbestätigung, noch ein gutes Klassenklima, noch das stufenweise Fortschreiten vom Konkreten zum Abstrakten, noch die Möglichkeit, die Richtung und die Geschwindigkeit des Lernens selbst zu bestimmen, noch der Einsatz von multimedialen Stimuli garantieren einen erfolgreichen Unterricht.

Dies ist keine erfreuliche Perspektive; aber es ist meiner Ansicht nach keine Übertreibung. Gewiß haben wir hierüber einige Kenntnisse; doch gibt es mehr Probleme, über die wir nichts Genaues wissen. Meiner Meinung nach können wir zur Zeit die Effektivität eines Unterrichts nicht zu-

verlässig voraussagen, auch wenn die philosophischen Grundlagen, der Stil, die Methoden und die Verfahren des Unterrichts bekannt sind.

Wenn dies richtig ist, stellt sich folgende Frage: Wie sollen die Finanzen der Geldgeber und die Zeit und Mühe der pädagogischen Forscher eingesetzt werden, um die Effektivität des Unterrichts heute und in der Zukunft zu maximieren? Eine Antwort, die ich unterstützen würde, ist die Investition in pädagogische und verhaltenswissenschaftliche Grundlagenforschung. Man sollte jedoch die Wirkung der Grundlagenforschung auf die Unterrichtspraxis realistisch beurteilen.

Innerhalb der Verhaltenswissenschaften gibt es gegenwärtig eine stark ausgeprägte empiristische Tendenz (Conant 1952). Dies gilt insbesondere für die angewandten Wissenschaften, die sich von der Verhaltenswissenschaft Anregungen erhoffen. Pädagogische Grundlagenforschung sollte uns immer mehr dazu befähigen, ohne vorherige Erprobung die Unterrichtsverfahren und die Organisation von Curriculummaterial genau zu bestimmen, die mit großer Wahrscheinlichkeit das Lernen der Schüler fördern. Mit einem allmählichen Fortschritt kann man rechnen. Aber es wäre unrealistisch, zu erwarten, daß wir jemals eine effektive Unterrichtsgestaltung mit mehr als geringer Wahrscheinlichkeit vorhersagen können. Ich zweifle nicht daran, daß die Curriculumentwicklung immer teilweise auf Regeln beruhen wird, die über den Daumen gepeilt sind. Ich zweifle auch nicht daran, daß viele Versuche nach dem Prinzip des »Trial and Error« immer notwendig sein werden, um erfolgreichen Unterricht zu *gewährleisten.*

Bisher habe ich dargelegt, daß wir nur in geringem Maße die Merkmale des Unterrichts vorhersagen können, die den Lernerfolg der Schüler maximieren, und daß man von der Grundlagenforschung erwarten kann, daß sie auch nur bescheidene Verbesserungen in dieser Hinsicht leisten kann. Doch etwas sollten wir jetzt tun: Wir können in Vorversuchen und Felduntersuchungen Unterrichtseinheiten unterscheiden, die sich gut oder schlecht für den Unterricht eignen. Daher sollten wir von dieser Möglichkeit, den Unterricht zu verbessern, Gebrauch machen. Aus diesem Grunde sollte Unterricht durch Schülerleistungen evaluiert werden, und jeder einzelne Schritt in der Entwicklung des Curriculummaterials sollte die Schülerleistungen berücksichtigen. Auf der Grundlage des vorhandenen Wissens ist es nicht möglich, Unterrichtsmethoden zu evaluieren, aber es ist möglich, dies bei einzelnen Unterrichtsstunden, Unterrichtseinheiten oder Curricula zu tun.

Zufriedenstellenden Unterricht kann man durch die systematische Anwendung des Prinzips des »Trial and Error« entwickeln. Dieser Prozeß erfordert die Bestimmung der Lernziele, die Vorbereitung von Curriculummaterialien, die diesen Lernzielen (hoffentlich) entsprechen und schließ-

lich die Erprobung der Curriculummaterialien mit den Adressaten. Auf der Grundlage erfolgreicher Versuche werden die Materialien dann überarbeitet. Der Prozeß von Versuch und Überarbeitung wird so lange fortgesetzt, bis die Lernziele erreicht sind oder die Entscheidung getroffen wird, daß es unmöglich ist, sie im Rahmen der zur Verfügung stehenden Zeit und Mittel zu erreichen.

Die Funktion der Felduntersuchung

Wenn die Ergebnisse der Vortests erkennen lassen, daß die Schüler die Lernziele erreichen, ist es an der Zeit, das gesamte zusammengehörende Curriculummaterial einem Feldtest zu unterziehen. Das gesamte Curriculummaterial umfaßt nicht nur die Materialien, die direkt den Schülern gegeben werden, und Ausführungen und Anleitungen, die Hinweise für den Lehrer darüber enthalten, wie Diskussionen, Laborübungen und das Lösen von Aufgaben und Problemen zu leiten sind; sondern es kann auch Lehrerhandbücher, die Organisation von Lehrerseminaren und die Anleitung zu einem angemessenen Unterricht miteinschließen. Ein Ziel einer Felduntersuchung ist es festzustellen, ob sich unter verschiedenen Anwendungsbedingungen das gesamte Curriculummaterial als erfolgreich erweist. Der Vortest kann für die gesamte Schülerpopulation, die mit diesem Material arbeiten soll, repräsentativ sein; er muß es aber nicht sein. Die Vortests wurden unter Umständen von jemandem durchgeführt und beaufsichtigt, der von dem Projekt angetan war und der über die richtige Benutzung des Materials Bescheid wußte. Was geschieht aber, wenn die Materialien in die Hände von Lehrern gegeben werden, die ihnen gegenüber interesselos oder gar abweisend eingestellt sind? Müssen die Materialien auf eine bestimmte Weise benutzt werden, oder sind sie auch bei unterschiedlichen Anwendungsbedingungen einigermaßen erfolgreich? Wenn das Curriculum in einer bestimmten Weise benutzt werden muß, ist dann für Lehrerhandbücher oder für Lehrerseminare gesorgt? Und bringen die Handbücher oder Seminare die Lehrer mit Erfolg auf den angestrebten Weg? Dies sind einige der Fragen, die in einer Felduntersuchung beantwortet werden können.

Wenn man die von Scriven (1967) eingeführten Begriffe verwendet, so ist das Ziel von Vortests formative Evaluation, um Mängel im Verständnis oder in der Leistung der Schüler aufzuzeigen, so daß Herausgeber, Autoren oder Lehrer die Curriculummaterialien und die Unterrichtsmethoden überarbeiten und vermutlich verbessern können.

Es ist für die Felduntersuchung nur ein sekundäres Ziel, den Curricu-

lumentwicklern die Ergebnisse ihrer Arbeit vor Augen zu führen. Das Hauptziel ist summative Evaluation. Dabei werden Daten gesammelt, um möglichen Adressaten – wie Erziehungsinstitutionen, Beamten der Schulverwaltung, Lehrern und Schülern – bei der Entscheidung zu helfen, ob ein bestimmtes Curriculum benutzt werden soll oder nicht.

Einige Befürworter der empirischen Validierung von Curriculummaterialien scheinen die Ansicht zu vertreten, die Effektivität der erzielten Verhaltensänderungen bei Schülern sei bei der Beurteilung des Unterrichts das einzige Kriterium. Ich möchte betonen, daß dies nicht mein Standpunkt ist. Unterrichtsstunden, Unterrichtseinheiten und Curricula sollten danach beurteilt werden, in welchem Ausmaß sie ihre Ziele erreichen; aber dies sollte nicht das einzige Kriterium sein. Andere Kriterien sind die Kosten des Unterrichtsablaufs in Form von Zeit, die die Schüler und Lehrer aufwenden müssen, die Billigung des Unterrichtsablaufs seitens der Schüler und Lehrer und alle Nebeneffekte (Stake 1967a). Genauigkeit, Modernität und Einfallsreichtum der Lehrinhalte waren die wichtigen Kriterien der bekannten Curriculum-Reformprojekte. Ein sehr wichtiges Kriterium ist der Wert der Ziele, die der Unterricht zu erreichen anstrebt. Wie Scriven (1967) bemerkt hat, »ist es offensichtlich uninteressant, wie gut die Lernziele erreicht werden, wenn sie wertlos sind.« Die Umkehrung dieser Aussage ist ebenfalls richtig: Unabhängig davon, wie wertvoll die Ziele sind, kann ein Unterricht nicht positiv bewertet werden, wenn er so uneffektiv ist, daß er diese Ziele nicht erreicht. Effektivität sollte als Kriterium für die Beurteilung des Unterrichts weder über- noch unterschätzt werden.

Manchmal sollte die Felduntersuchung des gesamten Curriculummaterials eine vergleichende Untersuchung sein. Diese Schlußfolgerung ist unvermeidlich, wenn die Felduntersuchung die Entscheidungen der Adressaten mitbestimmen soll. Es gibt in den Bereichen des schulischen Gesamtcurriculum verschiedene alternative Curricula zur Auswahl. Für den Fall, daß sich die Lernziele und -inhalte verschiedener Curricula überschneiden, ist für die Entscheidung in der Praxis durchaus die Frage angebracht, welches das effektivste ist.

Cronbach (1963) und Scriven (1967) haben zum Wert von vergleichenden Untersuchungen gegensätzliche Positionen bezogen. Bis auf eine Einschränkung stimme ich mit Scriven überein. Vergleichende Untersuchungen haben sehr wohl eine wertvolle Funktion. Aber Scriven scheint für die Adressaten umfangreiche Vergleichsuntersuchungen von Curricula in jedem Fachbereich vor Augen zu haben. Hierzu hat Cronbach zu Recht die Gegenposition vertreten, daß die meisten Vergleiche wahrscheinlich keine Unterschiede von statistischer Signifikanz oder praktischer Bedeutung ergeben würden.

Vergleichende Untersuchungen sind kostspielig. Sie können nicht wahllos durchgeführt werden. Ein Kriterium für die Entscheidung über die Durchführung einer vergleichenden Untersuchung ist folgendes: Es muß eine erhebliche Wahrscheinlichkeit dafür bestehen, daß eines der Curricula in der Tat effektiver ist als das andere. Vermutungen haben in der Grundlagenforschung durchaus ihren Platz. Für eine vergleichende pädagogische Untersuchung kann dies aber nicht gelten. Aus der Sicht dessen, der eine vergleichende Untersuchung durchführt, sollte lediglich bewiesen werden, daß eines der Curricula besser ist als das andere.

In einer vergleichenden Untersuchung haben Ergebnisse, die keine Unterschiede zeigen, einen sehr geringen gesellschaftlichen Nutzen. Wenn man mit Nachdruck die Vorstellung zurückweist, daß eine vergleichende Untersuchung den generellen Wert einer Unterrichtsmethode zeigen kann, und die Vorstellung akzeptiert, daß die wichtigste Begründung für eine vergleichende Untersuchung darin liegen muß, zu bestimmen, welches von zwei oder mehreren Curriculummaterialien das effektivste ist, dann ist es offensichtlich sinnlos, Curriculummaterial auf die bloße Möglichkeit hin zu vergleichen, daß das eine besser als das andere sein könnte; es sei denn vielleicht, man glaube, es gäbe viele gute Curricula, die unbeachtet herumliegen und darauf warten, entdeckt zu werden. Vielleicht ist die Feststellung von einem gewissen Wert, daß eine groß propagierte curriculare Innovation nicht effektiver ist als ein anderes Curriculum. Im allgemeinen jedoch können ergebnislos verlaufende vergleichende Untersuchungen die Entscheidung der Adressaten nicht erleichtern. Daher muß ein Irrtum in der Beurteilung vorgelegen haben, wenn eine vergleichende Untersuchung keine Unterschiede aufzeigt. Zeit und Geld, die in die Curriculumentwicklung und in die formative Evaluation hätten investiert werden sollen, sind so zu einem voreiligen Vergleich verschwendet worden.

Es mag eingewandt werden, daß Forschung nicht damit gerechtfertigt werden kann, bloß zu beweisen, was ohnehin mit hoher Wahrscheinlichkeit vermutet wird. Das Gegenargument basiert auf der These, die bereits zuvor in diesem Beitrag entwickelt wurde. Es gibt von vornherein keine Tests, die verläßlich die Effektivität eines Unterrichts vorhersagen können; gleiches gilt für Experten, deren Fähigkeiten bei der Beurteilung der Unterrichtseffektivität anerkannt sind. Kurz gesagt, es gibt keine akzeptablen Gründe für Aussagen über die Effektivität eines Unterrichts außer Ergebnissen, die tatsächlich die Effektivität beweisen.

Die Notwendigkeit relativer Normen

Die Auffassung, daß Curriculumeinheiten in bezug auf absolute Effektivitätsnormen evaluiert werden sollten, ist weit verbreitet. In der Tat ist dies die Auffassung, die ich im Hinblick auf Voruntersuchungen von Curriculumeinheiten vertrete. Bei Felduntersuchungen von Curriculummaterialien gibt es Gründe, sich nur mit Vorsicht ausschließlich auf absolute Normen zu verlassen. Vor allem existieren in der Pädagogik im Gegensatz zu anderen Bereichen – von der Landwirtschaft bis zur Automobilindustrie – keine übereinstimmend akzeptierten Leistungsnormen.

Angenommen, die Pädagogen könnten sich auf irgendeine allgemeine Norm einigen, wie auf die bekannte 90–90 Norm, die vom Air Force Training Command unter der Leitung von Colonel Gabriel Ofiesh vorgeschlagen wurde [2], was würde es bedeuten, wenn die Schüler durchschnittlich 90 % einer kriteriumsbezogenen Norm erreichten? Offensichtlich würde das nicht bedeuten, daß die Schüler 90 % all jenes Wissens beherrschten, das über ein Thema bekannt ist. Es würde bedeuten, daß sie 90 % von dem gelernt haben, was jemand für den Unterricht und für den Test ausgewählt hat. Hier liegt das Problem. Ungeachtet jüngster Fortschritte bei der Formulierung von Lernzielen, können immer noch bedeutsame Unterschiede in dem beabsichtigten oder in dem impliziten intellektuellen Niveau auftreten, mit dem ein Begriff entwickelt wird, obwohl angeblich die gleichen Ziele zugrunde liegen. Ein weiteres Problem liegt darin, daß das Leistungsniveau von den Testmethoden abhängig ist; ein Beispiel hierfür ist die Attraktivität von Distraktoren bei Tests mit Auswahl-Antwort-Aufgaben. Endlich schließt die Tatsache, daß ein Curriculum eine bestimmte Effektivitätsnorm erreicht, die Möglichkeit nicht aus, daß ein konkurrierendes Curriculum diese Norm mit weniger Zeitaufwand und mit geringeren Kosten besser erfüllt. Deshalb sind relative Normen und damit verbunden auch vergleichende Untersuchungen notwendig, um die Effektivität von Curriculummaterialien zu beurteilen.

Ich möchte nicht mißverstanden werden: Meiner Meinung nach sind absolute Effektivitätsnormen im Prinzip gut. Ich hoffe, es wird möglich sein, die Theorie und die Technik der Bestimmung absoluter Normen zu verbessern. In Anbetracht unserer Unzulänglichkeit, absolute Normen zu definieren und Leistung in bezug auf sie zu messen, sollten für die nächste Zukunft absolute Normen durch relative Normen ergänzt werden. Zum gegenwärtigen Zeitpunkt ist der direkte Vergleich der einzige verläßliche Weg, zu bestimmen, welches von zwei Curricula effektiver ist.

Vergleichende Untersuchungen haben eine eindeutige Funktion, wenn verschiedene Unterrichtsstunden (Unterrichtseinheiten, Curricula) im we-

sentlichen die gleichen Ziele haben. Ist dies der Fall, dann ist das effektivste Unterrichtsprogramm das beste, vorausgesetzt, daß andere Faktoren wie z. B. die Kosten vergleichbar sind. Die Adressaten können bei der Auswahl unter verschiedenen Curricula ihre Aufmerksamkeit hauptsächlich auf die Ergebnisse einer vergleichenden Untersuchung richten. Überdies – und dies ist einer der Gründe, warum ich für vergleichende Untersuchungen eintrete – wird der Wettbewerb, bessere Curriculummaterialien zu erstellen, auch dazu beitragen, effektiveren Unterricht zu schaffen.

Mir erscheint es nicht so einsichtig, daß vergleichende Untersuchungen sinnvoll sind, wenn die Lernziele der Curricula verschieden sind. Eine andere ungeklärte Frage ist: Wer sollte vergleichende Untersuchungen durchführen, die Entwickler von neuen Curricula oder unabhängige Evaluatoren? Ebenso gibt es Fragen über die geeignete Planung und Durchführung vergleichender Untersuchungen. Ehe ich mich zu diesen Fragen ganz allgemein äußere, werde ich lieber versuchen, sie an Hand eines Beispiels aus der Praxis zu erläutern. Der Rest dieses Beitrags beschreibt eine vergleichende Felduntersuchung, die durchgeführt wurde, um die Effektivität von neuem Curriculummaterial zu beweisen.

Die Felduntersuchung eines Unterrichtsprogramms in Populationsgenetik

Die Entwicklung des experimentellen Curriculummaterials

Mit der Unterstützung der Biological Sciences Curriculum Study (BSCS) wurde ein Programm in Populationsgenetik zum Selbstunterricht erstellt, das im Fach Biologie in der Sekundarstufe verwendet werden sollte (Faust/Anderson/Guthrie/Drantz 1967). Bei der Entwicklung des Programms wurde, wie oben kurz beschrieben, vorgegangen. Als erster Schritt wurden die Lernziele definiert. Hierbei diente die Behandlung der Populationsgenetik in den Lehrbüchern der Biological Sciences Curriculum Study als Richtlinie. Zunächst wurde eine Versuchsfassung eines Teils des Programms erstellt. Dieser Programmteil wurde mit einer Reihe einzelner Schüler der Sekundarstufe und einem der Programmautoren erprobt, wobei dieser die Arbeit der einzelnen Schüler überprüfte. Nach Versuchen mit einigen Schülern wurden dann jeweils Überarbeitungen vorgenommen. Die restlichen Teile des Programms wurden ebenso entwickelt. Schließlich wurde das vollständige Programm mit kleinen Schülergruppen getestet. Erneut wurden Überarbeitungen vorgenommen. Während der gesamten Entwicklung des Programms wurde ein sehr ausführlicher kriteriumsbe-

zogener Leistungstest benutzt; dieser bestand in der Hauptsache aus offen formulierten Fragen, bei denen Probleme gelöst, Begriffe und Gesetze definiert und erläutert werden mußten. Die Schüler, die an den Voruntersuchungen teilnahmen, mußten für die Durchführung des kriteriumsbezogenen Tests fast ebenso viel Zeit aufwenden wie für die Durcharbeitung des Programms selbst. Im allgemeinen wurde ein Programmteil als zufriedenstellend betrachtet, wenn alle an der Voruntersuchung beteiligten Schüler 90 % oder mehr der kriteriumsbezogenen Testaufgaben dieses Abschnitts richtig lösten. Die Fassung des Programms, die in dem Experiment verwendet wurde, enthielt in 234 Abschnitten, ohne Gleichungen und graphische Darstellungen, 14 000 Wörter.

Der Unterricht in der Kontrollgruppe

Das Programm über Populationsgenetik wurde verglichen mit der Behandlung der Populationsgenetik in dem Lehrbuch »Biological Science: An Inquiry Into Life«, das von der Biological Sciences Curriculum Study verfaßt worden war; inoffiziell ist dieses Buch bekannt als »BSCS yellow version«. Der Text enthält etwa 7 900 Wörter, die sich unmittelbar auf Populationsgenetik beziehen. Das Lehrbuchmaterial wurde durch Laborübungen ergänzt, die ebenfalls von der Biological Sciences Curriculum Study vorbereitet worden waren; der Unterricht wurde von einem Biologielehrer einer Sekundarstufe gegeben. Es wäre falsch, den Unterricht in der Kontrollgruppe als konventionellen Unterricht zu bezeichnen. Dieses Material wurde von einem Team von Biologen und Biologielehrern erarbeitet. Das Lehrbuch wurde einer größeren Revision unterzogen, die teilweise auf systematisch gesammelten Äußerungen vieler Lehrer aus allen Teilen des Landes beruhte, die die experimentelle Fassung des Lehrprogramms benutzten. Es ist offensichtlich, daß es für die Schüler in der Sekundarstufe kein besseres Unterrichtsmaterial für Populationsgenetik gibt als das Lehrbuch BSCS yellow version und die dazu gehörenden Hilfsmittel.

Anlage der Untersuchung

An dem Experiment nahmen annähernd 750 Schüler der Sekundarstufe teil; sie wurden von 9 Lehrern in 30 Klassen in zwei in Vororten gelegenen Schulen unterrichtet. Alle 9 Lehrer unterrichteten zwischen 2 und 4 Klassen. Die Klassen wurden nach dem Zufallsprinzip ausgewählt und die Programme mit der Auflage verteilt, daß nach Möglichkeit die Hälfte der Klassen eines jeden Lehrers das Programm erhalten sollte und die andere Hälfte nicht. Ferner wurden zwei Parallelformen des Leistungstests ent-

wickelt. Innerhalb jeder Klasse erhielt die Hälfte der Schüler, die ebenfalls zufällig ausgewählt wurde, eine Form als Vortest und die andere als Nachtest. Für die verbleibende Hälfte der Probanden wurde umgekehrt verfahren. Diese Untersuchungsanlage lieferte Grunddaten und Informationen auf der Basis einer relativ großen Anzahl von Testaufgaben mit einem relativ geringen Zeitaufwand seitens der Schüler; auf diese Weise wurde auch der typische Wiederholungseffekt vermieden, der auftreten kann, wenn Schüler genau den gleichen Test wiederholen.

Durchführung der Untersuchung

Die beteiligten Lehrer waren bereit, den Vor- und Nachtest zu bestimmten Zeitpunkten durchzuführen. In der Zwischenzeit erklärten sie sich damit einverstanden, das Programm in den dazu bestimmten Klassen und nicht in anderen Klassen zu verwenden. Den Lehrern wurde gesagt: »Setzen Sie bitte das Programm so ein, wie es Ihrer Unterrichtserfahrung am besten entspricht.« Ein Mitglied des Projektteams sprach kurz mit den Lehrern über die Art und Weise der Testdurchführung und über die Aufzeichnung der Ergebnisse; er unternahm jedoch keinen Versuch, das Programm zu loben oder Empfehlungen zu geben, wie es benutzt werden sollte.

Die recht unstrukturierten Lehreranweisungen können im Hinblick auf die Fragestellung der Untersuchung verstanden werden. Ist ein Programm über Populationsgenetik zum Selbstunterricht eine nützliche Ergänzung zu anderen BSCS-Materialien zu diesem Thema? Wir wollten sehen, ob das Programm unter den Bedingungen des alltäglichen Unterrichts nützlich ist, da dies die Umstände sind, unter denen die Lehrer Curriculummaterial verwenden müssen.

Im nachhinein gibt es keinen Zweifel darüber, daß mit dem Programm bessere Gesamtergebnisse erzielt worden wären, wenn für die Lehrer als Teil des gesamten Curriculummaterials ein Lehrerhandbuch beigegeben worden wäre. Zu jener Zeit jedoch – und ich denke, es war gut so – entschieden wir, kein Handbuch zur Verfügung zu stellen. Ein Handbuch ist nur in dem Maße sinnvoll, wie die Lehrer die Anleitungen, die darin enthalten sind, befolgen. Unsere Erfahrung hat gezeigt, daß die Lehrer die Handbücher, die dem Curriculummaterial beigegeben sind, oft nicht lesen. Es wurde als wichtig erachtet, herauszufinden, wie anfällig das Curriculummaterial unter ungünstigen Anwendungsbedingungen ist. Ein Handbuch für die Lehrer wird gegenwärtig erstellt; die Ergebnisse des Versuchs werden mitbenutzt, um die Lehrer zu überzeugen, daß die im Handbuch enthaltenen Empfehlungen beachtenswert sind.

Stake (1967a) hat überzeugend nachgewiesen, daß eine gute Evaluation des Unterrichts eine vollständige Beschreibung seiner Implementation beinhalten muß. Eine solche Beschreibung war in unserem Falle besonders wichtig, da den Lehrern sehr viel Spielraum gelassen wurde. Die Lehrer machten sowohl in den Versuchsklassen als auch in den Kontrollklassen Aufzeichnungen, mit denen alle Aktivitäten und ihre zeitliche Dauer zwischen Vor- und Nachtest beschrieben wurden. Alle Laborübungen, alle sonstigen Übungen und alle Anweisungen zum Lesen wurden genau aufgezeichnet. Ebenso füllten die Lehrer einen Fragebogen aus, der sowohl offene als auch geschlossene Fragen enthielt, die sich auf die Einstellung der Lehrer und Schüler zu dem Programm, auf Techniken der Programmbenutzung und deren Verhältnis zur anderen Unterrichtsarbeit bezogen; ebenso wurde nach Stärken und Schwächen des Programms gefragt. Die Schüler beantworteten einen Fragebogen, der sich mit ähnlichen Themen beschäftigte.

Gesamtanalyse der Ergebnisse des Leistungszuwachses

Da ganze Klassen nach dem Zufallsprinzip für den Unterricht mit und ohne Programm ausgewählt wurden, war die Klasse die Beobachtungseinheit. Die Grundlage für die Varianzanalyse war der mittlere Leistungszuwachs der einzelnen Klassen. In die Klassendurchschnitte gingen die Punktwerte aller Schüler ein, die sich dem Vor- und Nachtest in der jeweils vorgeschriebenen Form unterzogen hatten. Eine Reihe einzelner Schüler und eine vollständige Klasse wurden aus der Analyse ausgeschieden, da sie diese Kriterien nicht erfüllten.

Es wurde eine Varianzanalyse mit ungewogenen Mittelwerten durchgeführt. Dabei waren die Verwendung oder Nichtverwendung des Programms und die Schule die Faktoren. Lediglich der Unterschied zwischen dem Unterricht mit und ohne Programm war signifikant [$F\,(1{,}25) = 20{,}59$, $p<(0{,}01)$]. Dabei wurde ein w^2 von .39 erreicht. Mit anderen Worten: Es waren 39% der Varianz des Leistungszuwachses in den Klassen dem Unterricht mit dem Programm zuzuschreiben. Der tatsächliche Leistungszuwachs betrug bei der Verwendung des Unterrichtsprogramms 4,62 Items; ohne Unterrichtsprogramm wurde ein Zuwachs von 3,01 Items erreicht. Relativ bedeutet dies, daß die Schüler, die mit dem Programm unterrichtet worden waren, gegenüber den anderen einen um 53% höheren Leistungszuwachs erzielten. Die absolute Differenz war jedoch nicht so groß, wie wir erwartet hatten. Die Gründe für das Fehlen einer größeren absoluten Differenz werden später erörtert.

Leistung als eine Funktion der Herkunft der Testaufgaben

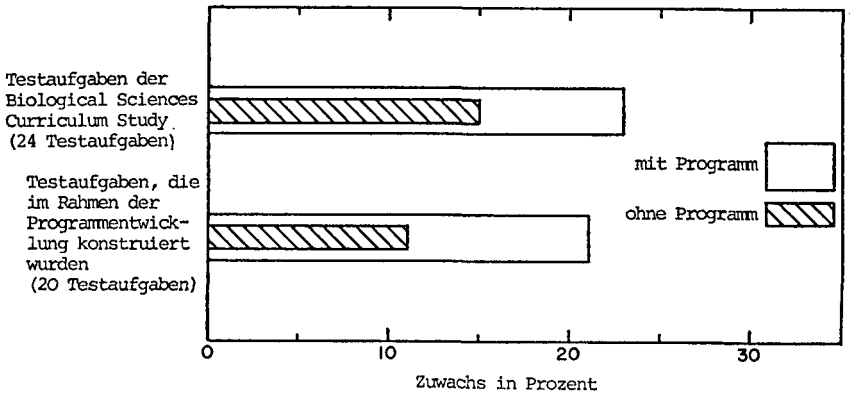
Während in der Felduntersuchung ein normenbezogener Test nach dem Auswahl-Antwort-Verfahren verwendet wurde, war der in den Voruntersuchungen des Programms verwendete Test kriteriumsbezogen; er war gekennzeichnet durch eigens konstruierte Testaufgaben. Für diese Änderung gab es zwei Gründe. Der erste war eine einfache Zweckmäßigkeitsüberlegung. Wir wollten nämlich die Schulen, die mit uns zusammenarbeiteten, nicht um die Zeit für einen längeren Test bitten. Der zweite und gewichtigere Grund war, die Glaubwürdigkeit der Ergebnisse in den Augen der Adressaten zu sichern, deren Entscheidung, das Programm zu verwenden oder nicht zu verwenden, diese Untersuchung beeinflussen sollte. In dem hier vorliegenden Fall ist die Biological Sciences Curriculum Study der unmittelbare Adressat. Diese Organisation hat viel Zeit und Geld darauf verwendet, Leistungstests zu Curriculumeinheiten zu entwickeln, die neben anderen Themen auch die Populationsgenetik zum Gegenstand haben. Da das Unterrichtsprogramm dafür vorgesehen war, die gleichen Lernziele zu erreichen, die sich auch die anderen BSCS-Materialien auf diesem Gebiet gesetzt hatten, konnte kaum ein überzeugender Einwand dagegen vorgebracht werden, diese Testaufgaben nicht zu verwenden, von denen Biologen und Biologielehrer annahmen, daß sie die Schülerleistung, bezogen auf diese Lernziele, gültig messen. Kriteriumsbezogene Tests sind die einzigen sinnvollen Tests für eine Unterrichtsevaluation; aber in diesem Falle war es von großer Wichtigkeit, die normenbezogenen Testaufgaben der BSCS zu verwenden, um den Verdacht zu vermeiden, die Überlegenheit dieses Programms beruhe lediglich auf eigens zugeschnittenen Testaufgaben.

In dem Leistungstest, der bei unserer Untersuchung Verwendung fand, wurden 24 Testaufgaben der BSCS-Tests, die sich mit Populationsgenetik befassen, aufgenommen. Es sollte betont werden, daß ein Schwierigkeitsgrad von fast 50 % nach der Durchführung des Unterrichts eines der Kriterien war, nach denen die Testaufgaben in die BSCS-Tests aufgenommen wurden. Zusätzlich wurden 20 Testaufgaben nach dem Auswahl-Antwort-Verfahren konstruiert, um eine noch größere Differenzierung zu erreichen. Abbildung 1 zeigt den Leistungszuwachs für den Unterricht mit und ohne Programm in Abhängigkeit von der Herkunft der Testaufgaben [3].

Leistung als eine Funktion der Unterrichtsinhalte

Die Lernziele des Programms können in drei Hauptgebiete klassifiziert werden:

Abbildung 1
Leistungszuwachs als Funktion der Herkunft der Testaufgaben



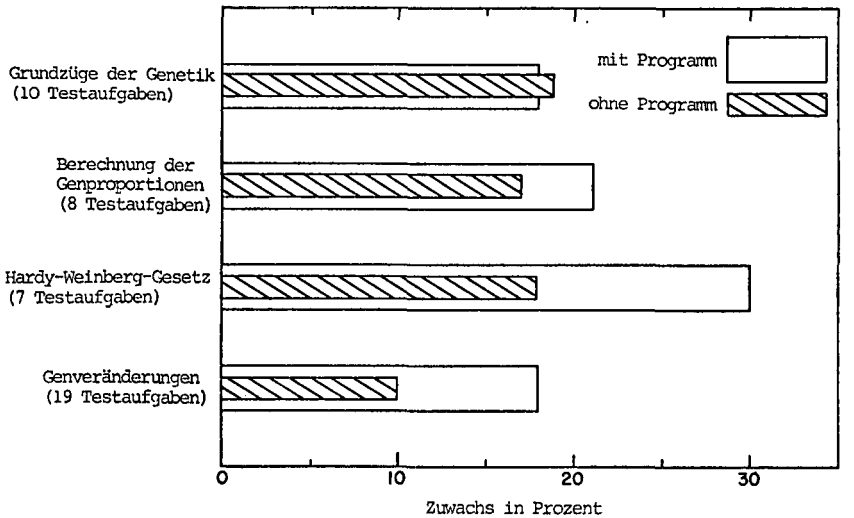
(1) die Berechnung der Genproportionen auf der Grundlage von ausgewählten Daten;
(2) die Logik des Hardy-Weinberg-Gesetzes;
(3) Faktoren, die eine Genveränderung bewirken (Mutation, Adaptation – Selektion, Wanderungssiebung, durch Zufall verursachter »genetic drift«, Paarungssiebung, Isolation).

Das Programm selbst mußte außerdem noch einen vierten Inhaltsbereich behandeln. Die Beherrschung der Mendelschen Gesetze ist für das Verstehen der Populationsgenetik von wesentlicher Bedeutung. Von dem Schüler wird angenommen, daß er die Grundzüge der Genetik beherrscht, bevor er mit dem Programm zu arbeiten beginnt. Da man sich auf die Zulänglichkeit des vorangegangenen Unterrichts nicht verlassen wollte, wurden zu Beginn des Programms die Grundzüge der Genetik durchgenommen. Abbildung 2 zeigt den Leistungszuwachs in den vier inhaltlichen Hauptbereichen.

Leistung als eine Funktion der Art der Testaufgaben

Eine der möglichen Schwächen in dem Verfahren, den Unterricht soweit zu verbessern, bis die Ergebnisse eines kriteriumsbezogenen Tests ein befriedigendes Niveau erreicht haben, ist die, daß dieses Verfahren zu einem einfachen Lehren auf den Test hin führen kann. Folgendes kann nämlich geschehen: Wenn eine Testaufgabe schlecht gelöst wird, so nehmen der Autor oder der Curriculumentwickler Sätze in den Unterricht auf, die die Antwort auf die Frage liefern. Oder er stellt vielleicht während des Un-

Abbildung 2
Leistungszuwachs als Funktion der Unterrichtsinhalte



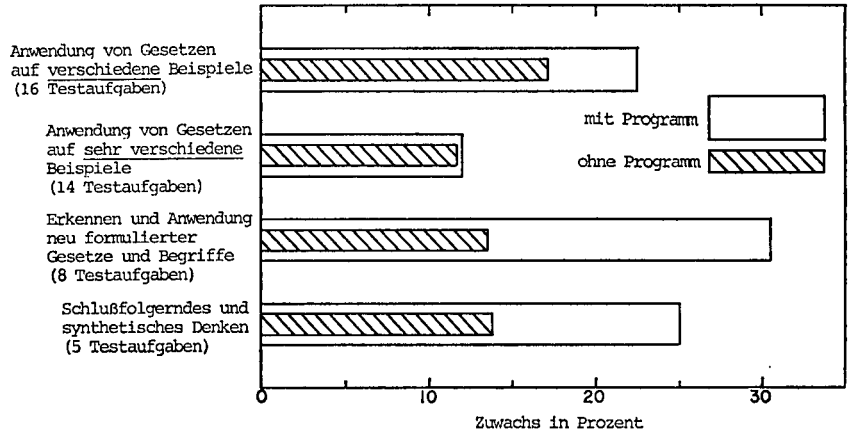
terrichtsverlaufs die Frage in einem Zusammenhang, in dem der Schüler die richtige Antwort finden muß. So muß sich auch bei einer schwierigen Testaufgabe das Ergebnis verbessern. Dessen ungeachtet, muß die Frage, was gelernt wurde, beantwortet werden. Es mag sehr wohl sein, daß die Schüler lediglich gelernt haben, eine Reihe von Wörtern zu wiederholen oder wiederzuerkennen. Definitionsgemäß versteht jemand einen Begriff oder ein Gesetz, wenn er alle möglichen Beispiele, die sich auf diesen Begriff oder auf dieses Gesetz beziehen, angemessen bearbeiten kann[4]. Wenn in einer Testaufgabe ein Beispiel verwendet wird, das während des Unterrichts gegeben wurde, kann dies lediglich ein verbales Wiederholen oder Wiedererkennen bedeuten. Wenn ein Schüler jedoch Testaufgaben richtig lösen kann, in denen Beispiele verwendet werden, die von denen *verschieden* sind, die im Unterricht gegeben wurden, ist die Folgerung durchaus angebracht, daß der Begriff von den Schülern verstanden wurde. Die Beispiele in den Testaufgaben können hinsichtlich ihrer Ähnlichkeit mit den Unterrichtsbeispielen skaliert werden. Kann jemand Fragen beantworten, die Beispiele enthalten, die sich nur wenig von den im Unterricht verwendeten unterscheiden, dann läßt sich sagen, daß er etwas von diesem Begriff oder Gesetz verstanden hat, während jemand, der Testaufgaben lösen kann, die im Vergleich zum Unterricht *sehr verschiedene* Beispiele enthalten, ein tiefes oder umfassendes Verständnis zeigt.

Begriffe können allgemein definiert werden; Gesetze können in abstrakter Sprache angegeben werden. Wenn ein Test im wesentlichen die Unterrichtssprache wiederholt, ist wiederum nur verbales Erkennen für eine richtige Antwort notwendig. Wenn ein Schüler jedoch angemessen mit Formulierungen eines Begriffs oder eines Gesetzes umgehen kann, die zwar wortmäßig verschieden sind, der Darstellung im Unterricht jedoch inhaltlich gleichen, deutet dies auf ein gewisses Verständnis.

Es ist ein Zeichen für synthetisches Denken, wenn ein Schüler eine Testaufgabe beantworten kann, deren Lösung die Anwendung von Begriffen und Gesetzen erfordert, die zu weit auseinanderliegenden Zeitpunkten im Unterricht behandelt wurden. Anderseits können diese Testaufgaben manchmal richtig gelöst werden, wenn der Schüler Schlüsse aus Aussagen zieht, die an einer Stelle während des Unterrichts gemacht wurden. Unter Verwendung der gerade beschriebenen Unterscheidungen wurde eine Inhaltsanalyse des Unterrichts und der Testaufgaben durchgeführt. Jede Testaufgabe wurde einer von fünf Kategorien zugeordnet. Zuordnungskriterien waren dabei die Ähnlichkeit der verwendeten Ausdrucksweise und die Ähnlichkeit der Aufgabenstellung zwischen Testaufgaben und den Aufgaben in den Programmen. Dabei wurde weder auf das Lehrbuch noch auf die Übungen noch auf den mündlichen Unterricht der Lehrer Rücksicht genommen. Ich muß darauf hinweisen, daß ich für die Verläßlichkeit der Zuordnung der Testaufgaben zu den einzelnen Kategorien nicht einstehen kann. Dies muß als ein grober, anfänglicher Versuch betrachtet werden, die Vorstellungen zu operationalisieren, die Pädagogen seit der Arbeit von Bloom und Mitarbeitern (1956) als bedeutsam ansehen (vgl. Anderson 1970 u. Anderson/Faust 1972). Abbildung 3 zeigt den Leistungszuwachs in der Versuchs- und Kontrollgruppe in Abhängigkeit von der Art der Testaufgaben. Da nach unserer Beurteilung nur eine Testaufgabe verbales Wiedererkennen maß, wurde diese Kategorie nicht in die Graphik aufgenommen.

Wie ich oben darlegte, können Testaufgaben Aufschluß geben über tiefes und umfassendes Verständnis, wenn sie Beispiele enthalten, die sehr verschieden von denen sind, die im Unterricht verwendet wurden. Wie man weiß, können die Anforderungen solcher Testaufgaben über die Lernziele eines bestimmten Curriculum hinausgehen. Während sie vermutlich in Leistungstests aufgenommen werden sollten, um Verständnisgrenzen feststellen zu können, ist Vorsicht bei der Beurteilung von ganzen Curriculummaterialien hinsichtlich ihrer Effektivität angebracht, sofern die Testaufgaben Beispiele enthalten, die von den Unterrichtsbeispielen sehr verschieden sind. Anders gesagt: Solche Testaufgaben erfassen eine weiterreichende Transferwirkung, die man nicht mit Sicherheit von einem Unterricht erwarten kann.

Abbildung 3
Leistungszuwachs als Funktion der Art der Testaufgaben



Leistung als eine Funktion des Lehrers

Es gab große Unterschiede darin, wie die Lehrer das Programm benutzten. Einige Lehrer billigten den Schülern überhaupt keine Unterrichtszeit zu, sich mit dem Programm zu befassen, während es auf der anderen Seite Lehrer gab, die die Populationsgenetik ausschließlich nach dem Programm lehrten. Tabelle 1 gibt die Anzahl der Minuten in den einzelnen Klassen wieder, die zwischen dem Vor- und Nachtest auf die verschiedenen Aktivitäten verwendet wurden. Diese Zahlen beruhen auf den Aufzeichnungen der Lehrer. Wir schlugen den Schulen einen zweiwöchigen Zeitraum zwischen Vor- und Nachtest vor. An einer Schule stimmten die Lehrer zu. Die Lehrer an der anderen Schule sagten: »Wir können dieses Curriculummaterial unmöglich in weniger als einem Monat durchnehmen«; sie erhielten deshalb einen Monat Zeit. Alle Lehrer in der Schule B berichteten, daß sie mit oder ohne Programm die gleiche Zeit für den Unterricht in Populationsgenetik aufgewendet hätten. Die Lehrer in der Schule A verwendeten bei der Benutzung des Programms für Populationsgenetik durchschnittlich etwa 10 % weniger Unterrichtszeit. Durchschnittlich gaben die Lehrer in den Klassen, in denen das Programm benutzt wurde, etwas weniger Seiten zu lesen auf als in Klassen, in denen das Programm nicht verwendet wurde.

Tabelle 1
Durchschnittliche Unterrichtszeit in Minuten (für die behandelten Themen) nach Schule und Art des Unterrichts

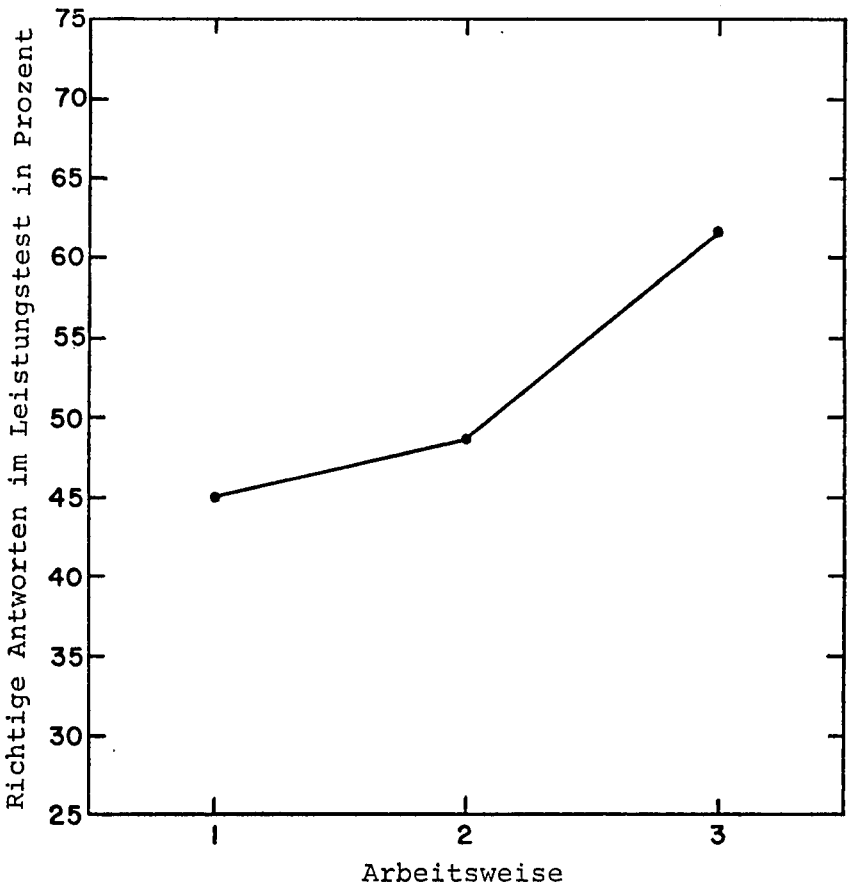
	Mit Programm		*Ohne Programm*	
	Schule A	Schule B	Schule A	Schule B
Durchschnittliche Unterrichtszeit zwischen Vor- und Nachtests	454	1031	454	1031
Zeit für Populationsgenetik				
mit Programm	151	0	0	0
andere Arbeit in Populationsgenetik . . .	52	451	228	451
insgesamt	203	451	228	451
Zeit für nicht populationsgenetisches Material	251	580	226	580

Die Lehrer wurden danach klassifiziert, wie sie das Programm den Schülern zuwiesen. Die erste Gruppe der Lehrer, wie in Abbildung 4 gezeigt wird, sorgte dafür, daß das Programm verfügbar war; sie forderten von den Schülern aber nicht, es durchzuarbeiten; auch war es nicht erlaubt, während der Unterrichtszeit damit zu arbeiten. Von der zweiten Gruppe wurde die Arbeit mit dem Programm verlangt, aber wiederum wurde keine Unterrichtszeit zur Verfügung gestellt, um damit zu arbeiten. Die Lehrer in der dritten Gruppe berichteten, daß sie das Programm zum festen Unterrichtsbestandteil gemacht hätten und für die Arbeit damit bis zu drei Unterrichtsstunden vorgesehen hätten. Die Ergebnisse zeigen jedoch, daß durchschnittlich etwa vier Stunden erforderlich sind, um das Programm durchzuarbeiten. Weniger als 20 % der Schüler berichteten, sie hätten das Programm in drei oder weniger Stunden bewältigt. Deshalb bearbeiteten die meisten Schüler, sofern sie es überhaupt taten, das Programm nicht in der Klasse.

Von großer Bedeutung sind die Durchschnittsergebnisse und deren Streuung. Ein F-Test ergab eine signifikant geringere Varianz im Leistungszuwachs bei den Lehrern von Klassen, die das Programm erhalten hatten, im Vergleich zu Lehrern von Klassen, bei denen dies nicht der Fall war. [$F(7,7) = 6{,}68$; $p < 0{,}05$, zweiendiger Test]. Überdies erreichte, wie man aus Abbildung 5 entnehmen kann, *jeder* Lehrer mit dem Programm mehr als ohne Programm [5].

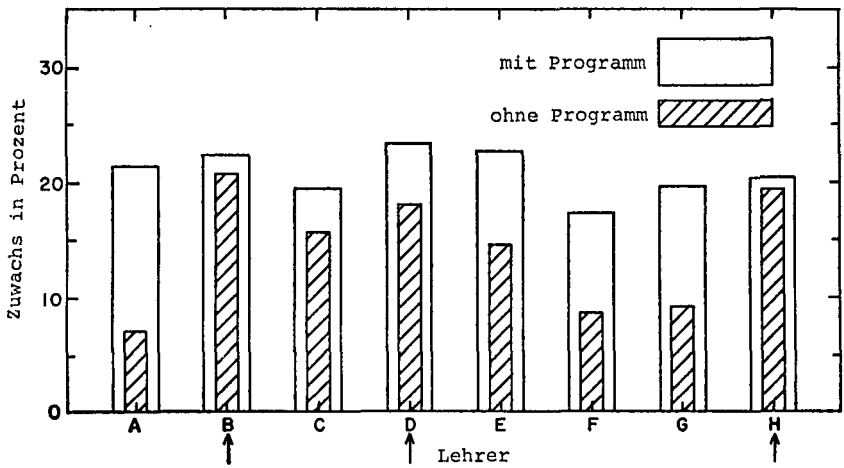
Die Lehrer wurden gefragt, ob das Programm ihren Unterricht ohne Programm beeinflußt hätte. Drei Lehrer (in Abbildung 5 mit einem Pfeil

Abbildung 4
Leistung als Funktion der Arbeitsbedingungen



markiert) gaben eine zustimmende Antwort. Lehrer D sagte: »Die Gliederung, die die Lehrer ihrem Unterricht zugrunde legten, und die Darstellung der Probleme in dem Programm wurden auch bei dem Unterricht in den Klassen verwendet, die das Programm nicht erhalten hatten.« Lehrer H äußerte sich folgendermaßen dazu: »Ich kann sagen, daß mir das Programm für alle meine Klassen geholfen hat, einen besseren Unterricht in Populationsgenetik zu geben. Ich verwertete viele Teile des Programms und fand, daß sie einen leichteren Zugang zu einem Thema ermöglichten, das andernfalls für viele Schüler schwierig gewesen wäre.«

Abbildung 5
Leistungszuwachs der Schüler bei den acht Lehrern



Leistung als eine Funktion der Schüler

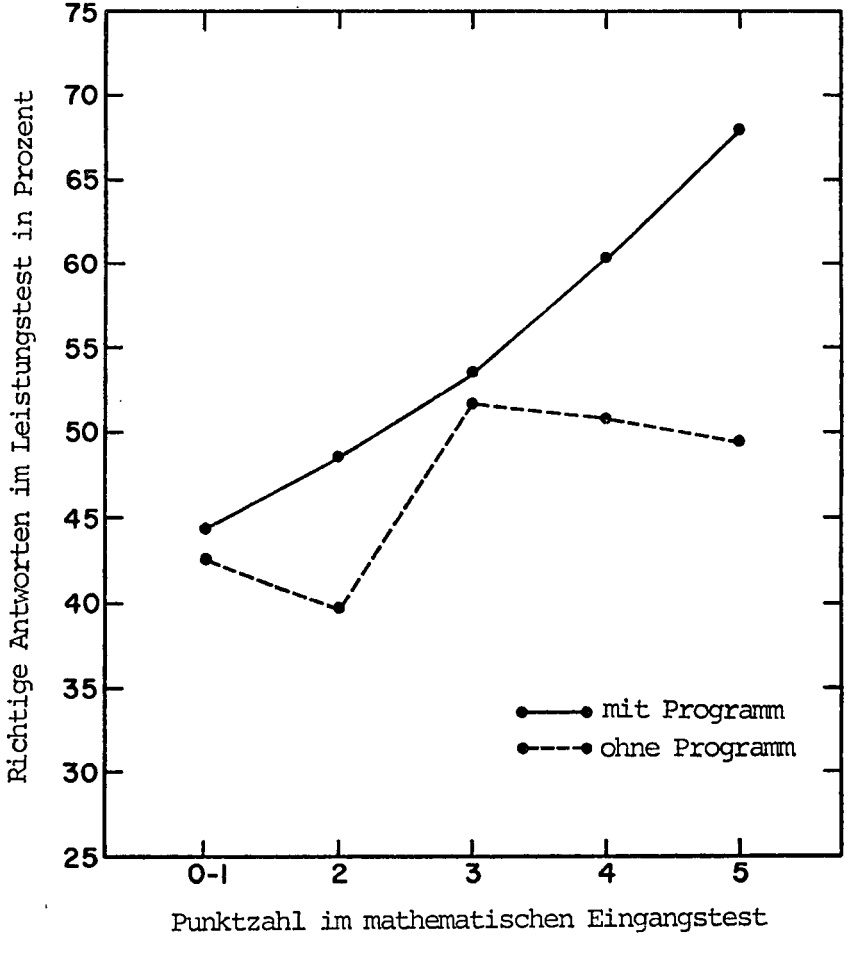
Jeder, der mit Unterrichtsplanung betraut ist, muß Voraussetzungen über den vorhandenen Kenntnisstand der Schüler machen, für die der Unterricht bestimmt ist. Das Programm in Populationsgenetik basiert auf der Voraussetzung, daß die Schüler, die damit arbeiten, mit Verhältniszahlen rechnen können und fähig sind, ein Binom zu quadrieren.

Nach Meinung vieler Pädagogen haben Programme zum Selbstunterricht bestenfalls den Wert, langsamen Schülern technisches Vokabular beizubringen. Eines der ursprünglichen Ziele dieses Projekts war es zu zeigen, daß Programme effizient benutzt werden können, um den besten Schülern eine Reihe von Gesetzen mit den dazugehörenden Begriffen zu vermitteln. Die Vermutung lag nahe, daß fast alle guten Schüler die erforderlichen mathematischen Fertigkeiten besaßen. Später wurde jedoch festgestellt, daß sehr viele gute Schüler, die die Hälfte der Stichprobe in der vergleichenden Untersuchung ausmachen sollten, zu der Zeit nicht erreichbar waren, zu der die Untersuchung durchgeführt werden sollte.

Mit den Schülern, die an der Untersuchung teilnahmen, wurde ein Eingangstest durchgeführt, der fünf Aufgaben enthielt. Zu unserer Bestürzung entdeckten wir, daß nur 40 % der Schüler in der Stichprobe die erforderlichen mathematischen Fertigkeiten besaßen, d. h. nur 40 % der Schüler beantworteten mindestens vier der Testaufgaben richtig. Abbil-

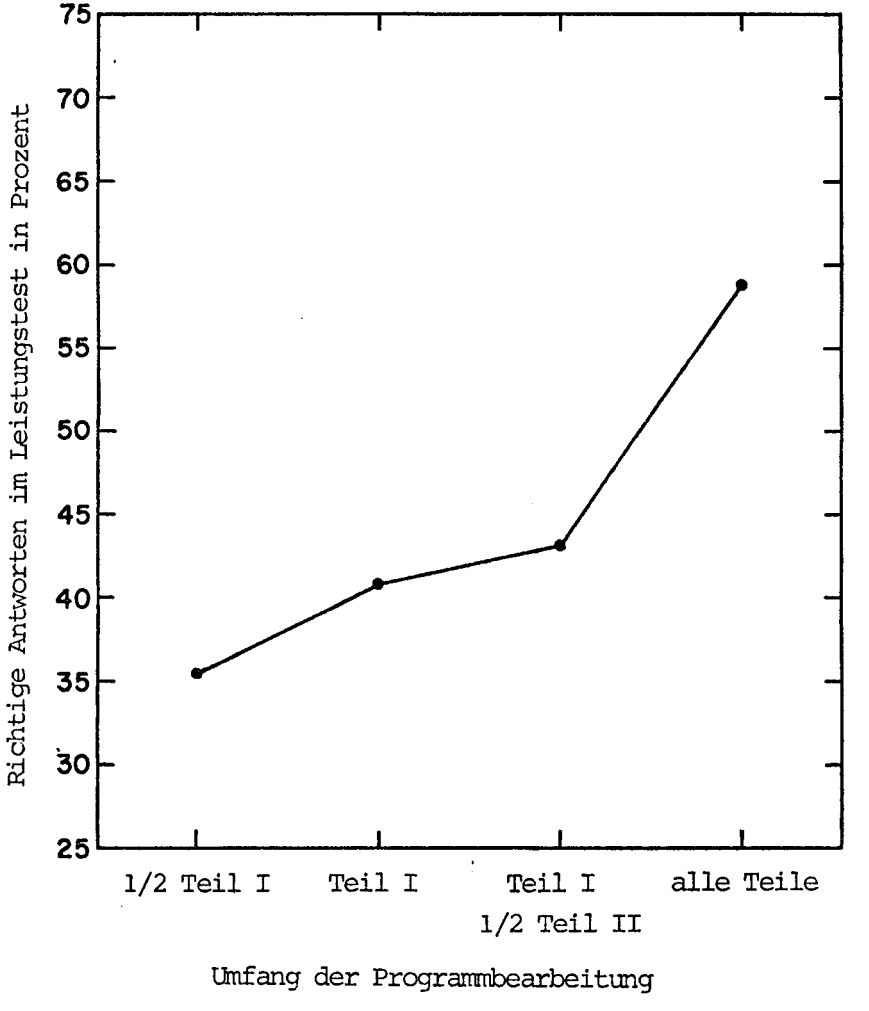
dung 6 zeigt die Leistung im Nachtest als eine Funktion der im Eingangstest festgestellten mathematischen Fertigkeiten. Das Programm war stets etwas effektiver, aber der Vorteil des Programms wirkte sich erheblich nur bei jenen Schülern aus, die in dem Eingangstest gute mathematische Fertigkeiten gezeigt hatten. Man konnte einen geringeren Leistungszuwachs bei den Schülern feststellen, die das Programm nicht erhielten, sogar bei denen, die die erforderlichen mathematischen Fertigkeiten besaßen.

Abbildung 6
Leistung als Funktion des mathematischen Eingangstests



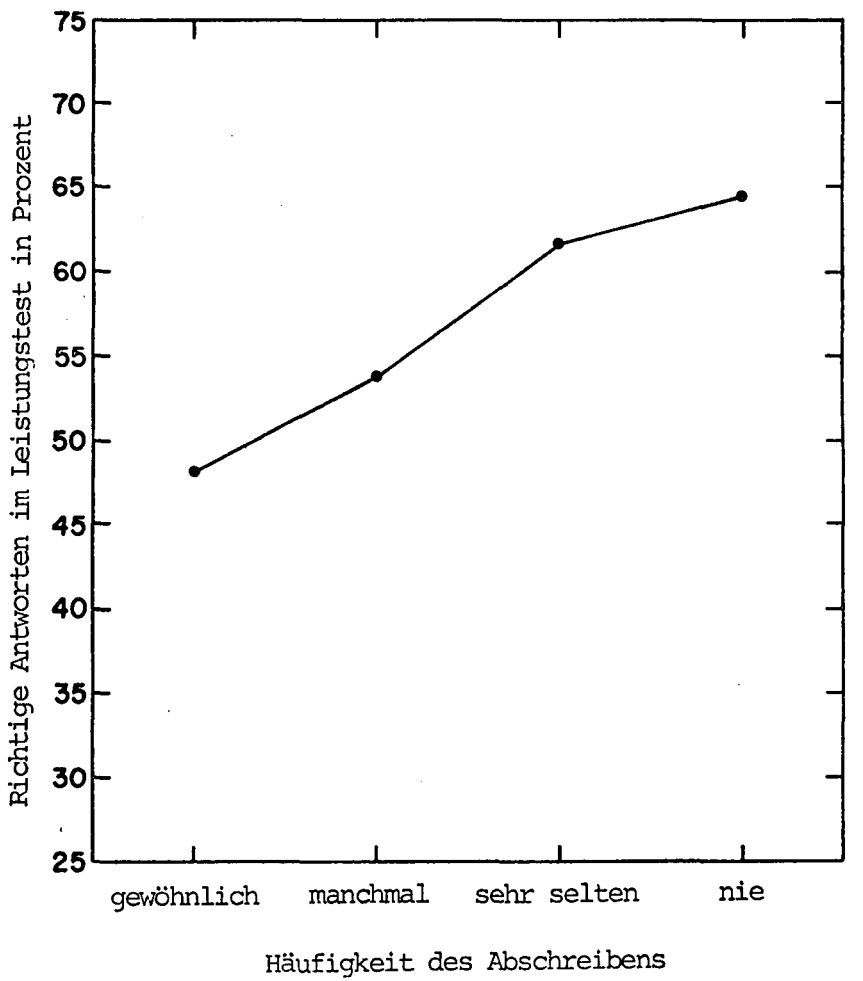
Im Fragebogen wurden die Schüler gebeten anzugeben, wie weit sie das Programm tatsächlich durchgearbeitet hatten. Etwa 75 % gaben an, das ganze Programm durchgearbeitet zu haben. Natürlich war die Leistung um so besser, je weiter das Programm durchgearbeitet worden war. Dieser Zusammenhang wird in Abbildung 7 gezeigt.

Abbildung 7
Leistung als Funktion des Umfangs der Programmbearbeitung



Wir wissen genau, daß Schüler von Programmen nicht viel lernen, wenn sie einfach die richtigen Antworten abschreiben (Faust/Anderson/Guthrie/Drantz 1967; Anderson/Faust 1968; Brown 1966; Kemp/Holland 1966). Die Schüler wurden danach gefragt, wie oft sie bei einer für sie schwierigen Frage die Seite umgedreht und die richtige Antwort abge-

Abbildung 8
Leistung als Funktion der berichteten Häufigkeit des Abschreibens richtiger Antworten bei schwierigen Abschnitten



schrieben hätten. Obwohl die Schüler in den Anweisungen zur Bearbeitung des Programms ermahnt wurden, jede Frage, *bevor* sie nach der richtigen Antwort schauten, schriftlich zu beantworten, gaben mehr als 40 % an, manchmal bei schwierigen Fragen Antworten abgeschrieben zu haben; 20 % gaben an, daß sie dies im allgemeinen taten. Abbildung 8 zeigt die Leistung in Abhängigkeit von der angegebenen Häufigkeit des unerlaubten Abschreibens richtiger Antworten.

Tabelle 2
Evaluation des Programms durch die Schüler
(N = 377)

Schüler in Prozent	Fragen
	1. Wenn ich die Wahl hätte,
71,6	(A) würde ich gerne öfter Programme benutzen, die dem Programm Populationsgenetik ähnlich sind
12,2	(B) wäre es mir gleich, welche Materialien benutzt würden
14,9	(C) würde ich es bevorzugen, keine Programme zu benutzen
1,3	keine Antwort
	2. Bei einem Vergleich eines Programms gleich dem in Populationsgenetik mit einem Lehrbuch meine ich, daß ich mit dem gleichen Aufwand an Zeit und Mühe
36,3	(A) mit dem Programm sehr viel mehr lernen würde
42,2	(B) mit dem Programm etwas mehr lernen würde
8,0	(C) gleich viel lernen würde
10,3	(D) mit dem Lehrbuch etwas mehr lernen würde
3,2	(E) mit dem Lehrbuch viel mehr lernen würde
	3. Wie sehr interessierte Dich das Programm in Populationsgenetik?
22,0	(A) Ich war sehr interessiert daran
45,9	(B) Ich war einigermaßen interessiert daran
22,3	(C) Ich verlor manchmal das Interesse
8,5	(D) Ich langweilte mich sehr
1,3	Keine Antwort
	4. Inwieweit verlangte das Programm in Populationsgenetik sorgfältiges Denken?
27,9	(A) Viele Seiten erforderten sorgfältiges Denken zur richtigen Beantwortung der Fragen
63,1	(B) Einige Seiten erforderten sorgfältiges Nachdenken
5,3	(C) Wenig Nachdenken erforderlich
1,9	(D) Das Programm war lächerlich einfach und verlangte fast kein Nachdenken
1,9	Keine Antwort

Die Einstellung der Lehrer und Schüler

Alle Lehrer empfanden das Programm als eine wertvolle Ergänzung des vorhandenen BSCS-Curriculummaterials über Populationsgenetik. Die Frage, ob sie das Programm wieder einsetzen würden, bejahten fünf von neun Lehrern; zwei Lehrer antworteten, daß sie es wahrscheinlich wieder benutzen würden; einer antwortete mit »wahrscheinlich nein«. Von einem Lehrer wurde diese Frage nicht beantwortet. Die Lehrer waren von dem Inhalt und der Organisation des Programms angetan; sie waren auch zufrieden mit dem Interesse, das das Programm bei den Schülern hervorrief. Zwei Lehrer gaben unaufgefordert die Auskunft, daß das Programm einen so guten Ruf hatte, daß sich einige Schüler in Klassen, die ohne das Programm unterrichtet wurden, Exemplare des Programms von ihren Schulkameraden entliehen.

Tabelle 2 faßt die Antworten der Schüler auf vier Fragen zusammen. Die meisten Schüler äußerten sich dahingehend, daß sie gerne wieder ein Programm wie das populationsgenetische benutzen würden, vorausgesetzt, daß sie mit diesem Programm mehr als mit einem Lehrbuch lernen und daß dieses Programm sie interessiert und sorgfältiges Denken verlangt.

Zusammenfassende Erörterung

Ein Ziel dieses Beitrags bestand darin, die Gültigkeit eines gesamten Curriculum nachzuweisen. Es galt zu zeigen, daß das neue Curriculummaterial, das in diesem Falle ein Programm zum Selbstunterricht in Populationsgenetik beinhaltete, effektiver als ein weit verbreitetes und sehr anerkanntes vergleichbares Curriculum ist.

Die Unterrichtseffektivität sollte sowohl in absoluten als auch in relativen Normen beurteilt werden. Der sich aus dem Test ergebende Durchschnitt der Gesamtleistung der Schüler, die das Programm erhalten hatten, betrug 53,6 % – kaum ein befriedigendes Ergebnis. (Der Durchschnitt für die Schüler, die das Programm nicht erhalten hatten, betrug 43,5 %).

Unter sehr günstigen Bedingungen jedoch führt das Programm zu besseren Leistungen. Alle Schüler, die den Eingangstest in Mathematik bestanden hatten und berichteten, sie hätten das Programm vollständig durchgearbeitet und vor der Beantwortung einer Frage nie oder selten nach der richtigen Antwort geschaut, erzielten einen Durchschnittswert von 70,5 %. Es ist wahrscheinlich, daß der Gesamtleistungsdurchschnitt höher als der in dieser Untersuchung festgestellte sein würde, wenn allen Schülern im Unterricht genügend Zeit gegeben würde, das Programm vollstän-

dig durchzuarbeiten; gleiches gilt für den Fall, daß das Durcharbeiten des Programms vom Lehrer gefordert, anstatt nur in das Belieben der Schüler gestellt wird; oder wenn die Lehrer ihren Forderungen mit den ihnen zur Verfügung stehenden Maßnahmen und Mitteln zur Lernmotivierung Nachdruck verleihen; oder wenn die Schüler dazu gebracht werden können, eine Antwort zu jeder Frage zu formulieren, bevor sie die richtigen Antworten nachschlagen; und ebenso gilt dies natürlich, wenn das Programm nur beim Unterricht mit den Schülern verwendet wird, die die erforderlichen mathematischen Fertigkeiten besitzen.

Zugegebenermaßen ist ein Leistungsniveau von 70 % (der maximal erreichbaren Punktzahl) unter optimalen Bedingungen kein überwältigendes Ergebnis [6]. Bei der Bewertung der erzielten Leistung ist jedoch zu berücksichtigen, daß nicht weniger als 25 % der Aufgaben in dem kriteriumsbezogenen Leistungstest über die Lernziele des Programms hinausgehen und daß fast 20 % der Aufgaben ein Thema betreffen (Grundzüge der Genetik), das zwar in dem Programm berücksichtigt, aber nicht explizit gelehrt wurde. Wenn man dies alles bedenkt, so ist das mit diesem Programm erzielte Leistungsniveau nicht schlecht, gleichgültig, ob man es relativ im Hinblick auf den Vergleichsunterricht oder in absoluten Maßstäben betrachtet.

Das Programm wird gegenwärtig auf der Grundlage der in der Felduntersuchung erzielten Ergebnisse und auf der Grundlage der Kritik der Genetiker und Biologielehrer überarbeitet.

Das Ziel dieses Beitrags war es, den Wert und die Bedeutung der vergleichenden Felduntersuchung nachzuweisen. Eine angebrachte Zurückhaltung bei der Erörterung des gegenwärtigen Wissensstands der Erziehungs- und Verhaltenswissenschaft, eine angemessene Beachtung der Komplexität des menschlichen Lernens und des Unterrichts und eine realistische Einschätzung der Möglichkeiten der Grundlagenforschung, unsere Fähigkeiten zu verbessern, eine effektive Unterrichtsgestaltung im vorhinein bestimmen zu können, lassen die Anwendung einer praxisbezogenen Strategie für die Entwicklung von Curriculummaterial als sinnvoll erscheinen. Der letzte Schritt in diesem Entwicklungsprozeß sollte eine Felduntersuchung sein, um empirisch die Effektivität des gesamten neuen Curriculummaterials zu beweisen. Es gibt keinen anderen Weg, um Effektivität gewährleisten zu können. Die Hauptaufgabe einer Felduntersuchung ist es, Ergebnisse zu liefern, aufgrund derer die Adressaten eine Entscheidung über die Annahme oder Ablehnung von Curricula fällen können. Wenn zwei Curricula die gleichen Lernziele haben (oder die gleichen Themen behandeln), sollte die Felduntersuchung aus einem Vergleich bestehen. Es genügt nicht zu zeigen, daß ein neues Curriculum die von irgendjemand gesetzten absoluten Effektivitätsnormen erfüllt, weil konkurrie-

rende Curricula diese Normen übertreffen oder die gleichen Normen mit weniger Zeitaufwand oder mit geringeren Kosten erfüllen können oder weil sie von Schülern und Lehrern vorgezogen werden.

Man hat oft gefordert, Unterricht empirisch zu validieren. Zum gegenwärtigen Zeitpunkt gibt es nur wenig Anzeichen, daß jemand hiervon überzeugt worden ist. Mein letztes Wort richtet sich an Autoren, Herausgeber und Verleger, die sagen, sie hätten besseres Curriculummaterial entwickelt. Warum sollte man ihnen glauben? Wo ist der Beweis für ihre Behauptungen? Das Erziehungswesen würde einen sehr großen Schritt vorankommen, wenn die Produzenten von Curriculummaterial es sich zur Regel machten, ihre Produkte empirisch zu validieren, und wenn es üblich wäre, daß die Adressaten eine solche Validierung als Voraussetzung für die Verwendung des Curriculummaterials verlangen würden.

William W. Cooley

Methoden der Evaluation von Schulinnovationen

Es gibt viele Aufsätze über Evaluationsmodelle, Evaluationsstrategien und Handlungsrezepte. Es gibt auch zahlreiche Versuche, Evaluationstaxonomien zu entwickeln. Aber es fehlen gut zugängliche Veröffentlichungen, die über die Verfahren und Ergebnisse wirklicher Evaluationsuntersuchungen berichten. Entweder werden die Ergebnisse niemals gedruckt, oder sie haben, wenn sie gedruckt sind, das Format großer Telefonbücher, die nur in geringer Zahl aufgelegt werden können und die in der Regel als Wanddekorationen im Erziehungsministerium enden.

Solange diese Berichte nicht allgemein zugänglich gemacht und von anderen Forschern kritisch untersucht werden können, lassen sich Evaluationsuntersuchungen kaum verbessern. Bei den Vorarbeiten für diesen Beitrag bin ich daher zur Überzeugung gekommen, daß man *nicht* einen weiteren Aufsatz *über* Evaluation, sondern die Beschreibung *einer* Evaluation braucht. Aus ihr müßte hervorgehen, wie ein Forscher sich bemüht, Daten zu erheben, aus denen sich eindeutige Informationen über den Wert neuer Unterrichtsmaterialien und pädagogischer Verfahren gewinnen lassen.

Über Evaluation läßt sich lediglich sagen, daß die Evaluation von Schulinnovationen insofern gute Forschung sein muß, als Forschung der Prozeß ist, in dessen Verlauf die Gültigkeit einer Hypothese bewiesen werden muß. Nach meiner Überzeugung unterscheidet sich evaluative Forschung von Grundlagenforschung und von weiten Bereichen angewandter Forschung nur in der Art der Hypothesen und darin, wie diese zu Beginn der Untersuchung formuliert werden. In der Grundlagenforschung beruhen die Hypothesen, die untersucht werden sollen, auf einer Theorie und einem entsprechenden System aufeinander bezogener Gedankengänge. In der angewandten Forschung stammen die Hypothesen, die untersucht werden sollen, aus der Anwendung der Wissenschaft und werden formuliert, wenn die abgesicherten Prinzipien, die diese Wissenschaft hervorgebracht hat, sich bei einer bestimmten Anwendung als inadäquat erweisen. Evaluative Forschung als eine Form der angewandten Forschung versucht, eher die

Gültigkeit der Hypothesen hinsichtlich *besonderer* Programme und Verfahren als die Gültigkeit der Hypothesen hinsichtlich allgemeiner, in vielen Programmen gleicher Variablen einzuschätzen. Den Bezugsrahmen für meine Ausführungen bildet das Learning Research and Development Center (LRDC) an der Universität von Pittsburgh und die Unterrichtsmaterialien und -verfahren, die hier entwickelt worden sind. Daher befaßt sich die hier beschriebene evaluative Forschung mit spezifischen Bildungsprogrammen, die Unterricht an individuelle Unterschiede anzupassen versuchen. Das Ziel der Forschung besteht darin, Informationen in bezug auf die Gültigkeit der Hypothesen über die pädagogischen Programme des Learning Research and Development Center zu gewinnen und verfügbar zu machen. Die Hypothesen und die Daten über ihre Gültigkeit sollen über den Nutzen der neuen Programme informieren und den Programmentwicklern Informationen über die relativen Stärken und Schwächen der Programmkomponenten geben.

Es gibt vier Institutionen, in denen man die Ergebnisse der Bemühungen des Learning Research and Development Center untersuchen kann. Am bekanntesten ist wahrscheinlich die Oakleaf-Schule, eine kleine Grundschule in einem Vorort von Pittsburgh, in der vor sieben Jahren der »individualisierte Unterricht« (Individually Prescribed Instruction, IPI) eingeführt wurde (Lindvall und Bolvin, 1967). Eine zweite Institution ist das System der Versuchsschulen, das das Regional Laboratory »Research for Better Schools« (RBS), in Philadelphia, aufgebaut hat. Das Institut »Research for Better Schools« disseminiert seit 1966 die Produkte des Learning Research and Development Center, die in der Oakleaf-Schule entwickelt worden sind. Eine dritte Institution ist die Frick-Schule, eine große Stadtschule im Zentrum von Pittsburgh, in der das Learning Research and Development Center in den letzten vier Jahren Programme entwickelt hat. Nachdem die Programme in der Frick-Schule entwickelt und getestet worden waren, wurden sie, viertens, in weiteren Schulen benutzt, die in einem Netzwerk zusammengefaßt sind. Im vergangenen Jahr entschieden sich vier Schulsysteme für die Programme, die wir in der Frick-Schule und in den Grundschulen der Schulsysteme entwickelt hatten, und implementierten sie. Das Learning Research and Development Center arbeitet mit diesen Schulen zusammen, so daß es auch den Prozeß der Dissemination pädagogischer Innovationen untersuchen kann. Lindvall und Cox (1970) und das Institut »Research for Better Schools« veröffentlichten Evaluationsuntersuchungen, die in der Oakleaf-Schule bzw. in den vom Institut »Research for Better Schools« betreuten Schulen gemacht worden waren. Ich werde meine Ausführungen daher auf die Evaluation, die in der Frick-Schule und den Schulen, die in dem Netzwerk zusammengefaßt sind, beschränken.

In der Frick-Schule entwickelte das Learning Research and Development Center ein individualisiertes Programm. Es bestand (1) aus einem Unterrichtsplan für jeden Schüler, der auf Grund der Ergebnisse in individuell eingesetzten Kriteriumstests entwickelt wurde; es enthielt (2) Hinweise und Vorschriften für die tägliche Anwendung des individuellen Unterrichtsplans. Es umfaßte (3) eine Neubestimmung der Lehrerrolle, bei der das Testen, die Tutorenarbeit und die Beweglichkeit des Lehrers besonders wichtig waren. Als Ergebnis sollte ein strukturiertes Curriculum in den grundlegenden Wahrnehmungs-, Lese- und Rechenfertigkeiten entstehen; es wurde durch ein wenig strukturiertes Curriculum ergänzt, in dem das Kind im bildnerischen und sprachlichen Gestalten, im sozio-dramatischen Spiel, in den Naturwissenschaften und in der Sozialkunde selbständig offene Lernaktivitäten wählen konnte.

Das Programm der Frick-Schule begann 1968/69 in Vorschulen und Kindergärten, wurde 1969/70 durch die erste Klasse, im vergangenen Jahr durch die zweite Klasse ergänzt und wird im nächsten Schuljahr von der Vorschule bis zur dritten Klasse reichen. Das Netzwerk begann 1969/70 mit drei Schulsystemen, zu denen im vergangenen Jahr ein viertes hinzukam, und das bis zum Herbst 1971 auf sieben Schulsysteme anwachsen soll. Wir beabsichtigen, das Netzwerk auf diese *sieben* Systeme zu beschränken, das für die Untersuchung des Disseminationsprozesses und die Evaluation unserer Curricula groß genug ist, ohne jedoch so groß zu sein, daß es als ein System, in dem Forschungs- und Entwicklungsarbeit geleistet werden soll, nicht mehr funktionsfähig ist. Die meisten Evaluationsuntersuchungen, die Daten über Schüler berücksichtigten, versuchten nachzuweisen, daß das Innovationsprojekt besser als ein vergleichbares anderes Programm ist. Welches Projekt als besser bezeichnet werden konnte, wurde auf Grund standardisierter Leistungsmessungen oder einer Reihe anderer Messungen bestimmt. Dazu wurden einige Kontrollschulen oder -klassen gebildet und dann die Mittelwerte verglichen. Wenn sich keine Unterschiede ergaben, waren die Ergebnisse nach Auffassung der Innovatoren nicht valide, und man bemühte sich weiterhin zu zeigen, inwiefern die Innovationen den bisherigen Bildungsprogrammen überlegen waren. Wenn sich aus den Ergebnissen des Vergleichs ergab, daß die Innovation einem anderen Programm überlegen war, waren die Innovatoren mit ihrer Arbeit und dem Evaluator zufrieden. Diejenigen jedoch, die der Innovation skeptisch gegenüberstanden, fanden irgendwelche Fehler im Innovations- und Evaluationsplan und bezweifelten die Gültigkeit der Ergebnisse.

Um diesen Punkt zu veranschaulichen, möchte ich einige Ergebnisse aus der Frick-Schule schildern. Um das Programm des Learning Research and

Development Center mit den bisherigen Schulprogrammen zu vergleichen, wurden in der Frick-Schule Kontrollgruppen eingerichtet, wobei uns die jährliche Erweiterung unseres Versuchs um ein neues Schuljahr zugute kam. Tabelle 1 veranschaulicht den allgemeinen Versuchsplan, der von

Tabelle 1
Versuchs- (E) und Kontroll- (K) Gruppen für die Frick-Schule

Jahr	Vor-schule	Kinder-garten	*Klasse* Erste	Zweite	Dritte	Vierte	Fünfte
1968–69	E	E	K	K	–	–	–
1969–70	E	E	E**	K*	K	–	–
1970–71	E	E	E	E	K	K	–
1971–72	E	E	E	E	E	K	K

* Gegensatz ist in Tab. 2 dargestellt
** Gegensatz ist in Tab. 3 dargestellt

Wang, Resnick und Schuetz (1970) entwickelt worden ist. Um Kontrollgruppen zu haben, untersuchten wir Klassen, die dem Programm um zwei Jahre voraus waren, während es selbst sich jährlich um ein Schuljahr erweiterte. Es konnten von einem Jahr zum anderen keine signifikanten Leistungsunterschiede zwischen den Evaluationsergebnissen eines bestimmten Schuljahres festgestellt werden. Es wurden auch bei den Variablen keine Unterschiede gefunden, die nach unserer Kenntnis Einfluß auf die Leistungen haben, ohne jedoch durch das Programm beeinflußbar zu sein wie etwa der sozioökonomische Status der Familie. Deshalb kann man zu Recht annehmen, daß in jedem Jahr die Kinder eines bestimmten Schuljahres Zufallsstichproben einer Grundgesamtheit waren.

Die in Tabelle 2 dargestellten Ergebnisse zeigen, daß das neue Programm statistisch signifikante Verbesserungen in allen drei Leistungsbereichen erbrachte, die in der zweiten Klasse mit dem Wide Range Achievement Test (WRAT) (Jastak, Bijou und Jastak 1965) gemessen wurden. Die Rechtschreibleistungen waren für unser Leseprogramm besonders interessant, weil wir die Rechtschreibung nicht direkt zu lehren versuchten, sondern sie als ein Nebenprodukt des Lesenlernens erwarteten.

Die Informationen, die auf Grund der Testnormierung gewonnen wurden, halfen uns, eine Vorstellung davon zu gewinnen, wieviel Zeitgewinn der Leistungszuwachs bedeutete. Die Unterschiede zeigten eine Verbesserung der Leseleistung um sieben, der Rechtschreib- und Rechenleistung um vier Monate an.

Tabelle 2
Vergleich im zweiten Schuljahr vor und nach dem LRDC-Programm
(Wide Range Achievement Test)

	»Vor« (Frühjahr 1970) (N = 98)	»Nach« (Herbst 1971) (N = 116)
Lesen		
Mittelwert (Rohwert)	41.45	49.91
Standardabweichung (Rohwert)	9.69	13.80
entsprechender Schuljahrswert *	2;2	2;9
	$F = 25.96$; df = 1 und 212; $p < .001$	
Rechtschreibung		
Mittelwert (Rohwert)	26.20	28.72
Standardabweichung (Rohwert)	5.08	5.44
entsprechender Schuljahrswert	1;9	2;3
	$F = 8.51$; df = 1 und 212; $p < .001$	
Rechnen		
Mittelwert (Rohwert)	23.40	25.22
Standardabweichung (Rohwert)	2.85	3.42
entsprechender Schuljahrswert	2;2	2;6
	$F = 17.62$; df = 1 und 212; $p < .001$	

* Der Wert 2;2 z. B. bedeutet: Die Leistung entspricht der Durchschnittsleistung nach zwei Monaten im zweiten Schuljahr.

Die Ergebnisse in Tabelle 3 verdeutlichen die Wirkung der Veränderungen zwischen der ersten und zweiten Version unseres Programms für die erste Klasse. Die Evaluation des Programms bei den ersten Klassen der Frick-Schule, die 1969/70 erfolgte, führte zu den Modifikationen, die im Herbst 1970 vorgenommen wurden. Die Gegenüberstellung der Ergebnisse des ersten und des zweiten Jahres gibt uns nützliche Informationen für die Kontrolle der Programmentwicklung. Programmveränderungen können so lange nicht als Verbesserungen dargestellt werden, als ihre Auswirkungen nicht bekannt sind. Die hier erzielten signifikanten Verbesserungen bestärkten die Programmkonstrukteure in ihrer Überzeugung, daß sie auf dem richtigen Weg waren. Außer dem Leistungszuwachs von einem Versuchsjahr zum anderen, erreichen die Schüler der ersten Klasse jetzt genauso gute Leistungen wie die Schüler der zweiten Klasse vor Beginn des Programms (vgl. hierzu die Mittelwerte der zweiten Spalte von Tab. 3 mit den Mittelwerten der ersten Spalte von Tab. 1).

Für den Programmentwickler sind diese Ergebnisse ohne Zweifel ermu-

Tabelle 3
Schulleistungen im ersten Schuljahr nach Veränderungen im LRDC-Programm (Wide Range Achievement Test)

	Nach dem 1. Jahr (Frühjahr 1970) (N = 143)	Nach dem 2. Jahr (Frühjahr 1971) (N = 124)
Lesen		
Mittelwert (Rohwert)	34.27	41.37
Standardabweichung (Rohwert)	10.32	11.85
entsprechender Schuljahrswert	1;7	2;2
	$F = 27.41$; $df = 1$ und 265; $p < .001$	
Rechtschreibung		
Mittelwert (Rohwert)	20.64	25.53
Standardabweichung (Rohwert)	4.65	5.77
entsprechender Schuljahrswert	1;3	1;7
	$F = 58.89$; $df = 1$ und 265; $p < .001$	
Rechnen		
Mittelwert (Rohwert)	22.36	23.98
Standardabweichung (Rohwert)	3.24	2.58
entsprechender Schuljahrswert	2;1	2;4
	$F = 20.03$; $df = 1$ und 265; $p < .001$	

tigend. Doch können sie auch anderen, nicht an dem Programm beteiligten Personen helfen, den Wert unseres neuen Programms zu beurteilen? Innovationen führen nicht immer zu einer Verbesserung der Mittelwerte, obgleich man nur selten negative Ergebnisse in der Literatur findet. Können nun diese Ergebnisse jemanden davon überzeugen, daß dieses Programm in die Grundschule seiner Gemeinde gehört? Sicherlich nicht.

Viele Unzulänglichkeiten solcher Ergebnisse werden sofort deutlich:

1. Da die Ergebnisse nur aus einer Versuchsschule stammen, geben sie keine Auskunft darüber, wie das Programm sich in anderen Schulen bewähren würde.
2. Die Beschränkung des Leistungsvergleichs auf die Ergebnisse eines Leistungstests verringert bei skeptischen Adressaten ihre Aussagekraft.
3. Ein statistischer Beweis allein hat niemals jemanden von irgend etwas überzeugt.

Der Innovator hat die Aufgabe, nachzuweisen, wie gut das neue Programm sich bewährt. Es gibt keine sicheren Verfahren, jemanden von etwas zu überzeugen, und auch statistische Ergebnisse besitzen keinen sicheren Überzeugungswert. Die Auseinandersetzung um die Schädlichkeit des

Zigarettenrauchens ist dafür ein klassisches Beispiel. Die mit statistischen Verfahren ermittelte Tendenz, eine Verbindung zwischen dem Zigarettenrauchen und Krebs herzustellen, war seit langem vorhanden und bekannt. Solange man jedoch nicht zeigen konnte, *wie* Zigarettenrauchen Krebs erzeugt, haben nur wenige diese Ergebnisse ernst genommen. Dennoch war die anfängliche Tendenz wichtig, weil sie die entsprechende Forschung anregte.

Um die Unzulänglichkeit zu überwinden, die sich aus der Beschränkung der Evaluation auf eine Versuchsschule ergibt, können wir unser Netzwerk in die Evaluation miteinbeziehen. Wenn die neuen Programme aus der Versuchsschule auf andere Schulen übertragen werden, entstehen jedoch auch neue Probleme. Wie können wir Gewißheit erhalten, daß unser Modell wirklich im Unterricht realisiert wird? Sobald ein Lehrer mit den neuen Verfahren vertraut gemacht worden ist und die neuen Materialien in seiner Klasse sind, macht er seinen Unterricht, ohne daß man weiß, inwieweit er wirklich dabei nach den Intentionen des neuen Programms handelt. Man braucht Methoden, um festzustellen, in welchem Ausmaß das Unterrichtsmodell in jeder Klasse implementiert wird, und um die Daten über das Ausmaß der Implementation mit den Schülerleistungen in jeder Klasse in Verbindung zu setzen. Wählt man die Klasse als Analyse-Einheit, kann man dieses Problem vielleicht lösen und die grundlegenden Merkmale des Unterrichtsmodells besser verstehen.

Viele Evaluationsuntersuchungen neuer Curricula oder neuer Unterrichtsmodelle haben sich vor allem der Varianzanalyse als statistischen Hilfsmittels bedient. Neuere Bemühungen haben auch multivariate Modelle verwendet. Der allgemeine Versuchsplan ist dabei derselbe geblieben. Nach experimenteller oder statistischer Kontrolle der anfänglichen Unterschiede zwischen den Schülern werden zwei oder mehr grob definierte pädagogische Programme oder Programmvarianten anhand eines oder mehrerer Leistungskriterien verglichen. Weder die Programmentwickler noch der potentielle Adressat haben aus solchen Untersuchungen viel gelernt.

Da eine überzeugende Evaluation in zahlreichen unterschiedlichen Klassen stattfinden muß und da diese Klassen sich in dem Ausmaß unterscheiden, in dem die verschiedenen Aspekte des Unterrichtsmodells realisiert werden, müssen Dimensionen bestimmt werden, mit denen das Ausmaß der Implementation gemessen werden kann; außerdem muß die Klasse als Analyse-Einheit in einem Korrelationsmodell verwendet werden.

Drei Arten von Variablen müssen berücksichtigt werden:

1. Das Anfangsverhalten der Schüler (Input)
2. Die Dimensionen des Unterrichts (Prozeß)

3. Die Schülerleistungen am Ende des Jahres (Output).
Der Hauptgrund für die Verwendung der Klasse als Analyse-Einheit liegt darin, daß Prozeßwerte für die Klasse charakteristisch sind. Ein weiterer wichtiger Aspekt dieses Verfahrens liegt darin, daß man die Auswirkungen erfassen kann, die eine unterschiedliche Verteilung beim Input auf den Output hat. Außerdem kann man feststellen, inwieweit das Programm bzw. die Programmvarianten unterschiedliche Outputs zur Folge haben. Dies wird dadurch erreicht, daß alle Werte des (Schüler-) Inputs oder Outputs auf vier statistische Maßzahlen für jede Klasse reduziert werden: Mittelwert (M), Standardabweichung (s), Schiefe (g_1) und Exzeß (g_2). Abb. 1 zeigt die Häufigkeitsverteilung und die vier statistischen Maßzahlen für eine der Klassen des Frick-Programms. Die Informationen über negative Schiefe, Hyperexzeß der Verteilung, ihre Lokalisation und ihre Streuung werden in diesen vier Werten beschrieben. Wiley (Wittrock / Wiley 1970) hat die Brauchbarkeit dieses Ansatzes behauptet; Lohnes (1971) bietet in einer Reanalyse der Daten der Cooperative Reading Study eine gute Illustration dafür. Ich möchte die bisherigen Ausführungen mit Hilfe wirklicher Daten aus den Klassen der Frick-Schule und des Netzwerks veranschaulichen.

Eine Dimension des Schüler-Inputs ist der Einstufungs-Test in unserem Rechencurriculum (vgl. Resnick, Wang und Kaplan, 1970). Ein ähnlicher Wert des (Schüler-)Outputs ist der Rechenwert im WRAT. Die Werte dieser zwei Messungen von 1500 Schülern können in acht Werte von 57 Klassen umgewandelt werden. Die vier statistischen Maßzahlen jeder Klasse basieren auf der Einstufung im Rechencurriculum als Inputmaß und den vier WRAT-Maßen als Outputmaß.

Bevor wir die Meßwerte über die unterschiedliche unterrichtliche Realisation des Programms in den Klassen miteinbeziehen, sollte man die Beziehungen zwischen diesen 8 Input- und Output-Werten untersuchen. Anstatt auf eine Korrelationsmatrix von 64 Elementen zu starren, bietet die kanonische Korrelation eine gute Zusammenfassung davon, wie die Inputwerte auf den Output bezogen werden. Tabelle 4 faßt die Ergebnisse einer kanonischen Korrelationsanalyse zwischen den vier Inputwerten und den vier Outputwerten zusammen.

Nur eine der vier möglichen kanonischen Beziehungen war auf dem 5 %-Niveau signifikant. Die kanonische Struktur und die Koeffizienten für diese größte Beziehung zeigen, daß ein Faktor, der zur Zeit des Inputs auf den Mittelwerten und den Standardabweichungen positiv und auf der Schiefe negativ geladen ist, mit einem Faktor korrelierte, der primär durch die Mittelwerte zur Outputzeit definiert ist. So scheint also die Form und der Mittelwert der Verteilung der Schüler im Herbst die mittleren Leistungen

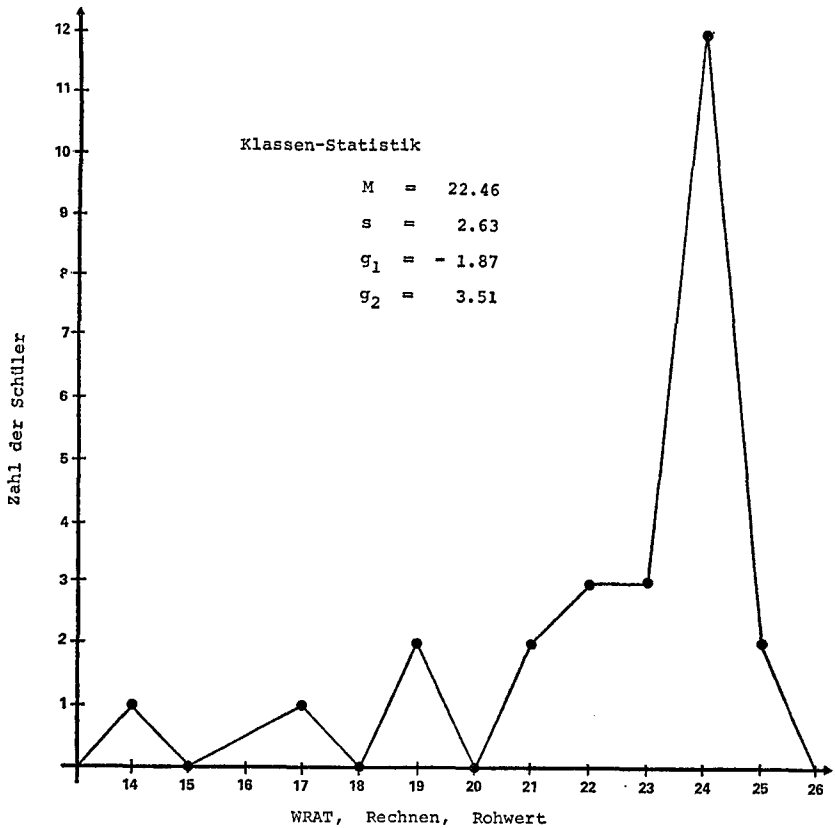


Abbildung 1
WRAT Rechnen, Verteilung für Klasse 1114
(N = 26)

der Klasse im Frühjahr zu beeinflussen. Jedoch besteht zwischen der Leistungsverteilung im Frühjahr und den Inputwerten im Herbst eine geringe Beziehung, d. h. das Ausmaß an Streuung, Schiefe und Exzeß im Frühjahr bezieht sich nur insoweit auf die Herbstwerte, als es von den Mittelwerten des Frühjahrs abhängt. Daher gibt es neben den Inputunterschieden noch andere Gründe für den Verlauf der Verteilungen im Frühjahr.

Der erste kanonische Faktor extrahiert ungefähr ein Drittel der Varianz jeder der zwei Gruppen der Variablen (.37 und .33). Die Varianz, die zusammen mit den kanonischen Relationen extrahiert wird, gestattet uns, die Redundanz des Output bei gegebenem Input einzuschätzen. Ein Redun-

danzkoeffizient von .18 zeigt an, daß 82 % der totalen Outputvarianz nicht durch diesen ersten Inputfaktor erklärt werden [1]. Daher muß ein Teil der Outputvarianz anders als durch die Inputvarianz erklärt werden.

Obwohl kanonische Analysen zwischen Input und Output interessant sein können, muß man die Prozeßdimensionen als eine dritte Art von Werten berücksichtigen und in die Analyse einbeziehen. Daher will ich zunächst beschreiben, wie die Prozesse gemessen werden, die wir auch als unterrichtliche Realisation oder Implementation bezeichnen.

Um den Prozeß der unterrichtlichen Implementation zu messen, müssen wir die Variablen identifizieren, die für das Unterrichtsmodell des Learning Research and Development Center besonders wichtig sind. Sieben Variablen scheinen für das Unterrichtsmodell relevant zu sein und lassen Unterschiede zwischen den Klassen erwarten:

1. Testverfahren

Tabelle 4
Kanonische Korrelationen zwischen den Werten im Herbst und den Werten im Frühjahr (N = 57 Klassen)

Klassenstatistik	Arithmet. Mittel	Standardabweichung	Kanonische Struktur	Kanonische Koeffizienten	
		INPUT			
Herbst-Quantifikation					
Mittelwert	7.12	8.12	.82	.92	erklärte
Standardabweichung	5.90	5.84	.53	−.29	Varianz = .37
Schiefe	1.11	1.11	−.66	−.85	Redun-
Exzeß	1.84	3.75	−.25	.66	danz = .20
		OUTPUT			
Frühjahrs-WRAT-Rechenwert					
Mittelwert	19.92	3.28	.99	.93	erklärte
Standardabweichung	3.17	1.01	−.57	−.12	Varianz = .33
Schiefe	−.49	.61	−.11	−.22	Redun-
Exzeß	.59	1.48	.09	−.16	danz = .18
		Kanonische Korrelation = .73			
		Chi-Quadrat = 50.12			
		df = 16			
		$p < .001$			

Andere mögliche kanonische Beziehungen sind nicht signifikant auf dem 5 %-Niveau

2. Unterrichtsanweisungen
3. Beweglichkeit des Lehrers (wie der Lehrer seinen Unterricht gestaltet und auf das Schülerverhalten angemessen reagiert)
4. Art des wirklich verwendeten Unterrichtsmaterials
5. Zeiteinhaltung
6. Ausnutzung des Klassenraums
7. Das curriculare Wissen des Lehrers und seine Kenntnis der ihm anvertrauten Kinder.

Um von diesen Bereichen zu meßbaren Dimensionen zu gelangen, bieten sich zwei Verfahren an. Im Bereich der Tests könnte man z. B. folgende Verfahren entwickeln, mit denen die Lehrer ihre Testpraktiken verbessern können:

1. Häufiges individuelles Testen der Schüler
2. Genaue Auswertung und Darstellung der Testergebnisse
3. Bestimmung eines festen Platzes, an dem im Klassenzimmer Tests bearbeitet werden
4. Verwendung des Mastery Level [2]
5. Testen aller Lernziele.

Ein Mitglied des Projektteams des Learning Research and Development Center (Champagne 1971) hat eine solche Liste entwickelt, die aus 108 Items für 7 Komponenten des Modells besteht, die alle von einem Unterrichtsbeobachter kontrolliert werden können. Ihre Erprobung im vergangenen Frühjahr zeigte, daß sie als ein Mittel für die Beurteilung der Effektivität des Fortbildungsprogramms für die im Netzwerk arbeitenden Lehrer geeignet war. In jedem Bereich müssen jedoch einige Haupt-Variablen identifiziert werden, wenn Datensammlung und -analyse im Rahmen der Evaluation durchführbar sein soll. Mehr als 150 Klassen könnten zur Evaluation herangezogen werden, jedoch müssen die Kosten für die Unterrichtsbeobachtung niedrig gehalten werden.

Reynolds (1971) hat ein gutes Beispiel für ein entsprechendes Verfahren gegeben. Seine Untersuchungen einiger Klassen der Oakleaf-Schule haben ergeben, daß die Korrelation zwischen der Einstufung des Schülers und den standardisierten Leistungswerten um so höher ist, je mehr die Einstufung und die Testverfahren mit dem Unterrichtsmodell übereinstimmen. Eine zentrale Voraussetzung unseres Unterrichtsmodells besagt, daß Lernen dann am wirksamsten ist, wenn ein Schüler in einem hierarchisch organisierten Curriculum an der Stelle arbeitet, die ein wenig über seinen bisherigen Leistungen liegt, jedoch unter dem, was er nicht mehr leisten kann. Die häufige Verwendung von Kriteriumstests [3] ist das Mittel, mit Hilfe dessen diese Einstufung fortwährend modifiziert werden kann. Wenn es jedoch nachlässig gehandhabt wird, verschwendet der Schüler seine Zeit

mit Aufgaben, die er bereits bewältigt hat oder für deren Bewältigung er keine Voraussetzungen hat.

Für eine bestimmte Klasse wird die Korrelation zwischen der Einstufung der Schüler im Curriculum und dem allgemeinen Leistungsniveau niedrig sein, wenn:

(1) die Schüler das ganze Curriculum durcharbeiten können oder sogar dazu ermuntert werden, ohne jede einzelne curriculare Einheit wirklich zu beherrschen;
(2) die Schüler im Curriculum unter ihrem Leistungsniveau arbeiten;
(3) Lehrer die Einstufung der Schüler dadurch beschränken, daß sie sie mehr oder weniger an der gleichen Stelle im Curriculum zusammenhalten.

Somit würde eine Korrelation innerhalb einer Klasse zwischen den im Herbst in standardisierten Tests erreichten Schülerleistungen und der Einstufung der Schüler im Herbst gute Aufschlüsse darüber erlauben, wie gut ein Lehrer Tests im Rahmen des Programms verwendet. Die anderen sechs Bereiche werden ähnlich behandelt, um festzustellen, welche Hauptvariablen man benutzen könnte, um den Grad der Implementation jedes Bereichs zu erfassen.

Nachdem nun die dritte Gruppe von Variablen behandelt worden ist, gilt es das Problem der Definition eines analytischen Schemas zu reflektieren, mit dessen Hilfe Prozeßwerte in Verbindung mit Input und Output untersucht werden können. Es gibt zahlreiche mögliche Ansätze, dieses Problem zu lösen. Vier davon sollen hier genannt werden:

1. Kanonische Korrelation zwischen Input und Output, um die Residuen der Outputfaktoren auf Prozeßwerte zu beziehen.
2. Multiple Korrelationen vom Input mit jedem Output, Berechnung der Residuen für jeden Outputwert und Verbindung dieser mit den Prozeßwerten.
3. Zunächst wurde eine Auspartialisierung des Inputs aus dem Output vorgenommen; sodann wurde eine kanonische Korrelation zwischen Output-Residuen und Prozeß berechnet.
4. Es erfolgte eine Auspartialisierung des Inputs aus dem Output und aus den Prozeßvariablen; sodann wurde eine kanonische Korrelation zwischen den Residuen des Outputs und der Prozeßvariablen bestimmt.

Ob die mit dem Input zusammenhängende Varianz vom Output und dem Prozeß oder nur vom Output getrennt werden soll, bedarf sorgfältiger Überlegung. Man kann zu Recht erwarten, daß die Inputwerte den Prozeß beeinflussen, d. h. die unterrichtliche Realisierung kann als eine Funktion der Lokalisation und der Form der Klassenverteilung im Input verschieden sein. Daher wäre es sicher nützlich, die Art solcher Beziehun-

gen zu kennen, obwohl wir vor allem wissen wollen, wie die wirklich verwendeten Unterrichtsverfahren die Varianz im Output, die nicht zum Input in Beziehung steht, erklären.

Um in dieser Frage einen ersten Schritt zur Lösung zu machen, wurde eine multiple Korrelation zwischen den vier Inputwerten im Herbst und den Mittelwerten im Frühjahr (*Tabelle 5*) bestimmt; darauf folgte eine Berechnung der Restwerte für die Mittelwerte des Frühjahrs, was eine Variation in den Mittelwerten des Klassen-Outputs erbrachte, die nicht durch die vier Inputwerte erklärt werden kann. Wegen der Dominanz der Mittelwerte des Frühjahrs bei der Definition des im Frühjahr ermittelten kanonischen Faktors in Tabelle 4 ist die multiple Korrelationsstruktur des Herbstes identisch mit der kanonischen Korrelationsstruktur des Herbstes, was die frühere Aussage über den Mangel an zusätzlicher Information bei den Verteilungen im Frühjahr bestätigt. Abb. 2 zeigt die Beziehung zwischen vorhergesagten und beobachteten Mittelwerten auch für die 57 Klassen. Die Restwerte sind die vertikalen Abstände jeder Klasse von der in der Mitte liegenden Regressionslinie.

Tabelle 5
Vorhersage der durchschnittlichen Rechenleistung vom Frühjahr aus den statistischen Maßzahlen im Herbst
(N = 57 Klassen)

Herbst Rechnen Prädiktor	Kriteriums-Korrelation	Standardisierte Partielle Regressions-Koeffizienten	Struktur
Mittelwert	.59	.64	.82
Standardabweichung	.39	−.19	.54
Schiefe	−.49	.63	−.68
Exzeß	−.20	.46	−.28
		Multiple Korrelation = .72	

Um von Mitarbeitern, die mit den Klassen vertraut waren, einige Vorschläge bezüglich der für die Klassenunterschiede wichtigen Dimensionen zu bekommen, entwickelte ich zwei Listen, von denen eine die Klassen mit hohen positiven Restwerten (Region A in Abb. 2), die andere die Klassen mit den hohen negativen Restwerten (Region B) enthält. Die beiden Listen wurden nicht als solche identifiziert. Anfangs ergaben sich Schwierigkeiten bei der Differenzierung der Unterschiede, weil Klassen, in denen der Lehrer sich offensichtlich bei der Dimension der Beweglichkeit und

den anderen Hauptdimensionen des Unterrichtsmodells richtig verhalten hatte, zusammen mit weniger wirksamen Klassen auf beiden Listen vertreten waren. Dennoch entstand eine störende Konsistenz. In Region A neigten die Lehrer dazu, den Einstufungstest vorzeitig abzubrechen, und unterbewerteten damit das allgemeine Niveau des Eingangsverhaltens ihrer Klasse. In Region B neigten sie dazu, die Einstufung der Schüler im Rechencurriculum des vergangenen Frühjahrs für die Plazierung im Herbst zu verwenden, und überbewerteten damit ihre Schüler, da sie die Sommerpause nicht berücksichtigten.

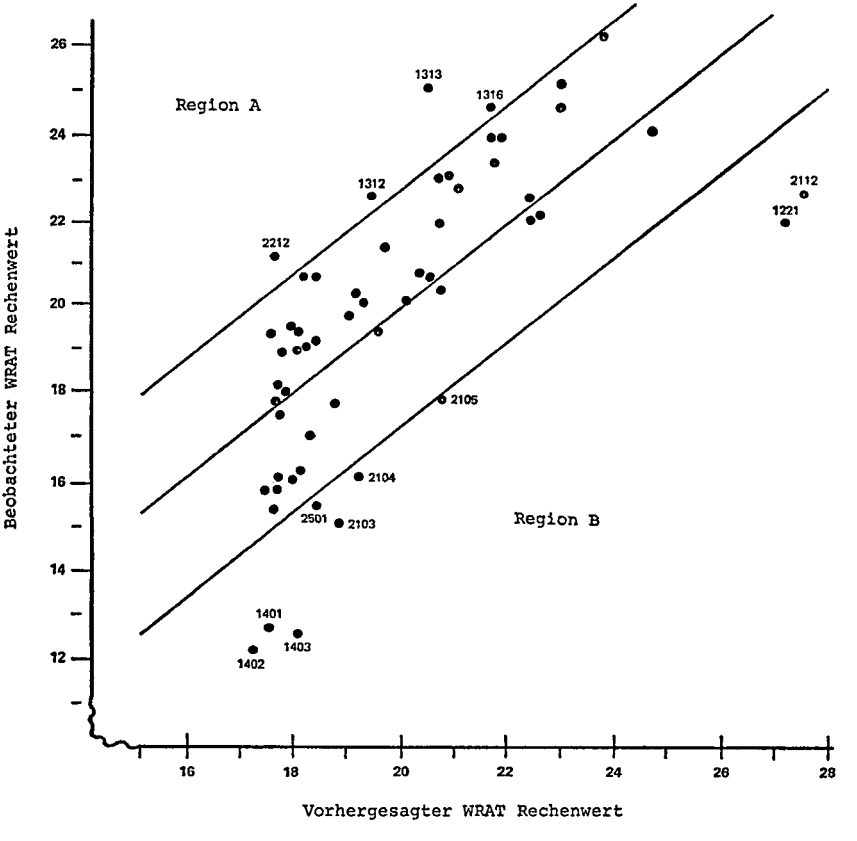


Abbildung 2

Stellung von 57 Klassen in einem zwei-dimensionalen Raum definiert durch eine lineare Funktion von vier Quantifikationswerten (Input) im Herbst und WRAT Rechenwerten (Output) im Frühjahr.

Dieser erste Schritt der Durchführung des Evaluationsplans teilte mir mehr darüber mit, wie sich die Klassen in bezug auf das Testen für die Einstufung der Schüler unterschieden, als über die Beziehungen zwischen den Unterrichtsverfahren und den Ergebnissen. Der Einstufungstest ist natürlich Teil des Unterrichtsmodells, und sein Einsatz steht unter der Kontrolle des Lehrers. Wenn aber erst Unterschiede bei der Realisierung dieses Aspekts des Modells entdeckt werden, kann aus ihrer Existenz bei diesem Regressionsansatz über das Unterrichtsmodell nichts mehr erfahren werden.

Wenn ein Forscher entdeckt, daß einer seiner Hauptwerte wie ein Gummiband ist, muß er bessere Untersuchungsverfahren entwickeln. Glücklicherweise kam zu dieser Zeit jemand auf ein besseres Verfahren. Lohnes (1971) überzeugte mich, daß eine Theorie der Input- und Outputmessungen notwendig ist, die den Forschungsprozeß stärker systematisch machen würde. Dies ist besonders wichtig, wenn man für jeden Versuch ein Schuljahr benötigt. Lohnes hat aber nicht nur deutlich gemacht, daß eine Theorie dieser Daten notwendig ist, er hat auch eine solche Theorie entwickelt. Um das zu verdeutlichen, muß ich auf einige Jahre zurückliegende Erfahrungen zurückgreifen. Lohnes und ich haben gemeinsam die Daten des Projekts TALENT bearbeitet, eine nationale Längsschnittuntersuchung, die mit über 400 000 Schülern der 9. bis 12. Klasse 1960 begann (Flanagan u. a. 1962). Eine Batterie von Tests und Fragebogen, deren Einsatz zwei Tage lang dauerte, wurde damals verwendet; ihre Daten wurden später durch Längsschnittwerte ergänzt, die an zentralen Stellen nach dem Sekundarabschluß erhoben wurden. Bei dieser Untersuchung überraschte uns die Vorhersagekraft einer kleinen Gruppe orthogonaler Faktoren, die Lohnes (1966) von der großen Batterie der TALENT-Prädiktoren abgeleitet hatte. Elf Faktoren für die Fähigkeiten und Motive schienen alle Informationen zu enthalten, die für die Vorhersage des von uns untersuchten Verhaltens nach dem Sekundarschulabschluß verfügbar waren. (Cooley/ Lohnes 1968).

Als ich Mitarbeiter am Learning Research and Development Center wurde, war ich über die mangelnde Berücksichtigung dieser grundlegenden allgemeinen Dimensionen individueller Unterschiede enttäuscht. Glaser (1968) und anderen Mitarbeitern gelang es, mich schließlich zu überzeugen, daß solche allgemeinen Einstellungen oder Motive nur wenig oder keine Relevanz für Unterrichtsentscheidungen haben. Die grundlegenden Dimensionen von TALENT, die sich als Prädiktoren für Erfolg und Befriedigung in unserer Gesellschaft so gut eigneten, sind nutzlos, um in der Praxis einen angemessenen Unterricht für einen Schüler zu entwickeln.

Lohnes überzeugte mich jedoch unlängst von der Notwendigkeit, die

TALENT-Dimensionen noch einmal nicht als *Prädiktoren* im Unterrichtsmodell, sondern als *Kriterien* für das Modell zu untersuchen. Nach seiner Auffassung müsse ein wertvolles Unterrichtsmodell auch dazu beitragen, die Wahrscheinlichkeit des Erfolgs und der Befriedigung eines Kindes nach seiner Schulzeit zu erhöhen. Aber auch wenn wir das Modell wiederholt definieren und modifizieren, können wir nicht zwanzig Jahre lang Längsschnittuntersuchungen durchführen, um festzustellen, welchen Fortschritt wir machen. Eine Möglichkeit bestand darin, diese Faktoren aus dem TALENT-Projekt, d. h. die Variablen in der Zeit vor der Sekundarschulerziehung und das Verhalten nach dieser Zeit, als Kriterien für die Wirksamkeit unseres Unterrichtsmodells zu verwenden. Die TALENT-Batterie selbst ist natürlich für Grundschulkinder nicht geeignet, aber die Primärfaktoren, die aus dieser Batterie hervorgingen, ließen sich auch in anderen Batterien finden.

Daher ist bei diesem Ansatz die Auswahl der Testbatterie für die Evaluation weit weniger willkürlich. Die Ergebnisse der Evaluation werden glaubwürdiger, wenn gezeigt werden kann, daß die Faktoren einen Übertragungswert auf das Erwachsenenleben haben. Es zeigt sich auch, wie man Grundschulpraktiken mit dem Prozeß der beruflichen Entwicklung in Beziehung setzen kann, woran kürzlich einige Beamte im Erziehungsministerium sehr interessiert waren.

Unter Evaluatoren ist die Frage umstritten, ob die Kriteriumsbatterie für die Evaluation aus standardisierten oder aus selbst angefertigten Tests bestehen soll, die sich auf die Items begrenzen, die eine Auswahl der Ziele des Curriculum repräsentieren, das evaluiert werden soll. Die Antwort darauf scheint mir jetzt klarer zu sein.

Unsere eigenen Tests sind wichtig, weil nur mit ihrer Hilfe die Frage beantwortet werden kann, ob unser Unterrichtsprogramm tatsächlich die Verhaltensweisen erreicht, die es erreichen soll. Eine umfassende Evaluation muß jedoch mehr leisten. Sie muß zeigen, wie Kinder durch dieses Programm befähigt werden, sich nach Abschluß der Schule im Leben zu bewähren. Wenn die Primärfaktoren für die Fähigkeiten und Motive gute Prädiktoren für den Erfolg und die Zufriedenheit junger Erwachsener sind, wenn sie eine Augenscheinvalidität (face validity) für die von ihnen vorausgesagten Kriterien besitzen und wenn solche Faktoren durch eine Verbindung zwischen Meßwerten aus der Verwendung des Unterrichtsmodells und standardisierten Tests gewonnen werden können, dann können und sollten diese Faktoren auch Kriterien für die Qualität unseres Programms sein.

Eine vollständige Beschreibung der Faktoren des TALENT-Projekts erfordert eine ganze Monographie (Lohnes 1966). Dennoch kann man we-

nigstens die Hauptfaktoren zusammenfassen, die in den Längsschnittsuntersuchungen Vorhersagekraft besaßen (Cooley/Lohnes 1968). Vier Kernfaktoren gingen aus 60 Eigenschaften des TALENT-Projekts hervor: Verbales Wissen, Englische Sprache, Mathematik und visuelles Erfassen. Der beste Prädiktor für die später erhobenen Kriterien und das wichtigste Konstrukt zur Erklärung der Interkorrelationen zwischen den 60 Eigenschaften des TALENT-Projekts ist der Faktor »verbales Wissen«. Lohnes (1966) sieht deutlich, daß dieser Faktor eine enge Approximation an die allgemeine Intelligenz darstellt. Er entschloß sich, ihn »verbales Wissen« zu nennen, weil »Intelligenz ein Begriff ist, der Mißverständnissen viel eher unterworfen ist als der Begriff Wissen«. Man sollte allmählich erkennen, daß ein Ergebnis des Unterrichts in der Maximierung der Punktwerte eines Schülers im allgemeinen Intelligenzfaktor[4] liegen kann und soll.

Von den 38 typischen Leistungswerten (Interessen und Bedürfnisse) leitete Lohnes 11 Motivfaktoren ab, von denen vier gute Prädiktoren dafür waren, wozu die Schüler nach dem Verlassen der Sekundarschule neigten. Drei dieser Faktoren waren sehr bekannte Interessendimensionen: Wirtschaft, Wissenschaft und Außenberufe. Der vierte Motivfaktor wurde mit »schulisches Interesse« bezeichnet. Lohnes (1966, 5–19) definiert diese Faktoren als »ein Motiv, das schulische Verhaltensweisen erklärt, denen die Gesellschaft zustimmt und die sie belohnt.«

Unsere evaluative Forschung in diesem Schuljahr wird von den Ergebnissen der Evaluation im vergangenen Jahr, der Lohnesschen Theorie über die Input- und Outputmessungen und dem Bedürfnis nach einer weiteren Erklärung des Ausmaßes der Implementation gesteuert. Im kommenden Herbst wissen wir über unser Unterrichtsmodell ein wenig mehr als in diesem Herbst. Evaluative Forschung kann und muß als integraler Teil der Curriculumentwicklung durchgeführt werden. Sie ist keine einmalige Handlung, die erst nach der Fertigstellung eines neuen Programms erfolgt. Evaluation kann nicht einfach in formative und summative Aktivitäten geteilt werden. Sie kann dem Programmentwickler Informationen liefern, während sie Informationen für potentielle Adressaten sucht. Sie ist Forschung. Sie wird durch Hypothesen gesteuert. Sie umfaßt eine Reihe von aufeinanderfolgenden Lösungsversuchen. Sie ist manchmal fehlerhaft, aber nie abgeschlossen.

Barry MacDonald

Informationen für Entscheidungsträger: Evaluation des Humanities Projects

Jede Evaluation zielt darauf ab, Informationen für Entscheidungsträger zu gewinnen; aber nicht alle Evaluatoren sind sich darüber einig, welches die wichtigsten Entscheidungsträger sind und welche Informationen sie brauchen. Eine These dieses Beitrags weist darauf hin, daß wenigstens in einigen Curriculumbereichen Evaluation sich stärker darum bemühen sollte, die Fragen der nicht unmittelbar an der Curriculumentwicklung beteiligten Entscheidungsträger zu entdecken und zu beantworten. Damit soll nicht der Wert formativer Evaluation in Frage gestellt werden. Zweifellos braucht man gute curriculare Materialien, aber ebenso empfiehlt es sich, wenn sie wirkungsvoll eingesetzt werden sollen, alle Einflüsse zu erforschen, die in den Schulen auf sie einwirken. Die folgende Darstellung des Humanities Projects und seiner Evaluation soll diese hier aufgestellte These unterstützen.

Wie soll eine demokratische Gesellschaft in ihren Schulen kontroverse Fragenkomplexe behandeln? Darin besteht kurz gesagt das Problem des Humanities Curriculum Projects, das die Nuffield Foundation und das Schools Council in Angriff nahmen. Nachdem drei Jahre an der Erforschung dieser Probleme gearbeitet worden war, wurden nunmehr einige Curriculummaterialien veröffentlicht und interessierte Schulen über die Entwicklung von Unterrichtsstrategien und die Behandlung der curricularen Probleme in diesem Bereich beraten.

Mit diesem Projekt wurde 1967 begonnen; es bildete einen Teil der Vorbereitungen für die Erhöhung der Pflichtschulzeit im Jahre 1972. Dem Projektteam wurde die Aufgabe gestellt, Materialien zu entwickeln und die Schulen zu beraten, wie man 14- bis 16jährigen Schülern mit durchschnittlichen oder unterdurchschnittlichen Fähigkeiten Probleme der Politischen Bildung (Humanities) vermitteln könne. Nach Auffassung des Teams besteht die entscheidende Aufgabe der Politischen Bildung in der Behandlung wichtiger humaner Problembereiche. Man entschloß sich daher, den Schwerpunkt der Bemühungen auf die Bereiche zu legen, in denen Wert-

konflikte auftreten. Die Teammitglieder waren der Überzeugung, daß sie so den Wunsch der Schulen nach einem für die Schüler relevanten Curriculum erfüllen könnten und dadurch den Schulen helfen würden, die kontroversen Problemkomplexe mit den Schülern offen und ehrlich zu erörtern.

Nach Auffassung des Teams bestand für die beteiligten Schulen das zentrale Problem darin, den Schülern Gelegenheit zu einer selbständigen Auseinandersetzung mit politischen Kontroversen zu geben, ohne dabei durch die politische Überzeugung der Lehrer gesteuert oder von anderen Schülern unter Druck gesetzt zu werden. Man wollte dieses Problem durch den Versuch lösen, ein Modell heuristischen Lehrens und einen entsprechenden Stil der Diskussionsführung im Unterricht zu entwickeln.

Als Grundlage für die Diskussion in den kontroversen Bereichen wurden Quellenmaterialien gesammelt. Lehrer übernahmen als Diskussionsleiter die Aufgabe, den Schülergruppen relevantes Material zur Untersuchung und Interpretation zur Verfügung zu stellen. Sie hielten ihre eigenen Anschauungen über die Probleme zurück und bemühten sich, die Schüler dazu zu bringen, widersprüchliche Ansichten zu formulieren. Obwohl der Gedanke der »Lehrerneutralität« in diesem Zusammenhang nicht neu ist, hat er während der Durchführung des Projekts viel Aufmerksamkeit auf sich gezogen. Manche sahen in dieser Neutralität das kennzeichnende Merkmal des Projekts. Dieses Urteil überbetont einen Aspekt in der Lehrerrolle, den das Projekt erforschen wollte. Das Team entwickelte zunächst Curriculummaterialien über Themen wie Krieg, Erziehung, Familie, Beziehungen zwischen den Geschlechtern, Mensch und Arbeit, Armut; (Materialien über das Rassenproblem, Recht und Ordnung, das Leben in den Städten sind in Vorbereitung.) Diese Materialien wurden in den Jahren 1968 bis 1970 von ungefähr 150 Lehrern in 36 Schulen in ganz England und Wales ausprobiert. Das Projektteam stellte Hypothesen über Lehrstrategien auf und legte es den Lehrern nahe, sie unter genauer Berücksichtigung der für sie charakteristischen Bedingungen zu prüfen. Zugleich bat man die Lehrer, ihre Urteile über die Brauchbarkeit des Materials abzugeben, alternative Verfahren oder Hypothesen vorzuschlagen und weitere für die Verbesserung der Diskussion notwendige heuristische Maßnahmen zu entwickeln.

Seit Ostern 1970 begann man damit, die Curriculummaterialien in den Handel zu bringen. Um den Anfragen einzelner Schulen und örtlicher Erziehungsbehörden nachzukommen, wurden überall in Großbritannien Kurse zur Einführung der Lehrer in die Benutzung des Materials eingerichtet. Während des laufenden Schuljahres haben etwa 600 Schulen das auf dem freien Markt erhältliche Material gekauft. Obwohl die meisten dieser Schulen die Absicht äußerten, die vom Projektteam entwickelten Lehrstrategien

zu übernehmen, sind bisher weniger als die Hälfte von ihnen bei den Fortbildungskursen vertreten gewesen.

Die Evaluation des Projekts begann 1968 mit meiner Einstellung. Meine Aufgabe war nicht genau bestimmt. Man wollte jedoch, daß ich die Entwicklung des Projekts untersuchte, um dem Projektteam über den Ablauf des Versuchs in den Schulen zu berichten und um ein geeignetes Evaluationsprogramm für die Implementation des Curriculum in den Jahren 1970 bis 1972 zu entwerfen. 1970 wurden drei weitere Kollegen für diese Aufgabe eingestellt; uns vier obliegt nun die Durchführung des Evaluationsprogramms.

Zu Beginn schien das Problem der Evaluation außerordentlich schwierig zu sein. Wenn ein Curriculum, wie in diesem Fall, in einem weitgehend unerforschten Bereich entwickelt wird, kann man nur auf geringe Erfahrungen zurückgreifen, um seine Auswirkungen oder etwa auftretende Probleme vorauszusagen. Erfahrungen aus früheren Innovationsprojekten lassen eine schlechte Prognose für dieses Projekt erwarten (vgl. Miles 1964). Auf den ersten Blick schien es durch die Mängel gekennzeichnet zu sein, die zu einem Fehlschlag früherer Innovationsprojekte geführt hatten. Es erforderte Einführungskurse für die Lehrer, war schwierig zu unterrichten und im Vergleich zu den in den Schulen verfügbaren Mitteln teuer. Außerdem stand es im Widerspruch zu vielen allgemein anerkannten Wertvorstellungen. Das Projekt zeigte zahlreiche Merkmale, die sein Scheitern durchaus möglich erscheinen ließen. Diese Ansicht beruhte auf der Voraussetzung, daß das Projekt danach beurteilt werden sollte, welche Lernergebnisse es in einer bestimmten Zeit bei den Schülern bewirkte. Das scheint mir heute ein unzureichendes Kriterium zu sein, um den Wert eines derartigen Curriculumprojekts zu beurteilen. Eine nähere Betrachtung einiger seiner wichtigen Merkmale soll dies deutlich machen und mir helfen, den Einfluß des Projektentwurfs auf die Entwicklung der Evaluation zu erklären. Dabei gilt es, drei Punkte zu erörtern:

Der erste Punkt betrifft den Entwurf des Projekts. Nach dem am meisten verbreiteten Modell der Curriculumentwicklung beginnt man damit, spezielle Lernziele zu formulieren, die das Endverhalten der Schüler angeben. Sodann werden die Inhalte und Lehrmethoden des Curriculum so lange modifiziert, bis das gewünschte Verhalten erreicht wird. Für den Evaluator stellen die curricularen Ziele die Erfolgskriterien dar. Seine Hauptaufgabe besteht darin, den Grad, bis zu dem sie erreicht werden, abzuschätzen.

Dieses Modell ist dann gut geeignet, wenn Lernziele einfach formuliert werden können, wenn bei ihrer Aufstellung leicht Konsens erreicht werden kann, wenn Nebeneffekte voraussichtlich unbedeutend und leicht feststell-

bar sind und wenn eine strenge Beachtung der Lernziele nicht dazu führt, die in ihnen nicht enthaltenen pädagogischen Werte zu verletzen. Das Projektteam war der Auffassung, daß die genannten Bedingungen auf das Humanities Curriculum nicht zutrafen, und hatte daher erhebliche Vorbehalte gegenüber diesem lernzielorientierten Modell. Man entschloß sich deshalb zu einem anderen Vorgehen (vgl. Stenhouse 1971), bei dem es folgende drei Fragen zu beantworten galt: Welche Inhalte sind von Bedeutung? Welches allgemeine Ziel ist für das Unterrichten dieser Inhalte angemessen? Welche Lernerfahrungen können dazu dienen, dieses Ziel zu erreichen? Zur Beantwortung dieser letzten Frage bedarf es eines umfassenden Unterrichtsversuchs. Durch die Verwendung von Hypothesen über die Auswirkungen des Curriculum an Stelle von Lernzielen hoffte das Team, eine gute Voraussetzung für die Entwicklung einer wirkungsvollen Lehrstrategie zu haben, die sich mit seinen besonderen Wertvorstellungen über die unterrichtliche Behandlung von Kontroversen deckte [1].

Dieser Auffassung lag die Überzeugung zugrunde, daß Lehrer, wenn sie sich an einem allgemeinen Ziel orientieren, eher wirkungsvolle Lehrstrategien entwickeln können. Daher sollte in diesem Modell der Versuch gemacht werden, das Ziel im Unterrichtsprozeß zu konkretisieren. Für den Lehrer bestand die Schwierigkeit darin, seine Unterrichtsführung mit dem Ziel in Übereinstimmung zu bringen und sie pädagogisch effektiv zu machen.

Bei diesem Modell, bei dem man von der Formulierung bestimmter Lernziele absieht, gibt es für den Evaluator kein eindeutiges Rezept. Er muß sorgfältig das Unterrichtsgeschehen beobachten und muß die verschiedenen Auswirkungen des Materials erforschen und die Beziehungen zwischen den Unterrichtsmodellen und ihren Auswirkungen aufdecken. Die *Ergebnisse* und der *Prozeß* bedürfen seiner Aufmerksamkeit. Ein besonderes Problem besteht darin, zu entscheiden, *welche* Auswirkungen untersucht werden sollen. In einem Evaluationsprogramm sollte man keine Antwort auf Fragen suchen, die keiner gestellt hat.

Der zweite Punkt betrifft die Zielsetzung des Projekts, Verständnis für gesellschaftliche Situationen, menschliche Handlungen und die sich daraus ergebenden Wertkonflikte zu entwickeln. Da für das Projektteam die wichtigste Aufgabe darin bestand, das Verständnis dieser Probleme zu vermitteln, lag es nahe, sich im Rahmen des Projekts eher um eine pädagogische Auseinandersetzung mit den kontroversen Fragen als um die Anpassung der Schüler an die bestehenden Verhältnisse und Normen zu bemühen. Die Folgen, die sich aus diesem Ziel für die Schüler- und Lehrerrollen und ihr Verhältnis zueinander ergaben, bestanden darin, daß in vielen Schulen Verhaltensweisen auftraten, die in Konflikt mit den allgemein ver-

breiteten Einstellungen und Gewohnheiten standen. Geht man davon aus, daß dieser Konflikt zweifellos Einfluß auf die Arbeit des Projekts hat und daß das Ausmaß des Konflikts von Schule zu Schule unterschiedlich ist, muß der Evaluator besonders den *Kontext* berücksichtigen, in dem das Curriculum realisiert wird.

Der dritte Punkt besteht in dem speziellen Einsatz, den das Team zur Lösung der Probleme der Curriculuminnovation gewählt hat. Während die meisten außerhalb der Lehrerschaft geplanten Ansätze zur Curriculumreform durch den Versuch charakterisiert sind, die Curricula von der Lehrfähigkeit der Lehrer unabhängig zu machen (teacher proof curricula), war das Projektteam davon überzeugt, daß es ohne eine sorgfältige Weiterbildung der Lehrer keine wirkungsvolle umfassende Curriculumentwicklung geben könne. Daher empfahl das Team den Lehrern, das Projekt eher als ein Mittel anzusehen, um die Probleme des Unterrichtens von Kontroversen besser selbständig handhaben zu können, als in ihm eine autoritäre, von Experten entwickelte Lösung zu erblicken. Für den Erfolg des Projekts war es wichtig, daß die Lehrer diese Auffassung verstanden und daß sie sich als Träger der Curriculumreform und nicht als Außenstehende fühlten.

Für die Evaluation folgte daraus, daß man die Kommunikation und die persönlichen Kontakte zwischen dem Projektteam und den Schulen untersuchen mußte, um Informationen über den Erfolg oder Mißerfolg dieser Bemühungen zu erhalten. Das heißt, man mußte in der Evaluation den *Input* des Projekts berücksichtigen.

Meine Aufgabe bestand darin, im Rahmen der Evaluation mit den verschiedenen Komponenten einer kreativen Curriculumentwicklung, zahlreichen möglichen Störfaktoren und einem neuen Curriculummodell fertig zu werden. In dieser Situation galt es für mich, die Entwicklung des Projekts so zu beschreiben, daß sie der Öffentlichkeit und dem professionellen Urteil zugänglich wurde. Angesichts der wahrscheinlich großen Bedeutung so vieler Aspekte des Projekts, fühlte ich mich anfänglich zu einer vollständigen Beschreibung seiner Auswirkungen und zur Beachtung aller relevanten Probleme verpflichtet. Evaluationsentwurf, -strategien und -taktiken würden sich, so hoffte ich, aus den Auswirkungen des Projekts auf die Struktur der Evaluationsprobleme ergeben.

Das Projekt in den Versuchsschulen (1968–1970)

Die 36 Schulen, die an dem Versuch im Herbst 1968 teilnahmen, wurden nicht mit den üblichen Stichprobentechniken ausgewählt. Statt dessen wurden sie von den zuständigen Verwaltungsbehörden empfohlen; die Verschiedenartigkeit der Schulen deutet auf interessante Unterschiede in den Beurteilungsmaßstäben und Prioritäten der örtlichen Schulbehörden hin. Nur in wenigen Fällen wurden die Kriterien der Nominierung explizit gemacht. In den meisten Fällen mußten sie erfragt oder erschlossen werden. Die Gründe für die Wahl der Local Education Authorities zu entdecken, war ein wichtiger Teil der Evaluation. Sie half mir, die Charakteristika der Schulstichprobe sowie die Politik und Strategie der örtlichen Schulbehörden in bezug auf die Curriculumentwicklung zu verstehen. Daraus ließen sich Schlüsse ziehen, wie gut die Local Education Authorities ihre Schulen kannten und auf Grund welcher Kriterien sie sie beurteilten. Ohne Zweifel bildeten zahlreiche unterschiedliche Kriterien die Grundlage für die Nominierung. In einigen Fällen wurden Schulen vorgeschlagen, die nach Meinung der Schulbehörden gut geeignet waren, an dem Versuch teilzunehmen; in anderen Fällen sollten bestimmte Schulen Anregungen zu neuen Ideen erhalten. Manchmal schien es, als habe man einer alten Schule die Teilnahme an dem Projekt als Entschädigung für die schlechten materiellen Verhältnisse, unter denen Lehrer und Schüler arbeiten mußten, zugedacht; dann wieder wurden vorbildliche Modellschulen empfohlen. Die Grundlagen für die Empfehlungen waren undurchsichtig. In einer Local Education Authority bestand ein wichtiges Kriterium für die Auswahl darin, ob die Schulleiter das Projekt dafür benutzen würden, größere finanzielle Anforderungen zu stellen oder nicht. Auf Initiative des Direktors und der Lehrerschaft hatten sich einige Schulen, die von dem Versuch gehört hatten, selbst beworben und ihr Anliegen bei der entsprechenden Schulbehörde vorgetragen. Im allgemeinen schienen die Local Education Authorities Schulen zu empfehlen, die ihrer Meinung nach für den Versuch geeignet waren. Aus einer näheren Kenntnis der Versuchsschulen wurde jedoch deutlich, daß einige Entscheidungsträger der Local Education Authorities ihre Schulen nicht genau kannten und folglich dazu neigten, ihr Urteil auf unzulängliche Kriterien zu gründen. Solche Beobachtungen über die Wahl der Local Education Authorities sollen nicht bedeuten, daß das Hauptkriterium für die meisten Empfehlungen nicht in der Geeignetheit der Schulen lag.

Die beteiligten Lehrer nahmen zusammen mit dem Projektteam an Regionalkonferenzen im Sommer 1968 teil. Auf ihnen wurde der Versuchsplan erläutert und die Aufgaben der Lehrer besprochen. Die meisten Leh-

rer verließen die Konferenz mit einigem Engagement für die ihnen gestellte Aufgabe.

Die Versuchsschulen waren über ganz England und Wales verteilt; sie lagen auf dem Lande, in Vororten, Städten und Großstädten. Die von allen Schulen ausgefüllten Fragebogen wiesen Unterschiede in der Art, Größe, Organisationsstruktur und in der formalen Beschreibung der Schüler auf. Aufgrund des Entschlusses, jeder Schule Entscheidungsbefugnisse zuzugestehen, vermehrten sich die bereits bestehenden Unterschiede noch durch die unterschiedlichen Entscheidungen über Einführung, Organisation und Verwirklichung des Versuchs. So standen z. B. in einer Schule vier Stunden, in einer anderen indessen 15 Stunden zur Verfügung. Die Zahl der in einer Schule beteiligten Lehrer schwankte zwischen 1 und 10. Einige Schulen ließen gute Schüler der 10. Klasse, andere die weniger guten Schulabgänger der 9. Klasse an dem Versuch teilnehmen. Die Situation wurde weiterhin durch Variablen wie Motivation, Verständnisfähigkeit und Erwartung der Versuchsteilnehmer erschwert, die erst im Laufe der Zeit deutlich in Erscheinung traten. Eine weitere Variable bestand im Ausmaß der Unterstützung, die eine Schule von ihrer Local Education Authority erhielt.

Die unmittelbaren Auswirkungen des Projekts waren im allgemeinen beunruhigend. Verwirrungen und Mißverständnisse waren so groß, daß ein Teil der Schulen nicht mehr den Empfehlungen des Projektteams angemessen nachkommen konnte. Es ergaben sich zahlreiche unerwartete Probleme und häufige Mißverständnisse über den Anspruch des Projekts wie z. B.:
(1) die Bedeutung der Schulleiter für die Realisierung von Innovationen wurde vom Projektteam unterschätzt, das anfangs das Ausmaß der Anforderungen, die an die wenig flexiblen Verwaltungen gestellt wurden, nicht richtig vorausgesehen hatte. Für die Schulen war es nicht einfach, die für den Versuch notwendigen Bedingungen zu schaffen; für die Lehrer war es nicht leicht, diese schwierigen und ungewohnten Aufgaben ohne Verständnis und Unterstützung durch den Schulleiter zu bewältigen.
(2) Die Lehrer waren sich nicht bewußt, daß die bisherigen Lernerfahrungen der Schüler die angestrebte Auseinandersetzung mit Kontroversen außerordentlich erschwerten und daß viele Schüler das Interesse für alle curricularen Inhalte verloren hatten. Auch vergegenwärtigte man sich nicht genügend, daß Lehrer und Schüler die traditionelle Rolle der Lehrerdominanz internalisiert hatten. Die Lehrer waren überrascht, daß die Schüler sich an den Diskussionen kaum beteiligten; sie waren darüber erstaunt, in welchem Ausmaß Schüler von der Initiative der Lehrer abhängig waren und mit welcher Zurückhaltung sie die Aufforderung, sich frei zu äußern, aufnahmen. Es schien, als seien fast alle Schulen und Lehrer autoritärer

eingestellt, als sie es selbst angenommen hatten. Die Auswirkung des Projekts auf die Autoritätsstruktur der Schule wurde immer deutlicher. Viele Lehrer kamen in schwierige Rollenkonflikte und versuchten vergeblich, das gestörte Vertrauensverhältnis zwischen sich und ihren Schülern zu überwinden. Das geht z. B. deutlich aus folgender Lehreräußerung hervor. »Ich bin sehr tolerant dem gegenüber, was die Gruppe in der Diskussion sagen möchte ... aber dann, wenn sie mich manchmal so ungezwungen oder direkt und unverfroren ansprechen, fühle ich doch, daß ich ein Erwachsener bin, und ich zeige ihnen meine Überlegenheit ... und meine Autorität, und gerade das gibt mir zu denken.«

(3) Offensichtlich hatte das Projektteam zunächst die Aufgaben des Projekts nicht klar und deutlich genug dargelegt. In den Augen der Lehrer hatte es eher einen moralischen als einen heuristischen Charakter. Die vorgeschlagenen Lehrstrategien wirkten eher so, als sollten sie nicht die Forschungshypothesen, sondern die Lehrfähigkeit der Lehrer prüfen. Beide Auffassungen führten dazu, daß die Lehrer aus ihren Erfahrungen wenig lernten und daher nur wenige Informationen von den Lehrern zum Projektteam gelangten.

Wenn die Bedingungen in den Schulen so einheitlich gewesen wären, wie die genannten drei Punkte darlegen, wäre die Evaluation anders verlaufen. Doch war das ganz und gar nicht so. Obwohl das Programm im allgemeinen sich als anspruchsvoll, schwierig und anregend herausstellte, gab es in manchen Fällen auch erhebliche Ausnahmen und Widersprüche. Während viele Schulen über ernsthafte Probleme wie den Schwierigkeitsgrad des Materials oder bestimmte Einstellungen der Schüler berichteten, waren andere über manche dieser Schwierigkeiten überrascht, die ihnen selbst überhaupt nicht begegnet waren. Daher können auch begrenzte Erklärungen des Erfolgs oder Mißerfolgs bei dem Versuch, die Schülerfähigkeiten, das Lehrerverhalten oder das Engagement eines Lehrerkollegiums zu verbessern, nicht ohne weiteres verallgemeinert werden. Es war nicht einfach, die Zahl der theoretisch zu berücksichtigenden Variablen aufgrund der Erfahrungen zu reduzieren. So legten z. B. Erfahrungen der Lehrer im Nordosten Englands die Vermutung nahe, daß das unterschiedliche Engagement von Jungen und Mädchen in kleinen Gruppendiskussionen nur durch starke Unterschiede zwischen den Geschlechtern erklärt werden könne, die in den Normen der Arbeiterklasse dieser Gegend ihren Ursprung hatten; während dagegen die Lehrer an einer wallisischen Schule halb scherzend behaupteten, sie kämen mit den Schülern in keine Diskussion, weil es in dieser Gegend von Wales keine kontroversen Fragenkomplexe gäbe. Selbst wenn solche regionalen Unterschiede außer Acht gelassen wurden, schienen die Probleme noch immer komplexer zu werden.

Während das Projektteam sich im ersten Jahr mit den Problemen der Schulen befaßte, um den Innovationsversuch funktionsfähig zu erhalten, konzentrierte ich mich darauf, die Vorgänge in den Schulen zu untersuchen und Informationen zu sammeln, die zur Erklärung unterschiedlicher Handlungs- und Reaktionsmuster beitragen konnten. Ich untersuchte die Aktivitäten der Teams, die Interaktionen zwischen seinen Mitgliedern, die Local Education Authorities und die Schulen. Ich sammelte Daten über die unterstützenden und hemmenden Einflüsse von außen, die bei der Implementation des Curriculum auftraten. Ich erarbeitete eine Liste mit empirisch abgesicherten und weniger abgesicherten Items, die ein institutionelles Profil für jede Schule ergaben. Mit Hilfe von Fragebogen versuchte ich abzuschätzen, wie weit die beteiligten Lehrer die Theorie des Projekts verstanden und wie sie dem Projekt gegenüber eingestellt waren. Ich organisierte mit Hilfe von Tonbandaufzeichnungen und ergänzenden schriftlichen Protokollen ein Feedback-System für die Lehrer und machte in mehreren Teilen des Landes Fernsehaufzeichnungen von Unterrichtsabläufen. Die Fragestellungen und das Erkenntnisinteresse des Projekts und des Evaluationsteams waren weitgehend identisch, so daß sich eine gute Basis für eine dauerhafte Zusammenarbeit ergab.

Ich begann, eine Reihe von Schulen zu besuchen, um sie unmittelbar zu untersuchen. Nachdem ich in etwa der Hälfte der Schulen gewesen war, gab ich diesen Plan zugunsten von Fallstudien einiger Schulen auf, weil ich die Gründe für das beobachtete Diskussionsverhalten der Gruppen nicht verstehen konnte. Warum waren in dieser Hinsicht die Unterschiede zwischen den Schulen größer als innerhalb der Schulen? Warum war eine Schülergruppe an den Problemen interessiert und eine andere mit ähnlichen Charakteristiken so desinteressiert und ablehnend? Weitere Fragen ergaben sich, als wir im Kontext der Schulen nach den Ursachen für die Unterschiede zu suchen begannen. Warum unterstützten einige Kollegen das Projekt, waren andere indifferent und wieder andere ablehnend? Warum reagierten Schulen auf ähnliche Probleme verschieden? Viele derartige Fragen stellten sich uns dabei, nachdem wir einen Überblick über die unterschiedlichen Reaktionen der Institutionen, Lehrer und Schüler gewonnen hatten.

Gegen Ende des ersten und im Verlauf des zweiten Jahres des Versuchs wurden in etwa sechs Schulen Felduntersuchungen durchgeführt. Der Auswahl der Schulen lagen zahlreiche unterschiedliche Kriterien zugrunde, die die verschiedenen Reaktionen der Schulen auf den Versuch berücksichtigten. Bei den Fallstudien wurden Unterrichtsbeobachtungen, Interviews mit dem Lehrerkollegium, den Schülern und den Eltern gemacht; man sammelte detaillierte Informationen über die verschiedenen innerhalb und

außerhalb der Schule entstehenden Einflüsse auf das Projekt. In diesem Zusammenhang kann jedoch über diese Untersuchungen nicht näher berichtet werden; ich will lediglich einige Einsichten erwähnen, die wir aus der Evaluation gewinnen konnten:

(1) Menschliches Verhalten in pädagogischen Situationen ist zahlreichen unterschiedlichen Einflüssen ausgesetzt. Dies ist zwar allgemein bekannt, wird jedoch manchmal bei der Curriculumevaluation übersehen, da man davon ausgeht, daß die Intentionen auch tatsächlich realisiert werden und daß sich die Unterrichtsereignisse nur geringfügig zwischen den Schulen unterscheiden.

(2) Die Bedeutung einer Innovation läßt sich nicht aus der Summe einzelner Wirkungen verstehen, sondern muß als ein System von Handlungen und Konsequenzen begriffen werden. Um eine einzelne Handlung zu verstehen, muß ihre Funktion innerhalb des Systems bestimmt werden. Daraus folgt, daß Innovationen viel mehr unerwartete Folgen haben, als man im allgemeinen bei der Innovations- und Evaluationsplanung annimmt.

(3) Nicht einmal zwei Schulen haben so ähnliche Bedingungen, daß Rezepte für Innovationshandlungen individuelle Entscheidungen ihrer Lehrerkollegien ersetzen könnten. Bereits aus den historisch verschiedenen Bedingungen der Schulen entstehen erhebliche Unterschiede im Innovationsverhalten, die beim Treffen von Entscheidungen berücksichtigt werden müssen.

(4) Ziele und Absichten der Curriculumentwickler entsprechen nicht immer denen der Adressaten des Curriculum. Wir stellten fest, daß das Projekt häufig in Machtkämpfen zwischen verschiedenen Gruppen im Lehrerkollegium als politisches Mittel eingesetzt wurde. Oder es wurde dazu benutzt, die Schüler besser kontrollieren zu können und das Ansehen der Institutionen zu verbessern, ohne jedoch die Unterrichtswirklichkeit innovierend zu verändern. Darin liegt ein Beispiel für eine Innovation ohne wirkliche Veränderung der Schulwirklichkeit.

Begründung und Bezugssystem des Evaluationsprogramms (1970–1972)

Um möglichen Mißverständnissen vorzubeugen, sollte ich darauf hinweisen, daß die Evaluation des Projekts nicht eine Aufgabe ist, die nur von besonderen Fachleuten ausgeführt werden kann. Alle Mitglieder des Projektteams haben viel Zeit für die Evaluation ihrer Arbeit aufgebracht, um sie besser zu verstehen und den Schulen besser behilflich sein zu können. Mit Hilfe des Projektteams haben viele Schulen Prüfungsprogramme und

Schülerbeurteilungsbogen erarbeitet. Ich bin lediglich für die Arbeit eines unabhängigen, in das Projekt integrierten Evaluationsteams verantwortlich, dessen Vorgehen ich jetzt beschreiben möchte.

Evaluation kann danach beurteilt werden, ob sie die richtigen Informationen den richtigen Leuten zur richtigen Zeit zur Verfügung stellt. Aber wer sind die richtigen Leute, was sind die richtigen Informationen und wann werden sie benötigt?

Da ich mit einem Projektteam zusammenarbeitete, das gegen den Gebrauch von Lernzielen war, mußte ich ein anderes geeignetes Konzept für die Evaluation entwickeln. Je stärker ich die Komplexität und Verschiedenartigkeit der Vorgänge in Versuchsschulen erkannte, desto skeptischer wurde ich dagegen, Evaluation auf die Messung des Erreichens von Lernzielen zu beschränken. Ich versuchte, meine Aufgabe im Hinblick auf die Adressaten meines Berichts zu bestimmen. Es erschien mir sinnvoll, die Evaluation an bestimmte Adressaten zu richten. Im Laufe der Zeit wurden sie als die Entscheidungsträger definiert. Vier Gruppen von Entscheidungsträgern ergaben sich: Geldgeber, örtliche Erziehungsbehörden (Local Education Authorities), Schulen und Prüfungsausschüsse. Als Aufgabe der Evaluation wurde die Beantwortung der Fragen der Entscheidungsträger angesehen. Bald erschien jedoch diese Aufgabendefinition als unbefriedigend, weil sie unterstellte, daß diese Gruppen im voraus wußten, welche Fragen wichtig waren. Solange über den Erziehungsprozeß so wenig bekannt ist, daß die Auswirkungen bestimmter Innovationen nicht voraussagbar sind, ist dies daher eine nicht gerechtfertigte Unterstellung.

Gegenwärtig sehen wir unsere Aufgabe darin, den Entscheidungsträgern behilflich zu sein, begründete Urteile zu fällen. Dazu liefern wir ihnen Informationen, die ihre Kenntnis der Faktoren, die auf curriculare Handlungen Einfluß haben, verbessern. Diese Aufgabenbestimmung enthält zwei wichtige Vorteile: Zunächst erhöht sich die Zahl der Personen, für die Evaluation wertvoll ist. Sodann berücksichtigt sie die oft geäußerten Einwände, daß die Daten der Evaluation zu spät verfügbar sind, um noch Entscheidungen über das Curriculum beeinflussen zu können. Solange Ergebnisse der Evaluation jedoch ausschließlich auf das Curriculum bezogen werden und nicht darüber hinaus generalisierbar sind, ist in vielen Fällen diese Kritik durchaus berechtigt. Unsere Ergebnisse sollten für die wiederholt auftretenden Probleme, die sich bei Entscheidungen über ein Curriculum ergeben, relevant sein, und sollten zu einem besseren Verständnis curricularer Innovationen beitragen.

Aufgrund dieser Überlegungen können wir die Ziele der Evaluation folgendermaßen bestimmen:

(1) Um sicher zu sein, welche Wirkungen das Projekt hat, müssen die Um-

stände, unter denen die Wirkungen auftreten, aufgezeichnet werden; sodann müssen die Informationen den Entscheidungsträgern so bearbeitet vorgelegt werden, daß sie ihnen helfen, die voraussichtlichen Folgen der Implementation des Curriculum zu beurteilen.
(2) Wir beabsichtigen, die gegenwärtige Situation und die Vorgänge in den Schulen so zu beschreiben, daß die Entscheidungsträger besser verstehen können, was sie zu verändern versuchen.
(3) Es gilt die Arbeit des Projektteams so zu beschreiben, daß es den Geldgebern und Bildungsplanern hilft, den Wert dieser Investition zu beurteilen und das geeignete Bezugssystem für die finanzielle Unterstützung, die Planung und Kontrolle genauer zu bestimmen.
(4) Um einen Beitrag zur Theorie der Evaluation zu leisten, müssen wir unsere Probleme klar formulieren, unsere Erfahrungen aufzeichnen und unsere Fehler öffentlich eingestehen.
(5) Es muß uns gelingen, zum Verständnis der allgemeinen Probleme einer innovativen Curriculumentwicklung beizutragen.

Nicht jeder würde der Wahl dieser fünf Punkte als Evaluationsziele eines Curriculumprojekts zustimmen. Jedoch sind Ziele nach meiner Ansicht zum Teil auch das Ergebnis bestimmter Situationen. Da Curriculumentwicklung immer mehr in den Aufgabenbereich neuer und relativ unerfahrener Institutionen fällt, sollten die Forscher, die Erfahrungen auf diesem Gebiet haben, sich bemühen, möglichst viel zum Verständnis der Probleme, die sich bei der Implementation von Innovationen ergeben, beizutragen.

Das Hauptproblem besteht noch immer darin, die *richtigen* Informationen auszuwählen und sie den Entscheidungsträgern zur Verfügung zu stellen. Die verschiedenen Gruppen der Entscheidungsträger unterscheiden sich danach, welche Daten sie benötigen. Lehrer sind hauptsächlich an der Erziehung der Schüler interessiert, Schulleiter an der Ausbildung der Lehrer, örtliche Erziehungsbehörden an der Verbesserung der Schulen, Curriculumplaner an Projektstrategien und Schulausschüsse an Prüfungen, die geeignet sind, die Leistungen der Schüler zu beurteilen. Die einzelnen Personen unterscheiden sich ferner darin, inwieweit sie den verschiedenen Daten vertrauen und wie hoch ihre Risikobereitschaft beim Handeln ist. Da wir unterschiedlichen Interessen gerecht werden müssen, versuchen wir eine umfassende Untersuchung des Projekts durchzuführen, in der wir für den Erwerb relevanter Informationen subjektive und objektive Verfahren verbinden (um eine einfache, wenn auch irreführende Dichotomie zu gebrauchen).

Unser Evaluationsplan enthält klinische, psychometrische und soziometrische Elemente. Wir versuchen, Informationen vor allem aus zwei sich

überschneidenden Schulstichproben, und zwar aus einer großen und einer kleinen Stichprobe zu gewinnen. Deshalb werden für eine bestimmte Zeit die Erfahrungen einer Zahl von Schulen genauer untersucht, während gleichzeitig aber auch genügend Informationen über die Unterrichtsabläufe einer größeren Zahl von Schulen gesammelt werden, damit Schlußfolgerungen von einer Stichprobe auf die andere gemacht werden können.

Der Entwurf sieht folgendermaßen aus:

a) *In der großen Stichprobe der Schulen (ca. 100):*
   (1) Mit Hilfe eines Fragebogens werden Daten über den Input, Kontext und die Implementation gesammelt.
   (2) Es werden Urteilsdaten von Lehrern und Schülern gesammelt.
   (3) Die Verhaltensänderungen der Lehrer und Schüler werden objektiv gemessen. (Wir haben zu Beginn dieses Jahres den Schülern Vortests gegeben, die nach dem gemeinsamen Urteil der Lehrer, der Schüler, des Projekt- und des Evaluationsteams die erwarteten Dimensionen der Änderung des Schülerverhaltens berücksichtigen. Das war ein ziemlich großer Aufwand, aber er ist gerechtfertigt, wenn er dazu beiträgt, die Auswirkungen des Curriculum auf die Schüler festzustellen und uns im nächsten Jahr die Verwendung einer kleinen, aber genauen Testbatterie ermöglicht.)
   (4) Die Veränderungen in der Lehrpraxis müssen durch den Gebrauch von speziell ausgearbeiteten Auswahl-Antwort-Aufgaben erfaßt werden, die nur einen geringen Zeitaufwand seitens des Lehrers erfordern und von den Schülern selbst gehandhabt werden können.
   (5) Die Wirkung auf die Schulen muß mit Hilfe von locker strukturierten Lehreraufzeichnungen festgehalten werden.

b) *In der kleinen Stichprobe der Schulen (ca. 20):*
   (1) Fallstudien über Art der Entscheidungsprozesse, über Kommunikationsprozesse, über Lehrerfortbildung und über das Ausmaß der Unterstützung in den örtlichen Schulbezirken.
   (2) Fallstudien an einzelnen Schulen innerhalb dieser Bezirke.
   (3) Erforschung der Dynamik einer Diskussion mit Hilfe eines Tonbands, Videorecorders und der Unterrichtsbeobachtung.

Wir müssen nun die Erfahrungen, die mehrere hundert Schulen mit dem Curriculummaterial des Humanities Projects gemacht haben, beschreiben und so darstellen, daß sie für die Entscheidungsträger nützlich sind. Unserer Meinung nach wurden die Probleme der Evaluation bisher zu sehr vereinfacht; oder aber man versuchte ausschließlich, sie mit Verfahren der empirischen Forschung zu lösen, wodurch die Funktion der Evaluation zu sehr eingeschränkt wurde. Vielleicht können bei unserem gegenwärtigen Verständnis komplexere Evaluationspläne mehr darüber Aufschluß geben,

was wir wirklich zu verändern versuchen und welche Mittel wir dazu brauchen. Deshalb haben wir einen so komplexen Evaluationsplan entwickelt. Mit dieser Evaluation wollen wir dazu beitragen, das Wechselspiel der Kräfte, die bei dieser Curriculuminnovation im Spiele sind, besser zu verstehen.

V Gesellschaftspolitische Aspekte der Evaluation

KLAUS NAGEL / ULF PREUSS-LAUSITZ

Thesen zur wissenschaftlichen Begleitung von Versuchen und Modellen im Bildungssystem[1]

I

Daß wissenschaftliche Begleitung als integraler Teil von Modellversuchen verstanden wird, ist in der Bundesrepublik neu. Aus der Entwicklung der geisteswissenschaftlichen Pädagogik zu einer empirisch-analytischen Erziehungswissenschaft folgt für diejenigen, die über den Strukturwandel des Bildungssystems zu entscheiden haben – nämlich Legislative und Exekutive –, die Erwartung, daß »wissenschaftliche« Kriterien für die politischen Entscheidungen vorliegen oder entwickelt werden können, was immer unter »wissenschaftlich« dabei zu verstehen ist.

Dieser Einsatz wissenschaftlicher Kompetenz bei der Erprobung von Modellen, der im Gesamtschulbereich schon institutionalisiert[2] und im Vorschulbereich mit einer Reihe von Projekten angelaufen ist[3], hat weniger wissenschaftsimmanente als politische und ökonomische Ursachen, – politische, weil als Antwort auf die Krisen im Bildungssektor, die den Bedarf an Arbeitskraft und ihre Qualifikation betreffen, die Struktur des Bildungssystems und nicht nur die Inhalte verändert werden müssen. Damit werden Institutionen der bürgerlichen Elitebildung bedroht, die jedoch ihren Charakter von ausgeprägten Eliteinstitutionen zu Aufstiegsinstitutionen hin verändert haben und damit von einer breiteren bürgerlichen Schicht getragen werden. Die Argumentation in den daraus resultierenden politischen Kämpfen wird deshalb nicht mehr nur auf der »ideologischen« Ebene ausgetragen; in wachsendem Maße spielen ökonomische und technologische Begründungen eine entscheidende Rolle. Dabei soll Wissenschaft solche Argumente liefern und politische Entscheidungen quasi-objektiv legitimieren. Die Ursachen für die Entwicklung liegen in der durch das sozioökonomische System bedingten gesellschaftlichen Notwendigkeit, die unproduktiven Kosten (sensu Marx) der Reformen so gering wie möglich zu halten, d. h. auch bei der Ausbildung einen Qualifikationsbegriff zugrunde zu legen, der einseitig auf Verwertung eben dieser erworbenen Qualifikationen abzielt. Da diese Zusammenhänge der Öffentlichkeit nicht bewußt sind, wird wissenschaftliche Rationalität als »objektive«, nicht-poli-

tische oder gar wertgebundene verstanden. Diese Trennung der Bereiche kann von Politikern benutzt werden, die eigenen Argumentationen gegenüber der Öffentlichkeit mit zusätzlichen »wissenschaftlichen« Argumenten zu stützen, ohne daß die Öffentlichkeit auch nur annähernd in der Lage wäre, hinter den wissenschaftlichen Argumenten die zugrunde liegenden Wertungen und die resultierenden politischen Implikationen zu entdecken (vgl. Fuchs 1970).

So ist vor allem die Ministerialbürokratie im Bildungssektor daran interessiert, Untersuchungen – zumeist als aufgabengebundene Auftragsforschung – im Gesamtschul- und Vorschulbereich durchzuführen (um die derzeit bedeutendsten Bereiche zu nennen), und bereit, erhebliche Mittel dafür zur Verfügung zu stellen. In dieser Situation griff die in der BRD in den sechziger Jahren entstandene empirisch-analytisch orientierte Sozial- und Erziehungswissenschaft das für sie einmalige Angebot staatlicher Instanzen auf, ihre Ressourcen und Arbeitsmöglichkeiten in ihren Fachbereichen zu erweitern, ohne die Problematik von Auftragsforschung zu berücksichtigen. Diese Problematik beinhaltet im Vorschulsektor, daß die Ziele entweder vorgegeben oder gar nicht expliziert sind, daß die Ergebnisse frühzeitig politischer Kontrolle unterworfen werden und daß die Versuche wegen bestimmter Auflagen und zu später Beteiligung von Wissenschaftlern an der Formulierung der Probleme den wissenschaftsmethodischen Ansprüchen oft nicht genügen (vgl. DIPF 1970, 68 u. 70).

Noch stärker als im Gesamtschulbereich wurde und wird im Vorschulbereich vor Beginn der Projekte die notwendige Reflexion über die *Fragestellungen*, die erst bildungspolitisch relevante Ergebnisse ermöglichen, verkürzt, ja versäumt. Dies ist nicht nur Ausdruck des von den Auftraggebern ausgeübten Handlungsdrucks, sondern auch des Fehlens einer öffentlichen Ziel- und Mitteldiskussion sowie der unpolitischen Haltung von Wissenschaftlern, die erst dann über die Komplexität von Maßnahmen zur Veränderung des Bildungssystems nachdenken, wenn sie die entsprechenden Aufträge erhalten. Soweit überhaupt Zielvorstellungen entwickelt wurden (Deutscher Bildungsrat 1970, 102–146), zeigt sich deutlich, daß zumeist gesamtgesellschaftliche Zusammenhänge nicht berücksichtigt werden und Vorschulerziehung Lernprozesse einleiten soll, die bei einem Strukturwandel anderer Sozialisationsagenten ebenfalls erreicht werden könnten. So wird Vorschulerziehung diskutiert, ohne die Veränderungsmöglichkeiten der Grundschule und der familialen Erziehungsbedingungen (und ihrer Grenzen) zu reflektieren, also ohne zu fragen, ob nicht eine radikal veränderte Grundschule zahlreiche Aufgaben der Vorschulerziehung überflüssig machte. Statt dessen kommt in zahlreichen Vorschulprojekten der Wille zum Ausdruck, Kinder »schulreifer« und »schulerfolgreicher« zu

machen, d. h. auf nicht hinterfragte schulische Bedingungen und Normen auszurichten.

Nach wie vor wird vor allem in der Diskussion um »kompensatorische« Erziehung von einer unzureichenden Sozialisation in der Unterschicht, also in der Arbeiterfamilie, ausgegangen. Dabei bleibt unberücksichtigt, daß, was in der Mittelschichtinstitution Schule, als die auch die Grundschule und auch der bisherige Kindergarten gelten können, dysfunktional ist, nicht nur aus der Sozialerfahrung der Arbeiterfamilie resultiert, sondern auch in ihrem Lebenszusammenhang angemessene, d. h. »richtige« Sozialisation ist. Um nur ein Beispiel zu nennen:

Die in der Arbeiterklasse häufig vorhandene »Unfähigkeit«, nicht langfristig planen zu können, ist Ergebnis ihrer Klassenerfahrung, nämlich nichts zu haben, was langfristige Planung notwendig und möglich machte (vgl. Rolff 1967). Kurzfristige Bedürfnisbefriedigung, das Gleich-haben-wollen, der Unwille zur Investition (soweit man etwas zu investieren hat) sind also nicht nur Ergebnisse einer bestimmten sozialen Erfahrung, sondern als verinnerlichter Verhaltensstil in *dieser Lage* angemessen, d. h. adäquat. Diejenige kompensatorische Erziehung ist daher falsch, die »deferred gratification patterns« den Arbeiterkindern zu vermitteln sucht, sie anpaßt an die Mittelschichtnorm langfristiger Investition (z. B. Bildung), ohne zu fragen, welche Folgen denn solches erworbene Verhalten in den unveränderten sozialen Verhältnissen hat.

Hier geht es nicht darum, einseitig Mittelschichtnormen ablehnend gegen Unterschichtnormen (diese glorifizierend) auszuspielen, sondern darauf hinzuweisen, daß solche Zusammenhänge sowohl in der Zieldiskussion der vorschulischen Erziehung als auch in den uns bekannten empirischen Projekten im Vorschulbereich nicht gesehen und berücksichtigt werden.

Damit hängt zusammen, daß die Bildungsinstitutionen, trotz anhaltender relativ allgemeiner Kritik, vor allem an Mittelschichtwerten orientiert sind. Daß Schule – auch vorschulische öffentliche Erziehung – in einer bürgerlichen Gesellschaft von bürgerlichen Normen bestimmt wird, ist weder verwunderlich noch im ganzen zu ändern. Sowohl für die Analyse wie für die Zielfindung einer emanzipatorischen Vorschulerziehung (und von diesem Selbstverständnis gehen zahlreiche Projekte aus) wäre es aber Grundbedingung, gerade den aus den gesamtgesellschaftlichen Bedingungen resultierenden widersprüchlichen Charakter von Unterschichtsozialisation und Mittelschichtcharakter öffentlicher Erziehung im Hinblick auf Emanzipation der Unterprivilegierten als *zentrales Untersuchungselement* zu verstehen und danach zu handeln. Was bislang angestrebt wird (im vorschulischen wie im Gesamtschul-Sektor), ist die individuelle »Emanzipation« – d. h. die Anpassung an bürgerliche Verhaltensstile und Lebenserwartungen – der individuelle Aufsteiger, der sich seiner Klasse entfremdet.

II

Trotz dieser noch ungeklärten Probleme und der fehlenden Zieldiskussion werden zahlreiche Projekte im Vorschulbereich durchgeführt [4], an die von den verschiedenen Bezugsgruppen unterschiedlichste Erwartungen geknüpft werden. Diese divergenten Interessen und Erwartungen sollen im folgenden kurz skizziert werden, um die potentiellen Konfliktsituationen, in denen Projekte wissenschaftlicher Begleitung stehen, zu kennzeichnen. Wir stützen uns bei diesen Thesen auf unsere Erfahrungen im Gesamtschulbereich, obwohl auch bei Projekten in der vorschulischen Erziehung einschlägige Erfahrungen vorliegen.

Politiker erwarten eine durch die Autorität von analytischer Wissenschaft abgesicherte Stützung ihrer sozial- und bildungspolitischen Argumentation. Sie brauchen alternativ strukturierte, publikumswirksame, parteigenehme Entscheidungshilfen. Häufig sind Bildungspolitiker gar nicht in der Lage, problemangemessene Entscheidungsalternativen zu erkennen, weil ihnen spezielle Kompetenz fehlt und deswegen detaillierte Informationen nicht verarbeitet werden können. Daraus resultiert auch ihr Bedürfnis, klare und einfach strukturierte Informationen zu erhalten, die das Wesentliche wiedergeben. Zum anderen wird häufig gefordert, daß diese Information »objektiv« und »neutral« ist. Darin steckt nicht nur die Furcht vor Manipulation durch die Wissenschaftler, sondern es kommt zugleich ein Wissenschaftsverständnis zum Ausdruck, das die historisch-gesellschaftlichen und unmittelbar politischen Implikationen aller wissenschaftlichen Tätigkeit ignoriert – ein Begriff von Wissenschaft, der deshalb um so leichter manipulativ benutzt werden kann. Dabei soll nicht unterschlagen werden, daß die Rationalität politischen Handelns eine auch an Machtpositionen und ihrer Erhaltung orientierte ist, was dazu führt, daß sich Argumente und Handlungen am Bewußtseinsstand der Öffentlichkeit orientieren. Dies erschwert den Kommunikationsprozeß zwischen Wissenschaftlern und Politikern auch auf der sprachlichen Ebene. In einer unaufgeklärt gelassenen Öffentlichkeit müssen für den Politiker Wählerstimmen immer wichtiger bleiben als Sachargumente, zumal wenn sie so komplexe Zusammenhänge wie im Bildungssektor betreffen.

Die *Verwaltung* ist die zweite relevante Bezugsgruppe. Dabei muß, trotz aller idealtypischen Vereinfachung, auf die Konflikte sowohl zwischen Verwaltungen unterschiedlicher Ressorts (Bau-, Finanz-, Sozial- u. Kulturverwaltung) als auch innerhalb der Kultusverwaltungen hingewiesen werden, die es schwierig machen, von einem einheitlichen Verwaltungsinteresse zu sprechen. Darüber hinaus ist sie gegenüber der politischen Spitze nicht ohne Eigeninteressen. Verwaltung, vor allem, wenn sie zugleich fach-

lich kompetent bzw. interessiert ist, steht in dem Konflikt zwischen den (eigenen) Zielvorstellungen und den ökonomischen, juristischen und politischen Begrenzungen. Sie ist daher an einer stetig, aber nicht zu schnell voranschreitenden begrenzten Reform interessiert und erwartet von wissenschaftlicher Begleitung sowohl die dafür adäquaten Implementationsstrategien als auch den meist ökonomisch orientierten »Effektivitätsnachweis«. Ihre Leitlinie ist die »kontrollierte Reform«, die bei den Adressaten (Lehrern, Eltern, Erziehern) nicht Anlaß zu herrschaftsgefährdender Unruhe [5], bei Öffentlichkeit und konkurrierenden Parteien nicht Anlaß zur Kritik bieten. Daher wird wissenschaftliche Begleitung seitens der Verwaltung als Kontroll- und Effektivitätsinstrument verstanden. Dabei genügen ihr auch oberflächliche Informationen, die dem Bewußtseinsstand der Öffentlichkeit über die Probleme angemessen sind (z. B. Übergangsquoten), ohne daß die zugrundeliegenden Prozesse oder Inhalte interessieren.

Die *Pädagogen* (Lehrer und Erzieherinnen) als diejenigen, die die strukturellen Reformen auch inhaltlich realisieren müssen, erwarten von wissenschaftlicher Begleitung zuallererst die konkrete Lösung ihrer dringendsten Praxisprobleme (Preuss Lausitz/Nagel/Hopf 1970). Diese Hilfe, das ist den Erziehern deutlich, kann nicht in Form von Kontroll-Untersuchungen oder in an Theoriegewinnung orientierten empirischen Arbeiten geleistet werden, sondern entweder in der Form regelmäßiger Beratung und Weiterbildung oder in einer engen Zusammenarbeit zwischen Forschungsgruppen und Erziehern (i. S. v. action research), die die gegebene Situation reflektiert und verändert. Diese Erwartungen werden – das zeigt sich unter den Gesamtschullehrern schon jetzt – häufig enttäuscht und zwar hauptsächlich aus zwei Gründen: Zum einen, weil sich der Sozial- und Erziehungswissenschaftler erst in den letzten Jahren verstärkt mit Sozialisationsprozessen innerhalb des komplexen Feldes Familie/Schule und vorschulische Erziehung/Bildungspolitik/gesamtgesellschaftliche Bedingungen zu beschäftigen beginnt und daher über handlungsrelevante komplexe Systemkenntnisse noch kaum verfügt; zum anderen, weil diese Hilfe sowohl von denen, die sie brauchen, wie von denen, die sie geben könnten, *technisch* verstanden wird, obwohl die Probleme gesellschaftliche Ursachen haben, die nicht durch einzelne Techniken zu ändern sind.

Das derzeit sehr dringende Problem in Gesamt- (wie anderen) Schulen, wie denn nun Apathie und Aggression bei Schülern zu beseitigen seien, wie man »intrinsische Motivation« freisetzen könne, ist nur nachgeordnet ein lerntheoretisches und didaktisches Problem. Zentral ist es ein Problem, das aus unzulänglichen Arbeitsbedingungen (zu große Lern- und Spielgruppen, mangelhafte Ausbildung der Pädagogen) und aus der Diskrepanz zwischen Curricula und Normen und den Erfahrungen, Erwartungen und

Verhaltensstilen der Jugendlichen aus unterschiedlichen sozialen Schichten resultieren (Liebel/Wellendorf 1969). Mit anderen Worten: Die Phänomene (z. B. Apathie, Aggression, Hospitalismus usw.) sind dann als Ausdruck gesellschaftlich-politischer Bedingungen zu begreifen, denn Klassenfrequenzen, Ausbildung und Curricula sind nicht innerhalb der Schule oder durch den einzelnen Erzieher änderbar, sondern können nur durch *politisches* (d. h. auch gemeinsames) Verhalten eben dieser Erzieher in Koalition mit anderen Gruppen (Eltern, Schülern, Bildungswissenschaftlern) verändert werden.

Wissenschaftler arbeiten – selbst wenn sie »rein theoretische« Interessen haben – in einem komplexen, widersprüchlichen Erwartungshorizont unterschiedlicher Bezugsgruppen, dessen sie sich kaum bewußt sind. Dieser zwingt sie, ihre Begleituntersuchungen, auch wenn sie dies nicht wünschen, als gesellschaftlich-politisch gebunden anzusehen und sich für bestimmte Erwartungen zu entscheiden. Darüber hinaus bringen sie eigene Interessen ein: Zum einen – »theoretisches« Interesse unterstellt – geben ihnen die Projekte Gelegenheit, in neuen Bereichen Qualifikationen zu erwerben; dies ist zumeist mit positiven Folgen für die Karriere verbunden. Zum anderen bietet sich aufgrund des politisch-öffentlichen Interesses an Begleituntersuchungen eine für Bildungsforschung günstige finanzielle Situation. Gerade soweit Projekte an Pädagogischen Hochschulen lokalisiert sind – was für den vorschulischen Bereich meist der Fall ist –, können dadurch zahlreiche Wissenschaftler zum ersten Male Forschung in größerem Ausmaß betreiben.

Betrachtet man die augenblicklichen Fragestellungen der Projekte, so wird deutlich, daß sich zahlreiche Wissenschaftler auf fachspezifische Ansätze ohne Einbeziehung der bildungspolitischen Konsequenzen und ohne handlungsrelevante Reflexion des Theorie-Praxis-Verhältnisses beschränken. Dabei fehlt es nicht selten sogar an Einsicht in die geringe Tragweite empirisch-analytischer Ergebnisse im Hinblick auf die angewandten Methoden. Darüber hinaus sind Wissenschaftler aufgrund der Karriereinteressen *mehr* an der theoretischen Klärung von Zusammenhängen als an praktisch relevanten Strategien zur Veränderung konkreter Sozialisationsbedingungen interessiert, weil sie nur so als *Wissenschaftler* anerkannt werden.

Andere Bezugsgruppen für wissenschaftliche Begleituntersuchungen von Modellversuchen lassen sich nur vereinzelt ausmachen. *Eltern* z. B. (oder eine diffus strukturierte, bildungspolitisch interessierte »Öffentlichkeit«) haben bislang kaum *artikulierte* Interessen außer dem vagen, daß Schulreform das »Beste« für *ihre* Kinder erreichen möge und die Wissenschaftler dazu beitragen sollen. Dabei entsteht nicht selten ein Widerspruch zwi-

schen dem Bewußtsein der Eltern und den von Wissenschaftlern (oder eventuell auch von der Verwaltung) als wünschenswert angesehenen Veränderungen. Dieser resultiert sowohl aus der unterschiedlichen Interessenlage der Eltern aus unterschiedlichen Klassen, wie auch aus der problembezogenen Unaufgeklärtheit der Eltern, was wiederum schichtspezifisch unterschiedlich ist. Deshalb bleibt es fraglich, ob sich pluralistische Konzepte der Beteiligung und Aufklärung aller als wirksam erweisen, wenn einzelne Gruppen über bessere Zugänge zu Informationen, bessere Voraussetzungen und Möglichkeiten der Einflußnahme und Mitsprache verfügen (vgl. Milberg 1970).

Die *Privatwirtschaft* ist bislang erst dann aktiv geworden (dann aber in massiver Weise), wenn ihre Profitinteressen bedroht waren oder die Gefahr bestand, daß in der Ausbildung zu kritische Inhalte vermittelt werden. Daher waren die Initiativen der Wirtschaft bislang auf die Lehrlingsausbildung (bzw. die Diskussion um die Integration der Berufsausbildung in der Sekundarstufe II), auf das 10. Hauptschuljahr, auf die Fächer Arbeitslehre und Gesellschaftskunde konzentriert (vgl. Baethge 1970). Vorschulische Erziehung, die nicht direkt, sondern nur vermittelt relevant ist für die Qualifikation der künftigen Arbeitskraft (und erst dadurch für die Verwertungsinteressen der Privatwirtschaft relevant wird), ist bisher nur für die Spielzeugindustrie bedeutsam geworden. Diese hat denn auch den Vorschulkongreß 1970 in Hannover finanziert, wobei offen bleiben kann, ob der Verlauf in ihrem Interesse war, wenn sie ohnehin nur an der Steigerung der Auflagenhöhe von Elternzeitschriften und Erweiterung des Absatzmarktes für didaktisches Spielzeug interessiert ist. Für die gesamte Wirtschaft ist die Vermehrung der Institutionen im vorschulischen Sektor aus *öffentlichen* Geldern nicht von den Kindern als zukünftigen Arbeitskräften, sondern primär von den entlasteten Müttern her wichtig: je früher die Schulpflicht, je mehr Kindergärten oder Vorschulen, desto mehr Mütter (und zunehmend die qualifizierter ausgebildeten) können berufstätig werden.

III

Wir haben versucht, das Bezugssystem wissenschaftlicher Begleitung von Modellversuchen als konfliktreiches Feld mit divergenten gesellschaftlichen Interessen zu skizzieren. Aus diesem Konfliktfeld selber lassen sich *keine* Zielsetzungen wissenschaftlicher Begleitung ableiten; diese müssen, so meinen wir, von einem emanzipatorischen (d. h. entschieden politischen) Wissenschaftsverständnis ausgehen. Dabei wird es einerseits um die Aufhellung der komplexen Zusammenhänge zwischen öffentlicher und familia-

ler Sozialisation und ihren Verbindungen mit anderen Faktoren gehen, wozu insbesondere die Soziallage und der Berufsstatus der Erzieher und Eltern, die Qualifizierung der Erzieher und die Struktur des Bildungssystems mit den in ihm verdinglichten Interessen gehören. Diese Analyse, die empirische wie hermeneutische Methoden gleicherweise erfordert, muß andererseits gekoppelt sein mit der Entwicklung von *Strategien* zur tatsächlichen Realisierung der Reformen. Deswegen muß wissenschaftliche Begleitung sich gemeinsam *mit den Betroffenen* (Lehrern und Erziehern, Eltern, z. T. Schülern, aber auch der Verwaltung und den Politikern) erst einmal um die Zielfindung im vorschulischen Bereich bemühen. Man könnte meinen, daß die politischen Entscheidungen für die strukturellen Reformen (Vorverlegung des ersten Schuljahres, also Einführung der Vorklassen und Ausbau der Kindergärten und deren Besuch auf freiwilliger Basis) schon gefallen seien und damit eine Diskussion der Ziele nicht mehr sinnvoll ist. Selbst wenn dies so wäre, fallen jedoch die Entscheidungen über die tatsächlichen *Inhalte* erst nach den institutionellen Lösungen.

Zum anderen sollte es die Funktion der Begleitung sein, statt wissenschaftliche »Kontrolle« auszuüben, erst einmal Lehrer und Erzieher in die Lage zu versetzen, die *gemeinsam* entwickelten Ziele zu realisieren. Dabei ist von den konkreten und unmittelbaren Erfahrungen der Erzieher *auszugehen*, um diese in einem komplexen Kommunikationssystem auf sozial- und erziehungswissenschaftliche Fragestellungen und Kenntnisse einerseits sowie auf gesellschaftlich-politische Zusammenhänge zu beziehen. Wissenschaftliche Begleitung übernimmt dabei Development-Hilfe (vgl. Fuchs 1970), trägt zur Rollen-Selbstreflexion der Erzieher bei, wobei auch die Wissenschaftler ihr Selbstverständnis ändern. Dabei müßten Innovationsstrategien entwickelt werden, die die beteiligten Lehrer und Kindergärtnerinnen befähigen, sich zunehmend zu qualifizieren, so daß sie mehr und mehr unabhängig von der Kompetenz der Wissenschaftler bei der Bewältigung ihrer Probleme werden. Damit ist nicht gemeint, daß Erziehungs- und Sozialwissenschaftler und die von ihnen durchgeführten Forschungsprojekte überflüssig werden. Sie erhalten allerdings einen anderen Charakter, sie sind nicht mehr isoliert (nicht nur in fachlicher, sondern auch sozialer Hinsicht), und die Prioritäten in den Fragestellungen sowie die Art der Fragen und die Kategorien ihrer Beantwortung ändern sich.

Hellmut Becker hat die zentrale Frage, wie Erzieher professionalisiert werden können und wie eine Vermittlung von Theorie in Praxis stattfinden kann, anders beantwortet: Er schlägt eine Vielzahl von »Pädagogischen Zentren« vor, die die Forschungsergebnisse der Grundlagenforscher und die Entwicklung der Bildungspolitik den Lehrern und Erziehern vermitteln sollen, um ihr Problembewußtsein zu schärfen (vgl. Becker 1971, 31–

34). Wir meinen demgegenüber, daß diese Arbeitsteilung und dieser einlinige Weg der Vermittlung die gewünschte Innovation gerade nicht ermöglicht, sondern daß Forschung – gerade bei Modellversuchen – von den Problemen und Erfahrungen der Praxis ausgehen muß und daß partiell die Trennung zwischen Forschern, Vermittlern und handelnden Erziehern aufgehoben werden muß. Daraus läßt sich auch ein gewandeltes Verhältnis der bisherigen Forschungsgruppen zur Publizierung ihrer Ergebnisse ableiten: Die Ergebnisse dürfen nicht nur für die Auftraggeber, den öffentlichen Geldgeber oder die wissenschaftliche Öffentlichkeit aufbereitet werden. Im Sinne von »action research« sollten Teilergebnisse sofort mit den Adressaten (in erster Linie Erziehern in der Schule und Vorschule und in der Familie) reflektiert werden (vgl. Fuchs 1970). Zum anderen werden die Ergebnisse anderen, nicht unmittelbar an der Interaktion beteiligten Betroffenen durch problemorientierte, den Sprachmustern der Adressaten angemessene Beiträge in Zeitungen, auf Versammlungen, ja auch in Flugblatt-Form oder in Form von »Briefen« durch die Wissenschaftler vermittelt werden [6]. Wenn die konkreten Probleme in Schule und Vorschule gesellschaftlich bedingt sind, dann müssen sich Wissenschaftler wie Erzieher an ihrem sozialen Ort politisch verhalten.

IV

Das skizzierte Modell wissenschaftlicher Beratung und Begleitung von Schulversuchen als Implementationsstrategie und Innovationshilfe für die Betroffenen hat Folgen für die *Institutionalisierung* der Projekte. Bei den bestehenden Forschungsgruppen für die Gesamtschulversuche läßt sich leicht nachweisen, daß sie der vorgeschlagenen Konzeption nicht entsprechen. Die Projekte werden entweder von den den Kultusbehörden nachgeordneten Institutionen oder als Auftragsforschung an Pädagogischen Hochschulen durchgeführt. Sie sind, den Interessen der Politiker und der geldgebenden Verwaltungen entsprechend, hauptsächlich als Leistungsvergleiche, methodisch problematische Systemvergleiche oder Kostenvergleiche oder andere Optimierungsfragen angelegt. Das hat eine geringe Aufklärungsfunktion der Untersuchungen (auch in wissenschaftlicher Hinsicht) zur Folge, weist auf die Abhängigkeit der Fragestellungen von den politischen Zielsetzungen der die Kulturpolitik der einzelnen Bundesländer bestimmenden Partei hin und findet als Herrschaftswissen für die Auftraggeber Verwendung. Diese Nachteile sind nicht nur Ausdruck der Lokalisierung solcher Projekte in staatsabhängigen Instituten, sondern auch des gesellschaftlichen Bewußtseins der beteiligten Forscher. Andererseits wird ge-

rade von solchen Instituten – im Gegensatz zu Universitäten, Pädagogischen Hochschulen und Fachhochschulen – der Zugang zu den Adressaten auf institutionellem Wege erleichtert. Nicht zuletzt muß berücksichtigt werden, daß Projekte an Hochschulen nach Ablauf der vertraglich festgesetzten Zeit aufgelöst werden können; ein »kritisches« Projekt, das der Verwaltung unangenehm sein könnte, wird danach kaum finanziele Unterstützung zur Fortführung finden. Der politische Druck auf Staatsinstitute ist zwar direkter, dafür aber auch deutlicher, so daß – vor allem, wenn die Interaktion mit den Adressaten stabil ist und ihre Bedürfnisse einbezieht – eher die Möglichkeit besteht, auch mit der Verwaltung Lernprozesse zu initiieren und langfristiger zu arbeiten.

Für die Vorschulprojekte gilt, daß sie nicht an den Fachhochschulen, wo die Ausbildung stattfindet, sondern an den Pädagogischen Hochschulen und Universitäten lokalisiert sind. Diese Form der Institutionalisierung muß zu einer Verschärfung der Rollentrennung von Lehre (Fachhochschulen) und Forschung führen. Das heißt, die Gefahr praxisferner und irrelevanter Untersuchungsansätze wird verstärkt. Darüber hinaus fehlen bislang Ansätze für eine institutionalisierte horizontale Kooperation zwischen den Projektgruppen, die zumindest die Diskussion von Methodenfragen und Forschungsansätzen erlauben würden.

Es ist u. E. notwendig, daß die an wissenschaftlicher Begleitung von Modellversuchen im Bildungssystem tätigen Erziehungs- und Sozialwissenschaftler mit ihrer Arbeit ihre eigene Rolle reflektieren, denn die Gefahr ist nicht von der Hand zu weisen, daß mit der politischen Durchsetzung der Gesamtschule wie der Vorschule die wissenschaftliche Begleitung der »rollenden Reform« weniger finanzielle Unterstützung erhält und wieder auf vereinzelte Projekte an Hochschulen reduziert wird. Damit würde die Konzeption einer wissenschaftlichen Begleitung als Forschung und Beratung in Verbindung mit einer Selbstaufklärung der Betroffenen im Keime erstickt, was die Veränderungsmöglichkeiten im Bildungssystem weiter vermindern dürfte.

Es soll nicht verkannt werden, daß die Parteinahme von Wissenschaftlern für die Interessen der jeweils Abhängigen und Unterdrückten dazu führen wird, daß sie selbst Pressionen und materiellen Sanktionen ausgesetzt werden. Dem ließe sich nur durch Solidarität zwischen Wissenschaftlern begegnen, die sich der gesellschaftlichen Bedeutung ihrer Arbeit bewußt sind.

Lee J. Cronbach und Evelore Parey
in Zusammenarbeit mit
Carol Codori und Michael Ravitch

VI Bibliographie Curriculumevaluation

Diese Bibliographie wurde ursprünglich für einen Kurs über Curriculumevaluation für Doktoranden an der Universität Stanford erarbeitet.

Die zahlreichen Veröffentlichungen über Evaluation lassen sich kaum noch überschauen. Es ist außerordentlich schwierig zu entscheiden, ob eine bestimmte Veröffentlichung wirklich wichtig ist. Man mag die Zuordnung des vorliegenden Materials zu den einzelnen Kategorien dieser Bibliographie als eine Art Gewichtung ansehen. Wir hoffen, daß dem deutschen Leser ein Einblick in einen relativ neuen Bereich der Unterrichtsforschung gegeben wird, in dem bisher kaum Veröffentlichungen in deutscher Sprache vorliegen. Bei der Überarbeitung der Bibliographie haben wir versucht, die für deutsche Verhältnisse relevanten und dem deutschen Leser zugänglichen Publikationen zu berücksichtigen. Wir unterscheiden folgende Kategorien:

I. Grundlegende Literatur

Diese Bücher sollte jeder kennen, der sich mit Evaluation beschäftigt. Es ist sicher nicht notwendig, jedes Buch von der ersten bis zur letzten Seite zu lesen. Mit großer Wahrscheinlichkeit aber wird man sich irgendwann im Zusammenhang mit einem bestimmten Problem auf ein Buch in dieser Kategorie beziehen müssen.

II. Spezielle Fragestellungen

Literaturhinweise zu speziellen Fragestellungen wurden in diese Kategorie aufgenommen. Oft handelt es sich um qualitativ hochstehende, wichtige Arbeiten; manchmal mußten wir uns jedoch damit begnügen, die besten Veröffentlichungen, die wir finden konnten, aufzuführen.

III. Schulfächer und Schulstufen

Diese Hinweise werden für die nützlich sein, die Evaluationsuntersuchungen für ein Unterrichtsfach oder eine Schulstufe planen.

IV. Beispiele für Evaluationsuntersuchungen

Als Beispiele werden in dieser Kategorie bereits durchgeführte Evaluationsuntersuchungen genannt.

V. Neuere Veröffentlichungen

In dieser Kategorie werden neue Veröffentlichungen aufgeführt, über deren Qualität man noch wenig aussagen kann. Einige werden später in eine der Kategorien I bis IV aufgenommen, andere durch bessere Arbeiten ersetzt.

I. Grundlegende Literatur

American Psychological Association, *Standards for educational and psychological tests and manuals.* 1966. American Psychological Association. (1200 Seventeenth St., N. W. Washington, D. C., 20036)

Anastasi, A. (Ed.), *Testing problems in perspective.* 1966. American Council on Education, (1785 Massachusetts Ave., N. W. Washington, D. C. 20005)

Bloom, B. S., et al., *Taxonomy of educational objectives. Part I. Cognitive domain.* New York: McKay 1956.

Bloom, B. S., Hastings, J. T., and Madaus, G., *Handbook of formative and summative evaluation of student learning.* New York: McGraw-Hill 1971.

Buros, O. K. (Ed.), *The sixth mental measurements yearbook.* Highland Park, N. J.: Gryphon Press 1965. Seventh yearbook in press [1972].

Caro, F. G., *Readings in evaluation research,* New York: Russel Sage Foundation 1971.

Cronbach, L. J., *Essentials of psychological testing.* New York: Harper & Row 1970 (3rd ed.).

Cronbach, L. J., Evaluation for course improvement. In: R. W. Heath (Ed.), *New curricula.* New York: Harper & Row 1964. (Also in *Teachers College Record,* 1963, 64, 672–683).

Dressel, P. L., and Mayhew, L., *General education – explorations in evaluation.* American Council on Education, 1954. (1758 Massachusetts Ave., N. W. Washington, D. C. 20005)

Dressel, P. L., et al., *Evaluation in higher education.* Boston: Houghton Mifflin 1961.

Ebel, R. L., *Measuring educational achievement.* New York: Prentice Hall 1965.

Ebel, R. L. (Ed.), *Encyclopedia of educational research.* New York: Macmillan 1969 (4th ed.).

Articles: Ammons, M.: *Objectives and outcomes,* 908–912; Coffman, W. E.: *Achievements tests,* 7–12; Ebel, R. L.: *Measurement in education,* 777–785; Heath, R. W.: *Curriculum evaluation,* 280–283; Tatsuoka, M. M.: *Experimental methods,* 474–481.

Educational Product Report, 1968 to date. (Educational Products Information Exchange Institute, 386 Park Avenue South, New York, N. Y. 10016)

ETS Conference Proceedings, *Proceedings of the Invitational Conference on Testing Problems.* Princeton: Educational Testing Service 1965 to date.

Evaluation Comment. University of California at Los Angeles, Center for the Study of Evaluation. (Periodical) 1968 to date. (145 Moore Hall, University of California, 405 Hilgard Ave., Los Angeles, California, 90024.)

Gage, N. L. (Ed.), *Handbook of research on teaching.* Chicago: Rand McNally 1963.

Glaser, R. (Ed.), *Teaching machines and programmed learning II.* 1965. National Education Association. (1201 Sixteenth St., N. W. Washington, D. C., 20036)

Grobman, H., *Evaluation activities of curriculum projects.* American Educational Research Association. Monograph Series on Curriculum Evaluation, No. 2, Chicago: Rand McNally 1968.

Henry, N. B. (Ed.), National Society for the Study of Education. *Measurement of understanding.* 45th yearbook, I, Chicago: University of Chicago Press 1946.

Krathwohl, D. R., et al., *A taxonomy of educational objectives II. Affective domain.* New York: McKay 1964.

Lindquist, E. F. (Ed.), *Educational measurement.* American Council on Education 1951. (1785 Massachusetts Ave., N. W., Washington, D. C., 20005)

Mehrens, W. A., and Ebel, R. L. (Eds.), *Principles of educational and psychological measurement.* Chicago: Rand McNally 1967.

Lindzey, G., and Aronson, E. (Eds.), *Handbook of social psychology.* Vol. 2: Research Methods. Reading, Mass.: Addison-Wesley, 1968 (2nd ed.).

Oppenheim, A. N., *Questionnaire design and attitude measurement.* New York: Basic Books 1966.

Review of Educational Research. *Educational evaluation,* 1970, 40, 2.

Scriven, M., The methodology of evaluation. Perspectives on Curriculum Evaluation. American Educational Research Association. *Monograph Series on Curriculum Evaluation,* No. 1, Chicago: Rand McNally 1967, 39–83.

Shaw, M. E., and Wright, J. M., *Scales for the measurement of attitudes.* New York: McGraw-Hill 1967.

Smith, E. R., and Tyler, R. W. (Eds.), *Appraising and recording student progress.* New York: Harper 1942.

Suchman, E. A., *Evaluative research: principles and practice in public service and social action programs.* New York: Russell Sage Foundation 1967.

Thorndike, R. L. (Ed.), *Educational measurement,* 1971. American Council on Education. (1785 Massachusetts Ave., N. W., Washington, D. C. 20005)

Tyler, R. W., *Constructing achievement tests.* Columbus: Ohio State University 1934.

–, *Basic principles of curriculum and instruction.* Chicago: University of Chicago Press 1950.

Wittrock, M. C., and Wiley, D. E. (Eds.), *The evaluation of instruction: issues and problems.* New York: Holt, Rinehart, and Winston 1970.

II. Spezielle Fragestellungen

A Historischer Überblick

Callahan, R. E., *Education and the cult of efficiency.* Chicago: University of Chicago Press 1962. (Paperback edition, Phoenix 1964)

Heath, R. W., Pitfalls in the evaluation of new curricula. *Science Education,* 1962, 46, 216.

Judd, C. H., et al., *Education as the cultivation of higher mental processes.* New York: Macmillan 1936.

B Funktion der Evaluation

Atkin, J. M., Some evaluation problems in a course content improvement project. *Journal of Research in Science Teaching,* 1, 1963, 129–132.

Bloom, B. S., The role of the educational sciences in curriculum development. *International Journal of the Educational Sciences.* 1966, 1, 5–16.

Forehand, G., The role of the evaluator in curriculum research. *Journal of Educational Measurement,* 1966, 3, 199–204.

Harris, C. W., Some issues in evaluation, *Speech Teacher,* 1963, 12, 191–199.

Stake, R. E., The countenance of educational evaluation. *Teachers College Record,* 1967/68, 523–40.

Timpane, P. M., Educational experimentation in national social policy. *Harvard Educational Review,* 1970, 40, 547–566.

Williams, W., and Evans, J. W., The politics of evaluation: the case of Head-start. *Annals of the American Academy of Political and Social Sciences,* 1969, 385, 118–182.

Zimiles, H., An analysis of current issues in the evaluation of educational programs. In: J. Hellmuth (Ed.), *The disadvantaged child,* II, New York: Brunner/Mazel 1968, 545–554.

C Ableitung und Formulierung von Bildungszielen

Broudy, H. S., The philosophical foundations of educational objectives. *Educational Theory,* 1970, 20, 3–21. Also in: Levit, M., *Curriculum.* Urbana: University of Illinois Press 1971.

Eisner, E. W., Educational objectives: help or hindrance. *School Review,* 1967, 75, 250–260, Comments by Ebel, Hastings, Payne 261–277; Response 277–282.

Evans, G., Behavioral objectives are no damn good. In: *Technology and Innovation in education.* Prepared by the Aerospace Educational Foundation. New York: Praeger 1968, 41–45.

Krathwohl, D. R., Stating objectives appropriately for program, for curriculum, and for instructional materials development. *Journal of Teacher Education,* 1965, 12, 83–90.

Lindvall, C. M., *Defining educational objectives.* Pittsburgh: University of Pittsburgh Press 1964.

Mager, R. F., *Preparing objectives for programmed instruction.* Palo Alto: Fearon 1962.

McGuire, Th. O., Decisions in curriculum objectives: a methodology for evaluation. *Alberta Journal of Educational Research,* 1969, 15, 17–30.

D Anlage und Analyse von Evaluationsuntersuchungen

Anderson, G. J., Walberg, H. J., and Welch, W. W., Curriculum effects on the social climate of learning. *American Educational Research Journal,* 1969, 6, 315–329.

Brownell, W. A., The evaluation of learning under dissimilar systems of instruction. *California Journal of Educational Research,* 1966, 17, 80–90.

Campbell, D. T., Administrative experimentation, institutional records, and nonreactive measures. In: J. C. Stanley (Ed.), *Improving experimental design and statistical analysis.* Chicago: Rand McNally 1967.

–, Reforms as experiments. *American Psychologist,* 1969, 24, 409–428.

–, Temporal changes in treatment-effect correlations: A quasi-experimental model for institutional records and longitudinal studies. In: *Proceedings of the 1970 Invitational Conference on Testing Problems.* Princeton: ETS 1971, 93–110.

Churchman, C. W. et al., *Introduction to operations research.* New York: Wiley 1957.

Cochran, W. G., The planning of observational studies of human populations. *Journal of the Royal Statistical Society,* 1965, 128, 234–266.

Cronbach, L. J., The logic of experiments on discovery. In: L. Shulman and E. Kieslar (Eds.), *Learning Discovery,* Chicago: Rand McNally 1966.

Dick, W., A methodology for the formative evaluation of instructional materials. *Journal of Educational Measurement,* 1968, 5, 99–102.

Edwards, A. L. and Cronbach, J. L., Experimental design for research in psychotherapy. *Journal of Clinical Psychology,* 1952, 8, 51–59.

Elashoff, J. D., Analysis of covariance: a delicate instrument. *American Educational Research Journal,* 1969, 6, 383–402.

Fischer, J. H., The question of control. In: *Proceedings of the 1965 Invitational Conference.* Princeton: Educational Testing Service [1966].

Gallagher, J. J., *Analyses of teacher classroom strategies associated with student cognitive and affective performance.* University of Illinois 1968. Educational Resources Information Center – ERIC (ED 02 1808).
ERIC Mikrofiche-Kollektion in Deutschland: Pädagogisches Zentrum 1 Berlin 31, Berliner Str. 40–41.

Gallagher, J. J., *Teacher variation in concept presentation im BSCS curriculum program,* University of Illinois 1966. ERIC (ED 020306).

Hyman, H. H. and Wright, Ch. R., Evaluating social action programs. In: P.

Lazarsfeld et al. (Eds.), *The Uses of Sociology*. New York: Basic Books 1967, 741–783.

Metfessel, N. S. and Michael, W. B., A paradigm involving multiple criterion measures for the evaluation of the effectiveness of school programs. *Educational and Psychological Measurement*, 1967, 27, 931–943.

Stanford Research Institute, *Toward master social indicators*. Research memorandum EPRC–67–47–2, 1969 (Educational Policy Research Center. Stanford Research Institute. Menlo Park, California, 94025).

Wallace, R. C. and Shavelson, R. J., *A systems analytic approach to evaluation: A heuristic model and its application*, ERIC, ED 058 202.

Wykstra, R. A., *Human capital formation and manpower management*. New York: Free Press 1970.

E Datensammlung und Interpretation

Belson, W. A., Respondent understanding of survey questions. *Polls*, 1968, 3, 1–10.

Bortmuth, J., *On the theory of achievement test items*. Chicago: University of Chicago Press 1970.

Cook, St. W. and Selltiz, C. A multi-indicator approach to attitude measurements. *Psychological Bulletin*, 1964, 62, 36–55.

Cox, R. C. and Graham, G. T., The development of a sequentially scaled achievement test. *Journal of Educational Measurement*, 1966, 3, 147–150.

Ebel, R. L., Content standard test scores. *Educational and Psychological Measurement*, 1962, 22, 15–25.

–, The value of internal consistency and classroom examinations. *Journal of Educational Measurement*, 1968, 5, 71–73.

Ferguson, R. L., *Computer-assisted criterion-referenced testing*. University of Pittsburgh, Learning Research and Development Center, Technical Report 1970.

Fiske, D. W., *Measuring the concepts of personality*. Chicago: Aldine 1971.

Goslin, D., *Guidelines for the collection, maintenance and dissemination of pupil records*. New York: Russell Sage Foundation 1970.

Hively, W. II, Patterson, H. L. and Page, S. H., A »universe-defined« system of arithmetic achievement tests. *Journal of Educational Measurement*, 1968, 5, 275–290.

Kropp, R. P. et al., The construction and validation of tests of cognitive processes as described in the Taxonomy of Education Objectives. *Journal of Educational Measurement*, 1964, 1, 39–42.

Levine, H. G. and McGuire, Ch. H., Role playing as an evaluative technique. *Journal of Educational Measurement*, 1968, 5, 1–8.

Levine, H. G. and McGuire, Ch. H., The validity and reliability of oral examinations in assessing cognitive skills in medicine. *Journal of Educational Measurement*, 1970, 7, 63–74.

Mischel, W., *Personality and assessment.* New York: Wiley 1968.
Popham, W. J. and Husek, T. R., Implications of criterion referenced measurement. *Journal of Educational Measurement,* 1969, 6, 1–9.
Simon, A. and Boyer, E. G. (Eds.), *Mirrors for behavior: an anthology of classroom observation instruments.* Philadelphia, Research for Better Schools, Vols. 1–14. 1967–1970; Suppl. Vols. A, B, 1970; Special Edition: Mirrors for behavior II, Vols. A, B, 1970. (Distributed by Classroom Interaction Newsletter c/o Research for Better Schools, Inc. 1700 Market Str., Philadelphia, Pennsylvania 19103).
Wallace, J., An abilities conception of personality. *American Psychologist,* 1966, 21, 132–138.
Webb, E. J. et al., *Unobtrusive measures,* Chicago: Rand McNally 1966.
Welch, W. W. and Walberg, H. J., Pretest and sensitization effects in curriculum evaluation. *American Educational Research Journal,* 1970, 7, 605–614.

F Sammelbände

Adelson, M. et al., Planning education for the future: comments on a pilot study. *American Behavioral Scientist,* 10, 1967, 7, Entire issue.
Glaser, R. (Ed.), *Training research and education.* New York: Wiley 1965. Ingenkamp, K. (Ed.), *Methods for the evaluation of comprehensive schools.* Weinheim: Beltz 1969.

III. Schulfächer und Schulstufen

A Geisteswissenschaften

Dressel, P. L. and Mayhew, L. B., *Critical analysis and judgment in the humanities.* Dubuque, Iowa: W. C. Brown 1956.

B Sprachunterricht

Braddock, R., Lloyd-Jones, R. and Schoer, L., *Research in written composition.* National Council of Teachers of English 1963. (508 S. Sixth Str., Champaign, Illinois 61820).
Conference on *Needed research in the teaching of English.* Cooperative Research Monograph No. 11, 1963. (U.S. Office of Education. Superintendent of Documents, U.S. Government Printing Office Washington 25, D. C. 20402).
Davis, F. B., Research in comprehension in reading. *Reading Research Quarterly,* 1968, 3, 499–545.
Godshalk, F. I., Swineford, F. and Coffman, W. E., The measurement of writing ability. *Research Monograph.* No. 6, College Entrance Examination Board, 1966. (Order from Educational Teaching service).

Maxwell, J. and Tovatt, A., *On writing behavioral objectives for English.* Commission on the English Curriculum. National Council of Teachers of English, 1970. (508 S. Sixth Str., Champaign, Illinois 61820).

C Fremdsprachen

Birkmaier, E. Evaluating the foreign language program. *North Central Association Quarterly,* 1966, 40, 263–271.

Lado, R.: *Language testing: the construction and use of foreign language tests.* New York: McGraw-Hill 1964.

D Sozialkundliche Fächer

Berg, H. D. (Ed.), *Evaluation in social studies.* 35th Yearbook, National Council for the Social Studies, 1965. (National Education Association of the United States. 1201 Sixteenth Str., N. W. Washington, D. C. 20036).

Dressel, P. L. and Mayhew, L. B., *Critical thinking in social science.* Dubuque, Iowa: W. C. Brown 1954.

Fair, J. and Shaftel, F. (Eds.), *Effective thinking in the social studies.* 37th Yearbook, National Council for the Social Studies, 1967. (National Education Association of the United States. 1201 Sixteenth Str., N. W. Washington, D. C. 20036).

Goolsby, T. M. Jr., Differentiating between measures of different outcomes in the social studies. *Journal of Educational Measurement,* 1966, 3, 219–222.

Jarolimek, J., *Guidelines for elementary social studies,* 1967. Association for Supervision and Curriculum Development, NEA, 1201 Sixteenth Str., N. W., Washington, D. C. 20036).

E Naturwissenschaften

The Biological Sciences Curriculum Study (B.S.C.S.) *Newsletter* No. 19, 1963; No. 24, 1965. (A Center at the University of Colorado, P. O. Box 930, Boulder, Colorado, 80302).

Karplus, R. (Ed.), *What is curriculum evaluation? Six answers.* 1968. Science Curriculum Improvement Study (Lawrence Hall of Science, University of California, Berkeley, California, 94720).

Klopfer, L. E. and McCann, D. C., Evaluation and unified science: measuring the effectiveness of the natural science course at the University of Chicago High School. *Science Education,* 1969, 53, 155–164.

Uricheck, M. J., Research proposal: An attempt to evaluate the success of the CBA and CAGMS Chemistry courses. *Science Education,* 1967, 51, 5–11.

Walbesser, H. H., *Science – A Process Approach.* An evaluation model and its

application. Second Report. Commission On Science Education. Miscellaneous Publication 68–4, 1968. (American Association for the Advancement of Science. 1515 Massachusetts Ave., N. W., Washington, D. C., 20005).

F Vorschulerziehung

Hess, R. D. and Baer, R. (Ed.), *Early education: current theory, research, and action.* Chicago: Aldine 1968.

G Grundschule

Kearney, N. C., *Elementary school objectives.* New York: Russell Sage Foundation 1953.

H Hochschule

Feldman, K. A., *Research strategies in studying college impact.* ACT Research Report No. 34, May 1970. (Research and Development Division. American College Testing Program, P. O. Box 168, Iowa City, Iowa 52240).

Menne, J. W., Techniques for evaluating the college environment. *Journal of Educational Measurement,* 1967, 4, 219–226.

Troyer, M. E. and Pace, C. R., *Evaluation in teacher education.* American Council on Education, 1944. (1758 Massachusetts Ave., N. W., Washington, D. C. 20005).

IV. Beispiele für Evaluationsuntersuchungen

Anderson, S. B., *Noseprints on the glass: how do we evaluate museum programs?* Princeton: ETS, 1966.

Astin, A., Learning mathematics: a survey of twelve countries. (Book review), *Science,* 30, 1967, 1721–1722.

Bach, G. L. and Saunders, Ph., Economic education: aspirations and achievements. *American Economic Review,* 1965, 55, 329–356.

Ball, S. and Bogatz, G. N., *The first year of Sesame Street: An evaluation* Princeton: Educational Testing Service, October 1970.

Barker, R. and Gump, P. V., *Big school, small school.* Stanford: Stanford University Press, 1964.

Charters, W. W., *Motion pictures and youth, a summary*. New York: MacMillan 1935.

Coleman, J. S. et al., *Equality of educational opportunity*. 1966. (U. S. Government Printing Office, Div. of Public Documents, Washington, D. C. 20402).

Fivars, G. and Gosnell, D., *Nursing education: the problem and the process. The critical incident technique*. New York: MacMillan 1966.

Fox, D. J., Conclusions of the Center's MES Evaluation. *Urban Review*, 1968, 2, 17–18.

Fox, D. J. et al., *More effective schools*, Evaluation of ESEA Title I Projects in New York City, 1967–68. Center for Urban Education, 1968. (Center for Urban Education, 105 Madison Ave., New York, N. Y. 10016).

Frankel, E., Evaluation of a curriculum for elementary science education. *Science Education*, 1968, 52, 284–290.

Grobman, A. B., *The changing classroom: the role of the Biological Sciences Curriculum Study*. New York: Doubleday 1969.

Grobman, H., The place of evaluation in the Biological Sciences Curriculum Study. *Journal of Educational Measurement*, 1966, 3, 205–212.

Herron, J. D., Evaluation and the new curricula. *Journal of Research in Science Teaching*, 1966, 4, 159–170.

Hill, Wm. F. et al., Group therapy for social impact: Innovation in leadership training. *American Behavioral Scientist*, 1967, 11, 1, Whole issue.

Husén, T., *International study of achievement in mathematics: A comparison of 12 countries*. New York: Wiley 1967.

Hyman, H. H. et al., *Application of methods of evaluation: four studies of the encampment for citizenship*. Berkeley: University of California Press 1962.

Karplus, R. and Thier, H. D., *A new look at elementary school science: science curriculum improvement study*. Chicago: Rand McNally 1967.

Klopfer, L. E., Effectiveness and effects of ESSP astronomy materials. *Journal of Research in Science Teaching*, 1969, 6, 64–75.

Lambert, W. E. and Tucker, G. R., *The bilingual education of children*. Rowley, Massachusetts: Newbury House [1972].

Maccoby, E. E. and Zellner, M., *Experiments in primary education. Aspects of Project Follow-Through*. New York: Harcourt Brace 1970.

May, M. and Lumsdaine, A. A., *Learning from films*. New Haven: Yale University Press 1958.

Michael, W. B., The realization of reliable and valid criterion measures for special undergraduate programs and courses aimed at the development and expression of creative behavior. *Journal of Educational Measurement*, 1964, 1, 97–102.

Morrison, E. J. (and R. M. Gagné), *Development and evaluation of an experimental curriculum for the New Quincy Vocational Technical School*. (8 reports) American Institute for Research, 1965–1966. (American Institutes for Research, 135 North Bellefield Ave., Pittsburgh, Pennsylvania 15213).

Newcomb, T. M. et al., *Persistence and change, Bennington College and its students after 25 years*. New York: Wiley 1967.

Shaycoft, M. F., *The high school years: Growth in cognitive skills*. Project TA-

LENT, 1967. (American Institutes for Research, 135 North Bellefield Ave., Pittsburgh, Pennsylvania, 15213).
Spaulding, R. L., *Durham Education Improvement Program.* Final Report. 3 Vols., School of Education, Duke University 1971.
Suppes, P. and Morningstar, M., Evaluation of three computer assisted instruction programs. *Science,* 1969, 166, No. 3903, 343 ff.
Warburton, F. W. and Southgate, V., *i. t. a.: an independent evaluation.* London: Murray 1969.
Welch, W. W. and Walberg, H. J., A design for curriculum evaluation. *Science Education,* 1968, 52, 10–16.
Wrightstone J. W., et al., *Evaluation of the higher horizons program for underprivileged children.* Cooperative Research Project No. 1124. (New York, Bureau of Educational Research, Board of Education of the City of New York, 1964).

V. Neuere Veröffentlichungen

Allen, L. R., An evaluation of certain cognitive aspects of the Material Objects unit of the Science Curriculum Improvement Study Elementary Science Program. *Journal of Research on Science Teaching,* 1970, 7, 277–281.
American Institutes for Research. *Evaluative research: Strategies and methods* 1970. (American Institutes for Research, 135 North Bellefield Ave., Pittsburgh, Pennsylvania 15213).
Anderson, R. C., How to construct achievement tests to assess comprehension. *Review of Educational Research,* 1972, 42, 145–170.
Astin, A., The methodology of research on college impact. Part I. *Sociology of Education,* 1970, 43, 223–254.
Beatty, W. H. (Ed.), *Improving educational assessment and an inventory of measures of effective behavior.* Association for Supervision and Curriculum Development. 1969. (National Education Association of the United States. 1201 Sixteenth St., N. W., Washington, D. C. 20036).
Belasco, J. A. and Trice, H. M., *The assessment of change in training and therapy.* New York: McGraw-Hill 1969.
Block, J. H., Criterion-referenced measurement: potential. *School Review,* 1971, 79, 289–297.
Block, J. H. (Ed.), *Mastery learning. Theory and practice.* New York: Holt, Rinehart and Winston 1971.
Bloom, B. S., Mastery learning and its implications for curriculum development. In: Eisner, E. W. (Ed.): *Confronting curriculum reform.* Boston, Mass.: Little, Brown and Co. 1971. See also Comment by Lee J. Cronbach, *Ibid.*
Boehm, A. et al., An exploratory technique for evaluating curricular interest. *Journal of Curriculum Studies,* 1970, 2, 59–66.
Boruch, R. F., Maintaining confidentiality of data in educational research. *American Psychologist,* 1971, 26, 413–430.

Caro, F. G., Issues in the evaluation of social programs. *Review of Educational Research*, 1971, 41, 87–114.

Center for the Study of Evaluation, *Elementary school hierarchical objectives charts*. 1970. (Center for the Study of Evaluation. 145 Moore Hall, University of California, Los Angeles, California, 90024).

–, *Elementary school test evaluations*. CSE, 1970.

–, *Preschool / Kindergarten Test evaluations*. CSE-ECRC, 1971.

Cronbach, L. J. and Furby, L., How we should measure »change« – or should we? *Psychological Bulletin*, 1970, 74, 68–80.

Cronbach, L. J. and Suppes, P. C. (Eds.), *Research for tomorrow's schools*. New York: MacMillan, 1969.

Curriculum Theory Network, Monograph Supplement Curriculum Evaluation: Potentiality and Reality, 1971/72, Vol. 8/9.

Drew, D. E., *A study of the NSF College Science Improvement Program*. American Council on Education Research Reports 1971. 6, 4. (Office of Research, American Council on Education, One Dupont Circle, Washington, D. C. 20036).

Ebel, R. L., Criterion-referenced measurements: limitations. *School Review*, 1971, 79, 282–288.

–, How to write true-false test items. *Educational and Psychological Measurement*, 1971, 31, 417–426.

Emrick, J. A., An evaluation model for mastery testing. *Journal of Educational Measurement*, 1971, 8, 321–326.

Ferguson, R. L., *Computer assistance for individualized measurement*. Learning Research and Development Center, University of Pittsburgh, March, 1971.

Flanders, N. A., *Analyzing classroom interactions*. Reading, Mass.: Addison-Wesley 1970.

Forehand, G. A., An evaluation system for curriculum innovation. *Teachers College Record*, 1971, 72, 577–591.

Gallagher, J. J., Nuthall, G. A. and Rosenshine, B., Classroom observation. American Educational Research Association. *Monograph Series on Curriculum Evaluation*, No. 6, Chicago: Rand McNally 1970.

Gibbs, G. L. and Ellis, P., *A selected, annotated bibliography of ERIC documents related to the evaluation of schools*. 1971. ERIC TN 5–71–04.

Ginther, R. J., Can a computer evaluate? *School Review*, 1971, 79, 602–613.

Glass, G. V., The growth of evaluation methodology. American Educational Research Association. *Monograph Series on Curriculum Evaluation*, No. 7, Chicago: Rand McNally [1971].

Gronlund, N. E., *Measurement and evaluation in teaching*. New York: Macmillan 1971 (2nd ed.).

Heinemann, H. N. and Sussna, E., Criteria for public investment in the two-year college: A program budgeting approach. *Journal of human resources*, 1971, 6, 171–184.

Holtzman, W. (Ed.)., *Computer-assisted instruction, testing, and guidance*. New York: Harper and Row 1970.

Jaeger, R. M., School testing to test the schools. In: *Proceedings of the 1970 Conference on Testing Problems.* Princeton: Educational Testing Service 1971, 39–52.

Jungwirth, E., An evaluation of the obtained development of the intellectual skills needed for »Understanding of the Nature of Scientific Enquiry« by B.S.C.S. pupils in Israel. *Journal of Research on Science Teaching,* 1970, 7, 277–281.

Levin, H. M., Cost-effectiveness analysis and educational policy: profusion, confusion, promise. *Journal of Human Resources,* 1970, 5, 1–33.

Lindvall, C. M., Cox, R. C. and Bolvin, J. O., Evaluation as a tool in curriculum development: the IPI evaluation program. American Educational Research Association. *Monograph Series on Curriculum Evaluation,* No. 5, Chicago: Rand McNally 1970.

Linn, R. L., Werts, Ch. E. and Tucker, L. R., The interpretation of regression coefficients in a school effects model. *Educational and Psychological Measurement,* 1971, 31, 83–93.

Lord, F. M., Robbins-Monro procedures for tailored testing. *Educational and Psychological Measurement,* 1971, 31, 3–31.

Meehl, P. E., Nuisance variables and the ex post facto design. In: Radner, M. and Winokur, S. (Eds.), *Minnesota Studies in the Philosophy of Science,* Vol. IV, Minneapolis: University of Minnesota Press 1970, 373–402.

Merkle, D. G., A leadership workshop on elementary school science: An in-depth evaluation. *Journal of Research in Science Teaching,* 1970, 7, 121–133.

Messick, S., *Evaluation of educational programs as research on educational process.* Princeton: Educational Testing Service, 1970 (RB 70–1).

Meyer, G. R., Reactions of pupils to Nuffield Science Teaching Project trial materials in England at the ordinary level of the General Certificate of Education. *Journal of Research in Science Teaching,* 1970, 7, 283–302.

National Assessment of Educational Progress. Womer, F. B., *What is National Assessment?* 1970. (National Assessment Office, Room 201 A, Huron Towers, 2222 Fuller Road, Ann Arbor, Michigan, 48105).

–, *Objectives.* (Separate booklets available in Art, 1971; Citizenship, 1969; Literature, 1970; Mathematics, 1970; Music, 1970; Reading, 1970; Science, 1969; Social Science, 1970; Writing, 1969.

–, Finley, C. J. and Berdie, F. S. , *The National Assessment approach to exercise development.* 1970.

–, *Report 5 1969–70 Writing: Group results for sex, region, and size of community.* April 1971. Superintendent of Documents. U. S. Government Printing Office, Washington, D. C. 20402.

–, *Report 6 1969–70 Citizenship: Group results for sex, region, and size of community.* July 1971.

–, *Report 7 1969–70 Science: Group and balanced group results for color, parental education, size and type of community and balanced group results for region of the country, sex.* December 1971.

Okey, J. R. and Gagné, R. M., Revision of a Science topic evidence of perfor-

mance on subordinate skills. *Journal of Research in Science Teaching*, 1970, 7, 321–325.

Phi Delta Kappa Commission on Evaluation, *Educational evaluation and decision making*. Bloomington, Ind.: PDK, 1971.

Popham, W. J. et al., Instructional objectives. American Educational Research Association. *Monograph Series on Curriculum Evaluation*, No. 3, Chicago: Rand McNally 1969.

Popham, W. J., Program fair evaluation: summative appraisal of instructional sequences with dissimilar objectives. *National Society for Programed Instruction Journal*, 1969, 8, 6–9.

Resnick, L. B. and Wang, M. C., *Approaches to the validation of learning hierarchies*. Learning Research and Development Center. University of Pittsburgh 1969.

Sawin, E. I., *Evaluation and the work of the teacher*. Belmont, Ca.: Wadsworth 1969.

Schutz, R. E., The role of measurement in education. *Journal of Educational Measurement*, 1971, 8, 141–146.

Smith, J. P., The development of a classroom observation instrument relevant to the Earth Science Curriculum Project. *Journal of Research in Science Teaching*, 1971, 8, 231–235.

Stufflebeam, D. L., The use of experimental design in educational evaluation. *Journal of Educational Measurement*, 1971, 8, 267–274.

Tyler, R. W. (Ed.), Educational evaluation: new roles, new means. *National Society for the Study of Education*, 68th Yearbook, Part II. Chicago: University of Chicago Press 1969.

Walberg, H. J., Curriculum evaluation: Problems and guidelines. *Teachers College Record*, 1970, 71, 557–570.

Wasik, J. L., A comparison of cognitive performance of PSSC and non-PSSC physics students. *Journal of Research in Science Teaching*, 1971, 8, 85–90.

Whitfield, R. C. and Kerr, J. F., Some problems in course evaluation. A British view. *Teachers College Record*, 1970, 72, 267–274.

Wick, J. W. and Beggs, D. L., *Evaluation for decision-making in the schools*. Boston: Houghton Mifflin 1971.

Quellenangaben und Anmerkungen *

(in der Reihenfolge des Bandes)

Christoph Wulf: Curriculumevaluation

Aus: Zeitschrift für Pädagogik 17, 1971, 2, 176–201. (Vom Autor leicht überarbeitete Fassung)

1 Für wertvolle Anregungen danke ich Dr. B. Bloom, University of Chicago, Dr. R. Stake, University of Illinois, Urbana, Prof. Dr. G. Priesemann, Christian-Albrechts-Universität Kiel, und Prof. Dr. D. Kamper, Universität Marburg.

2 Intensiv waren die Bemühungen um die Lehrplanrevision in Hessen, wo man das bestehende Schul- und Fächersystem durch die Entwicklung neuer Lehrpläne für die Klassen 5 bis 10 (Sekundarstufe I) überwinden wollte; sie sollten für die herkömmlichen Schulen und die Gesamtschulen gleichermaßen gültig sein (vgl. Klafki/Lingelbach 1972).

3 Übersicht über die unter dieser Rubrik zu nennenden Versuche zu erhalten, ist schwierig. Als Beispiel sei W. Teschner 1968 genannt.

4 Vgl. D. Knab, 1971; vgl. auch u. a.: Die Ausschreibung der Stiftung Volkswagenwerk für ein Curriculum institutionalisierter Elementarerziehung (CIEL); die Curriculumarbeit des Instituts für die Pädagogik der Naturwissenschaften in Kiel; die Laborschule und das Oberstufenkolleg in Bielefeld; das Schweizer Projekt EBAC, die Entwicklung und begleitende Analyse eines Curriculum; H. Blankertz, Strategie zur Entwicklung des Lehrplans für das Fach »Arbeitslehre«, in: Theorien und Modelle der Didaktik, 1969; vgl. auch Kollegstufe NW 1972.

5 Der entscheidende Grund für die Bedeutungszunahme der Evaluation liegt darin, daß die meisten der die Curriculumreform finanzierenden Institutionen Evaluation als Bestandteil des Projekts fordern, seit der Elementary and Secondary Education Act von 1965 Evaluation als notwendigen Bestandteil der Entwicklung von Curriculumprojekten deklarierte.

Das wachsende Interesse an Evaluation verdeutlicht die gewaltig steigende Zahl der Publikationen zu diesem Thema. B. Bloom setzte 1965 als Präsident der American Educational Research Association (AERA) eine Kommission für Curriculumevaluation ein. Ihr erstes Projekt war die AERA Monograph Series on Cur-

* Die mit einem * gekennzeichneten Anmerkungen stammen vom Herausgeber oder den Übersetzern.

riculum Evaluation; sieben Bände wurden zwischen 1967 und 1971 bei Rand McNally, Chicago, publiziert.

Curriculumevaluation erschien als Thema in drei von fünf Heften der 1969 Review of Educational Research (RER). Das Aprilheft 1970 der RER ist ganz dem Thema der »Educational Evaluation« gewidmet. Vgl. auch: R. W. Tyler (Ed.) 1969.

6 Ein Vorstellungsschema für den Bereich der Evaluation von Lernzielen könnte z. B. unter Heranziehung der Kategorien der Analyse der Lernzielproblematik erstellt werden, die de Corte für die vier Problemkreise (Formulierungsprobleme, Inventarisierungsprobleme, Klassifikationsprobleme, Wertprobleme) entworfen hat (1971); vgl. auch: Frey 1971; Hesse/Manz 1972; Meyer 1972.

Schwierig ist die Beantwortung der umstrittenen Frage, ob Evaluation ein eigener Bereich ist, der sich von Forschung unterscheidet. Gegen eine prinzipielle Unterscheidung unter Betonung ihrer wechselseitigen Beziehungen ist I. Westbury 1970, 252; ebenso dagegen: R. L. Baker 1969, dafür: R. E. Stake 1969. Stake sieht den Hauptunterschied zwischen Unterrichtsforschung, formativer Evaluation, summativer Evaluation und institutioneller Evaluation darin, daß der Grad an Verallgemeinerungsfähigkeit (generalizability) sich verringert. J. K. Hemphill (1969) bemüht sich ebenfalls, die Eigenständigkeit der Evaluation herauszuarbeiten. Entschieden für eine Unterscheidung zwischen Forschung und Evaluation treten A. G. Larkins/J. P. Shaver 1969 ein. Möglicherweise bietet die für die Evaluation zentrale Aufgabe der Wertung und Entscheidungsvorbereitung einen fruchtbaren Ausgangspunkt für die Unterscheidung von Evaluation und Forschung in theoretischer und methodischer Hinsicht.

7 Vgl. dazu die zahlreichen Publikationen des National Assessment of Educational Progress; eine Übersicht über das National Assessment Program bietet J. C. Merwin/F. B. Womer 1969. Vgl. dazu auch die Untersuchungen der International Association for the Evaluation of Educational Achievement (IEA), T. Husén 1969; vgl. auch W. Edelstein 1970.

8 Bloom (1968) hat versucht, einige Unzulänglichkeiten des Testens durch die Entwicklung einer synthetischen Theorie des Testens zu beheben. Sie integriert Messen (measurement) mit dem Ziel der Klassifikation, der Voraussage und des Experimentierens, Evaluation mit dem Ziel, Veränderungen in den Individuen festzustellen, Schätzung (assessment) mit dem Ziel, die Charakteristika von Individuen in bezug auf eine spezielle Umgebung, Aufgabe oder Kriteriensituation zu erfassen. Dabei wird Testen begriffen als »der Akt des Sammelns und Bearbeitens von Evidenz über menschliches Verhalten unter gegebenen Bedingungen zum Zweck des Verstehens, Voraussagens und Kontrollierens von zukünftigem menschlichen Verhalten« (a. a. O., 23). Die Grenze dieser Theorie liegt darin, daß sie nur eine Theorie des Testens ist. Es fehlt die Reflexion auf die verschiedenen Rollen der Evaluation, auf ihre Funktion für Entscheidungsprozesse, auf die für sie konstitutive Aufgabe, Wertungen und Urteile abzugeben. Evaluation ist umfassender als Testtheorie (vgl. Glass 1968).

9 Diese Monographie verdient trotz mehrerer kritischer Einwände Interesse, da sie den Versuch macht, ein System der Evaluation für den umfassenden inno-

vativen Schulversuch mit Individually Prescribed Instruction zu entwerfen. Dabei liegen die Schwerpunkte neben formativer und summativer Evaluation auf der Evaluation des Schülerverhaltens in IPI. Über IPI vgl. die zahlreichen Veröffentlichungen des Learning Research and Development Center der Universität Pittsburgh und des Research for Better Schools Laboratory in Philadelphia.

10 Vgl. als Beispiel für ein System intrinsischer Curriculumevaluation: I. Morrissett/W. W. Stevens 1967 and I. Morrissett/W. W. Stevens/C. P. Woodley 1969.

11 Zu dieser für die Curriculumentwicklung und Curriculumevaluation wichtigen Streitfrage vgl.: RER 1969; besonders dazu: Instructional Objektives, AERA Monograph Series on Curriculum Evaluation, 1969. Hierin der Streit zwischen E. W. Eisner auf der einen Seite, der außer für »instructional objectives« auch für »expressive objectives« eintritt, die die pädagogische Situation, Probleme und ähnliches angeben, aber ihre Ergebnisse nicht vorwegnehmen, und W. J. Popham und H. J. Sullivan auf der anderen Seite, die derartige Lernziele nicht als gleichwertig neben den operationalisierten »instructional objectives« anerkennen wollen. Wenn man »expressive objectives« als eine zulässige Art, Lernziele zu formulieren, anerkennt, ergibt sich als Aufgabe der Evaluation, »nicht eine gemeinsame Norm auf die hervorgebrachten Produkte anzuwenden, sondern die Aufgabe, das zu reflektieren, was produziert wurde, um seine Einzigartigkeit und Bedeutung zu enthüllen. Im expressiven Kontext ist das Produkt wahrscheinlich genauso überraschend für den Hersteller, wie es für den Lehrer ist, dem es begegnet« (Eisner, a. a. O., 16).

12 Z. B.: A. N. Whitehead 1929; Ph. E. Vernon / G. W. Allport 1931; D. W. Oliver/J. P. Shaver 1966; M. Scriven 1966.

13 Im angelsächsischen Bereich z. B.: W. Sellars/J. Hospers; P. H. Nowell-Smith 1954; R. B. Brandt 1959; J. Dewey 1960; M. Warnock 1960; M. Scriven 1966; D. W. Oliver/J. P. Shaver 1966; Ph. Foot 1967; R. S. Peters 51968.

14 Vgl.: F. F. Stephan/P. J. McCarthy 1958; M. Trow 1967; M. Trow 1969; O. F. Furno, 1966.

15 J. P. Guilford 1954; S. Messick 1961; C. H. Coombs, 1964; M. E. Shaw/J. M. Wright 1967; L. Gorlow/G. A. Noll 1967; D. Sjogren/G. W. England/R. Meltzer 1969; D. J. Dowd/S. C. West 1969.

16 W. Stephenson 1953; J. Nunnally 1959; L. W. Downey 1960; M. Sonntag 1968.

17 I. Pool 1959; B. Berelson 1952; D. P. Cartwright 1953.

18 The Center for the Study of Evaluation, University of California, Los Angeles, hat ein »Training Material Development Project« in Arbeit; es werden Materialien für verschiedene Adressaten zur Einführung in Probleme und Techniken der Evaluation entwickelt. – Joel Weiß/Jack Edwards entwickeln ein Evaluationssystem zur formativen Evaluation, Ontario Institute for Studies in Education, Toronto, Kanada.

LEE J. CRONBACH: Evaluation zur Verbesserung von Curricula

Übersetzung von Ines Graudenz (Dipl.-Psych.).

Originaltitel: Evaluation for course improvement; zugrunde gelegte Fassung aus: R. W. Heath, New curricula, Harper & Row 1964, benutzt und zitiert nach Abdruck in: N. E. Gronlund (Ed.): Readings in measurement and evaluation, London: The Macmillan Company ²1970. Erste Fassung in: Teachers College Record 64, 1963, 672–683.

1 Meine Ausführungen zu diesen Fragen konnten durch die Reaktionen, die ich auf die erste Fassung dieses Beitrags von einigen Leitern von Curriculumprojekten und von Kollegen erhielt, präzisiert werden.

MICHAEL SCRIVEN: Die Methodologie der Evaluation

Übersetzung von Gisela Spöhring und dem Herausgeber

Originaltitel: The methodology of evaluation

Erste Fassung: Publication 110 des Social Science Education Consortium, University of Colorado, Boulder 1966.

Zweite Fassung: American Educational Research Association, Monograph Series on Curriculum Evaluation, No. 1, Chicago: Rand McNally 1967.

Dritte Fassung: Sie wurde dem Herausgeber Anfang 1972 zugesandt und liegt der Übersetzung zugrunde. Zitiert wird nach der zweiten Fassung und nur bei Abweichungen in der dritten Fassung nach letzterer. Der Beitrag ist gekürzt.

1* Möbiussche Fläche (nach F. A. Möbius, 1790–1868): einseitige Fläche, veranschaulicht durch z. B. ein Papierband, das um 180° verdreht und zu einem Ring zusammengefügt wird (nach: dtv-Lexikon Bd. 12, München 1966).

ROBERT STAKE: Verschiedene Aspekte pädagogischer Evaluation

Übersetzung vom Herausgeber.

Originaltitel: The countenance of educational evaluation, Teachers College Record 68, 1967, 7, 523–40.

(Der Übersetzung wurde ein Nachdruck mit den Seiten 1–11 zugrunde gelegt, der auch den Zitaten zugrunde liegt).

1 Hier und an anderen Stellen dieses Beitrags beziehen wir uns, um die Darstellung zu vereinfachen, auf den Evaluator und den Pädagogen als zwei verschiedene Personen. Der Pädagoge ist oft sein eigener Evaluator oder ein Mitglied des Evaluationsteams.

2 Eine solche Liste bilden die *Evaluative Criteria,* die von der National Study of Secondary School Evaluation (1960) veröffentlicht worden sind. Es ist eine sehr sorgfältig erstellte Liste von Voraussetzungen und möglichen Prozessen, die im allgemeinen nach Unterrichtsfächern und -inhalten gegliedert ist. Sie ist als Liste wertvoll, da sie auf vernachlässigte Bereiche aufmerksam macht. Ihr Wert liegt auch darin, zur Verbesserung eines Curriculum, das im Entwicklungsstadium ist, beitragen zu können. Sie kann jedoch nur von begrenztem Wert für die

Evaluation sein. Denn sie hilft weder bei der Datenerhebung noch der Interpretation der Daten. Aufgrund ihrer Zielsetzung enthält sie Kriterien (zur Auswahl der Variablen) und überläßt die Beantwortung der Frage nach Normen (Welche »ratings« sind angemessen?) dem Beobachter.

* Die vierte Auflage der Evaluative Criteria erschien 1969; National Study of Secondary School Evaluation, 1785 Massachusetts Avenue, N. W. Washington D. C. 200 366.

Daniel L. Stufflebeam: Evaluation als Entscheidungshilfe

Übersetzung von Gudrun Eggert und dem Herausgeber.

Originaltitel: Evaluation as enlightment for decision making, in: Improving educational assessment and an inventory of measures of affective behavior, hg. W. H. Beatty, Association for Supervision and Curriculum Development, National Education Association, Washington D. C. 1969, 41–73.

1 Public Law 89–10: The Elementary and Secondary Education Act of 1965, Titel I.

2 Diese Kriterien sind auf den Seiten 70–71 der Titel – III – Richtlinien aufgeführt: A manual for project applicants and grantees, U.S. Office of Education, Washington D. C. 1967.

3 Egon G. Guba: Evaluation and the process of change, Notizen und Arbeitspapiere bezüglich der Administration von Programmen des Titel III des Public Law 89–10, The Elementary and Secondary Education Act of 1965, erweitert durch Public Law 89–750, April 1967, 312.

4 a. a. O.

5 Citizen Committee for Children of New York, Inc., Newsletter, Bericht von Mrs. Nathan W. Levin, Vorsitzende der Educational Services Section, vor dem Subcomitee on the Elementary and Secondary Education Act of the Education and Labor Committee of the House of Representatives, März 1967.

6 Egon G. Guba: Methodological Strategies for educational change, ein Referat, das auf der Tagung über Strategien zur pädagogischen Reform gehalten wurde, die vom 8.–10. November 1965 in Washington D. C. stattfand.

7 Das EPIE-Forum. Monatliche Veröffentlichung des Instituts für Informationsaustausch über Erziehungsprodukte, das von und für Pädagogen geschaffen wurde. New York: Educational Products Information Exchange Institute.

Marvin C. Alkin: Die Aufwands-Effektivitäts-Evaluation von Unterrichtsprogrammen

Übersetzung von Manfred Weiß (Dipl.-Kfm.) und Hans v. Schaper (cand. rer. pol.).

Originaltitel: Evaluating the cost-effectiveness of instructional programs (Referat, gehalten auf dem »Symposium on Problems in the Evaluation of Instruction, University of California, Los Angeles, Dezember 1967), Report No. 25, Los Angeles: University of California, Center for the Study of Evaluation of Instructional Programs 1969.

Abgedruckt auch in: M. C. Wittrock/D. E. Wiley, The evaluation of instruction, issues and problems, New York: Holt, Rinehart and Winston 1970, 221–238.

1 Sehr zum Leidwesen vieler unwilliger Beamter der Schulverwaltung gestehen wir jedoch daß dies ein tunlicher Ausgangspunkt wäre.

2 Manches spricht dafür, daß dies ein vernünftiges Verfahren ist. Vgl. J. S. Bekker 1962; Miller: Income and Higher Education, in: S. J. Muskin 1962; T. Schultz 1961.

Gene V. Glass: Die Entwicklung einer Methologie der Evaluation

Übersetzung von Hannes Graudenz (Dipl.-Psych.) und dem Herausgeber.

Originaltitel: The growth of evaluation methodology

Der deutschen Übersetzung liegt das dem Herausgeber vom Autor zugesandte Manuskript zugrunde.

1 Ein »allgemeines Phänomen« ist nachgewiesen oder kann entdeckt werden in einem weiten Feld von scheinbar verschiedenen Erscheinungen und wird als Kriterium zur Prüfung eines wissenschaftlichen Begriffs herangezogen. Ohne eine solche Qualifikation würde es bereits »Einschätzung wissenschaftlicher Wahrheit« bedeuten, empirisch festzustellen, daß man seine Schlüssel verloren hat. Der Begriff der Generalisierbarkeit von erwarteten Ergebnissen ist wichtig für die Unterscheidung von Evaluation und Forschung; er ist auch von großer praktischer Bedeutung beim Entwurf einer Evaluations-Untersuchung (vgl. Stake 1969).

2 In diesem Abschnitt beziehe ich mich weitgehend auf die Geschichte der North Central Association von Calvin O. Davis (1945).

3 Die Pionierarbeit von Joseph M. Rice mag manchem zu dieser Zeit bekannt gewesen sein, wurde aber wahrscheinlich eher als tendenzieller Journalismus denn als pädagogische Forschung angesehen.

Arno A. Bellack: Methoden zur Beobachtung des Unterrichtsverhaltens von Lehrern und Schülern

Übersetzung von Dorothea Szymanski (Dipl.-Psych.).

Originaltitel: Methods for observing classroom behaviour of teachers and students, in: K. Ingenkamp: Methods for the evaluation of comprehensive schools, Weinheim, Berlin, Basel 1969, 187–215.

1* Zum gegenwärtigen Zeitpunkt liegen bereits 16 Bände mit Instrumenten zur Unterrichtsbeobachtung (79 Instrumente) vor (vgl. Bibliographie).

2* Vgl. dazu auch Wulf 1972a.

Graham A. Nuthall: Ausgewählte neue Untersuchungen zur Unterrichtsinteraktion und zum Lehrverhalten: Ein kritischer Bericht

Übersetzung von J. Hermann (M. A.), G. Hermann (Dipl.-Psych.) und dem Herausgeber.

Originaltitel: A review of some selected recent studies of classroom interaction and teaching behavior.

American Educational Research Association, Monograph Series on Curriculum Evaluation, No. 6, Chicago: Rand McNally 1970, 6–29.

Dieser Beitrag erschien zuerst unter dem Titel: Types of research on teaching im New Zealand Journal of Educational Studies 3, 1968, 2.

1 Der Verfasser möchte sich bei den Professoren Nathaniel Gage (Stanford University) und Barak Rosenshine (University of Illinois, Urbana-Campaign), für verschiedene hilfreiche Hinweise bedanken.

2* Zum gegenwärtigen Zeitpunkt liegen bereits 16 Bände mit 79 Instrumenten zur Unterrichtsbeobachtung vor (vgl. Bibliographie).

3 Persönliche Mitteilung

4 Als dieser Artikel bereits geschrieben war, hat Professor Gage mich darauf hingewiesen, daß die Frage der zeitlichen Dauer der Stichprobe *irrelevant* sei, »insofern als das Kriterium der Effektivität tatsächlich unbeeinflußt ist oder entsprechend den Schülereigenschaften verändert wird, die durch andere Einflüsse als die, die in dieser Unterrichtsstichprobe wirksam sind, beeinflußt werden.« Professor Gage hat hier eine wichtige Frage angeschnitten, indem er darauf hinweist, daß die Gültigkeit der Ergebnisse solcher Untersuchungen nicht in erster Linie von dem Umfang der Unterrichtsstichprobe, sondern von der Gültigkeit und *Sensitivität* der erhaltenen Kriteriumswerte abhängig ist.

Samuel Ball/Gerry Ann Bogatz: Das erste Jahr von Sesame Street. Eine Evaluation.

Übersetzung und Anmerkungen von Josef Volk.

Der vorliegende Text ist eine Kurzfassung des Gesamtberichts: The first year of Sesame Street. An evaluation, Princeton, New Jersey, Educational Testing Service 1970. Er erschien unter dem Titel: A summary of the major findings. In: The first year of Sesame Street. An Evaluation, Princeton, New Jersey: ETS 1970.

Abdruck und Übersetzung erfolgte mit freundlicher Genehmigung des Children's Television Workshop.

1* Zur Verdeutlichung der angesprochenen Sachverhalte und Fertigkeiten soll aus dem Gesamtbericht (Appendix B) die *Beschreibung der Tests, Untertests und einiger Beispiele für Testaufgaben* wiedergegeben werden.

Körperteiletest

a) Zeigen – 10 Testaufgaben – Das Kind zeigt auf Teile seines Körpers, wenn diese vom Tester genannt werden. Fünf Testaufgaben sind im Nachtest weggelassen worden, da über 95 % der Kinder sie im Vortest richtig beantwortet hatten.

b) Benennen – 20 Testaufgaben – Das Kind nennt die Körperteile, die vom Tester gezeigt werden. Fünf Testaufgaben wurden im Nachtest weggelassen.

c) Funktion (zeigen) – 8 Testaufgaben – Das Kind zeigt auf Bilder von Körperteilen, die bestimmte Funktionen ausführen.

d) Funktion (nennen) – 4 Testaufgaben – Das Kind nennt den Namen des Körperteils, mit dem eine bestimmte Funktion ausgeführt wird. (Z. B. Du gehst mit deinen Füßen. Du riechst mit deiner Nase. Womit siehst Du?)

Buchstabentest

a) Erkennen von Buchstaben – 8 Testaufgaben – Das Kind muß aus vier Buchstaben, die ihm gezeigt werden, einen ihm genannten auswählen.

b) Benennen von Großbuchstaben – 16 Testaufgaben – Das Kind nennt jeden Großbuchstaben, auf den der Tester zeigt.

c) Benennen von Kleinbuchstaben – 8 Testaufgaben – Das Kind nennt jeden Kleinbuchstaben, auf den der Tester zeigt.

d) Vorgegebene Buchstaben in Wörtern finden – 4 Testaufgaben – Das Kind zeigt auf das von drei Wörtern, das den gezeigten Buchstaben enthält.

e) Erkennen von Buchstaben in Wörtern – 4 Testaufgaben – Das Kind zeigt auf das von den Wörtern, das den vom Tester genannten Buchstaben enthält.

f) Anfangslaute – 4 Aufgaben im Vortest und 6 im Nachtest – Das Kind wählt das Wort aus, das mit dem vom Tester genannten Buchstaben beginnt. Dazu werden dem Kind Wörter vorgesprochen und Wortbilder gezeigt.

(Z. B.: Das heißt Socke, Tisch, Auto, Ring. Welches Wort beginnt mit einem T?)

g) Wörter lesen – 6 Testaufgaben – Das Kind liest das Wort, das ihm gezeigt wird, vor.

h) Aufsagen des Alphabets – 1 Testaufgabe –.

Formentest

a) Erkennen von Formen – 4 Testaufgaben –

Dem Kind werden vier verschiedene Formen gezeigt. Es deutet auf die, die der Tester mit Namen nennt.

b) Benennen von Formen – 4 Testaufgaben –

Das Kind nennt den Namen der Form, die der Tester zeigt.

Zahlentest

a) Erkennen von Zahlen – 6 Testaufgaben –

Dem Kind werden vier verschiedene Zahlen gezeigt. Es deutet auf die vom Tester genannte Zahl.

b) Zahlen benennen – 15 Testaufgaben –

Das Kind nennt den Namen der Zahl, auf die der Tester zeigt.

c) Zahlverständnis – 6 Testaufgaben – Das Kind zeigt auf Dinge, die in einer bestimmten Anzahl vorliegen, oder es nimmt eine bestimmte Anzahl von Knöpfen von einem Stapel von 10 Stück weg.

d) Zählen – 9 Testaufgaben – Das Kind zählt eine unterschiedliche Anzahl von Bildern, Knöpfen oder Teilen seines Körpers.

e) Addieren und Subtrahieren – 7 Testaufgaben – Das Kind löst einfache Rechenaufgaben.

f) Zählen von 1 bis 20 – 1 Testaufgabe.

Parallelisierter Untertest für Buchstaben, Zahlen und Formen

– 11 Testaufgaben und eine Beispielaufgabe –

Dem Kind werden nacheinander vier Bilder, Buchstaben, Zahlen, geometrische Figuren oder Wörter gezeigt. Unter den vier gleichartigen Gegenständen muß es dann jeweils den herausfinden, der einer Vorlage entspricht.

Sortiertest

6 Testaufgaben – Das Kind wählt unter vier Bildern das aus, das nicht zu den anderen paßt, da es sich in Größe, Form, Zahl und Funktion unterscheidet.

Beziehungstest

17 Testaufgaben – Das Kind zeigt auf ein Bild, das ein Verhältnis in der Größe, der Stellung, der Menge oder der Entfernung zeigt. Die Kenntnis der Mengenbeziehung zeigt das Kind mit Hilfe von Spielknöpfen.

Klassifikationstest

24 Testaufgaben – Das Kind bekommt Bilder von drei Gegenständen gezeigt, die eine Eigenschaft gemeinsam haben (z. B. Größe, Form, Zahl oder Funktion). Es muß dann von vier anderen Bildern das auswählen, das zu den drei zuerst gezeigten paßt oder dasselbe zeigt.

Das Kind begründet auch, warum das Bild zu den anderen gehört. Es gibt ein Beispiel für ein bestimmtes Merkmal.

(Z. B.: Die Menschen tragen Schuhe. Die Menschen tragen Hemden. Was tragen die Menschen noch?)

Puzzle-Test

10 Testaufgaben, aber nur fünf gleich in Vor- und Nachtest –

Das Kind zeigt auf eines von vier Bildern mit dem gleichen Thema, auf dem etwas falsch ist oder etwas fehlt. Es erzählt dann dem Tester, was in dem Bild falsch ist oder fehlt.

2* Bei den Tabellen und Abbildungen handelt es sich um eine Auswahl aus dem Gesamtbericht, bei der auch die Ergebnisse der Untertests weggelassen wurden.

3* Zur Beschreibung der angewandten statistischen Methoden vergleiche den Gesamtbericht.

4* Im November 1971 wurden die Ergebnisse der Evaluation des zweiten Jahres veröffentlicht: Gerry Ann Bogatz/Samuel Ball: The second year of Sesame Street: an continuing evaluation, Princeton, New Jersey (Volume I [Darstellung], Volume II [Tabellen etc.]).

Dazu auch eine Kurzfassung: A summary of the major findings. In: The second year of Sesame Street: a continuing evaluation«.

Richard C. Anderson: Eine vergleichende Felduntersuchung Ein Beispiel vom Biologieunterricht in der Sekundarstufe

Übersetzung von Otto Itzel (Dipl.-Soz.)

Originaltitel: A comparative field experiment: An illustration from high school biology, in: J. Th. Hastings (Ed.), Proceedings of the 1968 invitational conference on testing problems, Princeton, New Jersey: Educational Testing Service 1969.

1 Der Autor ist Gerald Faust, John Guthrie und Veronica Drantz, die bei der Entwicklung der Curriculumeinheit mitgeholfen haben, zu großem Dank verpflichtet; gleiches gilt für Gerald Faust, Marianne Roderick und Phillip Zediker, die ihm bei der Erhebung und Auswertung der Daten geholfen haben. Zu großem Dank ist er auch Robert Stake verpflichtet, der einen Entwurf dieses Beitrags kritisch begutachtete. Die hier dargestellte Untersuchung wurde teilweise von der National Science Foundation finanziell unterstützt.

2* Vergleiche dazu auch Block 1971 und Wulf 1971 b.

3 Der prozentuale Zuwachs ergibt sich aus dem tatsächlichen Zuwachs, dividiert durch die maximal erreichbare Punktzahl.

4 Zu beachten ist, daß ein Curriculum sich darauf beschränken kann, einen begrenzten Geltungsbereich eines Begriffs oder Gesetzes zu vermitteln.

5 Ein Lehrer blieb infolge eines Versehens bei der Verteilung des Leistungstests bei der Analyse der Ergebnisse unberücksichtigt.

6 Zuvor nicht erwähnt wurden drei Klassen von besonders leistungsstarken Schülern (für die das Programm eigentlich bestimmt war), die die BSCS »Blue Version« benutzten. Die zwei Klassen, die das Programm erhielten, erreichten im Nachtest einen Wert von 83,5 %, während die Klasse, die das Programm nicht erhielt, 72,9 % erreichte.

William W. Cooley: Methoden der Evaluation von Schulinnovationen

Übersetzung von Gudrun Eggert und dem Herausgeber.

Originaltitel: Methods of evaluating school innovations, Manuskript eines Vortrags auf der 79. Annual Convention of the American Psychological Association, Washington, D. C., 3. Sept. 1971.

1 Zur Diskussion dieses Redunanzkoeffizienten vgl. Cooley/Lohnes (1971).

2* »Mastery Learning« kann nach Auffassung von Carroll und Bloom durch die Beeinflussung der folgenden fünf Variablen für 90 % der Schüler erreicht werden:

Eignung für bestimmte Arten des Lernens,

Qualität des Unterrichts,

Fähigkeit, Unterricht zu verstehen,

Ausdauer,

zum Lernen gewährte Zeit.

Zu wichtigen Beiträgen und relevanten Forschungsergebnissen zu diesem The-

ma vgl. Block 1971; zur Kritik: Lee Cronbach, in: Eisner 1971, 69–75; vgl. auch Wulf 1971 b.

3* Vgl. dazu Popham/Husek 1969.

4 Als ich diesen Vortrag fertiggestellt hatte, erfuhr ich zu meiner Freude, daß meine Kollegen Glaser und Resnick (1972) gerade ihren Entwurf für einen kritischen Bericht über den Stand der Unterrichtspsychologie für die Annual Review 1972 fertiggestellt hatten, in dem sie die Untersuchungen diskutierten, bei denen »Eignung statt als Kontrollvariable als abhängige Variable behandelt wird und bei denen versucht wird, diese Variable durch unterrichtliche Maßnahmen zu beeinflussen.« Es dürfte für uns beim Learning Research and Development Center ein aufregendes Jahr werden, wenn wir diese Lücke zwischen dem, was psychometrisch sinnvoll ist, und dem, was wir über die Unterrichtspsychologie wissen, auszufüllen versuchen.

Barry MacDonald: Informationen für Entscheidungsträger – Die Evaluation des Humanities Projects

Übersetzung von Mechthild Hagedorn (Dipl.-Psych.) und dem Herausgeber.

Originaltitel: Briefing decision-makers. The evaluation of the Humanities Curriculum Project, Center for Applied Research in Education, University of East Anglia, Norwich 1971.

1* Dieses Projekt basiert auf folgenden 5 Prämissen:
Im Unterricht sollen mit den Jugendlichen kontroverse Fragenkomplexe behandelt werden.
Der Lehrer ist in diesem Stadium der Erziehung bereit, sich beim Unterrichten der Kontroversen neutral zu verhalten, d. h. seine eigene Position nicht deutlich zu machen.
Im Mittelpunkt der Behandlung kontroverser Fragen sollte die Diskussion, nicht die Unterrichtung stehen.
In der Diskussion sollte die Verschiedenheit der Auffassungen der Teilnehmer erhalten bleiben; es sollte nicht versucht werden, einen Konsens herzustellen.
Als Diskussionsleiter sollte der Lehrer für die Qualität des Unterrichts verantwortlich sein (The Humanities Project, an introduction, S. 1, Heinemann, Educational Books, London 1970).

Klaus Nagel / Ulf Preuss-Lausitz: Thesen zur wissenschaftlichen Begleitung von Versuchen und Modellen im Bildungssystem

Aus: Zeitschrift für Pädagogik 17, 1971, 4, 453–462.

1 Der vorliegende Beitrag stellt die überarbeitete Fassung von Thesen dar, die die Verf. auf der Tagung »Probleme wissenschaftlicher Begleitung von Vorschulversuchen« zwischen dem 28. u. 30. 4. 1971 im Pädagogischen Zentrum Berlin vertreten haben.

2 Vgl.: Tagung über wissenschaftliche Begleitung und Beratung von Gesamt-

schulen. Berlin, 19.–21. 11. 1970, Protokolle. Hrsg.: Arbeitsgruppe Gesamtschulen im Pädagogischen Zentrum Berlin, Kurzinformationen/Arbeitspapiere 3/1971. – Weitere Beiträge zur wissenschaftlichen Begleitung von Gesamtschulen in: Gesamtschulen Informationsdienst 3/70, 4/70, PZ Berlin.

3 Tagungsbericht: Probleme wissenschaftlicher Begleitung von Vorschulversuchen, PZ Berlin 1971.

4 Vgl. die Angaben über die Projekte in den Bundesländern im Tagungsbericht a. a. O.

5 Z. B. die Bürgerinitiativen in Frankfurt/M. zur Senkung der Klassenfrequenzen in der Grundschule.

6 Vgl. hierzu die Praxis von Elternbriefen, insbesondere der Peter-Pelikan-Briefe d. Arbeitskreises Neue Erziehung, Berlin.

Literaturverzeichnis

ACHTENHAGEN, F., H. MEYER (Hrsg.): Curriculumrevision – Möglichkeiten und Grenzen. München 1971.

ADELSON, M., M. ALKIN, C. CAREY, O. HELMER: Planning education for the future: Comments on a pilot study. American Behavioral Scientist vom 10. April 1967 (Gesamtausgabe).

ALKIN, M. C.: Towards an evaluation model: A system approach. Working Paper No. 4, Los Angeles, Univ. of California: Center for the Study of Evaluation (CSE), 1967.

ALKIN, M. C.: Evaluation theory development. UCLA Evaluation Comment, vol. 2, No. 1, Los Angeles: Center for the Study of Evaluation 1969 a.

ALKIN, M, C.: The use of behavioral objectives in evaluation: Relevant or irrelevant? Paper presented to the Eighteenth Annual ETS Western Regional Conference on Testing Problems, San Francisco, Calif., May 9, 1968 b.

ALKIN, M. C.: Evaluating the cost-effectiveness of instructional programs. Report No. 25, Los Angeles: UCLA, Center for the Study of Evaluation, 1969; abgedruckt auch in: M. C. Wittrock/D. E. Wiley (Eds.) 1970

AMERICAN ASSOCIATION OF COLLEGES FOR TEACHER EDUCATION: Professional teacher education. Washington, D. C.: The Association 1968.

AMERICAN LIBRARY ASSOCIATION AND NATIONAL EDUCATION ASSOCIATION: Standards for school media programs. Chicago: American Library Association (50 E. Huron St., Chicago, Ill. 60611) 1969.

AMIDON, E. J., N. A. FLANDERS: The role of the teacher in the classroom. Minneapolis: Paul S. Amidon and Associates 1963.

AMIDON, E. J.: A. SIMON: Teacher pupil interaction. Review of Educational Research 35, 1965, 130–139.

AMIDON, E. J.: Interaction analysis: Recent development. Paper delivered at American Educational Research Association Annual Meeting, 1966.

ANDERSON, R. C.: Discussion of instructional variables and learning outcomes. In: M. Wittrock, D. Wiley (Eds.), The evaluation of instruction. New York: Holt, Rinehart & Winston 1960, 126–133.

ANDERSON, R. C., G. W. FAUST: The effects of strong formal prompts in programed instruction. American Educational Research Journal 4, 1967, 345–352.

ANDERSON, R. C., G. W. FAUST, M. RODERICK: Overprompting in programed instruction. Journal of Educational Psychology 59, 1968, 88–93.

ANDERSON, R. C., G. W. FAUST: Educational psychology. New York: Dodd-Mead 1972.

ASCHNER, M.: The analysis of verbal interaction in the classroom. In: A. Bellack (Ed.): Theory and research in teaching. New York, Columbia University: Teachers College Press 1963, 53–78.

ASCHNER, M., J. J. GALLAGHER: A system for classifying thought processes in the context of classroom verbal interaction. Urbana, University of Illinois: Institute for Resarch on Exceptional Children 1965.

ATKIN, J. M.: Some evaluation problems in a course content improvement project. Journal of Research in Science Teaching 1, 1963, 129–132.

AUSUBEL, D. P.: An evaluation of the BSCS approach to high school biology. American Biology Teacher 28, 1966, 176–186.

BAETHGE, M.: Ausbildung und Herrschaft. Frankfurt 1970.

BAKER, R. L.: Curriculum evaluation. Review of Educational Research 39, 1969, 3, 339–358.

BALL, S., G. A. BOGATZ: A summary of the major findings. In: The first year of Sesame Street. An evaluation, Princeton, N. J.: Educational Testing Service 1970.

BASSAM, H.: Teacher understanding and pupil efficiency in mathematics: A study of relationship. Arithmetic Teacher 9, 1962, 383–387.

BECKER, G. S.: Investment in human capital: Theoretical analysis. Journal of Political Economy, Oktober 1962.

BECKER, H.: Bildungsforschung und Bildungsplanung. Frankfurt 1971.

BECKER, H.: Bildungsforschung und Praxis. betrifft: erziehung 1971, 5, 31–34.

BELLACK, A. A.: Methods of observing classroom behaviour of teachers and students. In: K. Ingenkamp, Methods for the evaluation of comprehensive schools. Weinheim, Berlin, Basel 1969, 187–215.

BELLACK, A. A., H. M. KLIEBARD, R. T. HYMAN, F. L. SMITH: The language of the classroom. USOE Cooperative Research Project, New York, Columbia University: Teachers College Press 1966.

BERELSON, B.: Content analysis in communications research. Glencoe, Ill.: Free Press 1952.

BERICHT der Vorbereitenden Kommission unter Leitung von W. Klafki: Zur Lehrplanrevision für die Sekundarstufe in Hessen. Ohne Ortsangabe 1969.

BERLAK, H.: Comments. In: I. Morrissett (Ed.), Concepts and structure in the new social science curricula, New York: Holt, Rinehart and Winston 1966, 88–89.

BERLAK, H.: Values, goals, public policy and educational evaluation. Review of Educational Research 40, 1970, 2, 261–278.

BIDDLE, B. J., R. S. ADAMS: An analysis of classroom activities. Final Report, USOE Contract No. 3 – 20 – 002, Columbia, University of Missouri: Center for Research in Social Behaviour 1967.

BIDDLE, B. J.: Methods and concepts in classroom research. Review of Educational Research 37, 1967, 337–357.

BIDDLE, B. J.: Facets of Teacher Role Research. 1968. (Mimeo.)

BILLERBECK, K.: Kosten-Ertrags-Analyse. Berlin 1968.

Biological Sciences Curriculum Study. Biological science: An inquiry into life. New York: Harcourt, Brace & World 1963.

BLANKERTZ, H.: Theorien und Modelle der Didaktik. München 1969.

BLAUG, M.: Cost-benefit and cost-effectiveness in educational planning (Directorate for Scientific Affairs. Educational Management Techniques) OECD, Paris, 30th January 1968.

BLOCK, J. H. (Ed.): Mastery learning, theory and practice. New York: Holt, Rinehart and Winston 1971.

BLOOM, B. S. u. a.: Taxonomy of educational objectives, The classification of educational goals. Handbook I: Cognitive domain. New York: David McKay Company 1956.

BLOOM, B. S.: Quality control in education. Tomorrow's teaching. Oklahoma City: Frontiers of Science Foundation of Oklahoma, Inc. 1961, 54–61.

BLOOM, B. S.: Twenty-five years of educational research. American Educational Research Journal 3, 1966, 211–22.

BLOOM, B. S.: Toward a theory of testing which includes measurement-evaluation-assessment. Occasional Report 9, Los Angeles, University of California: Center for the Study of Evaluation 1968.

BLOOM, B. S., T. J. HASTINGS, G. F. MADAUS: Handbook on formative and summative evaluation of student learning. New York: McGraw-Hill 1971.

BOULDING, K. E.: The organizational revolution. New York: Harper & Brothers 1953.

BOULDING, K. E.: Beyond economics. Ann Arbor: University of Michigan Press 1968.

BOULDING, K. E.: Economics as a moral science. American Economic Review 59, 1969, 1–12.

BRACHT, G. H., G. V. GLASS: The external validity of experiments. American Educational Research Journal, 1968, 4, 437–474.

BRANDT, R. B.: Ethical theory. Englewood Cliffs, N. J.: Prentice-Hall 1959.

BROWN, R. W.: Format location of programed instruction confirmations. Journal of Programed Instruction 3, 1966, 1–4.

BRÜGELMANN, H.: Offene Curricula. Zeitschrift für Pädagogik 18, 1972, 1, 95–118.

BRUNER, J. S.: Towards a theory of instruction. Cambridge: Harvard University Press 1966.

BUROS, O. K.: The sixth mental measurements yearbook. Highland Park, N. J.: The Gryphon Press 1965.

CARROLL, J. B.: School learning over the long haul, Chapter 10. In: J. D. Krumboltz (Ed.), Learning and the educational process, Chicago: Rand McNally 1965.

CARTWRIGHT, D. P.: Analysis of qualitative material, Research methods in the behavioral sciences. Ed. L. Festinger, D. Katz, New York: Holt, Rinehart and Winston 1953.

CHAMPAGNE, D. W.: Assessment of the LRDC Follow-Through: Curriculum components and role performance. Pittsburgh: Learning Research and Development Center 1971.

CLARK, D. L., E. G. GUBA: An examination of potential change roles in education. Columbus: The Ohio State University 1965.

COLEMAN, J. S.: Equality of educational opportunity. Washington, D. C.: United States Department of Health, Education, and Welfare, Office of Education 1966.

CONANT, J. B.: Modern science and modern man. New York: Columbia University Press 1952.

COOLEY, W. W.: Methods of evaluating school innovations. Invited Address to 79th Annual Convention, Washington, D. C.: American Psychological Association, Sept. 3, 1971.

COOLEY, W. W., P. R. LOHNES: Predicting development of young adults. Palo Alto: American Institutes for Research 1968.

COOLEY, W. W., P. R. LOHNES: Multivariate data analysis. New York: Wiley 1971.

COOMBS, C. H.: A theory of data. New York: Wiley 1964.

COREY, S. M.: Action research to improve school practices. New York: Columbia University, Teachers College Press 1953.

CORTÉ, E. DE: Analyse der Lernzielproblematik. Zeitschrift für Pädagogik 17, 1971, 1, 75–89.

CRONBACH, L. J.: Course improvement through evaluation. Teachers College Record 64, 1963, 672–683 (vgl. S. 372, Anm. 1).

CRONBACH, L. J., P. SUPPES: Research for tomorrow's schools: Disciplined inquiry for education. New York: MacMillan 1969.

DAVIS, C. O.: A history of the North Central Association. Ann Arbor, Michigan: North Central Association of Colleges and Secondary Schools 1945.

DELL, D., J. HILLER: Computer analysis of teachers' explanations. Paper delivered at the American Educational Research Association Annual Meeting 1968.

DEUTSCHER BILDUNGSRAT, Empfehlungen der Bildungskommission: Strukturplan für das Bildungswesen, Bonn 1970.

DEWEY, J.: Theory of moral life. Ed. A. Isenberg, New York: Holt, Rinehart and Winston 1960.

DEUTSCHES INSTITUT FÜR INTERNATIONALE PÄDAGOGISCHE FORSCHUNG: Der hessische Schulversuch zur Früheinschulung. Mitteilungen und Nachrichten des DIPF, Frankfurt 1970.

DIEDERICH, J.: Fördern im Kernunterricht. Kontrollierte Beobachtungen und didaktische Überlegungen, Hannover (im Druck).

DOWD, D. J., S. C. WEST: An inventory of measures of affective behavior. In: Improving educational assessment and an inventory of measures of affective behavior. Ed. W. H. Beatty, Association for Supervision and Curriculum Development (ASCD), NEA, Washington, D. C. 1969, 90–158.

DOWNEY, L. W.: The task of public education: The perceptions of people. University of Chicago: Midwest Administrative Center 1960.

EDELSTEIN, W., Das »Projekt Schulleistung« im Institut für Bildungsforschung in der Max-Planck-Gesellschaft. Zeitschrift für Pädagogik 16, 1970, 4, 517–529.

EDUCATIONAL TESTING SERVICE: A long, hot summer of committee work on national assessment of education. ETS Developments, vol. XIII, November 1965.

EDWARDS, W., A. TVERSKY: Decision making. Harmondsworth, Middlesex: Penguin Books 1967.

EIGLER. G., H. G. SCHÖNWÄLDER, G. STRAKA, P. STRITTMATTER: Wissenschaftliche Begleituntersuchungen an Modellschulen. Eine Empfehlung zur Bildungsforschung. In: Bildung in neuer Sicht. Schriftenreihe des Kultusministeriums Baden-Württemberg zur Bildungsforschung, Bildungsplanung, Bildungspolitik. Villingen 1971.

EISNER, E.: Confronting curriculum reform. Boston: Little Brown Co. 1971.

EMMER, E. T.: The effect of teacher use of student ideas on student initiation. Paper delivered at American Educational Research Association Annual Meeting, 1968.

FAUST, G. W., R. C. ANDERSON, J. T. GUTHRIE, V. E. DRANTZ: Population genetics: A self-instructional program. I. Basic concepts. II. Genetic stability and change. Urbana, University of Illinois: Training Research Laboratory 1967 (Mimeo.).

FERGUSON, G. A.: On learning and human ability. Canadian Journal of Psychology 8, 1954, 95–112.

FERRIS, F. L., JR.: Testing in the new curriculum: Numerology, ›tyranny‹ or common sense? School Review 70, 1962, 112–131.

FLANAGAN, J. C. u. a.: Design for a study of American youth. Boston: Houghton Mifflin 1962.

FLANDERS, N. A.: Teacher influence, pupil attitudes and achievement. Minneapolis: University of Minnesota Press 1960.

FLANDERS, N. A.: Teacher influence, pupil attitudes, and achievement. United States Department of Health, Education and Welfare, Office of Education, Cooperative Research Monograph No. 12, Washington, D. C.: Government Printing Office 1965.

FLECHSIG, K.-H.: Leitfaden zum Kolleg »Theorie des Unterrichts«. Teil IV. Evaluation, Fachbereich Erziehungswissenschaft der Universität Konstanz, Konstanz 1970.

FOOT, PH.: Moral beliefs, theories of ethics. Oxford: Oxford University Press 1967.

FORTUNE, J. C., N. L. GAGE, R. E. SHUTES: A study of the ability to explain. Paper delivered at American Educational Research Association Annual Meeting, 1966.

FREY, K.: Theorien des Curriculums. Weinheim, Berlin, Basel 1971.

FUCHS, W.: Empirische Sozialforschung als politische Aktion. Soziale Welt 21/22, 1970, 1, 1–17.

FURNO, O. F.: Sample survey designs in education – Focus on administrative utilization. Review of Educational Research 36, 1966, 552–565.

FURST, N. F.: The multiple languages of the classroom. Paper presented at the American Educational Research Association Annual Meeting, 1967. (Erschienen auch als unveröffentlichte Dissertation. Temple University, Philadelphia 1967.)

GAGE, N. L. (Ed.): Handbook of research on teaching. The American Educational Research Association, Chicago: Rand McNally 1963.

GAGE, N. L.: Psychological conceptions of teaching. Lecture presented at Diamond Jubilee of School of Education, New York University 1966a.

GAGE, N. L.: Research on cognitive aspects of teaching. The way teaching is. Washington, D. C.: ASCD and NEA 1966b.

GAGE, N. L.: Teaching methods. In: R. Ebel (Ed.), Encyclopedia of Educational Research, London: MacMillan 41969, 1446–1458.

GAGE, N. L., W. R. UNRUH: Theoretical formulations for research in teaching. Review of Educational Research 37, 1967, 358–370.

GALLOWAY, C. M.: An exploratory study of observational procedures for determining teacher nonverbal communication. Unpublished dissertation, University of Florida 1962.

GALLAGHER, J. J.: Teacher variation in concept presentation in BSCS curriculum programs. Urbana, University of Illinois, Institute for Research on Exceptional Children, 1966.

GALLAGHER, J. J., F. SHAFFER u. a.: A system of topic classification. Urbana, University of Illinois, Institute for Research on Exceptional Children, 1966.

GEIGER, G.: Values and social science, The planning of change. Ed. W. G. Bennis u. a. New York: Holt, Rinehart and Winston 1961.

GEIS, F. JR.: The semantic differential technique as a means of evaluating changes in »affect«. Doctoral Diss. Harvard University, Cambridge, Mass. 1968.

GLASER, R.: Adapting the elementary school curriculum to individual performance. Proceedings of the 1967 invitational conference on testing problems. Princeton, N. J.: Educational Testing Service 1968, 3–36.

GLASER, R., L. B. RESNICK: Instructional psychology. Annual Review of Psychology, 1972, in press.

GLASS, G. V.: Comments on Professor Bloom's paper entitled »Toward a theory of testing which includes measurement-evaluation-assessment«. Occasional Report No. 11, Los Angeles, University of California: Center for the Study of Evaluation 1968. Auch in: Wittrock/Wiley 1971.

GLASS, G. V.: The growth of evaluation methology. Research Paper No. 17, Boulder, University of Colorado: Laboratory of Educational Research 1969; auch in: American Educational Research Association Monograph Seriens on Curriculum Evaluation, No. 7, Chicago: Rand McNally 1971.

GOOLER, D. D.: Data collection for educational decision-making, establishing priorities. Working Paper, Urbana, University of Illinois: Center for Instructional Research and Curriculum Evaluation o. J.

GORDON, T. J., O. HELMER: Report on a long-range forecasting study. Santa Monica, Cal.: Rand Corporation 1964.

GORLOW, L., G. A. NOLL: The measurement of empirically determined values. Educational and Psychological Measurement 27, 1967, 1115–1118.

GROBMAN, H.: Evaluation activities of curriculum projects. American Educational Research Association Monograph Series on Curriculum Evaluation, No. 2, Chicago: Rand McNally 1968.

GUBA, E. G., D. L. STUFFLEBEAM: Evaluation: The process of stimulating, aiding, and abetting insightful action. An address delivered at the Second National

Symposium for Professors of Educational Research. Boulder, Colorado 1968, November 21.

GUILFORD, J. P.: Psychometric methods. New York: McGraw-Hill 1954.

HAND, H. C.: National assessment viewed as the camel's nose. Phi Delta Kappan 47, 1965, 8–12.

HANSEN, W. L.: Total and private rates of return to investment in schooling. Journal of Political Economy, April 1963.

HARVEY, O. J. u. a.: Teachers' beliefs, classroom atmosphere and student behavior. American Educational Research Journal (AERJ) 5, 1968, 151–166.

HASTINGS, J. T.: Curriculum evaluation: The whys of the outcomes. Journal of Educational Measurement 3, 1966, 27–32.

HEMPHILL, J. K.: The relationships between research and evaluation studies. In: R. Tyler 1969a, 189–220.

HESSE, H. A., W. MANS: Einführung in die Curriculumforschung, Stuttgart 1972.

HEYNS, R. W., R. LIPPITT: Systematic observation techniques. In: G. Lindzey (Ed.), Handbook of social psychology. Cambridge, Mass.: Addison-Wesley Publishing Co. 1954, 370–404.

HILLER, J. H., G. FISHER, W. A. KAESS: A computer investigation of characteristics of teaching lecturing behavior. Paper delivered at American Educational Research Association Annual Meeting, 1968.

HIRSCH, W. Z., M. J. MARCUS: Some benefit-cost considerations of universal junior college education. National Tax Journal, June 1966.

HONIGMAN, F. K.: A three dimension system for analysing teacher-pupil interaction in the classroom. Paper delivered at American Educational Research Association Annual Meeting, 1968.

HORST, P.: Psychological measurement and prediction. Belmont, Calif.: Wadsworth Publishing Company 1966.

HUGHES, M.: Development of the means for the assessment of the quality of teaching in elementary schools. Salt Lake City: University of Utah Press 1959.

HUSÉN, T.: International impact of evaluation. In: R. W. Tyler (Ed.), Educational evaluation: New roles, new means. The sixty-eighth yearbook of the National Society for the Study of Education Part II. Chicago: University of Chicago Press 1969, 335–350.

JACKSON, P. W.: Teacher-pupil communication in the elementary classroom: An observational study. Unpublished paper presented at the American Educational Research Association Annual Meeting, 1965.

JACKSON, PH. W.: Life in classrooms. New York: Holt, Rinehart and Winston 1968.

JASTAK, J. F., S. W. BIJOU, F. R. JASTAK: Wide range achievement test. Wilmington, Del.: Guidance Association 1965.

JENSEN, G. E.: The validation of aims for American democratic education. Burgess, Minneapolis 1950.

JOYCE, B., B. HAROOTUNIAN: The structure of teaching. Chicago: Science Research Associates 1967.

KAMPER, D.: Geschichte und Menschliche Natur. Die Relevanz der gegenwärtigen Anthropologie-Kritik für eine Methodologie der Humanwissenschaften. Habilitationsarbeit, Fachbereich Gesellschaftswissenschaften, Marburg 1972.

KAPLAN, A.: The conduct of inquiry. San Francisco, Calif.: Chandler Publishing Company 1964.

KATZMAN, M. T.: Distribution and production in a big city elementary school system. Ann Arbor, Michigan: University Microfilms 1967.

KEMENY, J. G., J. L. SNELL: Mathematical models in the social sciences. Boston: Ginn & Co. 1962.

KEMP, F. D., J. G. HOLLAND: Blackout ratio and overt responses in programed instruction: Resolution of disparate results. Journal of Educational Psychology 57, 1966, 109–114.

KLAFKI, W.: Erziehungswissenschaft als kritisch-konstruktive Theorie: Hermeneutik – Empirie – Ideologiekritik. Zeitschrift für Pädagogik 17, 1971, 351–385.

KLAFKI, W., K. CH. LINGELBACH (Hrsg.): Probleme der Curriculumentwicklung. Entwürfe und Reflexionen. Frankfurt/M., Berlin, München 1972.

KLEIN, S. P., D. A. ROCK, F. EVANS: Using multiple moderators in the prediction of academic success. Princeton: Educational Testing Service 1967.

KNAB, D.: Ansätze zur Curriculumreform in der BRD. betrifft: erziehung 4, 1971, 2, 15–28.

Kollegstufe NW. In: Strukturförderung im Bildungswesen des Landes Nordrhein-Westfalen, Eine Schriftenreihe des Kultusminister, H. 17, Ratingen 1972.

KOMISAR, B. P.: Questions for research on teaching. Philadelphia: Temple University 1968. (Mimeo.)

KOUNIN, J. S.: An analysis of teachers' managerial techniques. Psychology in the Schools, 1967, 4, 221–227.

KRATHWOHL, D. R., B. S. BLOOM, B. B. MASIA: Taxonomy of educational objectives. The classification of educational goals. Handbook II: Affective domain. New York: David McKay Company 1964.

KRATHWOHL, D. R.: Stating objectives appropriately for program, for curriculum, and for instructional materials development. Journal of Teacher Education, 1965, 12, 83–92.

KUHN, T. S.: The structure of scientific revolution. Chicago: University of Chicago Press 1962.

LARKINS, A. G., J. P. SHAVER: Hard-nosed research and the evaluation of curricula. American Educational Research Association Annual Meeting 1969.

LA SHIER, W. S.: The use of interaction analysis in BSCS laboratory block classrooms. Journal of Teacher Education 18, 1967, 439–446.

LAWRENCE, PH. J.: The anatomy of teaching. Australian Journal of Education 19, 1966, 97–109.

LEWY, A.: The practice of curriculum evaluation (dem Herausgeber zugesandtes Manuskript). Jerusalem: The Israeli Curriculum Center of the Ministry of Education (ICCME) 1972.

LIEBEL, M., F. WELLENDORF: Schülerselbstbefreiung. Frankfurt/M. 1969.

LIENERT, G. A.: Testaufbau und Testanalyse, Weinheim 1967.

LINDQUIST, E. F. (Ed.): Educational Measurement. American Council on Education, Washington, D. C. 1951.

LINDVALL, C. M., J. O. BOLVIN: Programed instruction in the schools: An application of programming principles in individually prescribed instruction. Sixty-sixth yearbook of the National Society for the Study of Education, Part II. Chicago: Chicago University Press 1967, 217–254.

LINDVALL, C. M., R. C. COX: Evaluation as a tool in curriculum development. The IPI evaluation program, American Educational Research Association Monograph Series on Curriculum Evaluation, No. 5, Chicago: Rand McNally 1970.

LOHNES, P. R.: Measuring adolescent personality. Pittsburgh: American Institutes for Research 1966.

LOHNES, P. R.: Planning for evaluation of the LRDC instructional model. Pittsburgh: Learning Research and Development Center 1971.

LOHNES, P. R.: Statistical descriptors of school classes. Submitted to Journal of Educational Measurement 1971.

LORD, F. M.: Estimating norms by item-sampling. Educational and Psychological Measurement 22, 1962, 259–268.

LUMSDAINE, A. A.: Assessing the effectiveness of instructional programs. In: R. Glaser (Ed.), Teaching machines and programed learning II. Washington, D. C.: National Education Association 1965.

MACDONALD, B.: Briefing decision-makers. The evaluation of the Humanities Curriculum Project. Norwich, University of East Anglia: Center for Applied Research in Education 1971.

MACDONALD, J. B., E. ZARET: A study of openness in classroom interactions. Paper delivered at the American Educational Research Association Annual Meeting, 1967.

MAGER, R. F.: Preparing objectives for programed instruction. San Francisco: Fearon Publishers 1962.

MAGUIRE, T. O.: Value components of teacher's judgements of educational objectives. A – V Communication Review, 1968, 1, 63–86.

MAGUIRE, T. O.: Decisions and curriculum objectives. A methodology for evaluation. Alberta Journal of Educational Research 16, 1969, 17–30.

MCKEAN, R. N.: Efficiency in government through systems analysis. New York: John Wiley and Sons 1958.

MEDLEY, D. M., H. E. MITZEL: A technique for measuring classroom behavior. Journal of Educational Psychology 49, 1958, 86–92.

MEDLEY, D. M., H. E. MITZEL: Measuring classroom behavior by systematic observation. In: N. Gage (Ed.), Handbook of research on teaching. Chicago: Rand McNally 1963, 247–328.

MEDLEY, D. M., R. A. HILL: A comparison of two techniques for analyzing classroom behaviours. Paper delivered at American Educational Research Association Annual Meeting, 1968.

MERWIN, J. C., F. B. WOMER: Evaluation in assessing the progress of education to provide bases of public understanding and public policy. Ed. R. Tyler 1969 a, 305–334.

MESSICK, S.: The perceived structure of political relationships. Sociometry 24, 1961, 270–278.

MESSNER, R.: Funktionen der Taxonomien für die Planung von Unterricht. Zeitschrift für Pädagogik 16, 1970, 6, 755–779.

METFESSEL, N. S., W. B. MICHAEL: A paradigm involving multiple criterion measures for the evaluation of the effectiveness of school programs. Educational and Psychological Measurement 27, 1967, 931–936.

MEUX, M., B. O. SMITH: Logical dimensions of teaching behavior. Urbana, University of Illinois: Bureau of Educational Research 1961.

MEUX, M.: Studies of learning in the school setting. Review of Educational Research 37, 1967, 539–562.

MEYER, H.: Das ungelöste Deduktionsproblem in der Curriculumforschung. In: (Hrsg.) F. Achtenhagen, H. Meyer 1971, 106–132.

MEYER, H.: Zur wissenschaftlichen Begleitung. In: Kollegstufe NW 1972.

MEYER, H.: Einführung in die Curriculum Methodologie, München 1972.

MILBERG, H.: Schulpolitik in der pluralistischen Gesellschaft. Hamburg 1970.

MILES, M. B.: Innovation and education. New York: Columbia University 1964.

MODELL, W.: Hazards of new drugs. Science 139, 1963, 1180–1185.

MORRISSETT, I., W. W. STEVENS: Curriculum analysis. Social Education 31, 1967, 483–487.

MORRISSETT, I., W. W. STEVENS, C. P. WOODLEY: A model for analyzing curriculum materials and classroom transactions. Social Studies Curriculum Development, 39th Yearbook, Ed. D. Fraser, National Council for the Social Studies, NEA. Washington, D. C. 1969, 229–276.

MUSHKIN, S. J. (Ed.): Economics of higher education. Washington, D. C.: United States Department of Health, Education, and Welfare, Office of Education 1962.

NAGEL, K., U. PREUSS-LAUSITZ: Thesen zur wissenschaftlichen Begleitung von Versuchen und Modellen im Bildungssystem. Zeitschrift für Pädagogik 17, 1971, 4, 453–462.

NOWELL-SMITH, P. H.: Ethics. London: Penguin Books 1954.

NUNNALLY, J.: Tests and measurements, assessment and prediction. New York: McGraw-Hill 1959.

NUTHALL, G. A.: An experimental comparison of alternative strategies for teaching concepts. American Educational Research Journal 5, 1968, 561–584.

NUTHALL, G.: A review of some selected recent studies of classroom interaction and teaching behavior. In: Classroom Observation. American Educational Research Association Monograph Series on Curriculum Evaluation, No. 6. R. E. Stake (Ed.), Chicago: Rand McNally 1970, 6–29.

NUTHALL, G. A., P. J. LAWRENCE: Thinking in the classroom. Wellington: New Zealand Council for Educational Research 1965.

OLIVER, D. W., J. P. SHAVER: Teaching public issues in the high school. Boston: Houghton Mifflin 1966.

OSGOOD, CH. E., G. J. SUCI, P. H. TANNENBAUM: The measurement of meaning. Urbana: The University of Ill. Press 1957.

PERKINS, H. V.: A procedure for assessing the classroom behavior of students and teachers. American Educational Journal 1, 1964, 249–260.

PERKINS, H. V.: Classroom behavior and underachievement. American Educational Research Journal 2, 1965, 1–12.

PETERS, R. S.: Ethics and education. London: George Allen and Unwin, 51968.

PHI DELTA KAPPA, NATIONAL STUDY COMMITTEE ON EVALUATION: Educational evaluation and decision making. Bloomington, Indiana: Peacock 1971.

PODLOGAR, M., B. ROSENSHINE, N. L. GAGE: The teacher's effectiveness in explaining: Evidence on its generality and correlations with students' ratings. Paper delivered at American Educational Research Association Annual Meeting 1968.

POOL, I.: Trends in content analysis. Urbana: University of Illinois Press 1959.

POPHAM, W. J., T. R. HUSEK: Implications of criterion-referenced measurement. Journal of Educational Measurement 6, 1969, 1, 1–9.

POWELL, E. R.: Teacher behaviour and pupil achievement. Paper delivered at American Educational Research Association Annual Meeting, 1968.

PREUSS-LAUSITZ, U., K. NAGEL, W. HOPF: Erwartungen der Gesamtschulen an wissenschaftliche Begleitung und Beratung. Gesamtschulen Informationsdienst 3. Pädagogisches Zentrum Berlin 1970.

PRIESEMANN, G.: Zur Theorie der Unterrichtssprache. Düsseldorf 1971.

PROVUS, M.: Evaluation of ongoing programs in the public school system. In: R. W. Tyler (Ed.), Educational evaluation: New roles, new means. The sixty-eighth yearbook of the National Society for the Study of Education Part II. Chicago: University of Chicago Press 1969.

QUADE, E. S. (Ed.): Analysis for military decisions. Chicago: Rand McNally 1967.

REMMERS, H. H.: Rating methods in research on teaching. In: N. Gage (Ed.), Handbook of research on teaching. Chicago: Rand McNally 1963, 329–378.

REPORT OF THE NATIONAL ADVISORY COMMISSION ON CIVIL DISORDERS. New York: Bantam Books 1968.

RESEARCH FOR BETTER SCHOOLS. Progress Report II: Individually prescribed instruction. Philadelphia: Research for Better Schools 1971.

RESNICK, L. B., M. C. WANG, J. KAPLAN: Behavior analysis in curriculum design: A hierarchically sequenced introductory mathematics curriculum. Pittsburgh: Learning Research and Development Center 1970.

REYNOLDS, L. J.: A strategy for the evaluation of individualization. Pittsburgh: Learning Research and Development Center 1971.

ROBINSOHN, S. B.: Bildungsreform als Revision des Curriculum und Ein Struktur-Konzept für Curriculum-Entwicklung. Neuwied u. Berlin 1971.

ROLFF, H. G.: Sozialisation und Auslese durch die Schule. Heidelberg 1967.

ROSENSHINE, B.: Objectively measured behavioral predictors of effectiveness in explaining. In: N. L. Gage u. a., Explorations of the teacher's effectiveness in explaining, Technical Report No. 4. Stanford, California: Stanford Research and Development Center, School of Education 1968.

ROSENSHINE, B.: Evaluation of classroom instruction. Review of Educational Research 40, 1970, 2, 279–300.

RYAN, D. G.: Characteristics of teachers. Washington, D. C.: American Council on Education 1960.

SCHLAIFER, R.: Probability and statistics for business decisions. New York: McGraw-Hill 1959.

SCHULTZ, T.: Investment in human capital. American Economic Review 51, 1961, 1–16.

SCHWAB, J. J.: The practical: A language for curriculum. Washington: National Education Association, Center for the Study of Instruction (CSI) 1970.

SCHWAB, J. J.: The practical: Arts of eclectic. The School Review 79, 1971 a, 4, 493–542.

SCHWAB, J. J.: Praktische Legitimierung von Curricula. In: Bildung und Erziehung 24, 1971 b, 5, 334–341.

SCRIVEN, M.: The experimental investigation of psychoanalyses. In: S. Hook. (Ed.), Psychoanalysis, scientific method and philosophy. New York: New York University Press 1959, 226–251.

SCRIVEN, M.: Student values as educational objectives. Publication No. 124, Boulder, University of Colo.: Social Sciences Education Consortium 1966.

SCRIVEN, M.: Primary philosophy. New York: McGraw-Hill 1966.

SCRIVEN, M.: Value claims in the social sciences. Publication No. 123, Boulder, University of Colo.: Social Science Education Consortium 1966.

SCRIVEN, M.: The methodology of evaluation. In: R. E. Stake (Ed.), American Educational Research Association Monograph Series on Evaluation, No. 1, Chicago: Rand McNally 1967, 39–89.

SELLARS, W., J. HOSPERS: Readings in ethical theory. New York: Appleton-Century-Crofts 1952.

SHAW, M. E., J. M. WRIGHT: Scales for the measurement of attitudes. New York: McGraw-Hill 1967.

SIEGEL, L., L. C. SIEGEL: The instructional gestalt. In: L. Siegel (Ed.), Instruction: some contemporary viewpoints. San Francisco: Chandler Publishing 1967.

SIMON, A., E. G. BOYER: Mirrors for behavior. An anthology of observation instruments. Philadelphia: Research for Better Schools, vol. 1–6, 1967, vol. 7–14 und vol. 15 u. 16, Zusammenfassung, 1970.

SJOGREN, D. D.: Measurement techniques in evaluation. Review of Educational Research 40, 1970, 2, 303–307.

SJOGREN, D., G. W. ENGLAND, R. MELTZER: The development of an instrument for assessing the personal values of educational administrators. Colo. State University, Fort Collins 1969.

SMITH, B. O.: A concept of teaching. In: B. O. Smith, R. H. Ennis (Eds.), Language and concept in education. Chicago: Rand MacNally 1961.

SMITH, B. O., M. O. MEUX: A study of the logic of teaching. Urbana, University of Illinois: Bureau of Educational Research 1962.

SMITH, B. O., M. O. MEUX: The strategies of teaching. Urbana, University of Illinois: Bureau of Educational Research 1967.

SMITH, E. R., R. W. TYLER: Appraising and recording student progress. Adven-

ture in American education, vol. 3, New York und London: Harper and Brothers 1942 (8-year-study).

SMITH, L. M., W. GEOFFREY: The complexities of an urban classroom. New York: Holt, Rinehart and Winston, 1968.

SOAR, R. S.: An integrative approach to classroom learning. A final report. Philadelphia: Temple University. Public Health Service Grant No. 5-R11-MH 01096 and National Institute on Mental Health Grant No. 7-R11-MH 02045, 1966.

SOAR, R. S.: Pupil growth over two years in relation to differences in classroom process. Paper delivered at American Educational Research Association Annual Meeting 1967.

SOAR, R. S.: Teacher-pupil interaction and pupil growth. Paper delivered at American Educational Research Association Annual Meeting 1968.

SOLTIS, J. F.: Einführung in die Analyse pädagogischer Begriffe. Düsseldorf 1971.

SONNTAG, M.: Attitudes toward education and perception of teacher behavior. American Educational Research Journal (AERJ) 5, 1968, 385–402.

STAKE, R. E.: The countenance of educational evaluation. Teachers College Record 68, 1967 a, 523–540 (vgl. S. 372).

STAKE, R. E.: Toward a technology for the evaluation of educational programs. In: Perspectives of curriculum evaluation. American Educational Research Association Monograph Series on Curriculum Evaluation, No. 1, Chicago: Rand McNally 1967 b.

STAKE, R. E.: A research rationale for EPIE. The EPIE Forum 1, September 1967 c, 7–15.

STAKE, R. E.: Generalizability of program evaluation: the need for limits. Educational Product Report 2, 1969, 5, 39–41.

STAKE, R. E.: Objectives, priorities and other judgement data. Review of Educational Research 40, 1970, 2, 181–212.

STANLEY, J.: Benefits of research design. A pilot study. Final Report, Project No. X-005, Grant oE5-10-272, U. S.-Department of Health, Education and Welfare, U. S. Office of Education, Bureau of Research 1966.

STENHOUSE, L.: Some limitations of the use of objectives in curriculum research and planning. Paedagogica Europaea, Braunschweig 1971, 73–83.

STEPHAN, F. F., J. P. MCCARTHY: Sampling opinions. An analysis of survey procedure. New York: Wiley 1958.

STEPHENSON, W.: The study of behavior. Chicago: The University of Chicago Press 1953.

STEVENSON, CH. I.: Ethics and language. New Haven: Yale University Press 1944.

STUFFLEBEAM, D. L.: A depth study of the evaluation requirement. Theory into Practice 5, 1966, 121–133.

STUFFLEBEAM, D. L.: The use and abuse of evaluation in Titel III. Theory into Practice 6, 1967, 3, 126–33.

STUFFLEBEAM, D. L.: Evaluation as enlightment for decision making. In: Improving, educational assessment and an inventory of measures of affective beha-

vior. Ed. W. H. Beatty. Association for Supervision and Curriculum Development (ASCD), National Education Association (NEA), Washington, D. C. 1969, 41–73.

TABA, H.: Teaching strategies and cognitive function in elementary school children. San Francisco: San Francisco State College 1966.

TABA, H., S. LEVINE, F. ELZEY: Thinking in elementary school children. San Francisco: San Francisco State College 1964.

TAYLOR, P. A., T. O. MAGUIRE: A theoretical evaluation model. The Manitoba Journal of Educational Research, 1966, 1, 12–17.

TAYLOR, P. A., T. O. MAGUIRE: Perceptions of some objectives for a science curriculum. Science Education 51, 1967, 488–493.

TESCHNER, W.: Didaktik und Organisation des Deutschunterrichts an der Gesamtschule Berlin (BBR). Braunschweig 1968.

TESCHNER, W. P. (Hrsg.): Differenzierung und Individualisierung im Unterricht. Göttingen 1971.

TROW, M.: Education and survey research in the social sciences. Ed. C. Y. Glock, New York: Russel Sage Foundation 1967.

TROW, M.: Methodological problems in the evaluation of innovations. Report No. 31, Los Angeles, University of California: Center for the Study of Evaluation 1969.

TURNER, R. L.: Pupil influence on teacher behaviour. Classroom Interaction Newsletter 3, 1968, 5–8.

TYLER, R. W.: The functions of measurement in improving instruction. In: E. F. Lindquist (Ed.), Educational measurement, Washington, D. C.: American Council on Education 1951, 47–67.

TYLER, R. W.: Assessing the progress of education. Phi Delta Kappan 47, 1965, 13–16.

TYLER, R. W.: New dimensions in curriculum development. Phi Delta Kappan 48, 1966, 25–28.

TYLER, R. W. (Ed.): Educational evaluation: New roles, new means. The sixty-eighth yearbook of the National Society for the Study of Education Part II, Chicago: University of Chicago Press 1969 a.

TYLER, R. W.: Basic principles of curriculum and instruction. Chicago und London: The University of Chicago Press [29]1969 b.

VERNON, PH. E., G. W. ALLPORT: A test for personal values. Journal of Abnormal and Social Psychology 26, 1931, 231–248.

WALBESSER, H. H.: Curriculum evaluation by means of behavioral objectives. Journal of Research in Science Teaching, 1963, 1, 296–301.

WALBESSER, H. H.: Science curriculum evaluation: Observations on a position. The Science Teacher 33, 1966.

WANG, M. C., L. B. Resnick, P. R. SCHUETZ: PEP in the Frick elementary school: Interim evaluation report 1968–1969. Pittsburgh: Learning Research and Development Center 1970.

WARNOCK, M.: Ethics since 1900. London: Oxford University Press 1960.

WEBB, E. J., D. T. CAMPBELL, R. D. SCHWARTZ, L. SECHRIST: Unobtrusive mea-

sures: Nonreactive research in the social sciences. Chicago: Rand McNally 1966.

WEISS, J., J. EDWARDS: Formative curriculum evaluation. A manual of procedures. Toronto, OISE, in preparation.

WESTBURY, I.: Curriculum evaluation. Review of Educational Research 40, 1970, 2, 230–260.

WHITEHEAD, A. N.: The aims of education. New York: MacMillan 1929.

WHITHALL, J.: Development of a technique for the measurement of socio-emotional climate in classrooms. Journal of Experimental Education 17, 1949, 347–361.

WILLIAMS, G.: The sanctity of life and the criminal law. New York: Alfred A. Knopf 1968.

WITTROCK, M. C., D. E. WILEY (Eds.): The evaluation of instruction: Issues and problems. New York: Holt, Rinehart, and Winston 1970.

WRIGHT, E. M.: Development of an instrument for studying verbal behaviors in a secondary school mathematic classroom. Journal of Experimental Education 28, 1959, 103–121.

WRIGHTSTONE, W. u. a.: Evaluation of the higher horizons program for underprivileged children. Cooperative Research Project No. 1124. New York: Bureau of Educational Research. Board of Education of the City of New York o. J.

WULF, CH.: Curriculumevaluation. Zeitschrift für Pädagogik 17, 1971a, 2, 175–201.

WULF, CH.: Internationale Kooperation bei der Curriculumentwicklung. Zeitschrift für Pädagogik 17, 1971b, 5, 631–647.

WULF, CH.: Curriculumentwicklung in den New Social Studies in den USA. Entwicklungstendenzen und gegenwärtiger Stand. Aus: Politik und Zeitgeschichte, Beilage zur Wochenzeitschrift Das Parlament, 6B, 1972a.

WULF, CH.: Heuristische Lernziele – Verhaltensziele. Bildung und Erziehung 25, 1972b, 2, 14–23.

WULF, CH.: Evaluation. Ein kritischer Überblick. Neue Sammlung 12, 1972c, 3, 259–284.

ZAHORIK, J. A.: Classroom feedback behavior of teachers. Journal of Educational Research 62, 1968, 147–150.

Autorenverzeichnis

(In alphabetischer Reihenfolge)

Marvin C. Alkin

Professor für Pädagogik, Direktor des Center for the Study of Evaluation, University of California, Los Angeles; geb. 1934; zahlreiche Veröffentlichungen im Bereich der Evaluation, u. a.: Evaluation theory development, Evaluation Comment 2, 1969, 1, Los Angeles: University of California; Educational needs for the 1970's and beyond. Needs of elementary and secondary education for the seventies. General Subcommittee on Education, U. S. House of Representatives, Washington, D. C., U. S. Printing Office, 1970; A theory of evaluation, Working Paper 18, Los Angeles: University of California, Center for the Study of Evaluation 1971.

Richard C. Anderson

Professor für Pädagogik an der Universität of Illinois, Urbana-Champaign, Illinois; geb. 1934; zahlreiche Veröffentlichungen auf dem Gebiet der Pädagogischen Psychologie und des Programmierten Unterrichts, u. a.: Learning in discussions: A resumé of the authoritarian-democratic studies, Harvard Educational Review 29, 1959, 201–215; Can first graders learn an advanced problem solving skill?, Journal of Educational Psychology 56, 1965, 283–294; R. C. Anderson/D. P. Ausubel: Readings in the psychology of cognition, New York: Holt, Rinehart and Winston 1965.

Samuel Ball

Research Psychologist, Educational Studies, Educational Testing Service, Princeton, New Jersey, und Adjunct Associate Professor für Psychologie und Pädagogik, Teachers College, Columbia University, New York; geb. 1933; zahlreiche Veröffentlichungen in Zusammenhang mit der Evaluation von Sesame Street des Children's Television Workshop. U. a.: S. Ball/W. H. MacGinitie: Psychological foundations of education, New York: McGraw-Hill 1968; Educational psychology. In: Encyclopedia of education, MacMillan 1970.

Arno A. Bellack

Professor für Pädagogik am Teachers College, Columbia University, New York; Veröffentlichungen u. a.: (Ed.) 1956 Yearbook of Association for Supervision and Curriculum Development, National Education Association, What shall the High Schools teach?; (Ed.) Theory and research in teaching, New York: Teachers College Press, 1963; A. Bellack/H. Kliebard u. a.: The language of the classroom, New York: Teachers College Press 1966; A. Bellack/Ian Westbury: Research into classroom processes, New York: Teachers College Press 1971.

Gerry Ann Bogatz

Senior Research Assistant im Educational Testing Service in Princeton, New Jersey; zahlreiche Veröffentlichungen in Zusammenhang mit der Evaluation von Sesame Street des Children's Television Workshop.

William W. Cooley

Professor für Pädagogik und Computer Science, University of Pittsburgh; seit 1969 Kodirektor des Learning Research and Development Center; von 1964–1967 Direktor des Projekt TALENT, University of Pittsburgh und American Institutes for Research; geb. 1930. Zahlreiche Veröffentlichungen im Bereich der Datenverarbeitung und Innovationsforschung, u. a.: W. W. Cooley/R. Hummel: Systems approach in guidance, Review of Educational Research 39, 1969, 2, 251–262; Data processing and computing. In: R. L. Ebel (Ed.), Encyclopedia of educational research, Toronto: MacMillan 41969; W. W. Cooley/P. R. Lohnes: Multivariate data analysis, New York: Wiley 1971.

Lee J. Cronbach

Professor für Pädagodik an der Stanford University, geb. 1916; Präsident der American Educational Research Association, American Psychology Association, Psychometric Society. Autor und Herausgeber verschiedener Veröffentlichungen auf dem Gebiet der Pädagogischen Psychologie und Testpsychologie, u. a.: Essentials of psychological testing. New York: Harper and Row 31970; Educational Psychology. New York: Harcourt Brace and World 21964; Cronbach u. a.: Dependability of behavioral measurement, New York: Wiley 1972; L. Cronbach/G. Gleser: Psychological tests and personnel decisions. Univ. Illinois Press 21965; L. Cronbach/P. Suppes: Research for tomorrow's Schools, New York: MacMillan 1969; »Test validation«. In: R. L. Thorndike (Ed.), Educational measurement, American Council on Education 1971.

Gene V. Glass

Professor für Pädagogische Psychologie und Kodirektor des Laboratory of Educational Research an der Colorado University in Boulder; geb. 1940; zahlreiche Veröffentlichungen im Bereich der Datenverarbeitung und Evaluation, u. a.: G. V. Glass/J. C. Stanley: Statistical methods in education and psychology (Textbook), Englewood Cliffs, N. J.: Prentice-Hall 1970; G. V. Glass u. a.: Test items to accompany statistical methods in education und psychology, Englewood Cliffs, N. J.: Prentice-Hall 1970; G. V. Glass/P. A. Taylor: Factor analytic methodology, Review of Educational Research 36, 1966, 5, 566–587.

Barry MacDonald

Direktor des Evaluationsteams am Center for Applied Research in Education, University of East Anglia, Norwich, England; geb. 1933; zuvor Lehrer an Elementar- und Sekundarschulen; Autor zahlreicher Veröffentlichungen im Zusammenhang mit der Evaluation des Humanities Curriculum Project; für 1973 ist die Veröffentlichung zweier Bücher geplant: The experience of innovation und Teaching race: a pilot-study, beide: London, Heinemann.

Klaus Nagel

Dipl. Psych., Direktor am Pädagogischen Zentrum Berlin, Abt. Wissenschaftliche Begleitung und Beratung von Schulversuchen (Gesamtschulen); geb. 1937. Veröffentlichungen u. a.: »Motivation« und »Einführung in die Probleme der Testanwendung«. In: Kleinkindererziehung, Hrsg. G. Hundertmarck/H. Ulshoefer, München 1972; Gesamtschule zwischen Schulversuch und Strukturreform (mit G. Bühlow, W. Hopf, U. Preuss-Lausitz), Weinheim, Berlin, Basel 1972.

Graham Nuthall

Professor für Pädagogik an der University of Canterbury, Christchurch, New Zealand; geb. 1935; Veröffentlichungen u. a.: An experimental comparison of alternative strategies for teaching concepts, American Educational Research Journal 5, 1968, 4, 561–584; G. Nuthall/C. J. Wright: Relationships between teacher behaviors and pupil achievement in three experimental elementary science lessons, American Educational Research Journal 7, 1970, 4, 477–491; G. Nuthall/N. A. Flanders (Eds.): The classroom behavior of teachers, Hamburg: UNESCO International Institute for Education. (Special Number of the International Review of Education) (im Druck).

Evelore Parey

Wissenschaftlicher Oberrat am Pädagogischen Zentrum Berlin; z. Z. Studium der Erziehungswissenschaft, Spezialgebiet Curriculum, an der Stanford University; Veröffentlichungen u. a.: K. Ingenkamp in Zusammenarbeit mit E. Parey: Handbuch der Unterrichtsforschung, Bd. 1–3. (Deutsche Bearbeitung des Handbook of research on teaching von N. L. Gage.), Weinheim, Berlin, Basel 1970.

Ulf Preuss-Lausitz

Dipl.-Soz., wissenschaftlicher Referent am Pädagogischen Zentrum Berlin, Abt. Wissenschaftliche Beratung und Begleitung von Schulversuchen (Gesamtschulen); geb. 1940. Veröffentlichungen: Aufsätze in pädagogischen, bildungspolitischen und politischen Zeitschriften und in Sammelbänden. Mitautor von: Gesamtschule zwischen Schulversuch und Strukturreform, Weinheim, Berlin, Basel 1972.

Michael Scriven

Professor für Philosophie, University of California, Berkeley; geb. in Australien; PhD. Oxford, School of Literae Humaniores; zahlreiche Veröffentlichungen im Bereich der Philosophie, der Naturwissenschaften, Psychologie und Pädagogik, u. a.: Primary Philosophy, New York: McGraw-Hill 1966; Reasoning, Volume I: Argument analysis, Volume II: Scientific method, Volume III: Value issues, New York: McGraw Hill 1972 (in Vorbereitung); Explanation and prediction in evolutionary theory, Science 130, 1959, 477–482; Views of human nature, in: Behaviorism and phenomenology, ed. T. Wan, Chicago: University of Chicago Press 1964, 163–190; Philosophy of the social sciences and education, in: Philosophy and educational development, Boston: Houghton-Mifflin 1966; Environment education, The Journal 8, 1970, 4, 17–20; Objectivity and subjectivity in educational research, in: Yearbook of the National Society for Study of Education, Chicago: Chicago University Press 1972 (in Vorbereitung).

Robert Stake

Kodirektor des Center for Instructional Research and Curriculum Evaluation und Professor für Pädagogische Psychologie, University of Illinois, Urbana-Champaign; geb. 1929; zahlreiche Veröffentlichungen im Bereich der Evaluation, u. a.: Herausgeber der American Educational Research Association Series on Curriculum Evaluation, Bd. 1–7, Chicago: Rand McNally 1967–71; Generalizability of program evaluation. The need for limits, Educational product report 2, 1969, 39–40; Objectives, priorities, and other judgment data, Review of Educational Research 40, 1970, 2, 181–212.

Daniel L. Stufflebeam

Professor für Pädagogik und Direktor des Center of Evaluation an der Ohio State University, Columbus; geb. 1936; Veröffentlichungen im Bereich der Evaluation, u. a.: D. L. Stufflebeam u. a.: Educational evaluation and decision making, Itasca, Illinois: F. E. Peacock Publishers 1971; The use of experimental design in educational evaluation, Journal of Educational Measurement 8, 1971, 4, 267–274; The relevance of the CIPP evaluation model for educational accountability, Journal of Research and Development in Education 5, 1971, 1, 19–25.

Christoph Wulf

Wissenschaftlicher Mitarbeiter am Deutschen Institut für Internationale Pädagogische Forschung; geb. 1944; zahlreiche Veröffentlichungen im Bereich der Curriculumforschung, Evaluation und Friedenserziehung, u. a.: Curriculumevaluation, Zeitschrift für Pädagogik 17, 1972, 2, 175–201; Die Curriculumentwicklung in den New Social Studies, Entwicklungstendenzen und gegenwärtiger Stand, aus Politik und Zeitgeschichte 6, 1972; Heuristische Lernziele – Verhaltensziele, Bildung und Erziehung 25, 1972, 2, 14–24; Evaluation, ein kritischer Überblick, neue Sammlung 12, 1972, 3, 259–284; Auf dem Weg zu einer transnationalen Friedenserziehung, Bildung und Erziehung 25, 1972, 5, 58–68.

*Glossar**

Akkreditationsmodell: Evaluation nach diesem Modell zielt auf die Beurteilung und »Beglaubigung« von Bildungsinstitutionen aufgrund eines Satzes von Kriterien.

Aktionsparameter: Die in einem Modell zwischen vorgegebenen und zu erreichenden Größen stehenden Zwischenglieder nennt man »Aktionsparameter«, soweit sie vom Entscheidungsträger zielgerichtet beeinflußt werden können. Diese beeinflußbaren Merkmale des Modells haben nur instrumentalen Charakter, indem sie der Verwirklichung der Ziele dienen, jedoch nicht selbst Ziel des Prozesses sind.

Augenscheinvalidität s. Validität

Aufwands-Effektivitäts-Analyse (Cost-Effectiveness Analysis): Die Aufwands-Effektivitäts-Analyse ist eine Methode zur rationalen Auswahl unter verschiedenen Projekten. Das günstigste Aufwands-Ertrags-Verhältnis wird durch eine Gegenüberstellung des Aufwands mit dem Ertrag ermittelt, der zwar irgendwie meßbar sein muß, aber nicht in Geldeinheiten ausgedrückt wird.

Auspartialisieren s. Korrelation

Äußere Validität s. Validität

Board of Education: So wird in den USA die Institution bezeichnet, die im allgemeinen die Entscheidungen auf der Ebene der einzelnen Schule, des Schulbezirks oder des Staates trifft. Es stellt z. B. die Schulleiter und Lehrer ein und entscheidet über Neuanschaffungen usw.

Diskrepanz: Als Diskrepanz wird der Unterschied zwischen Intentionen und Ergebnissen bezeichnet.

Dissemination: Unter Dissemination wird die Verbreitung von Curricula in den Schulen und die Einführung der Lehrer in die Arbeit mit ihnen bezeichnet.

Emanzipation: Unter Emanzipation wird der organisierte Lernprozeß verstanden, »der dazu beiträgt, die Fähigkeit des Schülers zur Analyse gesellschaftlicher Zusammenhänge und damit sein Selbstverständnis und seine Handlungsfähigkeit in der jeweiligen historischen Situation zu fördern« (Bericht der vorbereitenden Kommission ... 1969, 11).

* Für Hilfe bei der Herstellung des Glossar danke ich H. Schlattmann (Dipl.-Psych.) und J. Wendeler (Dipl.-Psych.).

Evaluation: Evaluation richtet sich auf die Sammlung, Verarbeitung und Interpretation von Daten mit dem Ziel, bestimmte Fragen über Innovationen zu beantworten und Entscheidungen über sie zu treffen. Das schließt die Beschreibung und Bewertung der Angemessenheit von Zielen, Inhalten und Methoden und die Vorbereitung von Entscheidungen ein.

Kontextevaluation ist eine Evaluation der Voraussetzungen, die der Planung einer Innovation vorausgehen soll und die zur Erhebung der Bedürfnisse der Adressaten dienen soll. Die Aufgabe der *Inputevaluation* besteht darin, Informationen darüber zu liefern, wie Ressourcen im weiteren Sinne eingesetzt werden sollen, um die Ziele eines Bildungsprogramms zu realisieren. *Prozeßevaluation* findet während der Entwicklung eines Bildungsprogramms bzw. Curriculum statt und dient zur unmittelbaren Rückmeldung über den Realisierungsprozeß. Sie entspricht in vieler Hinsicht *formativer Evaluation,* mit der die Evaluation von Bildungsprogrammen während des Prozesses der Entwicklung (Formung) gemeint ist. *Produktevaluation* oder *Outputevaluation* zielt auf die Evaluation der Ergebnisse eines Curriculum oder Modellversuchs, sie wird auch als *Ergebnisevaluation* bezeichnet. Als Synonym wird häufig der Begriff *summative Evaluation* verwandt, mit der eine (vorläufig) abschließende Evaluation eines Bildungsprogramms gemeint ist. Eine wichtige Form der Evaluation ist die *intrinsische Evaluation,* mit der eine Analyse des Bildungsprogramms bezeichnet wird, bei der nicht seine Auswirkungen auf die Adressaten untersucht werden; eine ideologiekritische Untersuchung der Ziele eines Curriculum ist z. B. eine Form der intrinsischen Evaluation. Die Evaluation mit Hilfe von Kontrollgruppen wird als *vergleichende Evaluation* bezeichnet. Bei Verzicht auf Kontrollgruppen spricht man von *nicht-vergleichender Evaluation.*

Exzeß s. Verteilung

Faktorenanalyse: Die Faktorenanalyse ist eine Bezeichnung für eine Reihe statistischer Verfahren, mit deren Hilfe Zusammenhänge zwischen empirischen Daten beschrieben werden. Zusammenhänge zwischen den Daten, dargestellt in einer Matrix von Korrelationskoeffizienten, werden mit Hilfe möglichst weniger, voneinander weitgehend unabhängiger Faktoren beschrieben. Das empirische Ausgangsmaß ist eine Matrix von Korrelationskoeffizienten, und das Ziel der Faktorenanalyse besteht in der Extraktion möglichst weniger Faktoren, mit denen sich die in den Korrelationen zum Ausdruck kommenden Zusammenhänge möglichst einfach darstellen lassen.

Je nach Art des Korrelationsbereichs spricht man von R- oder Q-Technik. Bei der *R-Technik* drückt der Koeffizient die Ähnlichkeit von Meßinstrumenten aus, bei der *Q-Technik* die Ähnlichkeit von Personen. Bei der R-Technik lassen sich die extrahierten Faktoren als Grundfähigkeiten bzw. Grundeigenschaften interpretieren, bei der Q-Technik als Persönlichkeitstypen.

Feedback s. Rückmeldung

Felduntersuchung: Als Felduntersuchung bezeichnet man in den Sozial- und Verhaltenswissenschaften die Vorgehensweise, bei der das Verbleiben der untersuchten Individuen in der natürlichen Umgebung weitgehend gewährleistet ist. Hauptmethoden sind Beobachtung, Befragung, Interview und Tests.

Formative Evaluation s. Evaluation

F-Test: Ein F-Test ist ein statistisches Prüfverfahren, das feststellt, ob die Varianten von Untersuchungsstichproben nur zufällig voneinander abweichen, d. h. mit hoher Wahrscheinlichkeit derselben Population entstammen oder nicht.

Gewichtung: Bei der Zusammenfassung von Einzelmaßen in einem Gesamtmaß werden die Einzelmaße nach ihrer Bedeutung in diesem Zusammenhang jeweils mit einer Konstanten (ihrem Gewicht) multipliziert. Werden z. B. die Teiltests eines Intelligenztests in einem allgemeinen Intelligenzmaß zusammengesetzt, so kann man diejenigen Teiltests, die zur Bestimmung der allgemeinen Intelligenz besonders bedeutsam sind, höher als die restlichen gewichten.

Gültigkeit s. Validität

Implementation: Unter Implementation versteht man die Einführung von Innovationen in die Schule mit dem Ziel, durch sie die Unterrichtswirklichkeit zu verändern.

Innere Validität s. Validität

Innovation: Unter Innovation versteht man neuartige Ideen, Gegenstände und Interaktionsmuster, die von »Innovatoren« in ein bestehendes System eingeführt werden mit der Absicht, es zu verändern, d. h. unter Zugrundelegung bestimmter Kriterien zu verbessern.

Inputevaluation s. Evaluation

Intrinsische Evaluation s. Evaluation

Kanonischer Korrelationskoeffizient s. Korrelation

Kodierung: Als Kodierung bezeichnet man die Umwandlung einer Botschaft in ein Signal oder eines Signals in eine Botschaft nach einem vorher festgesetzten Kodierungsschlüssel. Im weiteren Sinne versteht man unter Kodierung jede Zuordnung von Einzelfällen zu Klassen eines Merkmales, die durch Symbole gekennzeichnet werden. Z. B. können bei einer Untersuchung alle Versuchspersonen, die der Unterschicht angehören, das Zahlensymbol »1«, alle der Oberschicht angehörenden Versuchspersonen das Zahlensymbol »5« erhalten, wodurch eine Kodierung erfolgt.

Kongruenz: Unter Kongruenz wird die Übereinstimmung von Intentionen und beobachteter Realität vor allem bei den Voraussetzungen, Prozessen und Ergebnissen von Bildungsprogrammen bezeichnet.

Konsistenzanalyse s. Reliabilität

Konstruktvalidität s. Validität

Kontextevaluation s. Evaluation

Kontingenz: Unter Kontingenz werden die (logischen oder empirisch feststellbaren) Beziehungen zwischen den Voraussetzungen eines Bildungsprogramms, den Realisierungsprozessen und den Ergebnissen bezeichnet.

Korrelation: Unter Korrelation versteht man eine Wechselbeziehung bzw. einen Zusammenhang. In der Statistik bezeichnet man als Korrelation eine Maßzahl, die den Grad des wechselseitigen Zusammenhangs von Merkmalen angibt. Die Stärke des Zusammenhangs wird durch den *Korrelationskoeffizienten* ausgedrückt, eine Maßzahl, die zwischen −1 und +1 liegt. Dabei bedeutet +1 vollständige, 0 keine und −1 vollständige umgekehrte Korrelation. Betrachtet

man z. B. die Merkmale Körpergröße und Körpergewicht bei einer Gruppe von Personen, so findet man, daß im allgemeinen höhere Körpergröße mit höherem Gewicht einhergeht; beide Merkmale sind positiv korreliert.

Der Zusammenhang zwischen *mehreren* Merkmalen einerseits und *einem* Merkmal andererseits wird mit einem *multiplen Korrelationskoeffizienten* ausgedrückt.

Bei der Bestimmung eines Zusammenhangs zwischen *zwei Gruppen* von Merkmalen erhält man einen *kanonischen Korrelationskoeffizienten.*

Hängt der Zusammenhang zwischen zwei oder mehreren Merkmalen noch von einem oder mehreren zusätzlichen Merkmalen ab, so läßt sich der Einfluß dieser zusätzlichen Merkmale durch *Auspartialisieren* ausschalten. Man erhält so als Ergebnis eine Partialkorrelation, die »reine Korrelation«, die die ursprünglichen Merkmale miteinander hätten, wenn die zusätzlichen Merkmale keinen Einfluß ausüben würden. Mißt man z. B. bei einer Gruppe von Versuchspersonen zwei motorische Leistungen, die beide mit der allgemeinen Intelligenz der Versuchspersonen zusammenhängen, so läßt sich durch Bestimmung einer Partialkorrelation der Zusammenhang zwischen den motorischen Leistungen bestimmen, wie er wäre, wenn die Intelligenz keinen Einfluß hätte.

Kosten-Nutzen-Analyse (Cost-Benefit Analysis): Die Kosten-Nutzen-Analyse dient der Auswahl von miteinander konkurrierenden Projekten im öffentlichen Sektor aufgrund des günstigsten Kosten-Nutzen-Verhältnisses. Das Kriterium des günstigsten Kosten-Nutzen-Verhältnisses hat die Kosten-Nutzen-Analyse mit der Investitionsrechnung eines privaten Unternehmens gemeinsam; sie unterscheidet sich von letzterer allerdings dadurch, daß die volkswirtschaftlichen Kosten (Aufwand) und der volkswirtschaftliche Nutzen (Ertrag) auch dann in die Analyse einbezogen werden, wenn sie nicht zu Marktpreisen bewertet werden können. Man ordnet ihnen »Schatten«-Preise zu.

Kovarianzanalyse s. Varianzanalyse

Kriteriumstest: Unter Kriteriumstest wird ein Test verstanden, in dem das Verhalten von Individuen in bezug auf ein Kriterium gemessen wird. Bei der Darstellung der Leistung in einem Kriteriumstest werden die Leistungen der übrigen Mitglieder der Lerngruppe nicht berücksichtigt. Bei der Bewertung orientiert man sich statt dessen nur an einem gesetzten Lernziel.

Management-System-Modell: Als Management-System-Modell wird ein Evaluationsmodell bezeichnet, in dem Evaluation vor allem die Aufgabe hat, den Entscheidungsträgern im Bildungssystem Informationen für Entscheidungen zur Verfügung zu stellen.

Markoff-Kette: Eine Markoff-Kette ist eine endliche oder unendliche Folge von Zuständen eines Systems. Markoff-Ketten können eine größere oder geringere Abhängigkeit haben. Z. B. ist eine Folge von Münzwürfen eine Markoff-Kette ohne jegliche innere Abhängigkeit. Die Abfolge der Buchstaben in einem geschriebenen Text kann als Markoff-Kette mit einer gewissen inneren Abhängigkeit aufgefaßt werden. Das Auftreten eines bestimmten Buchstabens ist nämlich in gewissem Grad, wenn auch nicht vollständig, von den vorausgegangenen Buchstaben der Kette abhängig.

Matrix: Unter einer Matrix wird die Anordnung von Elementen in einer rechteckigen Tabelle in Spalten und in Zeilen verstanden. Werden z. B. m Versuchspersonen in n Tests untersucht, so lassen sich die Ergebnisse in einer Daten-Matrix mit m Spalten und n Zeilen anordnen. D. h. in jeder Spalte stehen die Ergebnisse einer Versuchsperson in allen Tests und in jeder Zeile die Ergebnisse aller Versuchspersonen in einem Test. Mit Hilfe der in der Matrizenrechnung verwendeten Regeln lassen sich Rechenoperationen mit Matrizen in kompakter Form durchführen und übersichtlich darstellen.

Mittelwert s. Verteilung

Nicht-vergleichende Evaluation s. Evaluation

Parallelisieren: Unter Parallelisieren wird die Bildung von zwei oder mehreren Personengruppen verstanden, die im Hinblick auf ein oder mehrere Merkmale so ähnlich wie möglich sind.

Placebo: Bei pharmakologischen Untersuchungen bezeichnet man ein Leerpräparat als Placebo, das von einem Pharmakon, das untersucht werden soll, äußerlich nicht zu unterscheiden ist, jedoch keinerlei wirksame Substanzen enthält. Ein Vergleich der Wirkungsweise des Vollpräparates mit der des Placebos zeigt, ob ein über Suggestivwirkung hinausgehender Effekt eintritt.

Planning Programming Budgeting System (PPBS): Das Planning Programming Budgeting System dient der rationalen Verwendung der finanziellen Mittel der öffentlichen Haushalte. Im Gegensatz zu der oft praktizierten Methode der Fortschreibung der um gewisse Prozentsätze erhöhten Vorjahrestitel beginnt das PPBS mit der Festlegung der zu erreichenden Ziele (Planning); anschließend werden die zur Verwirklichung der Ziele notwendigen Programme ausgewählt (Programming), wofür dann die erforderlichen Haushaltsansätze gebildet werden (Budgeting).

Produktevaluation s. Evaluation

Prozeßevaluation s. Evaluation

Redundanz: Als Redundanz bezeichnet man den Unterschied zwischen der in einer bestimmten Menge von Zeichen maximal unterzubringenden Informationsmenge und der tatsächlich vorhandenen Informationsmenge. Beispielsweise ließe sich in hundert der in der deutschen Sprache verwandten Zeichen (Buchstaben) wesentlich mehr Information unterbringen, als es in der deutschen Sprache geschieht; die Redundanz ist also relativ groß.

Im allgemeinen Sprachgebrauch steht Redundanz für Überflüssiges oder Weitschweifiges in Aussagen und Mitteilungen.

Q-Technik s. Faktorenanalyse

Rationale: Unter »rationale« wird die Begründung eines Bildungsprogramms einschließlich seiner gesellschaftspolitischen Intentionen, seiner sonstigen Ziele, Methoden, Organisationsformen usw. bezeichnet.

Regressionskoeffizient s. Regressionsrechnung

Regressionslinie s. Regressionsrechnung

Regressionsrechnung: Wenn zwei Merkmale miteinander korreliert sind, läßt sich bei Kenntnis der Größe des einen mit Hilfe der Regressionsrechnung die Größe des anderen Merkmales abschätzen. Die geschätzten Werte befinden sich auf der

Regressionslinie, bei linearem Zusammenhang beider Merkmale auf einer Regressionsgeraden. In die Steigung der Regressionsgraden geht der Regressionskoeffizient ein. Die Differenzen zwischen den geschätzten und den tatsächlich erhaltenen Werten, Fehlerwerte oder *Residuen* genannt, sind ein Maß für die Exaktheit der Schätzung. Bei einer vollständigen Korrelation der Merkmale würden ihre Werte sämtlich o sein, bei einer nicht vollständigen Korrelation geben sie das Ausmaß an, in dem das geschätzte Merkmal unabhängig von dem Merkmal variiert, aufgrund dessen geschätzt wird.

Regional Laboratory: Als Regional Laboratory werden neueingerichtete regionale Zentren verstanden, von denen aus Curriculumentwicklung und -implementation mit der entsprechenden Lehrerfortbildung betrieben werden.

Reliabilität (Zuverlässigkeit): Als Reliabilität wird der Grad der Genauigkeit eines Maßes oder eines Tests bezeichnet. Zuverlässigkeit eines Tests liegt in seiner Eigenschaft und Qualität als »Meßinstrument«; sie bezeichnet den Grad der Genauigkeit, mit der ein Test das mißt, was er de facto mißt.

Zur Ermittlung des Maßes der Zuverlässigkeit, des *Reliabilitätskoeffizienten,* bedient man sich der Methode der Testwiederholung in Form desselben oder eines Paralleltests – diese Methode ergibt einen sogenannten *Stabilitätskoeffizienten* – oder der Methode der *Konsistenzanalyse,* die bei einmaliger Testvorgabe die Reliabilität aufgrund der inneren Beschaffenheit des Tests schätzt.

Research and Development Center: Die Research and Development Centers sind große Forschungs- und Entwicklungszentren, in denen die Ergebnisse der Forschung direkt in den Prozeß der Entwicklung von Bildungsprogrammen umgesetzt werden; sodann folgt abermals die Erforschung der durch die Innovationen bewirkten Veränderungen und die Modifizierung der Bildungsprogramme.

Residuen s. Regressionsrechnung

Ressourcen: Unter Ressourcen versteht man Hilfsquellen und Hilfsmittel. In der Ökonomie bezeichnet man als Ressourcen alle Mittel, die für die Produktion von Gütern und Dienstleistungen benötigt werden (Arbeitskräfte, Maschinen, Gebäude, Boden etc.). Analog wird der Begriff auch im Bildungswesen gebraucht.

Rückmeldung (Rückkoppelung; Feedback): Der Begriff wurde aus der Kybernetik übernommen und dient für jede Art der Rückmeldung über den Grad der Angemessenheit und der Wirksamkeit einer Handlung.

R-Technik s. Faktorenanalyse

Schätzskala: Eine Schätzskala ist ein Hilfsmittel bei der Abgabe von Schätzurteilen über ein Merkmal. Auf einer Skala, deren Stufen meist durch Zahlen und oft durch verbale Beschreibungen (z. B. sehr angenehm, angenehm, weniger angenehm usw.) gekennzeichnet sind, werden Schätzungen über die Ausprägung eines Merkmals angegeben (etwa durch Ankreuzen).

Schiefe s. Verteilung

Stabilitätskoeffizient s. Reliabilität

Standardabweichung s. Verteilung

Störvariable: Unter Störvariablen versteht man Variablen, die zwar einen Einfluß auf das Versuchsergebnis haben, für den Untersucher jedoch von geringem In-

teresse sind und deshalb leicht übersehen werden. Durch ihren Einfluß können die Ergebnisse von Untersuchungen verändert und verfälscht werden. Bei der Untersuchung der Langzeitwirkung eines Medikaments kann z. B. der Einfluß der Störvariable »jahreszeitliche Veränderung« das Ergebnis verfälschen, wenn er nicht entsprechend berücksichtigt wird.

Streuung s. Verteilung

Summative Evaluation s. Evaluation

Taxonomie: Unter Taxonomie wird im Bildungsbereich im allgemeinen die Klassifikation von Lernzielen verstanden.

Trade-off: Schließt sich die Erreichung von zwei Zielen zur gleichen Zeit mit einem bestimmten Niveau aus, so besteht der Trade-off darin, daß ein »Mehr bei einem Ziel« notwendigerweise ein »Weniger bei dem anderen Ziel« mit sich bringt.

Validität (Gültigkeit): Der Grad der Genauigkeit, mit dem ein Test »dasjenige Persönlichkeitsmerkmal oder diejenige Verhaltensweise, das (die) es messen soll oder zu messen vorgibt, tatsächlich mißt« (Lienert 1967). Z. B. kann ein Test für technisches Verständnis eine geringe Validität haben, wenn seine Aufgaben statt mit technischem Verständnis ebenso gut mit logisch-abstraktem Denken zu lösen sind. Man unterscheidet je nach den Kriterien, an denen man validiert, verschiedene Arten der Validität.

Bei der *Augenscheinvalidität* legt man ein subjektives Kriterium zugrunde, urteilt nach dem Eindruck oder beruft sich auf leicht sichtbare Eigenschaften des Meßverfahrens. Bei der *Konstruktvalidität* geht man von theoretischen Annahmen über die Beziehung des betreffenden Meßwertes zu anderen meßbaren Variablen aus. Treffen die Vorhersagen ein, so hält man die Konstruktvalidität für gegeben. Bei der *inneren Validität* stellt man eine Beziehung zu anderen Meßwerten derselben Art her, bei der *äußeren Validität* prüft man den Zusammenhang mit Kriterien anderer Art.

Varianzanalyse: Die Varianzanalyse ist ein statistisches Verfahren, das zu prüfen ermöglicht, ob sich Gruppen von Individuen bezüglich eines Kriteriums (eines Merkmals) voneinander signifikant unterscheiden. Z. B. könnte man untersuchen, ob sich Hauptschüler, Real- und Gymnasialschüler bezüglich ihrer Intelligenz durchschnittlich voneinander unterscheiden.

Bei der *Kovarianzanalyse* wird zusätzlich ein mit dem Kriterium korreliertes (ko-variierendes) Merkmal in die Analyse einbezogen und mit seinem Einfluß berücksichtigt.

Vergleichende Evaluation s. Evaluation

Verteilung: Eine empirische Verteilung ist die Aufstellung der Häufigkeiten von Meßwerten in bezug auf eine Skala. Eine empirische Verteilung kann mit Hilfe verschiedener Merkmale beschrieben werden:

Die zentrale Tendenz der Verteilung wird durch ihren *Mittelwert* dargestellt (meist arithmetisches Mittel).

Maße für die *Streuung* einer Verteilung, d. h. die Abweichung der Einzelwerte vom Mittelwert, sind die mittlere quadratische Abweichung oder *Standardabweichung*, die *Varianz* (gleich Quadrat der Standardabweichung) und der

Interquartilbereich.

Die Form der Verteilung kann durch ihre *Schiefe* – eine Verteilung kann rechts schief, links schief oder symmetrisch sein – sowie durch ihren *Exzeß*, d. h. ihre Steilheit gekennzeichnet sein.

Zuverlässigkeit s. Reliabilität

Zielkomplex-Modell: Als Zielkomplex-Modell wird ein Evaluationsmodell bezeichnet, in dem unter Evaluation in erster Linie die Bewertung von Innovationen verstanden wird. Dabei geht es vor allem darum, alle Ziele des gesamten Zielkomplexes zu bewerten und zu einem »zusammengesetzten Gesamturteil« über die Ziele zu gelangen.

Personenregister

Sachregister

Erziehung in Wissenschaft und Praxis

Beiträge zur Pädagogik der Gegenwart, herausgegeben von Andreas Flitner

Die Reihe setzt sich die Aufgabe, durch Monographien, durch Studientexte und Diskussionsbände einen weiteren Kreis von Studierenden sowie alle an pädagogischen Fragen interessierten Leser mit den Problemen und Ergebnissen der modernen deutschen und internationalen Erziehungswissenschaft bekanntzumachen.

Zu diesem Band: Eine demokratische Bildungsreform muß sich der Frage nach der Legitimität intendierter Innovationen stellen. Um Reformen in bildungspolitischer, pädagogischer und ökonomischer Hinsicht zu legitimieren, bedarf es wissenschaftlich gewonnener Erkenntnisse darüber, welche Auswirkungen sie auf die Schüler und darüber hinaus auf die Gesellschaft haben. Solche Kenntnisse dienen außer zur wissenschaftlichen Fundierung bildungspolitischer und pädagogischer Entscheidungen auch zur Revision und Verbesserung der Innovationen selbst. Daher ist es notwendig, neue Curricula, neue Formen der Unterrichtsorganisation, der Lehrerbildung und Schulversuche kritisch auf ihre Wirkung hin zu untersuchen, d. h. zu evaluieren. Eine solche Evaluation richtet sich auf die Sammlung, Verarbeitung und Interpretation von Daten mit dem Ziel, bestimmte Fragen über Innovationen zu beantworten und Entscheidungen über sie zu treffen. Das schließt die Beschreibung und die Bewertung der Qualität und Angemessenheit von Zielen, Inhalten, Methoden usw. ein. Mit Hilfe einer systematischen Evaluation, deren Theorie, Methodologie und Technologie in den Beiträgen dieses Bandes dargeboten wird, lassen sich viele Fragen beantworten, die sich Lehrern, Eltern, Beamten der Schulverwaltung, Bildungspolitikern und Erziehungswissenschaftlern im Rahmen der modernen Bildungsreform aufdrängen.

Der Band enthält Beiträge von: Marvin C. Alkin, Richard C. Anderson, Samuel Ball, Arno A. Bellack, Gerry Ann Bogatz, William W. Cooley, Lee J. Cronbach, Gene V. Glass, Barry MacDonald, Klaus Nagel, Graham A. Nuthall, Evelore Parey, Ulf Preuss-Lausitz, Michael Scriven, Robert E. Stake, Daniel L. Stufflebeam, Christoph Wulf.

Der Herausgeber: Christoph Wulf ist wissenschaftlicher Mitarbeiter am Deutschen Institut für Internationale Pädagogische Forschung (DIPF) in Frankfurt a. M.

ISBN 3-492-01985-4