II Grundfragen der Evaluation

Einführung

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

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

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

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

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

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

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

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

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

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

Lee J. Cronbach

Evaluation zur Verbesserung von Curricula

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

Die Funktion der Evaluation in Entscheidungsprozessen

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

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

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

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

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

Für die Verbesserung von Curricula war die Einführung der systema-

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

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

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

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

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

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

Viele Evaluationsexperten sahen in der Entwicklung von Tests eine

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Methoden der Evaluation

Spektrum der Methoden

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

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

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

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

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

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

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

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

Für die Curriculumentwickler sind unter den Ergebnissen des Curricu-

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

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

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

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

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

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

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

Leistungsmessung

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Zusammenfassung

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

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

Michael Scriven

Die Methodologie der Evaluation

Einführung

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

Überblick

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Professionelle und Amateur-Evaluation

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

Die unzulängliche Kommunikation zwischen Evaluatoren, Lehrern und

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

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

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

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

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

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

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

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

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

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

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

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

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

Evaluationsuntersuchungen und Prozeßuntersuchungen

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

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

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

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

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

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

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

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

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

Evaluation und Überprüfung der Zielerreichung

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

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

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

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

Intrinsische Evaluation und Ergebnisevaluation

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

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

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

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

Praktische Vorschläge für eine Mischform bei Evaluationsuntersuchungen (Hybrid Evaluation)

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Das Für und Wider einer reinen Ergebnisevaluation

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

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

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

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

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

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

Vergleichende und nicht-vergleichende Evaluation

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

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

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

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

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

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

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

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

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

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

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

Praktische Verfahrensweisen für die Evaluation mit Kontrollgruppen

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Robert E. Stake

Verschiedene Aspekte pädagogischer Evaluation

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

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

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

Der mögliche Beitrag formaler Evaluation

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

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

Voraussetzungen, den Unterrichtsaktivitäten und den schulischen Ergebnissen beachten.

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

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

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

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

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

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

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

Auf dem Weg zu einer vollständigen Beschreibung

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

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

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

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

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

Die Rolle des Urteils

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

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

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

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

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

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

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

Datenmatrizen

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

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

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

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

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

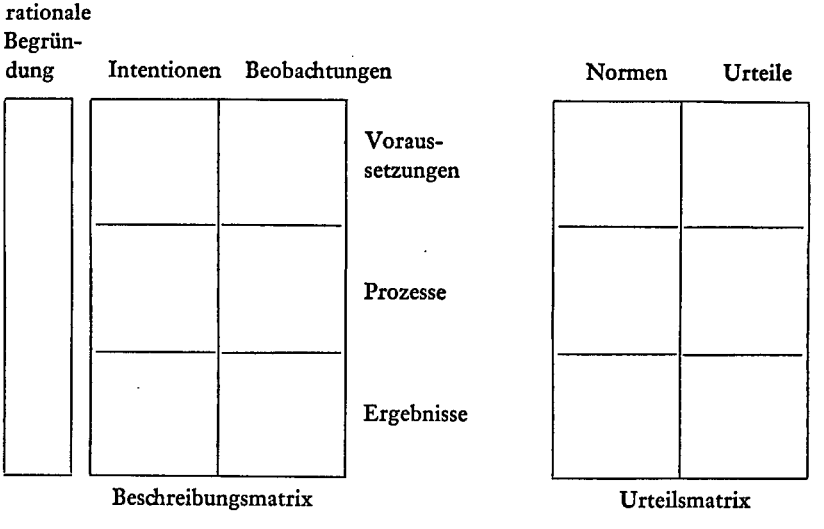

Abb. 1: Eine Matrix für Daten, die vom Evaluator eines Bildungsprogramms gesammelt werden sollen

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

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

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

Im Wissen, daß

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

Nicht einmal für die ferne Zukunft erwarten Pädagogen und Evaluatoren,

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

Ziele und Intentionen

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

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

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

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

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

Die Auswahl von Methoden der Beschreibung

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

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

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

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

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

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

Kontingenz und Kongruenz

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

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

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

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

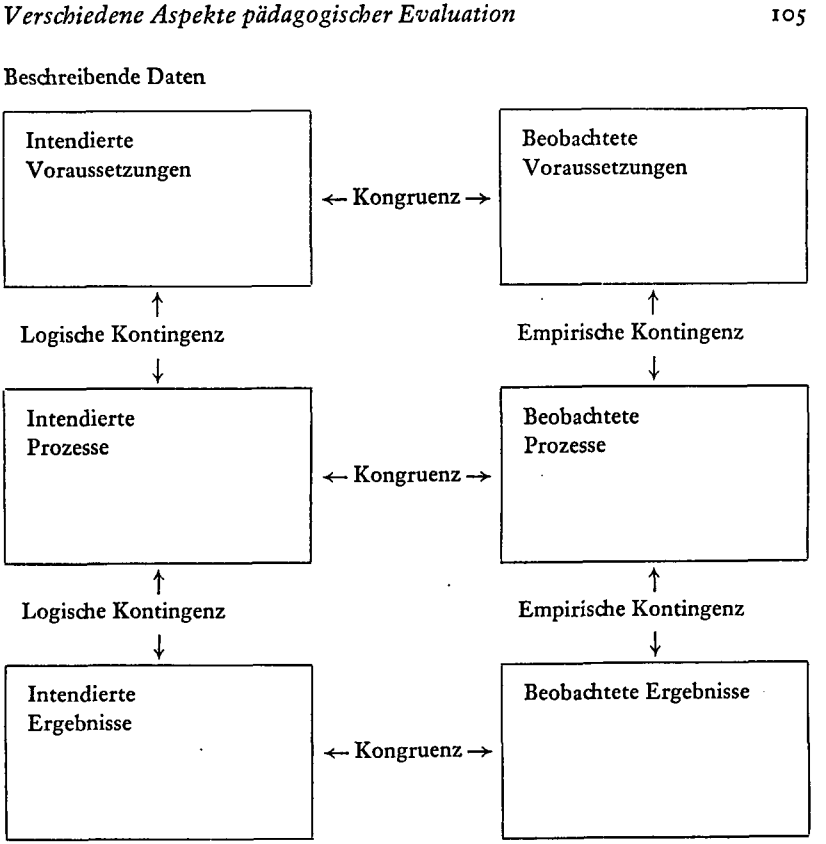

Abb. 2: Eine Darstellung des Prozesses der Verarbeitung von beschreibenden Daten

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

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

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

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

Normen und Urteile

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

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

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

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

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

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

Vergleichen und Urteilen

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

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

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

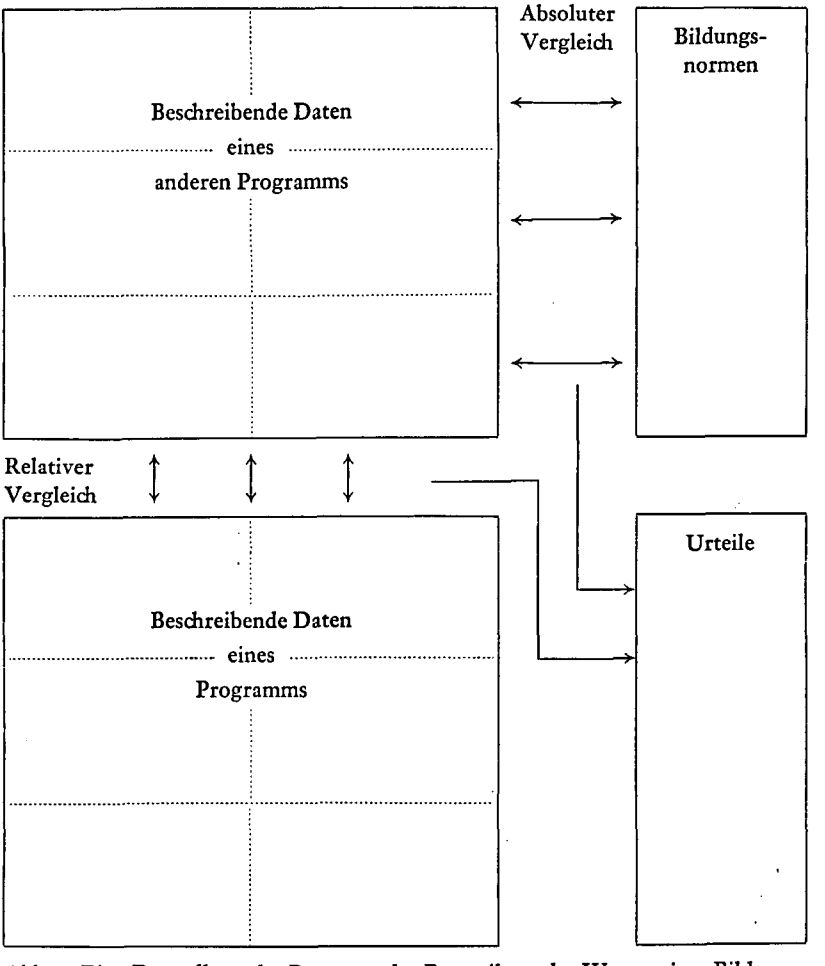

Abb. 3: Eine Darstellung des Prozesses der Beurteilung des Wertes eines Bildungsprogramms

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

Absolute und relative Evaluation

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

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

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

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

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

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

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

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

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

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

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

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

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