Cronbach, Lee J.

Evaluation zur Verbesserung von Curricula

Wulf, Christoph [Hrsg.]: Evaluation. Beschreibung und Bewertung von Unterricht, Curricula und Schulversuchen. München : R. Piper & Co. Verlag 1972, S. 41-59. - (Erziehung in Wissenschaft und Praxis; 18)

Evaluation

Beschreibung und Bewertung von Unterricht,
Curricula und Schulversuchen

Texte
herausgegeben von Christoph Wulf

R. Piper & Co. Verlag
München

ISBN 3-492-01985-4

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

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.

Lee J. Cronbach: Evaluation zur Verbesserung von Curricula

Übersetzung von Ines Graudenz (Dipl.-Psych.).

Originaltitel: Evaluation for course improvement; zugrunde gelegte Fassung aus: R. W. Heath, New curricula, Harper & Row 1964, benutzt und zitiert nach Abdruck in: N. E. Gronlund (Ed.): Readings in measurement and evaluation, London: The Macmillan Company ²1970. Erste Fassung in: Teachers College Record 64, 1963, 672–683.

1 Meine Ausführungen zu diesen Fragen konnten durch die Reaktionen, die ich auf die erste Fassung dieses Beitrags von einigen Leitern von Curriculumprojekten und von Kollegen erhielt, präzisiert werden.