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-

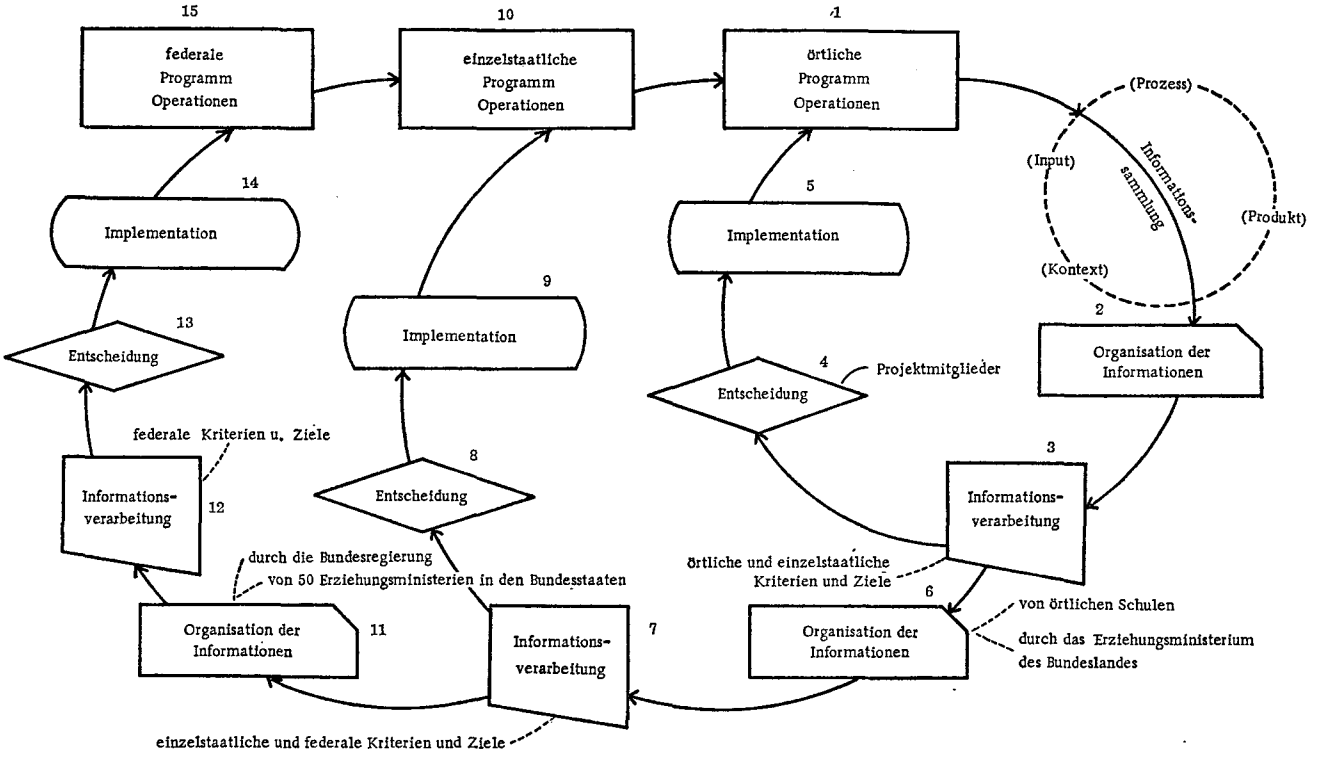


Abb. 1: Evaluation in vom Bund unterstützten Bildungsprogrammen (Stufflebeam 1967)

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.

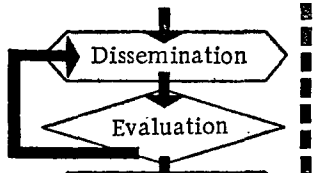
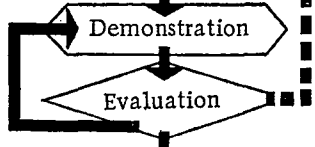
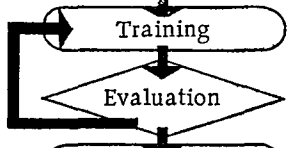
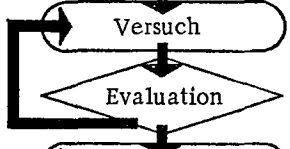
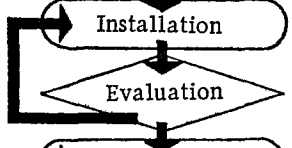
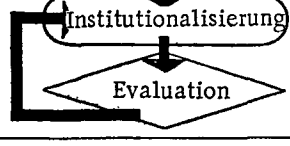
Definition des Plans

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.	Forschung → Evaluation	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.	Invention → Evaluation	Augenscheinvalidität (face validity); geschätzte Funktionsfähigkeit; Einfluß (relativer Beitrag).	Erzeugt die Innovation.
		Entwurf eines Plans zur Konstruktion der Innovation.	Plan → Evaluation	*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.	Konstruktion → Evaluation	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.	Versammlung → Evaluation	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.

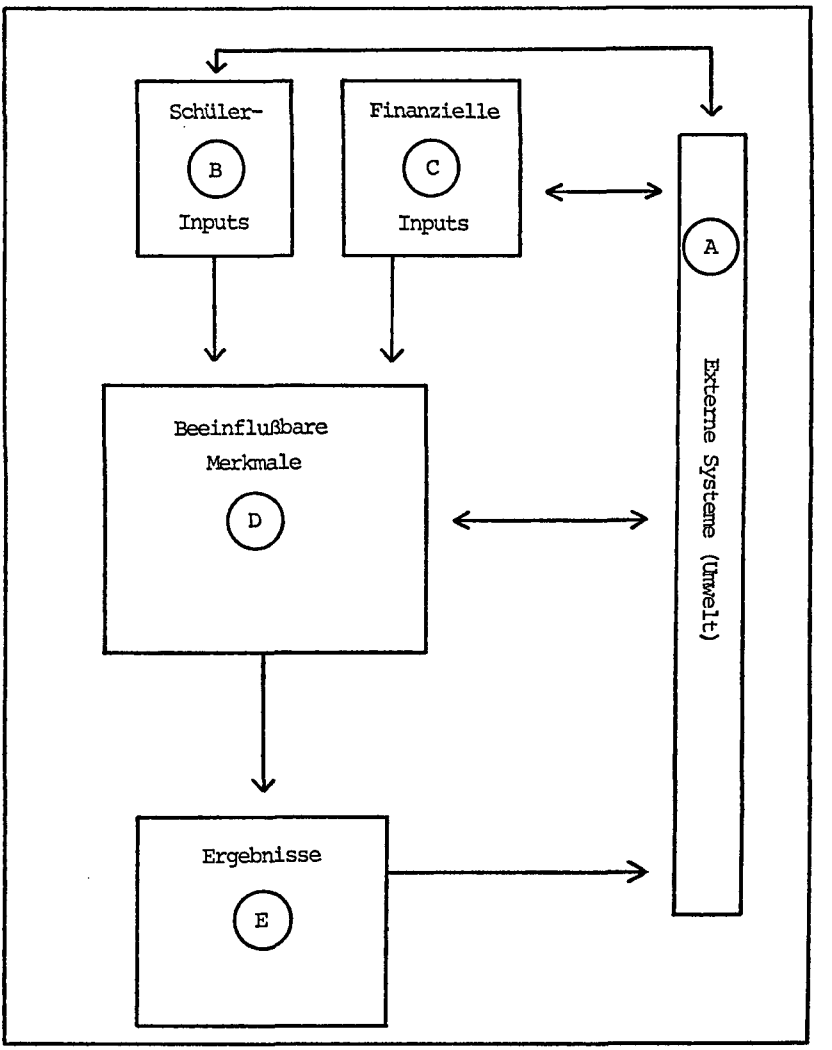
Kurz gesagt, ein Modell stellt einfach einen Versuch dar, die Hauptelemente einer Einheit oder eines Phänomens im Hinblick auf jeweilige Funktionen und gegenseitige Beziehungen zu klassifizieren, damit leichter beobachtet werden kann, wie die Elemente innerhalb der Einheit funktionieren, wie sie die 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

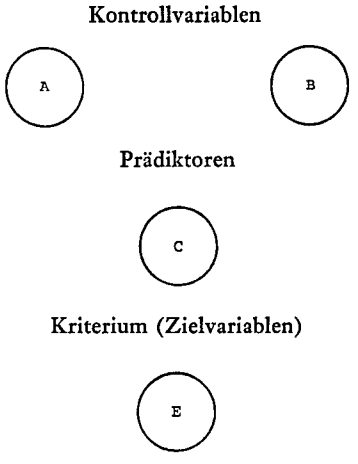
man davon ausgeht, daß unterschiedliche Schulen insgesamt oder zum Teil auf dieselben Ziele hinarbeiten, dann könnten die gesamten Programme dieser Schulen ebenfalls als Programmalternativen betrachtet werden. Unterschiedliche Schulen bieten unterschiedliche Programmalternativen. Folglich kann man den 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

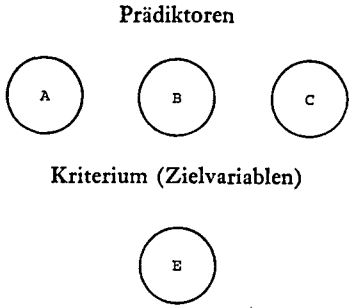


tutionen im Hinblick darauf evaluiert, was sie im Verhältnis zu den ihnen zur Verfügung stehenden menschlichen und finanziellen Ressourcen leisten, könnte das vorgestellte Modell für eine Aufwands-Effektivitäts-Analyse herangezogen werden. Mit anderen Worten, wenn man die finanziellen Inputs als vorgegebene 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. Erreichter Wert für $E_{l,i}$

$$\frac{\text{Erreichter Wert für } E_{l,i}}{\text{Erwartungswert für } E_{l,i}}$$

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.

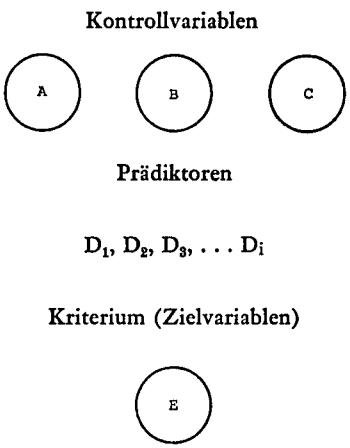
Es gibt natürlich von Ort zu Ort hinsichtlich der Qualität der verfügbaren bzw. zur Auswahl stehenden Ressourcen beträchtliche Unterschiede. Wenn der Ökonom z. B. Lehrer, Material usw. als Inputs des Systems betrachtet, rechnet er mit qualitativen Unterschieden der Inputs. In diesem Modell, das für Entscheidungen 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.

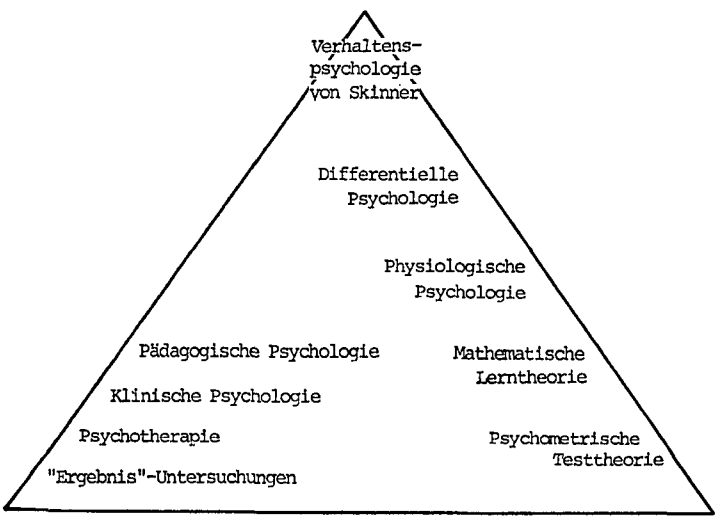
Forschung zielt auf die Abschätzung von drei unterschiedlichen Aspekten eines Gegenstands:

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-

Einschätzung der empirischen Überprüfbarkeit mit anerkannten Methoden (empirische Wahrheit)



Einschätzung des gesellschaftlichen Nutzens (reine Evaluation)

Einschätzung der logischen Konsistenz (rationale Wahrheit)

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.