Glass, Gene V.

Die Entwicklung einer Methodologie der Evaluation

Wulf, Christoph [Hrsg.]: Evaluation. Beschreibung und Bewertung von Unterricht, Curricula und Schulversuchen. München : R. Piper & Co. Verlag 1972, S. 166-206. - (Erziehung in Wissenschaft und Praxis; 18)

Evaluation

Beschreibung und Bewertung von Unterricht,
Curricula und Schulversuchen

Texte
herausgegeben von Christoph Wulf

R. Piper & Co. Verlag
München

ISBN 3-492-01985-4

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

Gene V. Glass

Die Entwicklung einer Methodologie der Evaluation

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

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

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

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

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

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

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

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

Pädagogische Forschung und Evaluation

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

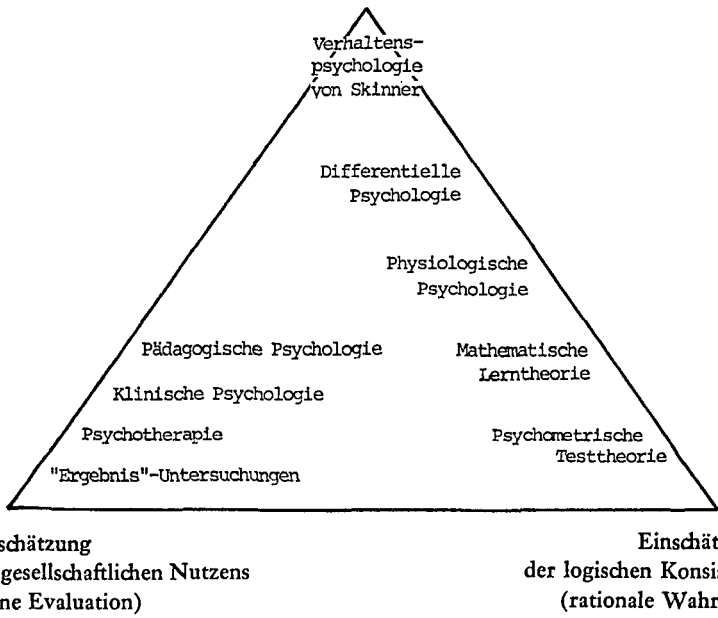

Abb. 1: Klassifikation psychologischer Forschungsansätze in bezug auf ihre Ziele.

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

Das Tylersche Evaluationsmodell

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

»Bestenfalls kann ein solcher Versuch zwei bereits bestehende Curricula

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Das Akkreditations-Modell

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Die Autoren der Normen für Medienprogramme wollten die Schulen

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

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

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

Das Management-System-Evaluationsmodell

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Das Zielkomplex-Modell

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

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

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

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

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

Die Weiterentwicklung des Zielkomplex-Modells der Evaluation

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

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

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

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

A. Sammlung von Daten

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

Allgemeinheitsgrad und Spezifitätsgrad von Daten

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

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

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

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

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

Prioritäten für Evaluationsdaten

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

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

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

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

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

(1) Finanzieller Aufwand der Sammlung verschiedener Daten;

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

(3) der finanzielle Aufwand der Implementation jeder Entscheidungsalternative.

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

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

Nehmen wir an, daß alle Faktoren mit Ausnahme der unterschiedlichen

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

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

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

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

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

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

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

Trotz des offenbar vielversprechenden Ansatzes solcher rudimentären

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

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

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

B. Die Gewichtung der Daten

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

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

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

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

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

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

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

(1) Rechtfertigung der Instrumente zur Datensammlung

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

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

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

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

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

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

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

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

3. Rechtfertigung der Auswahl von Zielen

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

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

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

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

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

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

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

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

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

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

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

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

Pädagogische Veröffentlichungen enthalten wertende Äußerungen über die einen oder anderen Curricula oder Unterrichtsmethoden. Die Bestimmung des relativen Wertes von »heuristischem Lehren« und »darstellendem Lehrervortrag« (discovery and dispository teaching) muß auf einer Analyse der Definitionen der beiden Begriffe und empirischen Längsschnittuntersuchungen der Wirkungen jeder Methode auf das Behalten von Wissen und auf die Entwicklung von Interesse, Motivation, Berufsplänen, Persönlichkeit usw. beruhen. Die gegenwärtigen Erörterungen über die Überlegenheit des heuristischen Lehrens über das darstellende Lehren verlangen nach ernsthaften Versuchen, die Begriffe logisch zu analysieren und aussagekräftige empirische Daten zu sammeln.

Zur Rechtfertigung von Bildungszielen bedarf es ohne Zweifel logischer und empirischer Analysen. Philosophen können wesentlich dazu beitragen, das Problem der Rechtfertigung der Auswahl von Zielen zu lösen, indem sie die logische Konsistenz zwischen curricularen Zielen und der Philosophie des Curriculum bzw. der Begründung des Curriculum und der Übereinstimmung mit den philosophischen Grundgedanken der Erziehung untersuchen. Man kann Wissenschaftler fragen, ob die für ihre Disziplin relevanten Ziele sich rechtfertigen lassen. So ist z. B. ein Biologe besonders kompetent, um zu beurteilen, ob Lysenkoismus wegen seines Wertes als Gegenstand wissenschaftlicher Forschung in einem Biologiekurs der Sekundarschule gelehrt werden sollte. Sozialwissenschaftler können von allen Wissenschaftlern wahrscheinlich am meisten zur Lösung der Probleme der Auswahl von Zielen beitragen.

Die Psychologie wird für die Rechtfertigung eines curricularen Ziels oft sehr relevant sein. Man betrachte als Beispiel das von der American Association for the Advancement of Science (AAAS) entwickelte naturwissenschaftliche Curriculum für die Primarstufe. Die Autoren dieses Curriculum betrachten Naturwissenschaft als Sammlung einer kleinen Zahl transferierbarer Prozesse und wissenschaftlicher Methoden.

Die AAAS-Materialien setzen sich zum Ziel, diese heuristischen Fertigkeiten dem Schüler zu vermitteln; der Kontext ihrer Anwendung, d. h. die

Inhalte des naturwissenschaftlichen Curriculum, wird für wesentlich weniger wichtig gehalten. Einige Kritiker haben das AAAS-Curriculum angegriffen; sie vertreten die Auffassung, daß es auf der Vermögenspsychologie des 19. Jahrhunderts beruht. Sie behaupten, psychologische Forschung habe gezeigt, daß das Gedächtnis nicht als eine Sammlung von Anlagen oder Fähigkeiten angesehen werden kann, die durch Gebrauch verbessert und sodann in einer Vielzahl von Situationen angewendet werden können. Die Frage, ob die AAAS-Materialien auf einer solchen Vorstellung vom Lernenden basieren und ob ein solches Konzept als eine Theorie des Verhaltens nutzlos ist, können nur Psychologen qualifiziert beantworten. Die Antworten könnten sicherlich Einfluß auf die Rechtfertigung des prozeßorientierten Charakters der AAAS-Curriculummaterialien haben.

Pädagogische Forschung, die die Auswahl von Bildungszielen rechtfertigen kann, ist dringend erforderlich. Es fehlen uns die elementarsten Daten – etwa aus Längsschnittuntersuchungen – darüber, wie Wissen behalten wird und Interessen entstehen. Wie sollen wir wissen, ob ein Curriculumentwickler gut beraten ist, wenn er sich mit der Förderung des Interesses an Mathematik beschäftigt, anstatt vielmehr mathematische Inhalte zu lehren? Wenn Längsschnittbefragungen zeigen, daß mathematische Inhalte innerhalb von 5 Jahren nach Beendigung des formalen Unterrichts vergessen werden, daß aber das Interesse an Mathematik fortbesteht und zu weiterer Beschäftigung und positiver Einstellung gegenüber den Wissenschaften führt, dann ist die Auswahl der Ziele der Curriculumentwickler wahrscheinlich gerechtfertigt. Offensichtlich wird pädagogische Evaluation auch von anderen Wissensgebieten abhängig sein, um mit ihrer Hilfe Fragen nach der Rechtfertigung der Auswahl von Zielen zu beantworten.

Schlußfolgerung

Wie jedes komplexe Werk des Menschen hat die Methodologie der Evaluation kein wirkliches Entwicklungspotential; das einzige Entwicklungspotential ist ein Plan für ihre zukünftige Entwicklung im Geist ihrer Schöpfer.

Abgedruckt auch in: M. C. Wittrock/D. E. Wiley, The evaluation of instruction, issues and problems, New York: Holt, Rinehart and Winston 1970, 221–238.

1 Sehr zum Leidwesen vieler unwilliger Beamter der Schulverwaltung gestehen wir jedoch daß dies ein tunlicher Ausgangspunkt wäre.

2 Manches spricht dafür, daß dies ein vernünftiges Verfahren ist. Vgl. J. S. Bekker 1962; Miller: Income and Higher Education, in: S. J. Muskin 1962; T. Schultz 1961.

Gene V. Glass: Die Entwicklung einer Methologie der Evaluation

Übersetzung von Hannes Graudenz (Dipl.-Psych.) und dem Herausgeber.

Originaltitel: The growth of evaluation methodology

Der deutschen Übersetzung liegt das dem Herausgeber vom Autor zugesandte Manuskript zugrunde.

1 Ein »allgemeines Phänomen« ist nachgewiesen oder kann entdeckt werden in einem weiten Feld von scheinbar verschiedenen Erscheinungen und wird als Kriterium zur Prüfung eines wissenschaftlichen Begriffs herangezogen. Ohne eine solche Qualifikation würde es bereits »Einschätzung wissenschaftlicher Wahrheit« bedeuten, empirisch festzustellen, daß man seine Schlüssel verloren hat. Der Begriff der Generalisierbarkeit von erwarteten Ergebnissen ist wichtig für die Unterscheidung von Evaluation und Forschung; er ist auch von großer praktischer Bedeutung beim Entwurf einer Evaluations-Untersuchung (vgl. Stake 1969).

2 In diesem Abschnitt beziehe ich mich weitgehend auf die Geschichte der North Central Association von Calvin O. Davis (1945).

3 Die Pionierarbeit von Joseph M. Rice mag manchem zu dieser Zeit bekannt gewesen sein, wurde aber wahrscheinlich eher als tendenzieller Journalismus denn als pädagogische Forschung angesehen.