IV Ausgewählte Beispiele zur Evaluation

Einführung

Häufig ist in der Evaluationsliteratur der letzten Jahre auf das Fehlen guter und leicht zugänglicher Evaluationsuntersuchungen hingewiesen worden (z. B. Westbury 1970; Cooley 1971). Im Unterschied zu den in den letzten Jahren zahlreich und sorgfältig entwickelten Evaluationsmodellen hat es wenige Berichte von Evaluationsuntersuchungen gegeben, die Eingang in Fachzeitschriften oder Veröffentlichungen über Evaluation gefunden hätten (vgl. Grobman 1968). So enthält eine von Baker (1969) zusammengestellte, 80 Titel umfassende Liste von Veröffentlichungen zur Curriculumevaluation nur sechs empirische Untersuchungen. Manchen Autoren ist das zum Anlaß geworden, in Anlehnung an Veröffentlichungen Schwabs (1970, 1971a, 1971b), die den Wert von Theorien für die Curriculumentwicklung radikal in Frage stellen, auch den Wert von Modellen und Theorien der Evaluation für die Praxis der Evaluation anzuzweifeln (z. B. Lewy 1972). Aus diesen Ausführungen geht deutlich hervor, daß es zwischen Modellen der Evaluation und der wirklichen Durchführung von Evaluationsuntersuchungen zahlreiche konzeptuelle, methodische und technologische Schwierigkeiten zu überwinden gibt, wozu die Modelle selbst wegen ihres hohen Generalisierungsgrades nur begrenzt beitragen können. Deshalb sollten in dieser Veröffentlichung dem Leser neben den anspruchsvollen Modellen aus Teil II auch konkrete Evaluationsuntersuchungen vor Augen geführt werden. Sie thematisieren als Ergänzung zu den Beiträgen der vorherigen Abschnitte neue Fragen und Probleme, die sich bei der Anlage einer Evaluationsuntersuchung und ihrer Durchführung unausweichlich ergeben. Auf Grund der Unterschiedlichkeit der Evaluationsberichte wird dabei das weite Spektrum der Möglichkeiten sichtbar, die es für eine sinnvolle Durchführung von Evaluationsuntersuchungen gibt. Sie werfen die Frage nach dem wissenschaftstheoretischen und bildungspolitischen Standort des Evaluators auf, ohne sie anders als für ihren eigenen Zusammenhang lösen zu können. Die hier ausgewählten Untersuchungen nehmen nicht in Anspruch, vorbildlich zu sein. Sie bieten durchaus Anlaß zur Kritik, die sie

z. T. in der offenen Darstellung der eigenen Fehler und Unzulänglichkeiten herausfordern. Jedoch gehören sie zu den wenigen interessanten, im Rahmen dieses Bandes reproduzierbaren Evaluationsuntersuchungen, die sich im angelsächsischen Bereich finden ließen.

Eine der bekanntesten Untersuchungen haben Ball und Bogatz mit ihrer Evaluation des ersten Jahres von Sesame Street vorgelegt, die als eine summative Evaluation die Leistungsfähigkeit dieser Sendungen und des Fernsehens als Unterrichtsmedium nachweisen konnte und die wertvolle Anregungen zur Modifikation der Sendereihe für das zweite Jahr brachte. Die Untersuchung ergab, (1) daß die Kinder, die am häufigsten fernsehen, auch am meisten lernen, (2) daß das, was in den Sendungen am stärksten behandelt wird, auch am besten gelernt wird und (3) daß das Programm keine besondere Beaufsichtigung der Kinder durch Erwachsene erforderlich macht. Diese Evaluationsstudie ist zugleich ein Beispiel für die Evaluation eines Bildungsprogramms, das nicht zum schulischen Bereich im engeren Sinne gehört.

Andersons Ausführungen zielen auf die Evaluation eines curricularen Programms. Dazu beginnen sie mit der Bestimmung des Standorts der Untersuchung im Kontext der Diskussion über die verschiedenen Ansätze zur Evaluation. Es folgt die exakte Einarbeitung dieser Ansätze in die Konzeptualisierung und Planung der Untersuchung. Sodann wird die Durchführung der Evaluation detailliert beschrieben und werden ihre Ergebnisse dargestellt. Die Untersuchung dient als Beispiel für die Evaluation eines curricularen Programms mit Hilfe einer Kontrollgruppe und soll die Brauchbarkeit dieser Form der Evaluation belegen.

In Cooleys Beitrag werden verschiedene in diesem Zusammenhang neue methodische und statistische Verfahren der Evaluation einer schulischen Innovation beschrieben, die für die Planung von Evaluationsuntersuchungen wichtig sein dürften. Sie wurden vom Autor und seinen Mitarbeitern für die Evaluation des umfassenden Schulversuchs des Learning Research und Development Center in Pittsburgh entwickelt, der mit dem Projekt Individually Prescribed Instruction durchgeführt wird.

Wesentlich neue Gesichtspunkte bringt auch der aus Großbritannien stammende Bericht über die Evaluation des Humanities Curriculum Project, die aufgrund der speziellen curricularen Vorstellungen des Projekts (vgl. Stenhouse 1971) vor besonderen Schwierigkeiten steht. Hier gilt es, eine Evaluation durchzuführen, ohne daß Lernziele im herkömmlichen Sinn als Kriterien der Evaluation verwendet werden. Das heißt, die Evaluation muß in Entsprechung zu der »Offenheit« des Projekts konzeptualisiert und durchgeführt werden. MacDonald gibt einen Erfahrungsbericht über die ersten zwei Jahre der Evaluation, in dem die zahlreichen Unzuläng-

lichkeiten offen dargelegt werden, und entwickelt einen Evaluationsplan für die nächsten zwei Jahre der Evaluation des Humanities Curriculum Project.

Das hier von MacDonald gewählte Verfahren der Evaluation erinnert in einigen Aspekten durchaus an Handlungsforschung (action research), die einen weithin neuen Bereich der pädagogischen Forschung darstellt. Handlungsforschung zielt auf die sofortige Lösung der untersuchten Probleme. Dafür ist sie bereit, den klassischen Forschungsplan aufzugeben und die entsprechenden Nachteile in Kauf zu nehmen. Sie beruht auf der Hypothese, daß Lehrer ihr Verhalten besonders dann verändern, wenn sie ihre Einstellungen ändern (Corey 1953). Dazu können sie im Rahmen der Lehrerfortbildung am besten gebracht werden, wenn sie – unter Beratung von Wissenschaftlern – sich der Erforschung ihrer eigenen Probleme zuwenden. Handlungsforschung ist u. a. durch zwei Aspekte gekennzeichnet:

(1) Ziel und Methode können nicht wie bei der klassischen empirischen Forschung in einer einfachen Zweck-Mittel-Relation gesehen werden. Die Ziele entstehen und verändern sich unter dem Einfluß der »Objekte« der Forschung im Laufe der Untersuchung, wodurch ebenfalls eine Veränderung der Forschungsverfahren bewirkt wird.

(2) Das Verhältnis von Subjekt und Objekt ist nicht durch die übliche Rollenverteilung gekennzeichnet. Die Distanz zwischen den Handelnden und den ihre Handlungen Erforschenden wird weitgehend aufgehoben. Das impliziert, daß die Validität und Reliabilität der Forschungsergebnisse nicht mehr gewährleistet ist und die Generalisierbarkeit der Ergebnisse in Frage gestellt werden muß.

Beim augenblicklichen Stand der Diskussion sollte man sich gegenüber der Handlungsforschung als einer Ausprägung der Evaluation offen zeigen und sich fragen, ob und in welchen Bereichen ihre Anwendung möglich und sinnvoll ist. Das neuerliche Interesse an dieser Form der Forschung in der BRD wird vielleicht bald zu ihrem besseren Verständnis führen.

Samuel Ball / Gerry Ann Bogatz

Das erste Jahr von Sesame Street

Eine Evaluation

Die Vorgeschichte der Untersuchung

Im Sommer 1968 begann das Children's Television Workshop (CTW), sein Programm Sesame Street zu planen. Alle Beteiligten stimmten überein, daß die Pläne eine unabhängige Evaluation der Auswirkungen des Programms einschließen sollten. Children's Television Workshop beauftragte das Educational Testing Service (ETS), eine gemeinnützige pädagogische Test- und Forschungsinstitution in Princeton, New Jersey, eine Evaluation durchzuführen, um festzustellen, in welchem Ausmaß die Fernsehsendung Sesame Street ihre gesetzten Ziele während des ersten Jahres erreicht hatte. Die Untersuchung versuchte u. a. folgende Fragen zu beantworten:

Was sind, im ganzen gesehen, die Auswirkungen von Sesame Street?

Was sind die modifizierenden Einflüsse von Alter, Geschlecht, vorausgehendem Leistungsstand und sozio-ökonomischem Status auf die Auswirkungen von Sesame Street?

Haben Kinder, die zu Hause Sesame Street sehen, im Vergleich zu Kindern, die es zu Hause nicht sehen, einen Gewinn davon?

Haben Kinder in Vorschulklassen, die Sesame Street als Teil ihres Schulcurriculum ansehen, einen Gewinn davon?

Haben Kinder aus spanisch-sprechenden Elternhäusern einen Gewinn von Sesame Street?

Wie beeinflussen die häuslichen Verhältnisse die Auswirkungen von Sesame Street?

Das innovative pädagogische Programm des Children's Television Workshop erhielt wesentliche Unterstützung von öffentlichen und privaten Stellen. Von Anfang an waren es die Carnegie Corporation New York, die Ford Foundation, das National Center for Educational Research and Development im U. S. Office of Education, das U. S. Office of Economic Opportunity und das National Institute of Child Health and Human Development. Unter den anderen Institutionen, die später für Unterstützung sorgten, wa-

ren die Corporation for Public Broadcasting, die National Foundation of Arts and Humanities und die John & Mary R. Markle Foundation.

Die Hauptergebnisse

In der ersten Sendeperiode von 26 Wochen zeigte Sesame Street, daß das Fernsehen ein wirkungsvolles Medium sein kann, um 3- bis 5jährigen Kindern wichtige einfache Sachverhalte und Fertigkeiten, wie z. B. das Erkennen und Benennen von Buchstaben und Zahlen, und komplexere höhere kognitive Fertigkeiten, wie das Klassifizieren und Sortieren nach einer Vielzahl von Kriterien, zu lehren. Die Forschungsergebnisse des Educational Testing Service erbrachten, daß Sesame Street Kinder aus sozial benachteiligten innerstädtischen Bezirken, aus mittelständischen Vororten und aus abgelegenen ländlichen Gebieten fördert. All diese Gruppen wurden in dieser Evaluation untersucht.

Die Leistungsfähigkeit des Bildungsfernsehens als Unterrichtsmedium wird durch drei Hauptergebnisse der Untersuchung deutlich:

1. Kinder, die am meisten zusahen, lernten auch am meisten. Das Ausmaß des Lernens – d. h. der Punktzuwachs, den ein Kind zwischen den Testergebnissen für bestimmte Fertigkeiten vor und nach dem Betrachten von Sesame Street zeigte – vergrößerte sich im Verhältnis zu dem Ausmaß der Zeit, die das Kind dem Programm zusah.

2. Die Fertigkeiten, die die meiste Zeit und Aufmerksamkeit durch das Programm erhielten, waren mit wenigen Ausnahmen auch die Fertigkeiten, die am besten gelernt wurden. Eine Analyse des Inhalts der Sendung zeigte z. B., daß mehr Zeit (13,9 %) als jedem anderen Gegenstand den Fertigkeiten, die mit Buchstaben in Beziehung stehen, gewidmet wurde. Auf diesem Gebiet der Buchstaben und Zahlen waren die Gewinne der Kinder am auffälligsten. Außer dem Erwerb von Fertigkeiten, die direkt und ausdrücklich gelehrt wurden, fand auch offensichtlich ein gewisser Lerntransfer statt, indem einige Kinder Dinge lernten, die in dem Programm nicht gelehrt wurden, z. B. ganze Wörter zu erkennen oder den eigenen Namen zu schreiben.

3. Das Programm erforderte keine ausdrückliche Beaufsichtigung durch Erwachsene, damit die Kinder auf den vom Programm umfaßten Gebieten lernten. Kinder, die Sesame Street zu Hause sahen, zeigten ebenso große Gewinne und in einigen Fällen sogar größere als Kinder, die in der Schule unter Aufsicht eines Lehrers das Programm sahen. Dieses Ergebnis hat besondere Bedeutung angesichts der Tatsache, daß mehr als vier Fünftel aller Drei- bis Vierjährigen und ebenso mehr als ein Viertel aller Fünfjährigen keinerlei Bildungseinrichtungen besuchen.

Das Hauptergebnis, daß Kinder, je länger sie die Sendung sehen, desto mehr auch lernen, gilt unabhängig von Alter, Geschlecht, geographischer Wohnlage, sozio-ökonomischem Status, Intelligenzalter und unabhängig davon, ob die Kinder zu Hause oder in der Schule das Programm sahen. In allen acht Bereichen, in denen die Kinder getestet wurden, vergrößerten sich die Lerngewinne mit der Häufigkeit des Zuschauens. Der Punktzuwachs war bei einigen Tests und Untertests jedoch höher, und einige Gruppen von Kindern zeigten einen höheren Punktzuwachs als andere. Die Dreijährigen erzielten die höchsten, die Fünfjährigen die geringsten Gewinne. D. h. dreijährige Kinder, die die Sendung sahen, hatten höhere Punktzahlen im Nachtest als diejenigen Vier- und Fünfjährigen, die die Sendung seltener sahen, selbst dann, wenn im Vortest die jüngeren Kinder niedrigere Punktzahlen als die älteren hatten. Dieses Ergebnis hat bedeutsame Folgen für die gesamte Erziehung, denn es legt nahe, daß dreijährige Kinder fähig sind, viele Fertigkeiten zu lernen, die gewöhnlich erst in späteren Jahren unterrichtet werden.

Ein ähnliches Ergebnis zeigte sich bei sozial privilegierten und sozial benachteiligten Kindern. Obwohl die sozial benachteiligten Kinder mit beträchtlich niedrigeren Leistungen in den Fertigkeiten, die gelehrt wurden, begannen, übertrafen diejenigen, die sehr oft und lange die Sendung sahen, die Kinder der Mittelschicht, die nur selten das Programm sahen. Diese Fernsehsendungen können offensichtlich die beträchtliche Kluft in der Bildung, die gewöhnlich sozial privilegierte und sozial benachteiligte Kinder trennt, schon bis zum Zeitpunkt des Eintritts in die erste Klasse verringern.

Ein auffallendes, obwohl sehr vorläufiges Ergebnis legt nahe, daß Sesame Street besonders effektiv sein könnte, den Kindern einige Fertigkeiten zu lehren, deren Muttersprache nicht Englisch ist und die in der Schule keine guten Leistungen erzielen. Eine sehr kleine Stichprobe von Kindern aus spanisch-sprechenden Elternhäusern im Südwesten erzielte bessere Gewinne als jede andere Untergruppe von Kindern.

Sesame Street hat einige seiner Ziele erfolgreicher verwirklicht als andere. Die Untersuchung liefert die Gründe und gibt Anhaltspunkte für die Verbesserung der Programmentwicklung. Es zeigte sich, daß in einigen Fällen der relative Mangel an Erfolg von einer anfänglichen Unterschätzung, in anderen Fällen von einer anfänglichen Überschätzung der Vorkenntnisse und Fertigkeiten der Kinder herrührte. Ein weiteres Ergebnis bestand darin, daß das Lernen erfolgreicher war, wenn die Fertigkeiten wie bei den Buchstaben direkt und nicht wie bei den Anfangslauten indirekt angesprochen wurden.

Stichprobe und Tests

Zu Beginn der Untersuchung wurden annähernd 1200 Kinder aus den fünf verschiedenen Regionen Boston (Massachusetts), Durham (North Carolina), Philadelphia (Pennsylvania), Phoenix (Arizona) und aus einem ländlichen Gebiet im Nordosten Kaliforniens ausgewählt. Die Stichprobe, die schließlich 943 Kinder zählte, bestand aus sozial benachteiligten Kindern aus innerstädtischen Bezirken, sozial privilegierten Kindern aus Vorortgebieten, Kindern aus ländlichen Gebieten und sozial benachteiligten spanischsprechenden Kindern. Im ganzen umfaßte die Stichprobe mehr Jungen als Mädchen und mehr Unterschicht- als Mittelschichtkinder. Unter den sozial benachteiligten waren mehr schwarze als weiße Kinder. Die meisten Kinder waren vier Jahre, einige waren drei und einige fünf Jahre alt. Die Mehrzahl der Kinder der Stichprobe sahen Sesame Street zu Hause und nicht im Vorschulunterricht.

Die Hersteller von Sesame Street hatten spezifische Lernziele für das Programm festgesetzt. Um den Lerngewinn im Hinblick auf diese Ziele und die Transferwirkungen zu bestimmen, wurden Meßinstrumente benutzt, die vom Educational Testing Service eigens für diese Evaluation entwickelt worden waren. Die acht Haupttests und ihre Untertests waren[1]:

Körperteiletest
- Auf die Körperteile zeigen
- Benennen der Körperteile
- Funktion von Körperteilen (zeigen)
- Funktion von Körperteilen (nennen)

Buchstabentest
- Erkennen von Buchstaben
- Benennen von Großbuchstaben
- Benennen von Kleinbuchstaben
- Vorgegebene Buchstaben in Wörtern finden
- Erkennen von Buchstaben in Wörtern
- Anfangslaute
- Wörter lesen

Formentest
- Erkennen von Formen
- Benennen von Formen

Zahlentest
- Erkennen von Zahlen
- Benennen von Zahlen
- Zahlverständnis (vgl. Beispielaufgabe 1)
- Zählen
- Addition und Subtraktion
- (Parallelisierter Untertest für Buchstaben, Zahlen und Formen)

Beziehungstest
- Umfangsbeziehungen
- Größenbeziehungen
- Positionsbeziehungen (vgl. Beispielaufgabe 2)

Sortiertest

Klassifikationstest (vgl. Beispielaufgabe 3)
- Klassifikation nach Größe
- Klassifikation nach Form
- Klassifikation nach Zahl
- Klassifikation nach Funktion

Puzzletest

Alle Tests waren nach dem gleichen Grundschema aufgebaut. Die Testmaterialien waren einfach, und die Tests wurden den Kindern einzeln von einem geschulten Erwachsenen aus ihrer Nachbarschaft gegeben. Ferner wurden Informationen über die familiären Verhältnisse eines jeden Kindes gesammelt und darüber, wie oft das Kind Sesame Street gesehen hatte. Die 943 Kinder der Stichprobe wurden in Quartilen aufgeteilt entsprechend der Länge der Zeit, die sie während der Dauer der Untersuchung Sesame Street gesehen hatten. Alle nachfolgenden Analysen sind auf diese Quartilen bezogen worden. Sie reichen von Quartil 1 (Q 1), in dem die Kinder Sesame Street selten oder nie sahen, bis Quartil 4 (Q 4), in dem die Kinder das Programm im Durchschnitt mehr als fünfmal in der Woche sahen. (Sesame Street war so populär, daß es nur wenige Kinder gab, die die Sendung wirklich nicht sahen; viele Kinder in Q 1 sahen das Programm gelegentlich.).

Gesamtergebnisse

Für die Stichprobe insgesamt gilt: Kinder aus den Quartilen, in denen die Sendung am meisten gesehen wurde, verhielten sich in allen Tests besser als Kinder aus den Quartilen, in denen sie weniger gesehen wurde. Kinder, die das Programm am meisten sahen (Q 4), hatten die höchsten Punktzahlen im Vortest (das bedeutet, daß sie schon mit einem Vorsprung anfingen), sie hatten die höchsten Punktzahlen im Nachtest, und sie erzielten den höchsten Punktzuwachs in der Zeit zwischen Vor- und Nachtest. Die allgemeine Tendenz, bei längerem und häufigerem Ansehen der Sendung einen höheren Punktzuwachs zu erzielen, war bei einigen Tests ausgeprägter als bei anderen. Besonders ausgeprägt war diese Tendenz bei den Buchstaben-, Zahlen- und Klassifikationstests; am wenigsten zeigte sie sich beim Körperteiletest.

Sozial benachteiligte Kinder

In der Gesamtstichprobe von 943 Kindern wurden 731 als sozial benachteiligt angesehen. Auch bei ihnen erhöhte sich der Punktzuwachs im Verhältnis zu der Häufigkeit, in der sie Sesame Street sahen. Im Hinblick auf die Gesamtpunktzahl für die 203 Testaufgaben, die im Vor- und Nachtest gleich war, gewannen die Kinder aus Q 1 19 Punkte, die Kinder aus Q 2 29 Punkte, die Kinder aus Q 3 38 Punkte und die Kinder aus Q 4 47 Punkte (siehe Tabelle 1)[2]. Ein Teil des Punktzuwachses, der von Kindern aus Q 1 erzielt wurde, muß weitgehend als eine Folge der Reifung angesehen werden, da viele von ihnen die Sendung niemals gesehen hatten. Die größeren Gewinne der Kinder in anderen Quartilen sind jedoch weitgehend eine Folge der Häufigkeit ihres Ansehens der Sendung. Dieselbe Beziehung ließ sich zwischen den verschiedenen Gesamtbeträgen für alle acht Haupttests beobachten. Den höchsten Punktzuwachs gab es bei den Buchstaben-, Zahlen- und Klassifikationstests (vgl. Tabelle 1).

Komplizierte statistische Analysen wurden durchgeführt, um zu bestimmen, ob die beobachteten Unterschiede sich zufällig eingestellt hatten, ob sie signifikant durch andere Faktoren herbeigeführt worden waren oder ob sie – wie es schien – weitgehend eine Folge der Häufigkeit des Ansehens der Sendung waren[3]. Die Häufigkeit des Zuschauens erwies sich bei weitem als die bedeutendste Variable; das bedeutet, ihr Einfluß schien gleichermaßen davon unabhängig zu sein, welches Geschlecht die Kinder hatten und ob sie das Programm zu Hause oder in der Schule sahen. Um die Auswirkungen der Häufigkeit des Zuschauens genau zu isolieren, wurde eine Spezialuntersuchung mit zwei Parallelgruppen von Kindern durch-

geführt (die »Age Cohorts Study« – die Altersgruppen-Untersuchung). Gruppe 1 war 53 bis 58 Monate zur Zeit des Vortests alt; Gruppe 2 war 53 Monate bis 58 Monate zur Zeit des Nachtests alt. Außer demselben Lebensalter zum Zeitpunkt des Vergleichs waren die beiden Gruppen auch von vergleichbarem Intelligenzalter und lebten in denselben Gemeinden. Es gab, kurz gesagt, keine beobachtbaren Unterschiede zwischen den beiden Gruppen in bedeutsamen Punkten wie Vorkenntnissen, IQ und häuslichen Verhältnissen. In jeder Gruppe waren mehr als 100 sozial benachteiligte Kinder, die keine Bildungseinrichtungen besuchten. Die Vortest-Punktzahlen von Gruppe 1 (bevor die Kinder Sesame Street gesehen haben konnten) wurden mit den Nachtest-Punktzahlen von Gruppe 2 verglichen, nachdem diese Kinder das Programm gesehen hatten. Die das Programm häufig sehenden Kinder in Gruppe 2 – Kinder aus Q 3 und Q 4 – erreichten über 40 Punkte mehr bei den 203 gemeinsamen Testaufgaben als die vergleichbaren Kinder in Gruppe 1, die die Sendung niemals gesehen hatten (vgl. Tabelle 2). In gleicher Weise signifikant ist die Tatsache, daß gelegentliche Zuschauer (Q 1) in Gruppe 2 sich nur um 12 Punkte von vergleichbaren Kindern in Gruppe 1, die Sesame Street nicht gesehen hatten, unterschieden. Zusammenfassend kann gesagt werden: Hielt man Auswirkungen der Reifung, IQ, Vorkenntnisse und häusliche Verhältnisse konstant, erzielten die häufigen Zuschauer große und bedeutsame Gewinne.

Obwohl die Häufigkeit des Zuschauens sich mit dem Alter der Kinder nicht auffallend veränderte, ergaben sich dennoch Änderungen in den Testergebnissen. Zum Zeitpunkt des Vortests schnitten Dreijährige, wie vorauszusehen war, weniger gut als Vierjährige und Vierjährige weniger gut als Fünfjährige ab. In bezug auf den Punktzuwachs im Nachtest waren die Ergebnisse jedoch gerade umgekehrt. Obwohl die Gruppe unter den Dreijährigen, die das Programm am häufigsten sahen, im Vortest mit einem niedrigeren Punktwert als irgendeine Gruppe der Fünfjährigen begann, erreichten die Dreijährigen, die die Sendung am häufigsten sahen, zum Zeitpunkt des Nachtests im Durchschnitt höhere Punktwerte als die Vierjährigen in Q 1, Q 2 und Q 3 und höhere als die Fünfjährigen in Q 1 und Q 2. Selbst Dreijährige, die das Programm nur zwei- oder dreimal in der Woche sahen, erzielten im Vergleich zu anderen Altersgruppen einen beachtlichen Punktzuwachs (vgl. Tabellen 3, 4, 5 und Abbildung 1).

Einige Testergebnisse hingen deutlich vom Alter ab. Unter den Kindern, die die Sendung häufig sahen, wurde beim Körperteiletest der höchste Punktzuwachs von den Dreijährigen erzielt; Drei- und Vierjährige erzielten bei den Zahlen einen höheren Punktzuwachs als Fünfjährige; und Fünfjährige erreichten beim Lesen von Wörtern (was einen Lerntransfer anzeigt) und bei den Anfangslauten (was in Sesame Street indirekt gelehrt wurde)

einen höheren Punktzuwachs als die anderen. Um es kurz zu sagen: Ziele, die indirekt gelehrt wurden, wurden von den älteren Zuschauern besser gelernt, und ein Lerntransfer zeigte sich bei ihnen deutlicher als erwartet werden konnte. Im allgemeinen galt: Wo spezifische Kenntnisse und Fertigkeiten direkt gelehrt wurden, erzielten die jüngeren Kinder einen höheren Punktzuwachs als die älteren.

Sozial privilegierte Kinder

169 Kinder in der Untersuchung wurden als sozial privilegiert angesehen. Sie erreichten im Vortest höhere Punktwerte als die anderen Gruppen und sahen im Durchschnitt einen größeren Teil der Sendungen als alle Gruppen der sozial benachteiligten Kinder. Eine relativ geringe Häufigkeit des Zuschauens brachte bei diesen Kindern relativ hohen Punktzuwachs (vgl. Tabelle 6 und Abbildung 2).

Spanisch-sprechende Kinder

Es wurden nur 43 spanisch-sprechende Kinder von der Untersuchung erfaßt. Sie unterschieden sich in dem Ausmaß, in dem sie vor dem Ansehen von Sesame Street mit der englischen Sprache in Berührung gekommen waren. Infolge dieser Unterschiede und dem geringen Umfang der Stichprobe können Schlußfolgerungen nur mit großer Vorsicht gezogen werden. Die größte Zahl der spanisch-sprechenden Kinder war in Q 1; lediglich ein Rest von 18 befand sich in der das Programm häufig sehenden Gruppe. Diese Kinder erzielten einen fast unglaublich hohen Punktzuwachs. Der Punktzuwachs der spanisch-sprechenden Kinder aus Q 3 war in der Tat so hoch wie der der anderen Kinder in Q 4. Beim Buchstabentest begannen die spanisch-sprechenden Kinder aus Q 4 mit den niedrigsten Punktwerten im Vortest und erreichten die höchsten Punktwerte im Nachtest. Andere Buchstaben-Untertests und die Tests über Zahlen, Formen, über das Sortieren, die Beziehungsverhältnisse und das Klassifizieren zeigten das gleiche Ergebnis: ein niedriger Start mit nachfolgendem sehr hohem Punktzuwachs für Kinder, die das Programm sehr häufig sahen.

Kinder aus ländlichen Gebieten

In der Untersuchung hatten die Kinder aus ländlichen Regionen in den Vortests relativ niedrige Punktwerte, erreichten aber in den Nachtests als Folge des Zuschauens einen hohen Punktzuwachs. Ihre Eltern hatten oft eine bessere Bildung als die Eltern der sozial benachteiligten Stadtkinder.

Ihre großen Gewinne legen nahe, daß Sesame Street als pädagogisches Medium für Kinder, die in abgelegenen Gegenden oder in kleinen Dörfern wohnen, sehr geeignet ist.

Sesame Street in den Schulen

Die Lehrer, deren Klassen Sesame Street im Rahmen der Untersuchung sahen, wurden gebeten, ihre Reaktionen auf das Programm anzugeben. Obwohl sie Sesame Street als Unterrichtsmittel für kleine Kinder anerkannten, waren sie über die Eignung des Programms für den Unterricht selbst geteilter Meinung. Einige vertraten nachdrücklich die Auffassung, daß die Sendung wertvolle Zeit verbrauche, die besser für andere Vorhaben verwendet werden könnte, andere meinten, daß sie eine wertvolle Bereicherung des Schultags sei.

Kinder, Eltern und Sesame Street

Die Kinder, die Sesame Street am meisten sahen und deshalb am meisten lernten, hatten Mütter, die oft mit ihnen zusammen die Sendung ansahen und oft zu ihnen über die Sendung sprachen. In diesen Familien hatten die Eltern in der Regel etwas höhere Erwartungen für ihre Kinder.

Schlußfolgerung

Hinsichtlich seiner selbst gesteckten Ziele war Sesame Street im allgemeinen sehr erfolgreich. Die Untersuchung des Educational Testing Service zeigt, daß drei- bis fünfjährige Kinder aus verschiedenen häuslichen Verhältnissen wichtige einfache und komplexe kognitive Fertigkeiten durch das Ansehen von Sesame Street erwarben. Die am meisten die Sendung sahen, erzielten auch die höchsten Gewinne. Die zusammenfassende Schlußfolgerung lautet: die Leistungsfähigkeit des Bildungsfernsehens als eines wirkungsvollen Mediums, um bestimmte Fertigkeiten sehr kleinen Kindern zu lehren, ließ sich durch Sesame Street nachweisen [4].

Abbildung 1

Vortestergebnis und Punktzuwachs im Gesamttest für alle sozial benachteiligten 3-, 4- und 5jährigen Kinder

(nach der Häufigkeit des Zuschauens in Quartilen unterteilt)

N = 127 3jährige
N = 433 4jährige
N = 159 5jährige

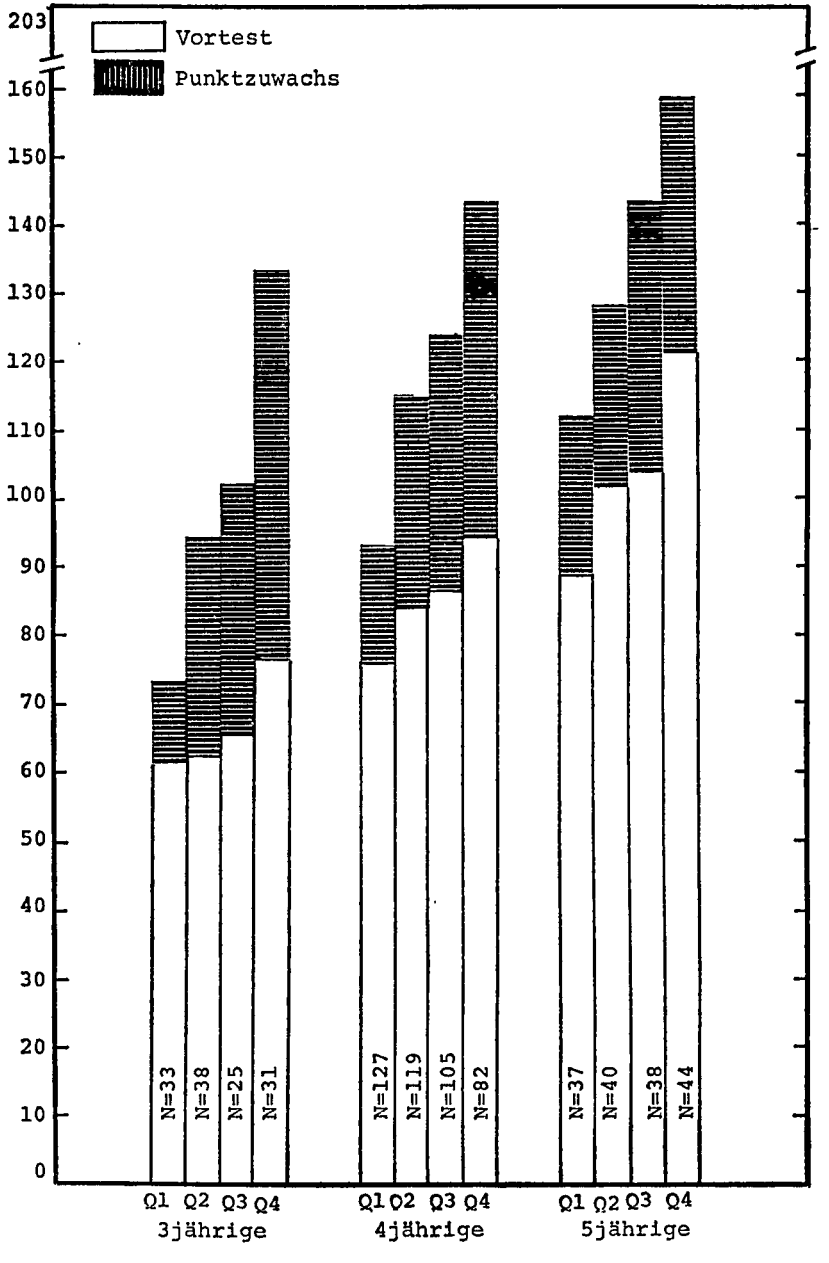


Abbildung 2

Vortestergebnis und Punktzuwachs im Gesamttest
für alle sozial privilegierten Kinder
(nach der Häufigkeit des Zuschauens in Quartilen unterteilt)
N = 169

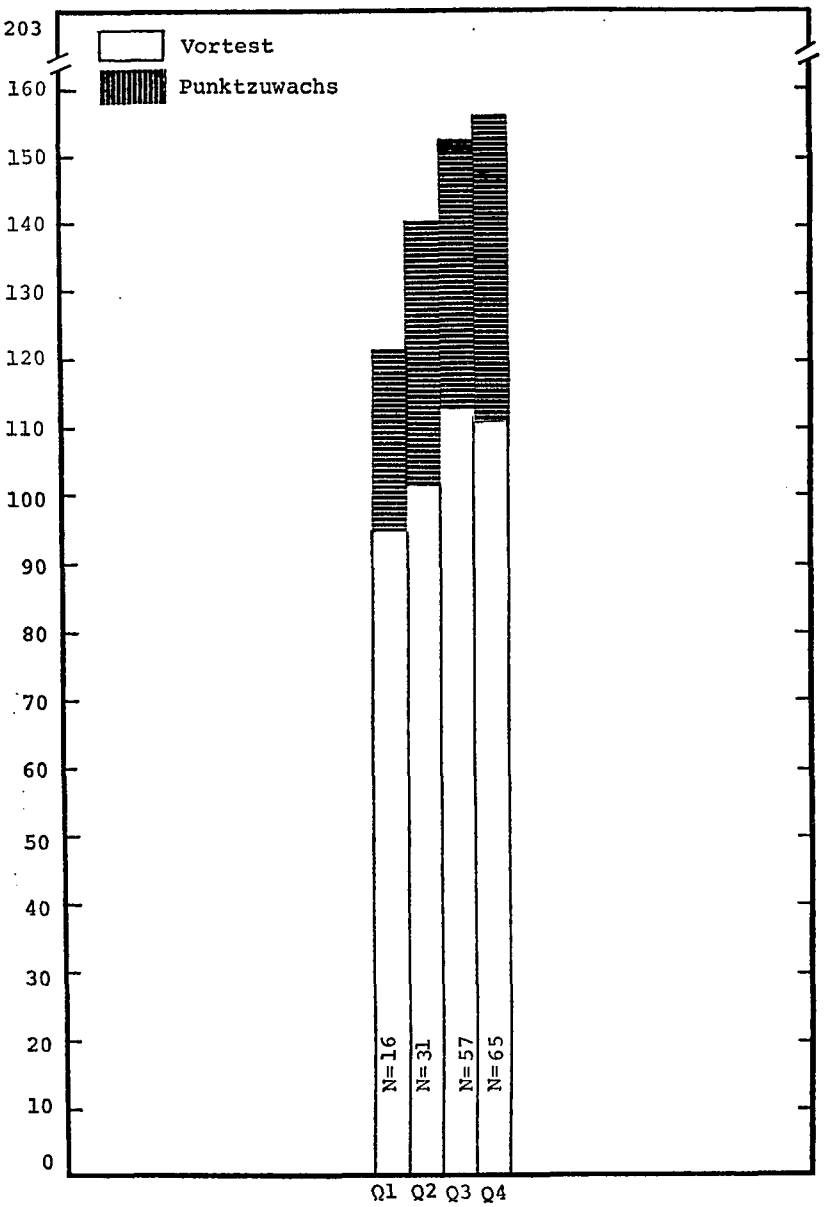
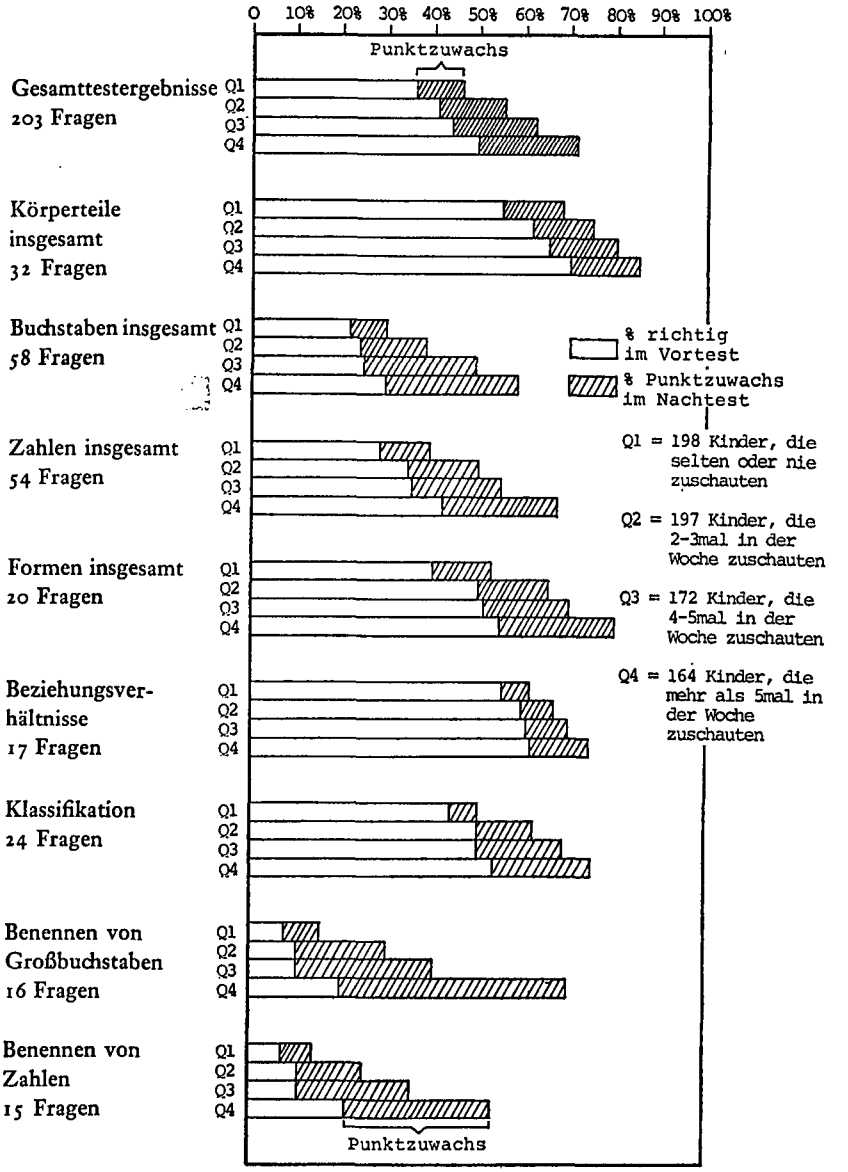


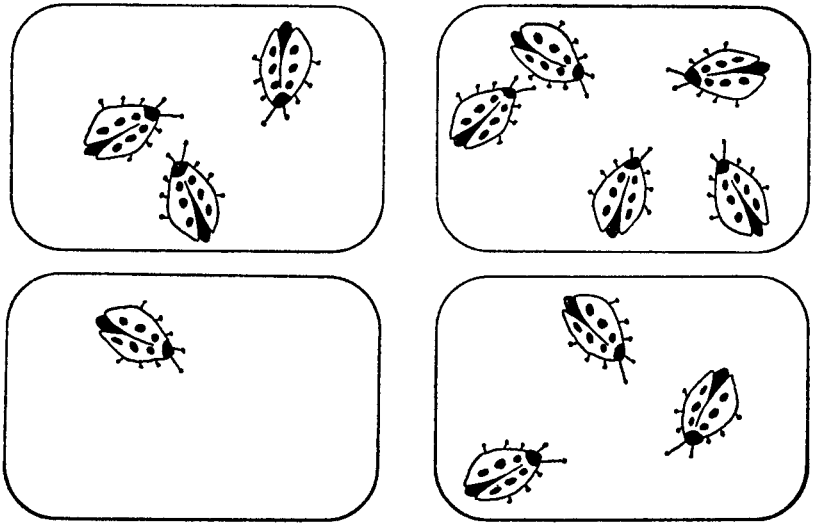
Abbildung 3

Zusammenstellung der Ergebnisse, die sozial benachteiligte Kinder in den verschiedenen Tests erzielten

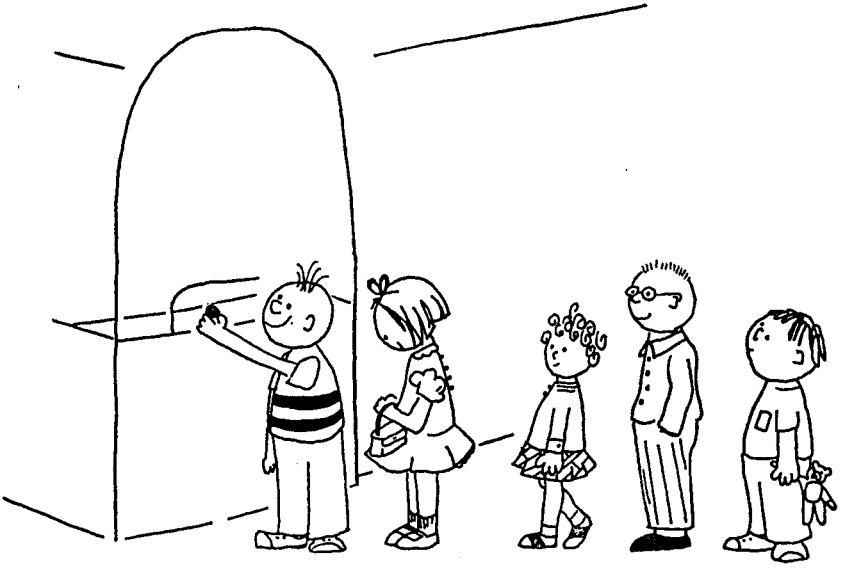
(Die von allen sozial benachteiligten Kindern in Vor- und Nachtest richtig beantworteten Testaufgaben sind in Prozenten angegeben.)



Testaufgabe (Beispiel 1)
Schau auf die Marienkäfer hier, hier, hier und hier. In welchem Kästchen sind fünf Marienkäfer?

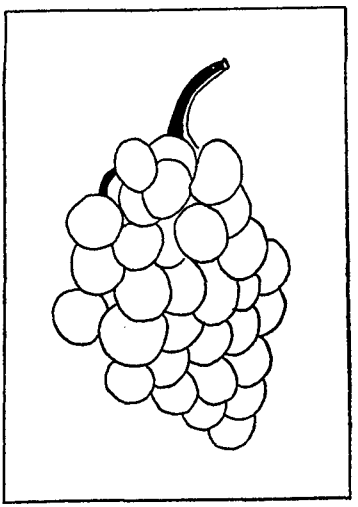


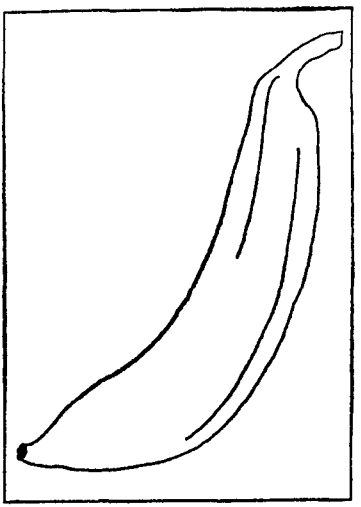
Testaufgabe (Beispiel 2)
Hier stehen Kinder in einer Reihe an. Sie warten, um in ein Kino gehen zu können. Welches Kind steht zuletzt in der Reihe?

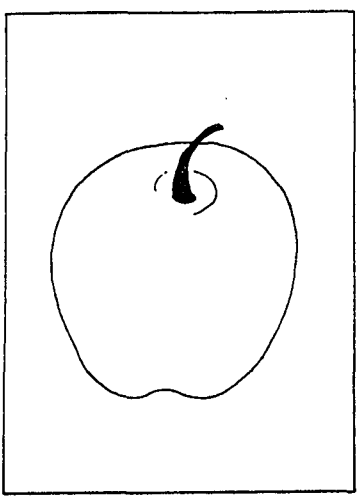


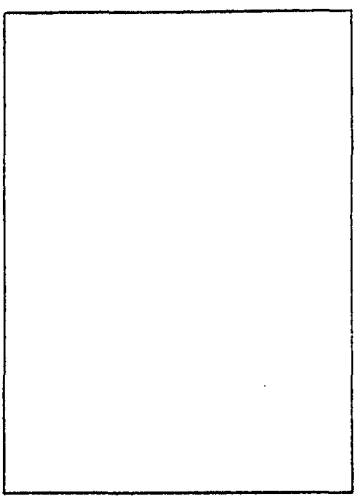
Testaufgabe (Beispiel 3)

Hier ist ein Bild von Weintrauben, von einer Banane und einem Apfel. Ein Bild fehlt. Wir wollen das Bild, das hierher paßt, herausfinden.



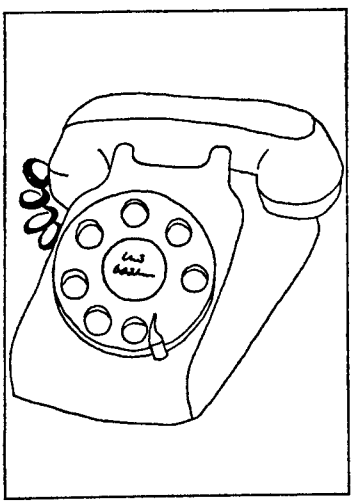


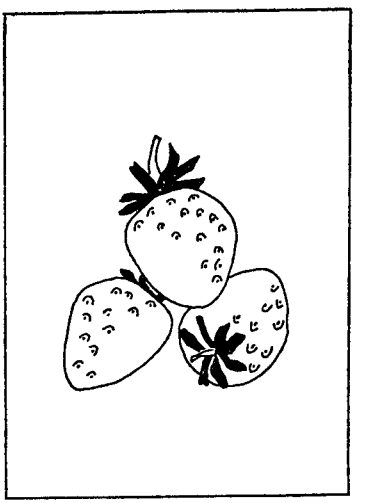


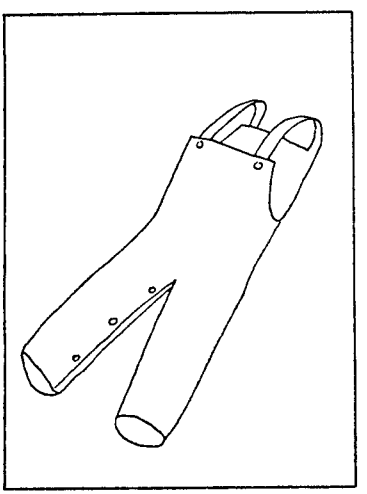


Testaufgabe (Beispiel 4)

. Hier siehst du ein Telephon, Erdbeeren, eine Hose und ein Buch.
Was davon paßt zu den Weintrauben, der Banane und dem Apfel?







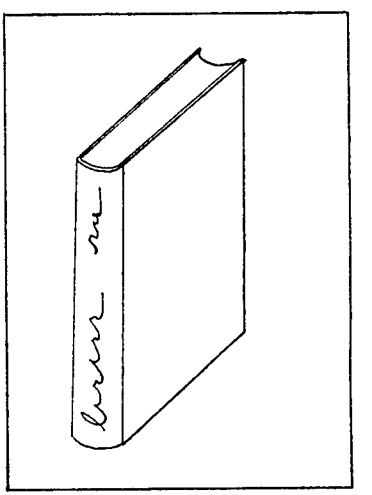


Tabelle 1

Vortestergebnis und Punktzuwachs für alle sozial benachteiligten Kinder

(in Quartilen unterteilt)

N = 731

Haupttests (ohne Untertests)	Maximal möglicher Punktwert *	Q 1 N = 198				Q 2 N = 197				Q 3 N = 172				Q 4 N = 164			
		Vortest		Punktzuwachs		Vortest		Punktzuwachs		Vortest		Punktzuwachs		Vortest		Punktzuwachs	
		$\overline{X}$	s	$\overline{X}$	s	$\overline{X}$	s	$\overline{X}$	s	$\overline{X}$	s	$\overline{X}$	s	$\overline{X}$	s	$\overline{X}$	s
Gesamttestergebnisse	203	75.62	24.73	18.63	20.04	84.42	27.60	29.11	22.51	87.74	27.63	37.97	25.29	97.54	32.16	47.36	26.15
Körperteile insgesamt	32	18.11	6.51	3.88	5.71	20.00	6.35	4.38	5.50	21.09	6.04	4.74	5.31	22.47	6.05	5.24	4.88
Buchstaben insgesamt	58	13.07	5.95	4.30	7.43	14.42	7.37	8.22	9.26	14.95	7.00	11.89	11.00	17.98	10.12	15.97	11.19
Formen insgesamt	20	8.43	3.50	2.29	3.77	9.89	4.01	3.15	4.05	10.04	3.64	4.29	4.07	10.64	3.50	5.49	3.52
Zahlen insgesamt	54	16.18	8.20	5.43	7.05	18.56	9.38	8.52	8.23	19.64	10.10	10.88	9.51	23.69	11.15	13.01	9.52
Parallelisierter Untertest	11	7.83	2.76	1.26	2.87	8.38	2.55	1.50	2.50	8.90	2.19	1.12	2.09	9.32	1.77	1.02	1.82
Beziehungsverhältnisse insges.	17	9.07	2.98	1.11	3.18	9.88	3.06	1.52	3.34	10.08	2.77	1.80	2.93	10.15	3.13	2.47	3.34
Sortieren insgesamt	6	2.30	1.33	0.47	1.85	2.54	1.44	0.81	1.82	2.52	1.50	1.38	1.76	2.73	1.39	1.64	1.71
Klassifikation insgesamt	24	10.57	4.15	1.67	4.41	11.98	4.63	2.96	4.78	12.06	4.68	4.56	4.97	12.88	4.60	5.32	4.67
Puzzles insgesamt	5	1.88	1.40	0.43	1.86	2.04	1.37	0.80	1.64	2.15	1.28	0.83	1.58	2.41	1.45	0.98	1.57

$\overline{X}$ = arithmetisches Mittel

s = Standard-Abweichung

* Die Differenz zwischen dem maximal möglichen Punktwert für den Gesamttest und der Summe der maximal möglichen Punktwerte der einzelnen Haupttests ergibt sich dadurch, daß einige Testaufgaben in mehreren Tests verwendet wurden.

Tabelle 2

Vortest- und Nachtestpunktwerte für sozial benachteiligte Kinder, die zu Hause der Sendung zuschauten

(nach der Häufigkeit des Zuschauens in Quartilen unterteilt)

Gruppe 1 = Kinder, die zum Zeitpunkt des *Vortests* 53–58 Monate alt waren

Gruppe 2 = Kinder, die zum Zeitpunkt das *Nachtests* 53–58 Monate alt waren

(Alterskohorten)

Haupttests (ohne Untertests)	Maximal möglicher Punktwert *	Q 1				Q 2				Q 3				Q 4			
		Gruppe 1 N = 31 Vortest		Gruppe 2 N = 26 Nachtest		Gruppe 1 N = 33 Vortest		Gruppe 2 N = 33 Nachtest		Gruppe 1 N = 27 Vortest		Gruppe 2 N = 18 Nachtest		Gruppe 1 N = 23 Vortest		Gruppe 2 N = 24 Nachtest	
		$\overline{X}$	s	$\overline{X}$	s	$\overline{X}$	s	$\overline{X}$	s	$\overline{X}$	s	$\overline{X}$	s	$\overline{X}$	s	$\overline{X}$	s
Gesamttestergebnisse	203	76.77	22.27	88.42	21.83	81.97	18.90	101.70	24.78	90.37	25.21	130.33	29.59	99.04	36.42	139.33	35.99
Körperteile insgesamt	32	17.87	6.49	21.04	6.01	20.24	5.74	22.91	5.84	21.93	5.57	26.83	3.73	22.87	5.51	26.75	4.58
Buchstaben insgesamt	58	14.06	6.45	14.65	3.91	13.09	3.65	18.24	6.82	14.81	5.90	26.83	11.89	18.52	11.33	31.92	14.18
Formen insgesamt	20	7.45	3.36	11.04	3.43	9.09	3.21	11.21	3.27	9.93	4.08	14.22	3.61	10.35	4.21	15.46	3.91
Zahlen insgesamt	54	16.77	7.06	19.00	7.64	17.97	7.10	23.76	9.63	20.37	9.42	32.67	10.67	23.96	12.42	35.54	11.77
Parallelisierter Untertest	11	7.97	2.93	9.31	1.85	8.45	1.99	9.97	1.16	8.78	2.28	10.33	0.59	9.17	1.67	10.00	1.50
Beziehungsverhältnisse insges.	17	9.61	2.35	10.65	2.78	10.33	2.98	11.30	2.27	10.81	2.32	12.39	2.48	10.26	3.77	18.00	2.52
Sortieren insgesamt	6	2.13	1.38	2.69	1.41	1.67	1.29	3.33	1.49	2.81	1.55	4.28	1.32	2.30	1.22	4.54	1.25
Klassifikation insgesamt	24	10.71	3.84	11.96	4.25	11.03	2.91	13.79	4.25	12.89	4.50	17.78	4.10	13.04	5.06	17.75	5.14
Puzzles insgesamt	5	2.03	1.56	2.31	0.93	2.55	1.37	2.55	1.39	2.26	1.02	3.44	1.38	2.52	1.44	2.92	1.35

$\overline{X}$ = arithmetisches Mittel

s = Standard-Abweichung

* Die Differenz zwischen dem maximal möglichen Punktwert für den Gesamttest und der Summe der maximal möglichen Punktwerte der einzelnen Haupttests ergibt sich dadurch, daß einige Testaufgaben in mehreren Tests verwendet wurden.

Tabelle 3

Vortestergebnis und Punktzuwachs für alle sozial benachteiligten 3jährigen Kinder

(nach der Häufigkeit des Zuschauens in Quartilen unterteilt)

N = 127

Haupttests (ohne Untertests)	Maximal möglicher Punktwert *	Q 1 N = 33				Q 2 N = 38				Q 3 N = 25				Q 4 N = 31			
		Vortest		Punktzuwachs		Vortest		Punktzuwachs		Vortest		Punktzuwachs		Vortest		Punktzuwachs	
		$\overline{X}$	s	$\overline{X}$	s	$\overline{X}$	s	$\overline{X}$	s	$\overline{X}$	s	$\overline{X}$	s	$\overline{X}$	s	$\overline{X}$	s
Gesamttestergebnisse	203	60.76	20.34	12.42	25.67	62.42	20.82	30.71	21.14	65.48	15.76	37.20	28.28	75.81	25.14	57.23	25.66
Körperteile insgesamt	32	13.88	5.21	3.03	6.26	15.76	5.77	4.79	5.91	16.72	5.44	6.64	6.94	18.84	6.26	8.00	5.52
Buchstaben insgesamt	58	10.73	5.99	3.79	9.20	10.18	4.95	7.35	8.99	11.32	3.99	10.52	9.71	11.91	6.65	20.13	12.14
Formen insgesamt	20	7.70	3.16	1.03	3.83	7.84	3.90	3.39	3.96	7.36	2.81	5.00	4.25	9.13	3.50	6.29	3.59
Zahlen insgesamt	54	11.21	6.40	2.94	9.34	11.37	6.08	9.34	7.53	13.00	5.39	8.08	10.02	16.38	8.39	14.13	9.79
Parallelisierter Untertest	11	6.94	2.70	0.94	3.43	6.53	3.33	3.05	3.04	7.00	2.68	2.40	2.72	8.25	2.53	2.03	2.74
Beziehungsverhältnisse insges.	17	7.42	2.46	1.39	3.55	8.45	3.13	1.79	3.46	8.24	2.62	1.76	3.44	8.72	2.39	3.23	2.70
Sortieren insgesamt	6	2.33	1.29	−0.12	1.73	2.21	1.36	0.42	1.73	2.44	1.26	0.92	1.85	2.41	1.10	1.52	1.59
Klassifikation insgesamt	24	8.67	3.53	1.27	3.59	8.50	4.43	4.53	4.69	9.12	3.48	4.44	4.81	10.56	4.66	5.71	3.68
Puzzles insgesamt	5	1.76	1.28	0.21	1.85	1.63	1.10	0.45	1.43	1.28	1.02	1.24	1.48	2.03	1.49	1.19	1.60

$\overline{X}$ = arithmetisches Mittel

s = Standard-Abweichung

* Die Differenz zwischen dem maximal möglichen Punktwert für den Gesamttest und der Summe der maximal möglichen Punktwerte der einzelnen Haupttests ergibt sich dadurch, daß einige Testaufgaben in mehreren Tests verwendet wurden.

Tabelle 4

Vortestergebnis und Punktzuwachs für alle sozial benachteiligten 4jährigen Kinder

(nach der Häufigkeit des Zuschauens in Quartilen unterteilt)

N = 433

Haupttests (ohne Untertests)	Maximal möglicher Punktwert *	Q 1 N = 127				Q 2 N = 119				Q 3 N = 105				Q 4 N = 82			
		Vortest		Punktzuwachs		Vortest		Punktzuwachs		Vortest		Punktzuwachs		Vortest		Punktzuwachs	
		$\bar{X}$	s	$\bar{X}$	s	$\bar{X}$	s	$\bar{X}$	s	$\bar{X}$	s	$\bar{X}$	s	$\bar{X}$	s	$\bar{X}$	s
Gesamttestergebnisse	203	75.13	22.21	18.24	18.40	84.09	23.25	30.60	24.35	86.63	23.64	38.50	25.44	93.79	29.50	49.01	24.62
Körperteile insgesamt	32	18.35	6.22	4.09	5.31	20.08	6.22	4.92	5.47	21.14	5.85	4.64	5.19	22.27	5.75	5.10	4.50
Buchstaben insgesamt	58	13.20	5.92	3.45	6.37	13.94	6.08	8.46	9.17	14.56	5.67	12.02	11.17	17.33	8.84	15.37	10.45
Formen insgesamt	20	8.21	3.42	2.55	3.82	9.87	3.67	3.31	4.40	9.94	3.59	4.32	4.13	10.38	3.39	5.63	3.72
Zahlen insgesamt	54	15.82	6.89	5.69	6.32	18.72	7.96	8.84	8.83	19.08	8.85	11.37	9.81	21.95	10.43	14.65	8.65
Parallelisierter Untertest	11	7.81	2.77	1.35	2.76	8.41	2.28	1.49	2.40	8.98	2.05	1.05	1.94	9.46	1.31	0.77	1.33
Beziehungsverhältnisse insges.	17	8.99	2.79	1.02	3.11	9.78	2.70	1.65	3.33	9.99	2.62	1.95	3.11	9.70	3.22	2.80	3.50
Sortieren insgesamt	6	2.05	1.28	0.62	1.91	2.48	1.40	0.95	1.84	2.47	1.46	1.36	1.81	2.52	1.28	1.87	1.59
Klassifikation insgesamt	24	10.52	3.84	1.17	4.60	12.01	3.98	2.91	4.66	11.83	4.33	4.86	5.02	12.33	4.38	5.77	5.07
Puzzles insgesamt	5	1.86	1.44	0.32	1.84	2.10	1.37	0.78	1.69	2.17	1.24	0.79	1.58	2.19	1.33	1.01	1.58

$\bar{X}$ = arithmetisches Mittel

s = Standard-Abweichung

* Die Differenz zwischen dem maximal möglichen Punktwert für den Gesamttest und der Summe der maximal möglichen Punktwerte der einzelnen Haupttests ergibt sich dadurch, daß einige Testaufgaben in mehreren Tests verwendet wurden.

Tabelle 5

Vortestergebnis und Punktzuwachs für alle sozial benachteiligten 5jährigen Kinder

(nach der Häufigkeit des Zuschauens in Quartilen unterteilt)

N = 159

Haupttests (ohne Untertests)	Maximal möglicher Punktwert *	Q 1 N = 37				Q 2 N = 40				Q 3 N = 38				Q 4 N = 44			
		Vortest		Punktzuwachs		Vortest		Punktzuwachs		Vortest		Punktzuwachs		Vortest		Punktzuwachs	
		$\overline{X}$	s	$\overline{X}$	s	$\overline{X}$	s	$\overline{X}$	s	$\overline{X}$	s	$\overline{X}$	s	$\overline{X}$	s	$\overline{X}$	s
Gesamttestergebnisse	203	88.68	29.20	23.08	19.14	101.23	30.69	26.75	17.30	104.13	30.82	38.97	25.73	120.91	29.78	37.32	26.37
Körperteile insgesamt	32	20.38	7.15	3.92	6.68	23.35	4.34	2.93	4.98	23.18	6.02	4.08	5.34	25.73	4.40	3.41	3.55
Buchstaben insgesamt	58	14.97	5.59	6.35	8.45	18.40	10.05	8.70	9.70	18.79	8.98	13.66	11.64	24.16	12.71	14.32	11.71
Formen insgesamt	20	9.35	3.74	2.81	3.06	11.08	4.15	3.30	3.04	11.97	3.15	3.39	3.58	12.20	3.15	4.64	3.25
Zahlen insgesamt	54	21.00	10.71	5.95	6.87	23.53	11.37	7.58	6.54	25.89	11.87	11.18	9.41	31.89	10.12	9.66	9.93
Parallelisierter Untertest	11	8.84	2.61	1.05	2.84	9.48	1.72	0.70	1.64	9.97	1.05	0.32	1.49	9.96	1.19	0.66	1.27
Beziehungsverhältnisse insges.	17	10.81	3.28	0.97	2.85	11.28	3.44	1.18	3.56	11.11	2.66	1.58	2.34	12.02	2.62	1.25	3.05
Sortieren insgesamt	6	2.89	1.33	0.62	1.74	2.83	1.50	0.95	1.85	2.74	1.67	1.71	1.63	3.27	1.57	1.36	1.87
Klassifikation insgesamt	24	12.05	5.07	3.19	4.08	14.28	4.74	2.45	5.08	14.05	4.98	4.13	4.64	15.49	4.24	4.18	4.66
Puzzles insgesamt	5	2.05	1.39	1.00	1.83	2.33	1.46	1.08	1.65	2.45	1.37	0.92	1.62	3.02	1.45	0.73	1.60

$\overline{X}$ = arithmetisches Mittel

s = Standard-Abweichung

* Die Differenz zwischen dem maximal möglichen Punktwert für den Gesamttest und der Summe der maximal möglichen Punktwerte der einzelnen Haupttests ergibt sich dadurch, daß einige Testaufgaben in mehreren Tests verwendet wurden.

Tabelle 6

Vortestergebnis und Punktzuwachs für alle sozial privilegierten Kinder

(nach der Häufigkeit des Zuschauens in Quartilen unterteilt)

N = 169

Haupttests (ohne Untertests)	Maximal möglicher Punktwert *	Q 1 N = 16				Q 2 N = 31				Q 3 N = 57				Q 4 N = 65			
		Vortest		Punktzuwachs		Vortest		Punktzuwachs		Vortest		Punktzuwachs		Vortest		Punktzuwachs	
		$\bar{X}$	s	$\bar{X}$	s	$\bar{X}$	s	$\bar{X}$	s	$\bar{X}$	s	$\bar{X}$	s	$\bar{X}$	s	$\bar{X}$	s
Gesamttestergebnisse	203	95.44	23.90	26.69	16.04	102.13	21.65	38.65	17.02	112.77	24.36	40.46	18.83	110.83	25.63	45.25	22.87
Körperteile insgesamt	32	24.13	5.77	3.19	4.97	25.74	4.90	2.52	4.31	26.37	5.64	2.35	4.28	25.71	4.79	3.14	4.50
Buchstaben insgesamt	58	15.19	8.79	8.06	9.26	16.81	7.03	12.48	10.10	19.25	10.21	17.09	9.99	18.62	8.86	19.63	11.46
Formen insgesamt	20	10.63	3.48	3.00	4.23	11.35	3,20	4.32	2.74	12.37	3.05	3.88	3.59	12.31	3.15	4.62	3.39
Zahlen insgesamt	54	22.13	10.37	8.69	5.38	24.13	8.65	12.06	6.79	28.07	9.80	12.16	8.17	27.50	10.83	12.40	7.68
Parallelisierter Untertest	11	9.31	1.45	0.81	1.17	9.90	1.01	0.39	1.20	9.67	1.09	0.65	1.11	9.32	1.60	1.05	1.74
Beziehungsverhältnisse insges.	17	10.63	2.58	1.56	2.85	10.48	2.34	2.10	2.69	11.58	1.96	1.19	2.19	11.71	2.57	1.38	2.64
Sortieren insgesamt	6	2.75	1.34	0.50	1.41	2.81	1.22	1.52	1.29	2.98	1.41	1.65	1.83	2.86	1.41	1.75	1.54
Klassifikation insgesamt	24	11.50	3.12	3.69	5.33	14.03	3.56	4.97	4.01	15.19	4.21	4.58	4.95	15.11	4.23	4.55	4.27
Puzzles insgesamt	5	2.75	1.18	0.13	0.96	2.23	1.15	1.23	1.41	2.93	1.42	0.79	1.59	3.15	1.21	0.48	1.60

$\bar{X}$ = arithmetisches Mittel

s = Standard-Abweichung

* Die Differenz zwischen dem maximal möglichen Punktwert für den Gesamttest und der Summe der maximal möglichen Punktwerte der einzelnen Haupttests ergibt sich dadurch, daß einige Testaufgaben in mehreren Tests verwendet wurden.

Richard C. Anderson

Eine vergleichende Felduntersuchung: Ein Beispiel vom Biologieunterricht in der Sekundarstufe[1]

Eine gebräuchliche, aber wenig sinnvolle Form der pädagogischen Forschung ist der Versuch, verschiedene Unterrichtsmethoden miteinander zu vergleichen. In den letzten Jahren gab es zahlreiche Vergleiche zwischen Vorträgen, die über das Fernsehen ausgestrahlt wurden, und Vorträgen, die direkt vor den Adressaten gehalten wurden, zwischen forschendem Lernen und darstellendem Lehrervortrag, zwischen schüler- und lehrerzentriertem Unterricht, zwischen programmiertem und Lehrbuch-Unterricht usw. Der Unterricht selbst war bei diesen Untersuchungen nur von geringer Bedeutung. Er war lediglich das Vehikel zur Evaluation einer Unterrichtsmethode; man nahm an, man könne die dabei erzielten Ergebnisse auf beliebige Unterrichtsinhalte übertragen. Gegenwärtig gibt es meiner Meinung nach eine allgemeine Übereinstimmung darüber, daß diese Annahme ungerechtfertigt war (Cronbach 1963; Lumsdaine 1965). Nichtsdestoweniger will ich darlegen, daß die vergleichende Untersuchung, wenn man sie anders einsetzt, einen Teil des Aufwands an Zeit und Mühe in der pädagogischen Forschung verdient.

Die Begründung lautet etwa so: Unsere Fähigkeit, die für die Schüler beste Unterrichtsart vorherzusagen, ist gering. Es gibt keine Unterrichtsmethoden, die sich gegenüber anderen Methoden stets als besser erwiesen haben. Es gibt keine Unterrichtsmerkmale, die notwendigerweise mit einer besseren Schülerleistung verknüpft sind. Weder kleine Lernschritte noch aktives Antworten, noch sofortige Leistungskontrolle und Erfolgsbestätigung, noch ein gutes Klassenklima, noch das stufenweise Fortschreiten vom Konkreten zum Abstrakten, noch die Möglichkeit, die Richtung und die Geschwindigkeit des Lernens selbst zu bestimmen, noch der Einsatz von multimedialen Stimuli garantieren einen erfolgreichen Unterricht.

Dies ist keine erfreuliche Perspektive; aber es ist meiner Ansicht nach keine Übertreibung. Gewiß haben wir hierüber einige Kenntnisse; doch gibt es mehr Probleme, über die wir nichts Genaues wissen. Meiner Meinung nach können wir zur Zeit die Effektivität eines Unterrichts nicht zu-

verlässig voraussagen, auch wenn die philosophischen Grundlagen, der Stil, die Methoden und die Verfahren des Unterrichts bekannt sind.

Wenn dies richtig ist, stellt sich folgende Frage: Wie sollen die Finanzen der Geldgeber und die Zeit und Mühe der pädagogischen Forscher eingesetzt werden, um die Effektivität des Unterrichts heute und in der Zukunft zu maximieren? Eine Antwort, die ich unterstützen würde, ist die Investition in pädagogische und verhaltenswissenschaftliche Grundlagenforschung. Man sollte jedoch die Wirkung der Grundlagenforschung auf die Unterrichtspraxis realistisch beurteilen.

Innerhalb der Verhaltenswissenschaften gibt es gegenwärtig eine stark ausgeprägte empiristische Tendenz (Conant 1952). Dies gilt insbesondere für die angewandten Wissenschaften, die sich von der Verhaltenswissenschaft Anregungen erhoffen. Pädagogische Grundlagenforschung sollte uns immer mehr dazu befähigen, ohne vorherige Erprobung die Unterrichtsverfahren und die Organisation von Curriculummaterial genau zu bestimmen, die mit großer Wahrscheinlichkeit das Lernen der Schüler fördern. Mit einem allmählichen Fortschritt kann man rechnen. Aber es wäre unrealistisch, zu erwarten, daß wir jemals eine effektive Unterrichtsgestaltung mit mehr als geringer Wahrscheinlichkeit vorhersagen können. Ich zweifle nicht daran, daß die Curriculumentwicklung immer teilweise auf Regeln beruhen wird, die über den Daumen gepeilt sind. Ich zweifle auch nicht daran, daß viele Versuche nach dem Prinzip des »Trial and Error« immer notwendig sein werden, um erfolgreichen Unterricht zu *gewährleisten.*

Bisher habe ich dargelegt, daß wir nur in geringem Maße die Merkmale des Unterrichts vorhersagen können, die den Lernerfolg der Schüler maximieren, und daß man von der Grundlagenforschung erwarten kann, daß sie auch nur bescheidene Verbesserungen in dieser Hinsicht leisten kann. Doch etwas sollten wir jetzt tun: Wir können in Vorversuchen und Felduntersuchungen Unterrichtseinheiten unterscheiden, die sich gut oder schlecht für den Unterricht eignen. Daher sollten wir von dieser Möglichkeit, den Unterricht zu verbessern, Gebrauch machen. Aus diesem Grunde sollte Unterricht durch Schülerleistungen evaluiert werden, und jeder einzelne Schritt in der Entwicklung des Curriculummaterials sollte die Schülerleistungen berücksichtigen. Auf der Grundlage des vorhandenen Wissens ist es nicht möglich, Unterrichtsmethoden zu evaluieren, aber es ist möglich, dies bei einzelnen Unterrichtsstunden, Unterrichtseinheiten oder Curricula zu tun.

Zufriedenstellenden Unterricht kann man durch die systematische Anwendung des Prinzips des »Trial and Error« entwickeln. Dieser Prozeß erfordert die Bestimmung der Lernziele, die Vorbereitung von Curriculummaterialien, die diesen Lernzielen (hoffentlich) entsprechen und schließ-

lich die Erprobung der Curriculummaterialien mit den Adressaten. Auf der Grundlage erfolgreicher Versuche werden die Materialien dann überarbeitet. Der Prozeß von Versuch und Überarbeitung wird so lange fortgesetzt, bis die Lernziele erreicht sind oder die Entscheidung getroffen wird, daß es unmöglich ist, sie im Rahmen der zur Verfügung stehenden Zeit und Mittel zu erreichen.

Die Funktion der Felduntersuchung

Wenn die Ergebnisse der Vortests erkennen lassen, daß die Schüler die Lernziele erreichen, ist es an der Zeit, das gesamte zusammengehörende Curriculummaterial einem Feldtest zu unterziehen. Das gesamte Curriculummaterial umfaßt nicht nur die Materialien, die direkt den Schülern gegeben werden, und Ausführungen und Anleitungen, die Hinweise für den Lehrer darüber enthalten, wie Diskussionen, Laborübungen und das Lösen von Aufgaben und Problemen zu leiten sind; sondern es kann auch Lehrerhandbücher, die Organisation von Lehrerseminaren und die Anleitung zu einem angemessenen Unterricht miteinschließen. Ein Ziel einer Felduntersuchung ist es festzustellen, ob sich unter verschiedenen Anwendungsbedingungen das gesamte Curriculummaterial als erfolgreich erweist. Der Vortest kann für die gesamte Schülerpopulation, die mit diesem Material arbeiten soll, repräsentativ sein; er muß es aber nicht sein. Die Vortests wurden unter Umständen von jemandem durchgeführt und beaufsichtigt, der von dem Projekt angetan war und der über die richtige Benutzung des Materials Bescheid wußte. Was geschieht aber, wenn die Materialien in die Hände von Lehrern gegeben werden, die ihnen gegenüber interesselos oder gar abweisend eingestellt sind? Müssen die Materialien auf eine bestimmte Weise benutzt werden, oder sind sie auch bei unterschiedlichen Anwendungsbedingungen einigermaßen erfolgreich? Wenn das Curriculum in einer bestimmten Weise benutzt werden muß, ist dann für Lehrerhandbücher oder für Lehrerseminare gesorgt? Und bringen die Handbücher oder Seminare die Lehrer mit Erfolg auf den angestrebten Weg? Dies sind einige der Fragen, die in einer Felduntersuchung beantwortet werden können.

Wenn man die von Scriven (1967) eingeführten Begriffe verwendet, so ist das Ziel von Vortests formative Evaluation, um Mängel im Verständnis oder in der Leistung der Schüler aufzuzeigen, so daß Herausgeber, Autoren oder Lehrer die Curriculummaterialien und die Unterrichtsmethoden überarbeiten und vermutlich verbessern können.

Es ist für die Felduntersuchung nur ein sekundäres Ziel, den Curricu-

lumentwicklern die Ergebnisse ihrer Arbeit vor Augen zu führen. Das Hauptziel ist summative Evaluation. Dabei werden Daten gesammelt, um möglichen Adressaten – wie Erziehungsinstitutionen, Beamten der Schulverwaltung, Lehrern und Schülern – bei der Entscheidung zu helfen, ob ein bestimmtes Curriculum benutzt werden soll oder nicht.

Einige Befürworter der empirischen Validierung von Curriculummaterialien scheinen die Ansicht zu vertreten, die Effektivität der erzielten Verhaltensänderungen bei Schülern sei bei der Beurteilung des Unterrichts das einzige Kriterium. Ich möchte betonen, daß dies nicht mein Standpunkt ist. Unterrichtsstunden, Unterrichtseinheiten und Curricula sollten danach beurteilt werden, in welchem Ausmaß sie ihre Ziele erreichen; aber dies sollte nicht das einzige Kriterium sein. Andere Kriterien sind die Kosten des Unterrichtsablaufs in Form von Zeit, die die Schüler und Lehrer aufwenden müssen, die Billigung des Unterrichtsablaufs seitens der Schüler und Lehrer und alle Nebeneffekte (Stake 1967a). Genauigkeit, Modernität und Einfallsreichtum der Lehrinhalte waren die wichtigen Kriterien der bekannten Curriculum-Reformprojekte. Ein sehr wichtiges Kriterium ist der Wert der Ziele, die der Unterricht zu erreichen anstrebt. Wie Scriven (1967) bemerkt hat, »ist es offensichtlich uninteressant, wie gut die Lernziele erreicht werden, wenn sie wertlos sind.« Die Umkehrung dieser Aussage ist ebenfalls richtig: Unabhängig davon, wie wertvoll die Ziele sind, kann ein Unterricht nicht positiv bewertet werden, wenn er so uneffektiv ist, daß er diese Ziele nicht erreicht. Effektivität sollte als Kriterium für die Beurteilung des Unterrichts weder über- noch unterschätzt werden.

Manchmal sollte die Felduntersuchung des gesamten Curriculummaterials eine vergleichende Untersuchung sein. Diese Schlußfolgerung ist unvermeidlich, wenn die Felduntersuchung die Entscheidungen der Adressaten mitbestimmen soll. Es gibt in den Bereichen des schulischen Gesamtcurriculum verschiedene alternative Curricula zur Auswahl. Für den Fall, daß sich die Lernziele und -inhalte verschiedener Curricula überschneiden, ist für die Entscheidung in der Praxis durchaus die Frage angebracht, welches das effektivste ist.

Cronbach (1963) und Scriven (1967) haben zum Wert von vergleichenden Untersuchungen gegensätzliche Positionen bezogen. Bis auf eine Einschränkung stimme ich mit Scriven überein. Vergleichende Untersuchungen haben sehr wohl eine wertvolle Funktion. Aber Scriven scheint für die Adressaten umfangreiche Vergleichsuntersuchungen von Curricula in jedem Fachbereich vor Augen zu haben. Hierzu hat Cronbach zu Recht die Gegenposition vertreten, daß die meisten Vergleiche wahrscheinlich keine Unterschiede von statistischer Signifikanz oder praktischer Bedeutung ergeben würden.

Vergleichende Untersuchungen sind kostspielig. Sie können nicht wahllos durchgeführt werden. Ein Kriterium für die Entscheidung über die Durchführung einer vergleichenden Untersuchung ist folgendes: Es muß eine erhebliche Wahrscheinlichkeit dafür bestehen, daß eines der Curricula in der Tat effektiver ist als das andere. Vermutungen haben in der Grundlagenforschung durchaus ihren Platz. Für eine vergleichende pädagogische Untersuchung kann dies aber nicht gelten. Aus der Sicht dessen, der eine vergleichende Untersuchung durchführt, sollte lediglich bewiesen werden, daß eines der Curricula besser ist als das andere.

In einer vergleichenden Untersuchung haben Ergebnisse, die keine Unterschiede zeigen, einen sehr geringen gesellschaftlichen Nutzen. Wenn man mit Nachdruck die Vorstellung zurückweist, daß eine vergleichende Untersuchung den generellen Wert einer Unterrichtsmethode zeigen kann, und die Vorstellung akzeptiert, daß die wichtigste Begründung für eine vergleichende Untersuchung darin liegen muß, zu bestimmen, welches von zwei oder mehreren Curriculummaterialien das effektivste ist, dann ist es offensichtlich sinnlos, Curriculummaterial auf die bloße Möglichkeit hin zu vergleichen, daß das eine besser als das andere sein könnte; es sei denn vielleicht, man glaube, es gäbe viele gute Curricula, die unbeachtet herumliegen und darauf warten, entdeckt zu werden. Vielleicht ist die Feststellung von einem gewissen Wert, daß eine groß propagierte curriculare Innovation nicht effektiver ist als ein anderes Curriculum. Im allgemeinen jedoch können ergebnislos verlaufende vergleichende Untersuchungen die Entscheidung der Adressaten nicht erleichtern. Daher muß ein Irrtum in der Beurteilung vorgelegen haben, wenn eine vergleichende Untersuchung keine Unterschiede aufzeigt. Zeit und Geld, die in die Curriculumentwicklung und in die formative Evaluation hätten investiert werden sollen, sind so zu einem voreiligen Vergleich verschwendet worden.

Es mag eingewandt werden, daß Forschung nicht damit gerechtfertigt werden kann, bloß zu beweisen, was ohnehin mit hoher Wahrscheinlichkeit vermutet wird. Das Gegenargument basiert auf der These, die bereits zuvor in diesem Beitrag entwickelt wurde. Es gibt von vornherein keine Tests, die verläßlich die Effektivität eines Unterrichts vorhersagen können; gleiches gilt für Experten, deren Fähigkeiten bei der Beurteilung der Unterrichtseffektivität anerkannt sind. Kurz gesagt, es gibt keine akzeptablen Gründe für Aussagen über die Effektivität eines Unterrichts außer Ergebnissen, die tatsächlich die Effektivität beweisen.

Die Notwendigkeit relativer Normen

Die Auffassung, daß Curriculumeinheiten in bezug auf absolute Effektivitätsnormen evaluiert werden sollten, ist weit verbreitet. In der Tat ist dies die Auffassung, die ich im Hinblick auf Voruntersuchungen von Curriculumeinheiten vertrete. Bei Felduntersuchungen von Curriculummaterialien gibt es Gründe, sich nur mit Vorsicht ausschließlich auf absolute Normen zu verlassen. Vor allem existieren in der Pädagogik im Gegensatz zu anderen Bereichen – von der Landwirtschaft bis zur Automobilindustrie – keine übereinstimmend akzeptierten Leistungsnormen.

Angenommen, die Pädagogen könnten sich auf irgendeine allgemeine Norm einigen, wie auf die bekannte 90–90 Norm, die vom Air Force Training Command unter der Leitung von Colonel Gabriel Ofiesh vorgeschlagen wurde [2], was würde es bedeuten, wenn die Schüler durchschnittlich 90 % einer kriteriumsbezogenen Norm erreichten? Offensichtlich würde das nicht bedeuten, daß die Schüler 90 % all jenes Wissens beherrschten, das über ein Thema bekannt ist. Es würde bedeuten, daß sie 90 % von dem gelernt haben, was jemand für den Unterricht und für den Test ausgewählt hat. Hier liegt das Problem. Ungeachtet jüngster Fortschritte bei der Formulierung von Lernzielen, können immer noch bedeutsame Unterschiede in dem beabsichtigten oder in dem impliziten intellektuellen Niveau auftreten, mit dem ein Begriff entwickelt wird, obwohl angeblich die gleichen Ziele zugrunde liegen. Ein weiteres Problem liegt darin, daß das Leistungsniveau von den Testmethoden abhängig ist; ein Beispiel hierfür ist die Attraktivität von Distraktoren bei Tests mit Auswahl-Antwort-Aufgaben. Endlich schließt die Tatsache, daß ein Curriculum eine bestimmte Effektivitätsnorm erreicht, die Möglichkeit nicht aus, daß ein konkurrierendes Curriculum diese Norm mit weniger Zeitaufwand und mit geringeren Kosten besser erfüllt. Deshalb sind relative Normen und damit verbunden auch vergleichende Untersuchungen notwendig, um die Effektivität von Curriculummaterialien zu beurteilen.

Ich möchte nicht mißverstanden werden: Meiner Meinung nach sind absolute Effektivitätsnormen im Prinzip gut. Ich hoffe, es wird möglich sein, die Theorie und die Technik der Bestimmung absoluter Normen zu verbessern. In Anbetracht unserer Unzulänglichkeit, absolute Normen zu definieren und Leistung in bezug auf sie zu messen, sollten für die nächste Zukunft absolute Normen durch relative Normen ergänzt werden. Zum gegenwärtigen Zeitpunkt ist der direkte Vergleich der einzige verläßliche Weg, zu bestimmen, welches von zwei Curricula effektiver ist.

Vergleichende Untersuchungen haben eine eindeutige Funktion, wenn verschiedene Unterrichtsstunden (Unterrichtseinheiten, Curricula) im we-

sentlichen die gleichen Ziele haben. Ist dies der Fall, dann ist das effektivste Unterrichtsprogramm das beste, vorausgesetzt, daß andere Faktoren wie z. B. die Kosten vergleichbar sind. Die Adressaten können bei der Auswahl unter verschiedenen Curricula ihre Aufmerksamkeit hauptsächlich auf die Ergebnisse einer vergleichenden Untersuchung richten. Überdies – und dies ist einer der Gründe, warum ich für vergleichende Untersuchungen eintrete – wird der Wettbewerb, bessere Curriculummaterialien zu erstellen, auch dazu beitragen, effektiveren Unterricht zu schaffen.

Mir erscheint es nicht so einsichtig, daß vergleichende Untersuchungen sinnvoll sind, wenn die Lernziele der Curricula verschieden sind. Eine andere ungeklärte Frage ist: Wer sollte vergleichende Untersuchungen durchführen, die Entwickler von neuen Curricula oder unabhängige Evaluatoren? Ebenso gibt es Fragen über die geeignete Planung und Durchführung vergleichender Untersuchungen. Ehe ich mich zu diesen Fragen ganz allgemein äußere, werde ich lieber versuchen, sie an Hand eines Beispiels aus der Praxis zu erläutern. Der Rest dieses Beitrags beschreibt eine vergleichende Felduntersuchung, die durchgeführt wurde, um die Effektivität von neuem Curriculummaterial zu beweisen.

Die Felduntersuchung eines Unterrichtsprogramms in Populationsgenetik

Die Entwicklung des experimentellen Curriculummaterials

Mit der Unterstützung der Biological Sciences Curriculum Study (BSCS) wurde ein Programm in Populationsgenetik zum Selbstunterricht erstellt, das im Fach Biologie in der Sekundarstufe verwendet werden sollte (Faust/Anderson/Guthrie/Drantz 1967). Bei der Entwicklung des Programms wurde, wie oben kurz beschrieben, vorgegangen. Als erster Schritt wurden die Lernziele definiert. Hierbei diente die Behandlung der Populationsgenetik in den Lehrbüchern der Biological Sciences Curriculum Study als Richtlinie. Zunächst wurde eine Versuchsfassung eines Teils des Programms erstellt. Dieser Programmteil wurde mit einer Reihe einzelner Schüler der Sekundarstufe und einem der Programmautoren erprobt, wobei dieser die Arbeit der einzelnen Schüler überprüfte. Nach Versuchen mit einigen Schülern wurden dann jeweils Überarbeitungen vorgenommen. Die restlichen Teile des Programms wurden ebenso entwickelt. Schließlich wurde das vollständige Programm mit kleinen Schülergruppen getestet. Erneut wurden Überarbeitungen vorgenommen. Während der gesamten Entwicklung des Programms wurde ein sehr ausführlicher kriteriumsbe-

zogener Leistungstest benutzt; dieser bestand in der Hauptsache aus offen formulierten Fragen, bei denen Probleme gelöst, Begriffe und Gesetze definiert und erläutert werden mußten. Die Schüler, die an den Voruntersuchungen teilnahmen, mußten für die Durchführung des kriteriumsbezogenen Tests fast ebenso viel Zeit aufwenden wie für die Durcharbeitung des Programms selbst. Im allgemeinen wurde ein Programmteil als zufriedenstellend betrachtet, wenn alle an der Voruntersuchung beteiligten Schüler 90 % oder mehr der kriteriumsbezogenen Testaufgaben dieses Abschnitts richtig lösten. Die Fassung des Programms, die in dem Experiment verwendet wurde, enthielt in 234 Abschnitten, ohne Gleichungen und graphische Darstellungen, 14 000 Wörter.

Der Unterricht in der Kontrollgruppe

Das Programm über Populationsgenetik wurde verglichen mit der Behandlung der Populationsgenetik in dem Lehrbuch »Biological Science: An Inquiry Into Life«, das von der Biological Sciences Curriculum Study verfaßt worden war; inoffiziell ist dieses Buch bekannt als »BSCS yellow version«. Der Text enthält etwa 7 900 Wörter, die sich unmittelbar auf Populationsgenetik beziehen. Das Lehrbuchmaterial wurde durch Laborübungen ergänzt, die ebenfalls von der Biological Sciences Curriculum Study vorbereitet worden waren; der Unterricht wurde von einem Biologielehrer einer Sekundarstufe gegeben. Es wäre falsch, den Unterricht in der Kontrollgruppe als konventionellen Unterricht zu bezeichnen. Dieses Material wurde von einem Team von Biologen und Biologielehrern erarbeitet. Das Lehrbuch wurde einer größeren Revision unterzogen, die teilweise auf systematisch gesammelten Äußerungen vieler Lehrer aus allen Teilen des Landes beruhte, die die experimentelle Fassung des Lehrprogramms benutzten. Es ist offensichtlich, daß es für die Schüler in der Sekundarstufe kein besseres Unterrichtsmaterial für Populationsgenetik gibt als das Lehrbuch BSCS yellow version und die dazu gehörenden Hilfsmittel.

Anlage der Untersuchung

An dem Experiment nahmen annähernd 750 Schüler der Sekundarstufe teil; sie wurden von 9 Lehrern in 30 Klassen in zwei in Vororten gelegenen Schulen unterrichtet. Alle 9 Lehrer unterrichteten zwischen 2 und 4 Klassen. Die Klassen wurden nach dem Zufallsprinzip ausgewählt und die Programme mit der Auflage verteilt, daß nach Möglichkeit die Hälfte der Klassen eines jeden Lehrers das Programm erhalten sollte und die andere Hälfte nicht. Ferner wurden zwei Parallelformen des Leistungstests ent-

wickelt. Innerhalb jeder Klasse erhielt die Hälfte der Schüler, die ebenfalls zufällig ausgewählt wurde, eine Form als Vortest und die andere als Nachtest. Für die verbleibende Hälfte der Probanden wurde umgekehrt verfahren. Diese Untersuchungsanlage lieferte Grunddaten und Informationen auf der Basis einer relativ großen Anzahl von Testaufgaben mit einem relativ geringen Zeitaufwand seitens der Schüler; auf diese Weise wurde auch der typische Wiederholungseffekt vermieden, der auftreten kann, wenn Schüler genau den gleichen Test wiederholen.

Durchführung der Untersuchung

Die beteiligten Lehrer waren bereit, den Vor- und Nachtest zu bestimmten Zeitpunkten durchzuführen. In der Zwischenzeit erklärten sie sich damit einverstanden, das Programm in den dazu bestimmten Klassen und nicht in anderen Klassen zu verwenden. Den Lehrern wurde gesagt: »Setzen Sie bitte das Programm so ein, wie es Ihrer Unterrichtserfahrung am besten entspricht.« Ein Mitglied des Projektteams sprach kurz mit den Lehrern über die Art und Weise der Testdurchführung und über die Aufzeichnung der Ergebnisse; er unternahm jedoch keinen Versuch, das Programm zu loben oder Empfehlungen zu geben, wie es benutzt werden sollte.

Die recht unstrukturierten Lehreranweisungen können im Hinblick auf die Fragestellung der Untersuchung verstanden werden. Ist ein Programm über Populationsgenetik zum Selbstunterricht eine nützliche Ergänzung zu anderen BSCS-Materialien zu diesem Thema? Wir wollten sehen, ob das Programm unter den Bedingungen des alltäglichen Unterrichts nützlich ist, da dies die Umstände sind, unter denen die Lehrer Curriculummaterial verwenden müssen.

Im nachhinein gibt es keinen Zweifel darüber, daß mit dem Programm bessere Gesamtergebnisse erzielt worden wären, wenn für die Lehrer als Teil des gesamten Curriculummaterials ein Lehrerhandbuch beigegeben worden wäre. Zu jener Zeit jedoch – und ich denke, es war gut so – entschieden wir, kein Handbuch zur Verfügung zu stellen. Ein Handbuch ist nur in dem Maße sinnvoll, wie die Lehrer die Anleitungen, die darin enthalten sind, befolgen. Unsere Erfahrung hat gezeigt, daß die Lehrer die Handbücher, die dem Curriculummaterial beigegeben sind, oft nicht lesen. Es wurde als wichtig erachtet, herauszufinden, wie anfällig das Curriculummaterial unter ungünstigen Anwendungsbedingungen ist. Ein Handbuch für die Lehrer wird gegenwärtig erstellt; die Ergebnisse des Versuchs werden mitbenutzt, um die Lehrer zu überzeugen, daß die im Handbuch enthaltenen Empfehlungen beachtenswert sind.

Stake (1967a) hat überzeugend nachgewiesen, daß eine gute Evaluation des Unterrichts eine vollständige Beschreibung seiner Implementation beinhalten muß. Eine solche Beschreibung war in unserem Falle besonders wichtig, da den Lehrern sehr viel Spielraum gelassen wurde. Die Lehrer machten sowohl in den Versuchsklassen als auch in den Kontrollklassen Aufzeichnungen, mit denen alle Aktivitäten und ihre zeitliche Dauer zwischen Vor- und Nachtest beschrieben wurden. Alle Laborübungen, alle sonstigen Übungen und alle Anweisungen zum Lesen wurden genau aufgezeichnet. Ebenso füllten die Lehrer einen Fragebogen aus, der sowohl offene als auch geschlossene Fragen enthielt, die sich auf die Einstellung der Lehrer und Schüler zu dem Programm, auf Techniken der Programmbenutzung und deren Verhältnis zur anderen Unterrichtsarbeit bezogen; ebenso wurde nach Stärken und Schwächen des Programms gefragt. Die Schüler beantworteten einen Fragebogen, der sich mit ähnlichen Themen beschäftigte.

Gesamtanalyse der Ergebnisse des Leistungszuwachses

Da ganze Klassen nach dem Zufallsprinzip für den Unterricht mit und ohne Programm ausgewählt wurden, war die Klasse die Beobachtungseinheit. Die Grundlage für die Varianzanalyse war der mittlere Leistungszuwachs der einzelnen Klassen. In die Klassendurchschnitte gingen die Punktwerte aller Schüler ein, die sich dem Vor- und Nachtest in der jeweils vorgeschriebenen Form unterzogen hatten. Eine Reihe einzelner Schüler und eine vollständige Klasse wurden aus der Analyse ausgeschieden, da sie diese Kriterien nicht erfüllten.

Es wurde eine Varianzanalyse mit ungewogenen Mittelwerten durchgeführt. Dabei waren die Verwendung oder Nichtverwendung des Programms und die Schule die Faktoren. Lediglich der Unterschied zwischen dem Unterricht mit und ohne Programm war signifikant [$F\,(1{,}25) = 20{,}59$, $p<(0{,}01)$]. Dabei wurde ein w^2 von .39 erreicht. Mit anderen Worten: Es waren 39 % der Varianz des Leistungszuwachses in den Klassen dem Unterricht mit dem Programm zuzuschreiben. Der tatsächliche Leistungszuwachs betrug bei der Verwendung des Unterrichtsprogramms 4,62 Items; ohne Unterrichtsprogramm wurde ein Zuwachs von 3,01 Items erreicht. Relativ bedeutet dies, daß die Schüler, die mit dem Programm unterrichtet worden waren, gegenüber den anderen einen um 53 % höheren Leistungszuwachs erzielten. Die absolute Differenz war jedoch nicht so groß, wie wir erwartet hatten. Die Gründe für das Fehlen einer größeren absoluten Differenz werden später erörtert.

Leistung als eine Funktion der Herkunft der Testaufgaben

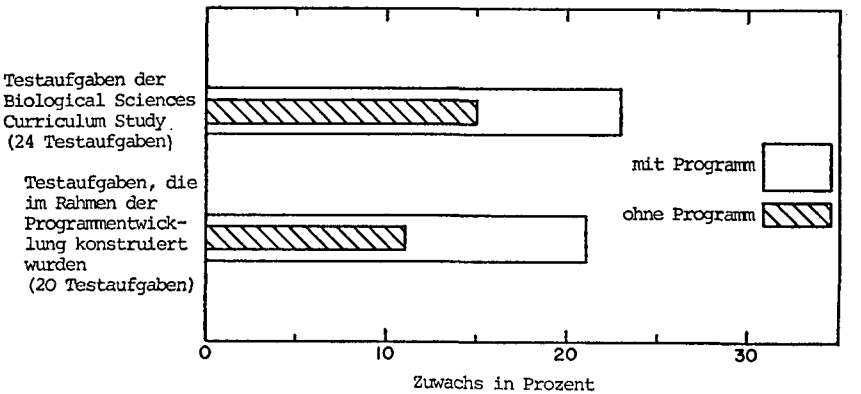
Während in der Felduntersuchung ein normenbezogener Test nach dem Auswahl-Antwort-Verfahren verwendet wurde, war der in den Voruntersuchungen des Programms verwendete Test kriteriumsbezogen; er war gekennzeichnet durch eigens konstruierte Testaufgaben. Für diese Änderung gab es zwei Gründe. Der erste war eine einfache Zweckmäßigkeitsüberlegung. Wir wollten nämlich die Schulen, die mit uns zusammenarbeiteten, nicht um die Zeit für einen längeren Test bitten. Der zweite und gewichtigere Grund war, die Glaubwürdigkeit der Ergebnisse in den Augen der Adressaten zu sichern, deren Entscheidung, das Programm zu verwenden oder nicht zu verwenden, diese Untersuchung beeinflussen sollte. In dem hier vorliegenden Fall ist die Biological Sciences Curriculum Study der unmittelbare Adressat. Diese Organisation hat viel Zeit und Geld darauf verwendet, Leistungstests zu Curriculumeinheiten zu entwickeln, die neben anderen Themen auch die Populationsgenetik zum Gegenstand haben. Da das Unterrichtsprogramm dafür vorgesehen war, die gleichen Lernziele zu erreichen, die sich auch die anderen BSCS-Materialien auf diesem Gebiet gesetzt hatten, konnte kaum ein überzeugender Einwand dagegen vorgebracht werden, diese Testaufgaben nicht zu verwenden, von denen Biologen und Biologielehrer annahmen, daß sie die Schülerleistung, bezogen auf diese Lernziele, gültig messen. Kriteriumsbezogene Tests sind die einzigen sinnvollen Tests für eine Unterrichtsevaluation; aber in diesem Falle war es von großer Wichtigkeit, die normenbezogenen Testaufgaben der BSCS zu verwenden, um den Verdacht zu vermeiden, die Überlegenheit dieses Programms beruhe lediglich auf eigens zugeschnittenen Testaufgaben.

In dem Leistungstest, der bei unserer Untersuchung Verwendung fand, wurden 24 Testaufgaben der BSCS-Tests, die sich mit Populationsgenetik befassen, aufgenommen. Es sollte betont werden, daß ein Schwierigkeitsgrad von fast 50 % nach der Durchführung des Unterrichts eines der Kriterien war, nach denen die Testaufgaben in die BSCS-Tests aufgenommen wurden. Zusätzlich wurden 20 Testaufgaben nach dem Auswahl-Antwort-Verfahren konstruiert, um eine noch größere Differenzierung zu erreichen. Abbildung 1 zeigt den Leistungszuwachs für den Unterricht mit und ohne Programm in Abhängigkeit von der Herkunft der Testaufgaben [3].

Leistung als eine Funktion der Unterrichtsinhalte

Die Lernziele des Programms können in drei Hauptgebiete klassifiziert werden:

Abbildung 1
Leistungszuwachs als Funktion der Herkunft der Testaufgaben



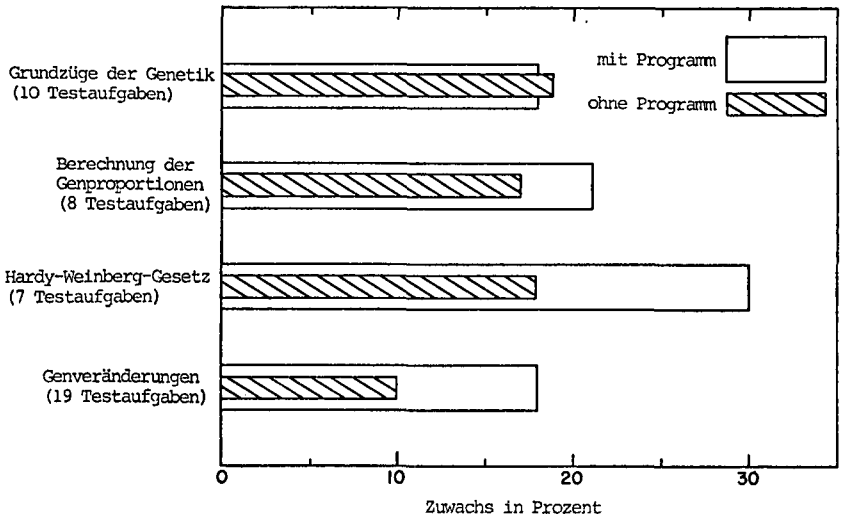
(1) die Berechnung der Genproportionen auf der Grundlage von ausgewählten Daten;
(2) die Logik des Hardy-Weinberg-Gesetzes;
(3) Faktoren, die eine Genveränderung bewirken (Mutation, Adaptation – Selektion, Wanderungssiebung, durch Zufall verursachter »genetic drift«, Paarungssiebung, Isolation).

Das Programm selbst mußte außerdem noch einen vierten Inhaltsbereich behandeln. Die Beherrschung der Mendelschen Gesetze ist für das Verstehen der Populationsgenetik von wesentlicher Bedeutung. Von dem Schüler wird angenommen, daß er die Grundzüge der Genetik beherrscht, bevor er mit dem Programm zu arbeiten beginnt. Da man sich auf die Zulänglichkeit des vorangegangenen Unterrichts nicht verlassen wollte, wurden zu Beginn des Programms die Grundzüge der Genetik durchgenommen. Abbildung 2 zeigt den Leistungszuwachs in den vier inhaltlichen Hauptbereichen.

Leistung als eine Funktion der Art der Testaufgaben

Eine der möglichen Schwächen in dem Verfahren, den Unterricht soweit zu verbessern, bis die Ergebnisse eines kriteriumsbezogenen Tests ein befriedigendes Niveau erreicht haben, ist die, daß dieses Verfahren zu einem einfachen Lehren auf den Test hin führen kann. Folgendes kann nämlich geschehen: Wenn eine Testaufgabe schlecht gelöst wird, so nehmen der Autor oder der Curriculumentwickler Sätze in den Unterricht auf, die die Antwort auf die Frage liefern. Oder er stellt vielleicht während des Un-

Abbildung 2
Leistungszuwachs als Funktion der Unterrichtsinhalte



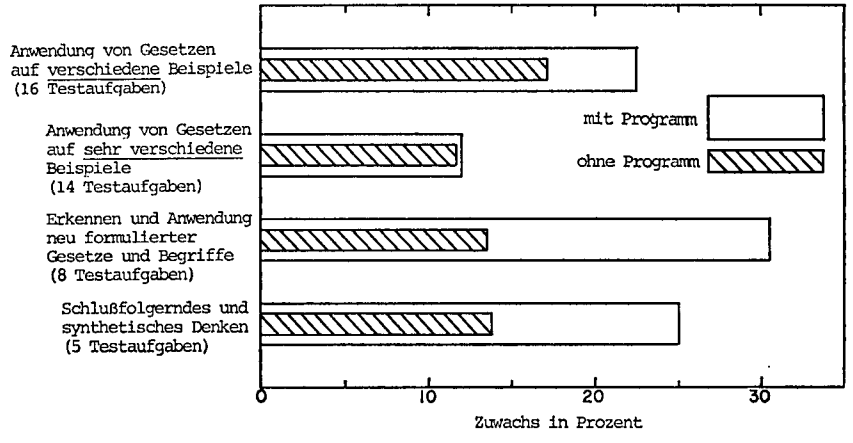
terrichtsverlaufs die Frage in einem Zusammenhang, in dem der Schüler die richtige Antwort finden muß. So muß sich auch bei einer schwierigen Testaufgabe das Ergebnis verbessern. Dessen ungeachtet, muß die Frage, was gelernt wurde, beantwortet werden. Es mag sehr wohl sein, daß die Schüler lediglich gelernt haben, eine Reihe von Wörtern zu wiederholen oder wiederzuerkennen. Definitionsgemäß versteht jemand einen Begriff oder ein Gesetz, wenn er alle möglichen Beispiele, die sich auf diesen Begriff oder auf dieses Gesetz beziehen, angemessen bearbeiten kann [4]. Wenn in einer Testaufgabe ein Beispiel verwendet wird, das während des Unterrichts gegeben wurde, kann dies lediglich ein verbales Wiederholen oder Wiedererkennen bedeuten. Wenn ein Schüler jedoch Testaufgaben richtig lösen kann, in denen Beispiele verwendet werden, die von denen *verschieden* sind, die im Unterricht gegeben wurden, ist die Folgerung durchaus angebracht, daß der Begriff von den Schülern verstanden wurde. Die Beispiele in den Testaufgaben können hinsichtlich ihrer Ähnlichkeit mit den Unterrichtsbeispielen skaliert werden. Kann jemand Fragen beantworten, die Beispiele enthalten, die sich nur wenig von den im Unterricht verwendeten unterscheiden, dann läßt sich sagen, daß er etwas von diesem Begriff oder Gesetz verstanden hat, während jemand, der Testaufgaben lösen kann, die im Vergleich zum Unterricht *sehr verschiedene* Beispiele enthalten, ein tiefes oder umfassendes Verständnis zeigt.

Begriffe können allgemein definiert werden; Gesetze können in abstrakter Sprache angegeben werden. Wenn ein Test im wesentlichen die Unterrichtssprache wiederholt, ist wiederum nur verbales Erkennen für eine richtige Antwort notwendig. Wenn ein Schüler jedoch angemessen mit Formulierungen eines Begriffs oder eines Gesetzes umgehen kann, die zwar wortmäßig verschieden sind, der Darstellung im Unterricht jedoch inhaltlich gleichen, deutet dies auf ein gewisses Verständnis.

Es ist ein Zeichen für synthetisches Denken, wenn ein Schüler eine Testaufgabe beantworten kann, deren Lösung die Anwendung von Begriffen und Gesetzen erfordert, die zu weit auseinanderliegenden Zeitpunkten im Unterricht behandelt wurden. Anderseits können diese Testaufgaben manchmal richtig gelöst werden, wenn der Schüler Schlüsse aus Aussagen zieht, die an einer Stelle während des Unterrichts gemacht wurden. Unter Verwendung der gerade beschriebenen Unterscheidungen wurde eine Inhaltsanalyse des Unterrichts und der Testaufgaben durchgeführt. Jede Testaufgabe wurde einer von fünf Kategorien zugeordnet. Zuordnungskriterien waren dabei die Ähnlichkeit der verwendeten Ausdrucksweise und die Ähnlichkeit der Aufgabenstellung zwischen Testaufgaben und den Aufgaben in den Programmen. Dabei wurde weder auf das Lehrbuch noch auf die Übungen noch auf den mündlichen Unterricht der Lehrer Rücksicht genommen. Ich muß darauf hinweisen, daß ich für die Verläßlichkeit der Zuordnung der Testaufgaben zu den einzelnen Kategorien nicht einstehen kann. Dies muß als ein grober, anfänglicher Versuch betrachtet werden, die Vorstellungen zu operationalisieren, die Pädagogen seit der Arbeit von Bloom und Mitarbeitern (1956) als bedeutsam ansehen (vgl. Anderson 1970 u. Anderson/Faust 1972). Abbildung 3 zeigt den Leistungszuwachs in der Versuchs- und Kontrollgruppe in Abhängigkeit von der Art der Testaufgaben. Da nach unserer Beurteilung nur eine Testaufgabe verbales Wiedererkennen maß, wurde diese Kategorie nicht in die Graphik aufgenommen.

Wie ich oben darlegte, können Testaufgaben Aufschluß geben über tiefes und umfassendes Verständnis, wenn sie Beispiele enthalten, die sehr verschieden von denen sind, die im Unterricht verwendet wurden. Wie man weiß, können die Anforderungen solcher Testaufgaben über die Lernziele eines bestimmten Curriculum hinausgehen. Während sie vermutlich in Leistungstests aufgenommen werden sollten, um Verständnisgrenzen feststellen zu können, ist Vorsicht bei der Beurteilung von ganzen Curriculummaterialien hinsichtlich ihrer Effektivität angebracht, sofern die Testaufgaben Beispiele enthalten, die von den Unterrichtsbeispielen sehr verschieden sind. Anders gesagt: Solche Testaufgaben erfassen eine weiterreichende Transferwirkung, die man nicht mit Sicherheit von einem Unterricht erwarten kann.

Abbildung 3
Leistungszuwachs als Funktion der Art der Testaufgaben



Leistung als eine Funktion des Lehrers

Es gab große Unterschiede darin, wie die Lehrer das Programm benutzten. Einige Lehrer billigten den Schülern überhaupt keine Unterrichtszeit zu, sich mit dem Programm zu befassen, während es auf der anderen Seite Lehrer gab, die die Populationsgenetik ausschließlich nach dem Programm lehrten. Tabelle 1 gibt die Anzahl der Minuten in den einzelnen Klassen wieder, die zwischen dem Vor- und Nachtest auf die verschiedenen Aktivitäten verwendet wurden. Diese Zahlen beruhen auf den Aufzeichnungen der Lehrer. Wir schlugen den Schulen einen zweiwöchigen Zeitraum zwischen Vor- und Nachtest vor. An einer Schule stimmten die Lehrer zu. Die Lehrer an der anderen Schule sagten: »Wir können dieses Curriculummaterial unmöglich in weniger als einem Monat durchnehmen«; sie erhielten deshalb einen Monat Zeit. Alle Lehrer in der Schule B berichteten, daß sie mit oder ohne Programm die gleiche Zeit für den Unterricht in Populationsgenetik aufgewendet hätten. Die Lehrer in der Schule A verwendeten bei der Benutzung des Programms für Populationsgenetik durchschnittlich etwa 10 % weniger Unterrichtszeit. Durchschnittlich gaben die Lehrer in den Klassen, in denen das Programm benutzt wurde, etwas weniger Seiten zu lesen auf als in Klassen, in denen das Programm nicht verwendet wurde.

Tabelle 1
Durchschnittliche Unterrichtszeit in Minuten (für die behandelten Themen) nach Schule und Art des Unterrichts

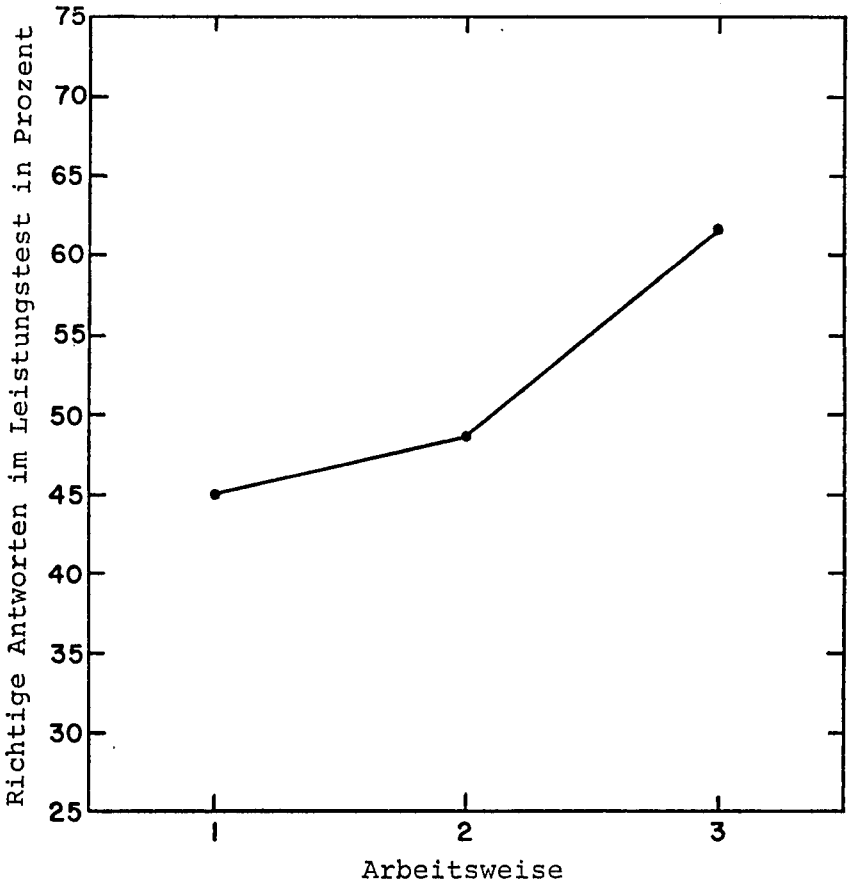
	Mit Programm		*Ohne Programm*	
	Schule A	Schule B	Schule A	Schule B
Durchschnittliche Unterrichtszeit zwischen Vor- und Nachtests	454	1031	454	1031
Zeit für Populationsgenetik mit Programm	151	0	0	0
andere Arbeit in Populationsgenetik . . .	52	451	228	451
insgesamt	203	451	228	451
Zeit für nicht populationsgenetisches Material	251	580	226	580

Die Lehrer wurden danach klassifiziert, wie sie das Programm den Schülern zuwiesen. Die erste Gruppe der Lehrer, wie in Abbildung 4 gezeigt wird, sorgte dafür, daß das Programm verfügbar war; sie forderten von den Schülern aber nicht, es durchzuarbeiten; auch war es nicht erlaubt, während der Unterrichtszeit damit zu arbeiten. Von der zweiten Gruppe wurde die Arbeit mit dem Programm verlangt, aber wiederum wurde keine Unterrichtszeit zur Verfügung gestellt, um damit zu arbeiten. Die Lehrer in der dritten Gruppe berichteten, daß sie das Programm zum festen Unterrichtsbestandteil gemacht hätten und für die Arbeit damit bis zu drei Unterrichtsstunden vorgesehen hätten. Die Ergebnisse zeigen jedoch, daß durchschnittlich etwa vier Stunden erforderlich sind, um das Programm durchzuarbeiten. Weniger als 20 % der Schüler berichteten, sie hätten das Programm in drei oder weniger Stunden bewältigt. Deshalb bearbeiteten die meisten Schüler, sofern sie es überhaupt taten, das Programm nicht in der Klasse.

Von großer Bedeutung sind die Durchschnittsergebnisse und deren Streuung. Ein F-Test ergab eine signifikant geringere Varianz im Leistungszuwachs bei den Lehrern von Klassen, die das Programm erhalten hatten, im Vergleich zu Lehrern von Klassen, bei denen dies nicht der Fall war. [$F(7,7) = 6{,}68$; $p < 0{,}05$, zweiendiger Test]. Überdies erreichte, wie man aus Abbildung 5 entnehmen kann, *jeder* Lehrer mit dem Programm mehr als ohne Programm [5].

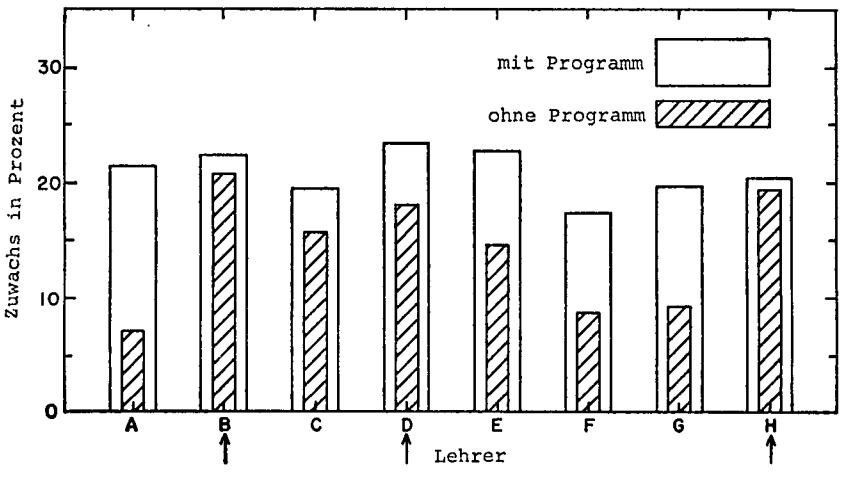
Die Lehrer wurden gefragt, ob das Programm ihren Unterricht ohne Programm beeinflußt hätte. Drei Lehrer (in Abbildung 5 mit einem Pfeil

Abbildung 4
Leistung als Funktion der Arbeitsbedingungen



markiert) gaben eine zustimmende Antwort. Lehrer D sagte: »Die Gliederung, die die Lehrer ihrem Unterricht zugrunde legten, und die Darstellung der Probleme in dem Programm wurden auch bei dem Unterricht in den Klassen verwendet, die das Programm nicht erhalten hatten.« Lehrer H äußerte sich folgendermaßen dazu: »Ich kann sagen, daß mir das Programm für alle meine Klassen geholfen hat, einen besseren Unterricht in Populationsgenetik zu geben. Ich verwertete viele Teile des Programms und fand, daß sie einen leichteren Zugang zu einem Thema ermöglichten, das andernfalls für viele Schüler schwierig gewesen wäre.«

Abbildung 5
Leistungszuwachs der Schüler bei den acht Lehrern



Leistung als eine Funktion der Schüler

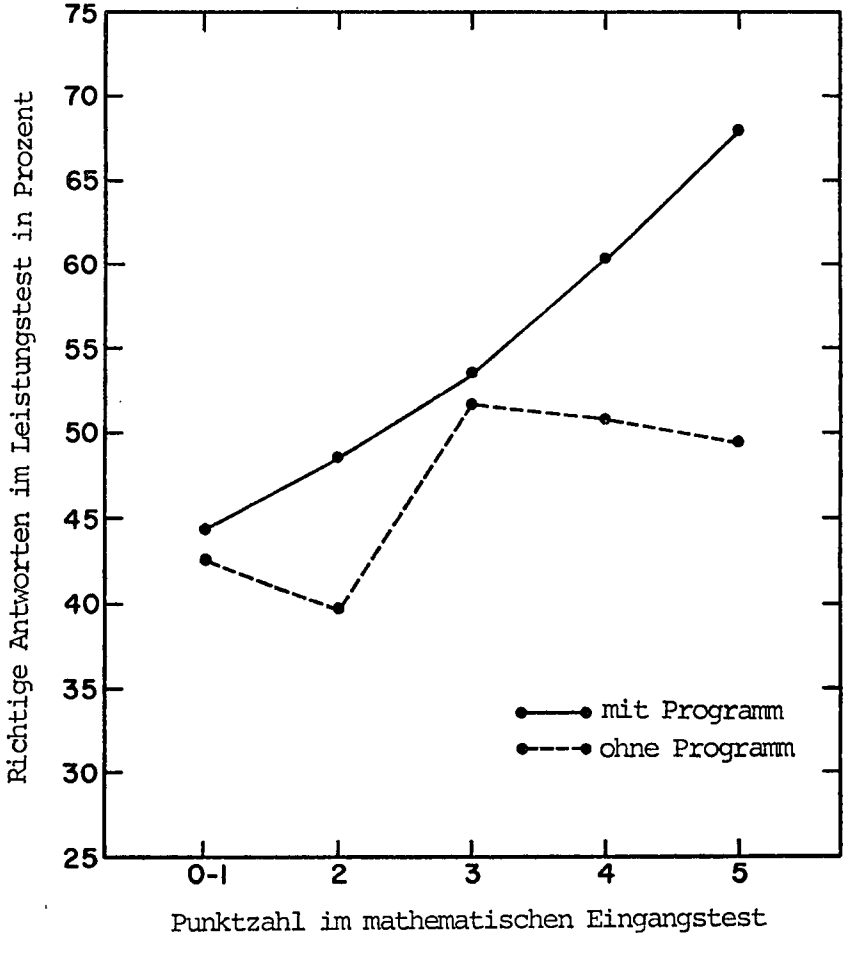
Jeder, der mit Unterrichtsplanung betraut ist, muß Voraussetzungen über den vorhandenen Kenntnisstand der Schüler machen, für die der Unterricht bestimmt ist. Das Programm in Populationsgenetik basiert auf der Voraussetzung, daß die Schüler, die damit arbeiten, mit Verhältniszahlen rechnen können und fähig sind, ein Binom zu quadrieren.

Nach Meinung vieler Pädagogen haben Programme zum Selbstunterricht bestenfalls den Wert, langsamen Schülern technisches Vokabular beizubringen. Eines der ursprünglichen Ziele dieses Projekts war es zu zeigen, daß Programme effizient benutzt werden können, um den besten Schülern eine Reihe von Gesetzen mit den dazugehörenden Begriffen zu vermitteln. Die Vermutung lag nahe, daß fast alle guten Schüler die erforderlichen mathematischen Fertigkeiten besaßen. Später wurde jedoch festgestellt, daß sehr viele gute Schüler, die die Hälfte der Stichprobe in der vergleichenden Untersuchung ausmachen sollten, zu der Zeit nicht erreichbar waren, zu der die Untersuchung durchgeführt werden sollte.

Mit den Schülern, die an der Untersuchung teilnahmen, wurde ein Eingangstest durchgeführt, der fünf Aufgaben enthielt. Zu unserer Bestürzung entdeckten wir, daß nur 40 % der Schüler in der Stichprobe die erforderlichen mathematischen Fertigkeiten besaßen, d. h. nur 40 % der Schüler beantworteten mindestens vier der Testaufgaben richtig. Abbil-

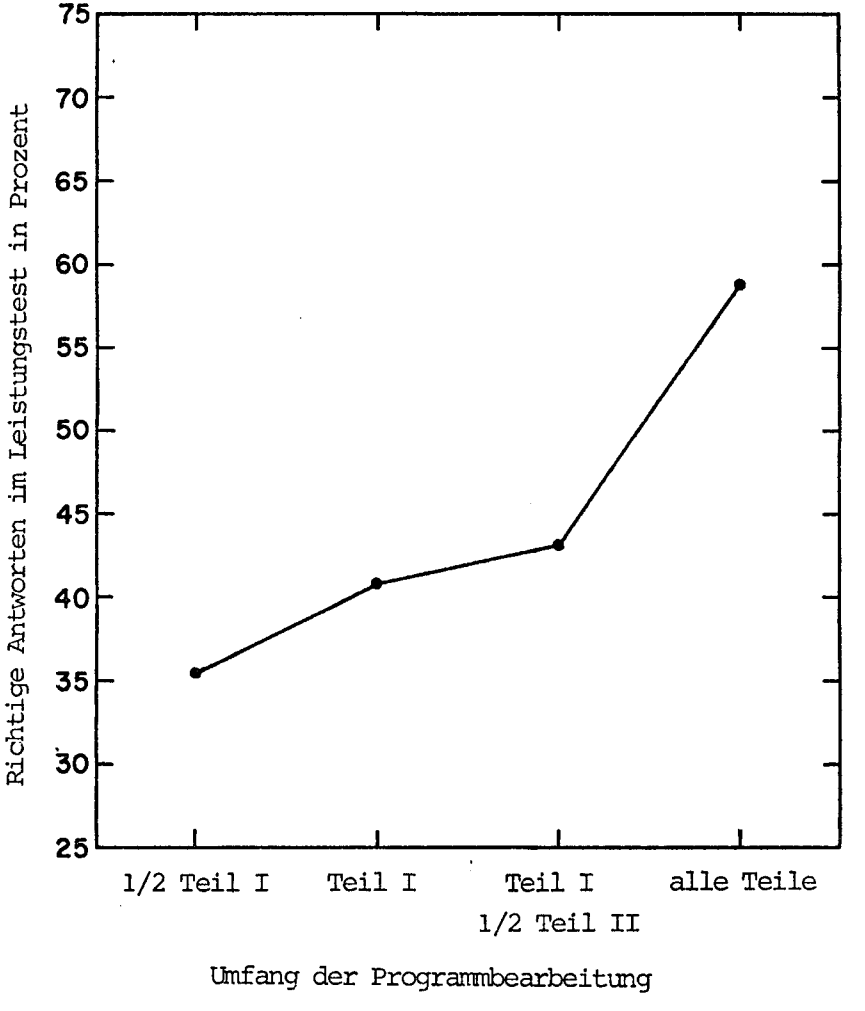
dung 6 zeigt die Leistung im Nachtest als eine Funktion der im Eingangstest festgestellten mathematischen Fertigkeiten. Das Programm war stets etwas effektiver, aber der Vorteil des Programms wirkte sich erheblich nur bei jenen Schülern aus, die in dem Eingangstest gute mathematische Fertigkeiten gezeigt hatten. Man konnte einen geringeren Leistungszuwachs bei den Schülern feststellen, die das Programm nicht erhielten, sogar bei denen, die die erforderlichen mathematischen Fertigkeiten besaßen.

Abbildung 6
Leistung als Funktion des mathematischen Eingangstests



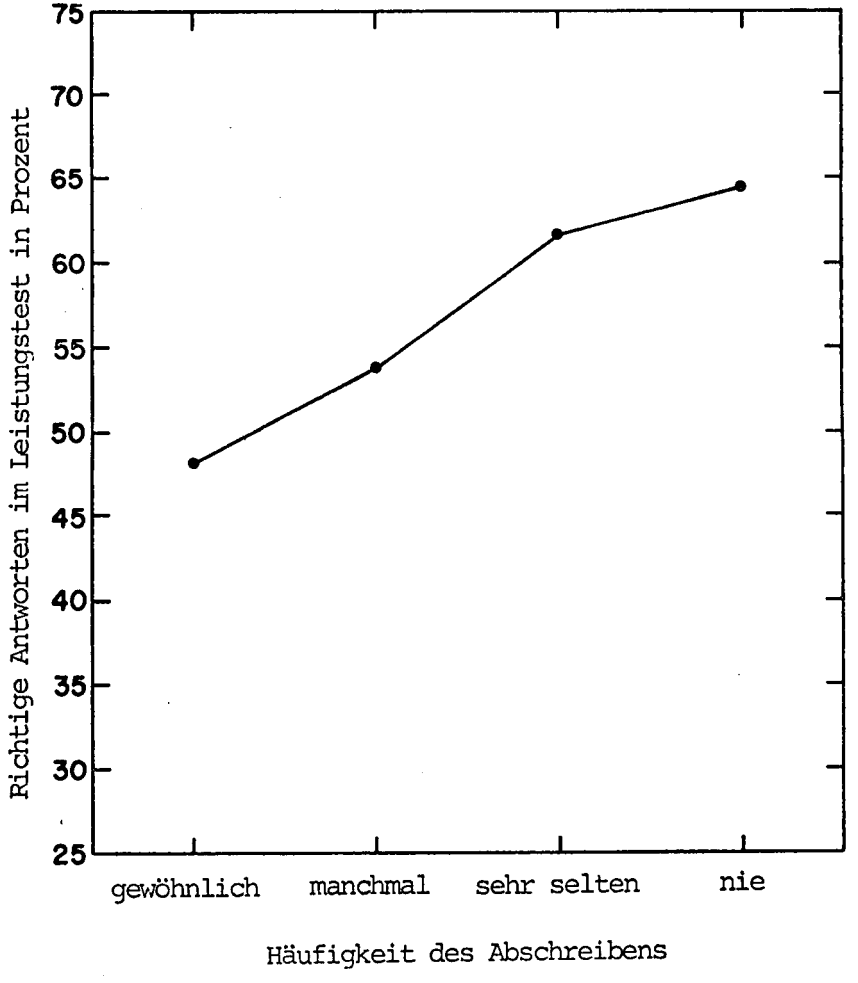
Im Fragebogen wurden die Schüler gebeten anzugeben, wie weit sie das Programm tatsächlich durchgearbeitet hatten. Etwa 75 % gaben an, das ganze Programm durchgearbeitet zu haben. Natürlich war die Leistung um so besser, je weiter das Programm durchgearbeitet worden war. Dieser Zusammenhang wird in Abbildung 7 gezeigt.

Abbildung 7
Leistung als Funktion des Umfangs der Programmbearbeitung



Wir wissen genau, daß Schüler von Programmen nicht viel lernen, wenn sie einfach die richtigen Antworten abschreiben (Faust/Anderson/Guthrie/Drantz 1967; Anderson/Faust 1968; Brown 1966; Kemp/Holland 1966). Die Schüler wurden danach gefragt, wie oft sie bei einer für sie schwierigen Frage die Seite umgedreht und die richtige Antwort abge-

Abbildung 8
Leistung als Funktion der berichteten Häufigkeit des Abschreibens richtiger Antworten bei schwierigen Abschnitten



schrieben hätten. Obwohl die Schüler in den Anweisungen zur Bearbeitung des Programms ermahnt wurden, jede Frage, *bevor* sie nach der richtigen Antwort schauten, schriftlich zu beantworten, gaben mehr als 40 % an, manchmal bei schwierigen Fragen Antworten abgeschrieben zu haben; 20 % gaben an, daß sie dies im allgemeinen taten. Abbildung 8 zeigt die Leistung in Abhängigkeit von der angegebenen Häufigkeit des unerlaubten Abschreibens richtiger Antworten.

Tabelle 2
Evaluation des Programms durch die Schüler
(N = 377)

Schüler in Prozent	Fragen
	1. Wenn ich die Wahl hätte,
71,6	(A) würde ich gerne öfter Programme benutzen, die dem Programm Populationsgenetik ähnlich sind
12,2	(B) wäre es mir gleich, welche Materialien benutzt würden
14,9	(C) würde ich es bevorzugen, keine Programme zu benutzen
1,3	keine Antwort
	2. Bei einem Vergleich eines Programms gleich dem in Populationsgenetik mit einem Lehrbuch meine ich, daß ich mit dem gleichen Aufwand an Zeit und Mühe
36,3	(A) mit dem Programm sehr viel mehr lernen würde
42,2	(B) mit dem Programm etwas mehr lernen würde
8,0	(C) gleich viel lernen würde
10,3	(D) mit dem Lehrbuch etwas mehr lernen würde
3,2	(E) mit dem Lehrbuch viel mehr lernen würde
	3. Wie sehr interessierte Dich das Programm in Populationsgenetik?
22,0	(A) Ich war sehr interessiert daran
45,9	(B) Ich war einigermaßen interessiert daran
22,3	(C) Ich verlor manchmal das Interesse
8,5	(D) Ich langweilte mich sehr
1,3	Keine Antwort
	4. Inwieweit verlangte das Programm in Populationsgenetik sorgfältiges Denken?
27,9	(A) Viele Seiten erforderten sorgfältiges Denken zur richtigen Beantwortung der Fragen
63,1	(B) Einige Seiten erforderten sorgfältiges Nachdenken
5,3	(C) Wenig Nachdenken erforderlich
1,9	(D) Das Programm war lächerlich einfach und verlangte fast kein Nachdenken
1,9	Keine Antwort

Die Einstellung der Lehrer und Schüler

Alle Lehrer empfanden das Programm als eine wertvolle Ergänzung des vorhandenen BSCS-Curriculummaterials über Populationsgenetik. Die Frage, ob sie das Programm wieder einsetzen würden, bejahten fünf von neun Lehrern; zwei Lehrer antworteten, daß sie es wahrscheinlich wieder benutzen würden; einer antwortete mit »wahrscheinlich nein«. Von einem Lehrer wurde diese Frage nicht beantwortet. Die Lehrer waren von dem Inhalt und der Organisation des Programms angetan; sie waren auch zufrieden mit dem Interesse, das das Programm bei den Schülern hervorrief. Zwei Lehrer gaben unaufgefordert die Auskunft, daß das Programm einen so guten Ruf hatte, daß sich einige Schüler in Klassen, die ohne das Programm unterrichtet wurden, Exemplare des Programms von ihren Schulkameraden entliehen.

Tabelle 2 faßt die Antworten der Schüler auf vier Fragen zusammen. Die meisten Schüler äußerten sich dahingehend, daß sie gerne wieder ein Programm wie das populationsgenetische benutzen würden, vorausgesetzt, daß sie mit diesem Programm mehr als mit einem Lehrbuch lernen und daß dieses Programm sie interessiert und sorgfältiges Denken verlangt.

Zusammenfassende Erörterung

Ein Ziel dieses Beitrags bestand darin, die Gültigkeit eines gesamten Curriculum nachzuweisen. Es galt zu zeigen, daß das neue Curriculummaterial, das in diesem Falle ein Programm zum Selbstunterricht in Populationsgenetik beinhaltete, effektiver als ein weit verbreitetes und sehr anerkanntes vergleichbares Curriculum ist.

Die Unterrichtseffektivität sollte sowohl in absoluten als auch in relativen Normen beurteilt werden. Der sich aus dem Test ergebende Durchschnitt der Gesamtleistung der Schüler, die das Programm erhalten hatten, betrug 53,6 % – kaum ein befriedigendes Ergebnis. (Der Durchschnitt für die Schüler, die das Programm nicht erhalten hatten, betrug 43,5 %).

Unter sehr günstigen Bedingungen jedoch führt das Programm zu besseren Leistungen. Alle Schüler, die den Eingangstest in Mathematik bestanden hatten und berichteten, sie hätten das Programm vollständig durchgearbeitet und vor der Beantwortung einer Frage nie oder selten nach der richtigen Antwort geschaut, erzielten einen Durchschnittswert von 70,5 %. Es ist wahrscheinlich, daß der Gesamtleistungsdurchschnitt höher als der in dieser Untersuchung festgestellte sein würde, wenn allen Schülern im Unterricht genügend Zeit gegeben würde, das Programm vollstän-

dig durchzuarbeiten; gleiches gilt für den Fall, daß das Durcharbeiten des Programms vom Lehrer gefordert, anstatt nur in das Belieben der Schüler gestellt wird; oder wenn die Lehrer ihren Forderungen mit den ihnen zur Verfügung stehenden Maßnahmen und Mitteln zur Lernmotivierung Nachdruck verleihen; oder wenn die Schüler dazu gebracht werden können, eine Antwort zu jeder Frage zu formulieren, bevor sie die richtigen Antworten nachschlagen; und ebenso gilt dies natürlich, wenn das Programm nur beim Unterricht mit den Schülern verwendet wird, die die erforderlichen mathematischen Fertigkeiten besitzen.

Zugegebenermaßen ist ein Leistungsniveau von 70 % (der maximal erreichbaren Punktzahl) unter optimalen Bedingungen kein überwältigendes Ergebnis [6]. Bei der Bewertung der erzielten Leistung ist jedoch zu berücksichtigen, daß nicht weniger als 25 % der Aufgaben in dem kriteriumsbezogenen Leistungstest über die Lernziele des Programms hinausgehen und daß fast 20 % der Aufgaben ein Thema betreffen (Grundzüge der Genetik), das zwar in dem Programm berücksichtigt, aber nicht explizit gelehrt wurde. Wenn man dies alles bedenkt, so ist das mit diesem Programm erzielte Leistungsniveau nicht schlecht, gleichgültig, ob man es relativ im Hinblick auf den Vergleichsunterricht oder in absoluten Maßstäben betrachtet.

Das Programm wird gegenwärtig auf der Grundlage der in der Felduntersuchung erzielten Ergebnisse und auf der Grundlage der Kritik der Genetiker und Biologielehrer überarbeitet.

Das Ziel dieses Beitrags war es, den Wert und die Bedeutung der vergleichenden Felduntersuchung nachzuweisen. Eine angebrachte Zurückhaltung bei der Erörterung des gegenwärtigen Wissensstands der Erziehungs- und Verhaltenswissenschaft, eine angemessene Beachtung der Komplexität des menschlichen Lernens und des Unterrichts und eine realistische Einschätzung der Möglichkeiten der Grundlagenforschung, unsere Fähigkeiten zu verbessern, eine effektive Unterrichtsgestaltung im vorhinein bestimmen zu können, lassen die Anwendung einer praxisbezogenen Strategie für die Entwicklung von Curriculummaterial als sinnvoll erscheinen. Der letzte Schritt in diesem Entwicklungsprozeß sollte eine Felduntersuchung sein, um empirisch die Effektivität des gesamten neuen Curriculummaterials zu beweisen. Es gibt keinen anderen Weg, um Effektivität gewährleisten zu können. Die Hauptaufgabe einer Felduntersuchung ist es, Ergebnisse zu liefern, aufgrund derer die Adressaten eine Entscheidung über die Annahme oder Ablehnung von Curricula fällen können. Wenn zwei Curricula die gleichen Lernziele haben (oder die gleichen Themen behandeln), sollte die Felduntersuchung aus einem Vergleich bestehen. Es genügt nicht zu zeigen, daß ein neues Curriculum die von irgendjemand gesetzten absoluten Effektivitätsnormen erfüllt, weil konkurrie-

rende Curricula diese Normen übertreffen oder die gleichen Normen mit weniger Zeitaufwand oder mit geringeren Kosten erfüllen können oder weil sie von Schülern und Lehrern vorgezogen werden.

Man hat oft gefordert, Unterricht empirisch zu validieren. Zum gegenwärtigen Zeitpunkt gibt es nur wenig Anzeichen, daß jemand hiervon überzeugt worden ist. Mein letztes Wort richtet sich an Autoren, Herausgeber und Verleger, die sagen, sie hätten besseres Curriculummaterial entwickelt. Warum sollte man ihnen glauben? Wo ist der Beweis für ihre Behauptungen? Das Erziehungswesen würde einen sehr großen Schritt vorankommen, wenn die Produzenten von Curriculummaterial es sich zur Regel machten, ihre Produkte empirisch zu validieren, und wenn es üblich wäre, daß die Adressaten eine solche Validierung als Voraussetzung für die Verwendung des Curriculummaterials verlangen würden.

William W. Cooley

Methoden der Evaluation von Schulinnovationen

Es gibt viele Aufsätze über Evaluationsmodelle, Evaluationsstrategien und Handlungsrezepte. Es gibt auch zahlreiche Versuche, Evaluationstaxonomien zu entwickeln. Aber es fehlen gut zugängliche Veröffentlichungen, die über die Verfahren und Ergebnisse wirklicher Evaluationsuntersuchungen berichten. Entweder werden die Ergebnisse niemals gedruckt, oder sie haben, wenn sie gedruckt sind, das Format großer Telefonbücher, die nur in geringer Zahl aufgelegt werden können und die in der Regel als Wanddekorationen im Erziehungsministerium enden.

Solange diese Berichte nicht allgemein zugänglich gemacht und von anderen Forschern kritisch untersucht werden können, lassen sich Evaluationsuntersuchungen kaum verbessern. Bei den Vorarbeiten für diesen Beitrag bin ich daher zur Überzeugung gekommen, daß man *nicht* einen weiteren Aufsatz *über* Evaluation, sondern die Beschreibung *einer* Evaluation braucht. Aus ihr müßte hervorgehen, wie ein Forscher sich bemüht, Daten zu erheben, aus denen sich eindeutige Informationen über den Wert neuer Unterrichtsmaterialien und pädagogischer Verfahren gewinnen lassen.

Über Evaluation läßt sich lediglich sagen, daß die Evaluation von Schulinnovationen insofern gute Forschung sein muß, als Forschung der Prozeß ist, in dessen Verlauf die Gültigkeit einer Hypothese bewiesen werden muß. Nach meiner Überzeugung unterscheidet sich evaluative Forschung von Grundlagenforschung und von weiten Bereichen angewandter Forschung nur in der Art der Hypothesen und darin, wie diese zu Beginn der Untersuchung formuliert werden. In der Grundlagenforschung beruhen die Hypothesen, die untersucht werden sollen, auf einer Theorie und einem entsprechenden System aufeinander bezogener Gedankengänge. In der angewandten Forschung stammen die Hypothesen, die untersucht werden sollen, aus der Anwendung der Wissenschaft und werden formuliert, wenn die abgesicherten Prinzipien, die diese Wissenschaft hervorgebracht hat, sich bei einer bestimmten Anwendung als inadäquat erweisen. Evaluative Forschung als eine Form der angewandten Forschung versucht, eher die

Gültigkeit der Hypothesen hinsichtlich *besonderer* Programme und Verfahren als die Gültigkeit der Hypothesen hinsichtlich allgemeiner, in vielen Programmen gleicher Variablen einzuschätzen. Den Bezugsrahmen für meine Ausführungen bildet das Learning Research and Development Center (LRDC) an der Universität von Pittsburgh und die Unterrichtsmaterialien und -verfahren, die hier entwickelt worden sind. Daher befaßt sich die hier beschriebene evaluative Forschung mit spezifischen Bildungsprogrammen, die Unterricht an individuelle Unterschiede anzupassen versuchen. Das Ziel der Forschung besteht darin, Informationen in bezug auf die Gültigkeit der Hypothesen über die pädagogischen Programme des Learning Research and Development Center zu gewinnen und verfügbar zu machen. Die Hypothesen und die Daten über ihre Gültigkeit sollen über den Nutzen der neuen Programme informieren und den Programmentwicklern Informationen über die relativen Stärken und Schwächen der Programmkomponenten geben.

Es gibt vier Institutionen, in denen man die Ergebnisse der Bemühungen des Learning Research and Development Center untersuchen kann. Am bekanntesten ist wahrscheinlich die Oakleaf-Schule, eine kleine Grundschule in einem Vorort von Pittsburgh, in der vor sieben Jahren der »individualisierte Unterricht« (Individually Prescribed Instruction, IPI) eingeführt wurde (Lindvall und Bolvin, 1967). Eine zweite Institution ist das System der Versuchsschulen, das das Regional Laboratory »Research for Better Schools« (RBS), in Philadelphia, aufgebaut hat. Das Institut »Research for Better Schools« disseminiert seit 1966 die Produkte des Learning Research and Development Center, die in der Oakleaf-Schule entwickelt worden sind. Eine dritte Institution ist die Frick-Schule, eine große Stadtschule im Zentrum von Pittsburgh, in der das Learning Research and Development Center in den letzten vier Jahren Programme entwickelt hat. Nachdem die Programme in der Frick-Schule entwickelt und getestet worden waren, wurden sie, viertens, in weiteren Schulen benutzt, die in einem Netzwerk zusammengefaßt sind. Im vergangenen Jahr entschieden sich vier Schulsysteme für die Programme, die wir in der Frick-Schule und in den Grundschulen der Schulsysteme entwickelt hatten, und implementierten sie. Das Learning Research and Development Center arbeitet mit diesen Schulen zusammen, so daß es auch den Prozeß der Dissemination pädagogischer Innovationen untersuchen kann. Lindvall und Cox (1970) und das Institut »Research for Better Schools« veröffentlichten Evaluationsuntersuchungen, die in der Oakleaf-Schule bzw. in den vom Institut »Research for Better Schools« betreuten Schulen gemacht worden waren. Ich werde meine Ausführungen daher auf die Evaluation, die in der Frick-Schule und den Schulen, die in dem Netzwerk zusammengefaßt sind, beschränken.

In der Frick-Schule entwickelte das Learning Research and Development Center ein individualisiertes Programm. Es bestand (1) aus einem Unterrichtsplan für jeden Schüler, der auf Grund der Ergebnisse in individuell eingesetzten Kriteriumstests entwickelt wurde; es enthielt (2) Hinweise und Vorschriften für die tägliche Anwendung des individuellen Unterrichtsplans. Es umfaßte (3) eine Neubestimmung der Lehrerrolle, bei der das Testen, die Tutorenarbeit und die Beweglichkeit des Lehrers besonders wichtig waren. Als Ergebnis sollte ein strukturiertes Curriculum in den grundlegenden Wahrnehmungs-, Lese- und Rechenfertigkeiten entstehen; es wurde durch ein wenig strukturiertes Curriculum ergänzt, in dem das Kind im bildnerischen und sprachlichen Gestalten, im sozio-dramatischen Spiel, in den Naturwissenschaften und in der Sozialkunde selbständig offene Lernaktivitäten wählen konnte.

Das Programm der Frick-Schule begann 1968/69 in Vorschulen und Kindergärten, wurde 1969/70 durch die erste Klasse, im vergangenen Jahr durch die zweite Klasse ergänzt und wird im nächsten Schuljahr von der Vorschule bis zur dritten Klasse reichen. Das Netzwerk begann 1969/70 mit drei Schulsystemen, zu denen im vergangenen Jahr ein viertes hinzukam, und das bis zum Herbst 1971 auf sieben Schulsysteme anwachsen soll. Wir beabsichtigen, das Netzwerk auf diese *sieben* Systeme zu beschränken, das für die Untersuchung des Disseminationsprozesses und die Evaluation unserer Curricula groß genug ist, ohne jedoch so groß zu sein, daß es als ein System, in dem Forschungs- und Entwicklungsarbeit geleistet werden soll, nicht mehr funktionsfähig ist. Die meisten Evaluationsuntersuchungen, die Daten über Schüler berücksichtigten, versuchten nachzuweisen, daß das Innovationsprojekt besser als ein vergleichbares anderes Programm ist. Welches Projekt als besser bezeichnet werden konnte, wurde auf Grund standardisierter Leistungsmessungen oder einer Reihe anderer Messungen bestimmt. Dazu wurden einige Kontrollschulen oder -klassen gebildet und dann die Mittelwerte verglichen. Wenn sich keine Unterschiede ergaben, waren die Ergebnisse nach Auffassung der Innovatoren nicht valide, und man bemühte sich weiterhin zu zeigen, inwiefern die Innovationen den bisherigen Bildungsprogrammen überlegen waren. Wenn sich aus den Ergebnissen des Vergleichs ergab, daß die Innovation einem anderen Programm überlegen war, waren die Innovatoren mit ihrer Arbeit und dem Evaluator zufrieden. Diejenigen jedoch, die der Innovation skeptisch gegenüberstanden, fanden irgendwelche Fehler im Innovations- und Evaluationsplan und bezweifelten die Gültigkeit der Ergebnisse.

Um diesen Punkt zu veranschaulichen, möchte ich einige Ergebnisse aus der Frick-Schule schildern. Um das Programm des Learning Research and

Development Center mit den bisherigen Schulprogrammen zu vergleichen, wurden in der Frick-Schule Kontrollgruppen eingerichtet, wobei uns die jährliche Erweiterung unseres Versuchs um ein neues Schuljahr zugute kam. Tabelle 1 veranschaulicht den allgemeinen Versuchsplan, der von

Tabelle 1
Versuchs- (E) und Kontroll- (K) Gruppen für die Frick-Schule

Jahr	Vor-schule	Kinder-garten	*Klasse* Erste	Zweite	Dritte	Vierte	Fünfte
1968–69	E	E	K	K	–	–	–
1969–70	E	E	[E]**	[K]*	K	–	–
1970–71	E	E	[E]	[E]	K	K	–
1971–72	E	E	E	E	E	K	K

* Gegensatz ist in Tab. 2 dargestellt
** Gegensatz ist in Tab. 3 dargestellt

Wang, Resnick und Schuetz (1970) entwickelt worden ist. Um Kontrollgruppen zu haben, untersuchten wir Klassen, die dem Programm um zwei Jahre voraus waren, während es selbst sich jährlich um ein Schuljahr erweiterte. Es konnten von einem Jahr zum anderen keine signifikanten Leistungsunterschiede zwischen den Evaluationsergebnissen eines bestimmten Schuljahres festgestellt werden. Es wurden auch bei den Variablen keine Unterschiede gefunden, die nach unserer Kenntnis Einfluß auf die Leistungen haben, ohne jedoch durch das Programm beeinflußbar zu sein wie etwa der sozioökonomische Status der Familie. Deshalb kann man zu Recht annehmen, daß in jedem Jahr die Kinder eines bestimmten Schuljahres Zufallsstichproben einer Grundgesamtheit waren.

Die in Tabelle 2 dargestellten Ergebnisse zeigen, daß das neue Programm statistisch signifikante Verbesserungen in allen drei Leistungsbereichen erbrachte, die in der zweiten Klasse mit dem Wide Range Achievement Test (WRAT) (Jastak, Bijou und Jastak 1965) gemessen wurden. Die Rechtschreibleistungen waren für unser Leseprogramm besonders interessant, weil wir die Rechtschreibung nicht direkt zu lehren versuchten, sondern sie als ein Nebenprodukt des Lesenlernens erwarteten.

Die Informationen, die auf Grund der Testnormierung gewonnen wurden, halfen uns, eine Vorstellung davon zu gewinnen, wieviel Zeitgewinn der Leistungszuwachs bedeutete. Die Unterschiede zeigten eine Verbesserung der Leseleistung um sieben, der Rechtschreib- und Rechenleistung um vier Monate an.

Tabelle 2
Vergleich im zweiten Schuljahr vor und nach dem LRDC-Programm
(Wide Range Achievement Test)

	»Vor« (Frühjahr 1970) (N = 98)	»Nach« (Herbst 1971) (N = 116)
Lesen		
Mittelwert (Rohwert)	41.45	49.91
Standardabweichung (Rohwert)	9.69	13.80
entsprechender Schuljahrswert *	2;2	2;9
	$F = 25.96$; df = 1 und 212; $p < .001$	
Rechtschreibung		
Mittelwert (Rohwert)	26.20	28.72
Standardabweichung (Rohwert)	5.08	5.44
entsprechender Schuljahrswert	1;9	2;3
	$F = 8.51$; df = 1 und 212; $p < .001$	
Rechnen		
Mittelwert (Rohwert)	23.40	25.22
Standardabweichung (Rohwert)	2.85	3.42
entsprechender Schuljahrswert	2;2	2;6
	$F = 17.62$; df = 1 und 212; $p < .001$	

* Der Wert 2;2 z. B. bedeutet: Die Leistung entspricht der Durchschnittsleistung nach zwei Monaten im zweiten Schuljahr.

Die Ergebnisse in Tabelle 3 verdeutlichen die Wirkung der Veränderungen zwischen der ersten und zweiten Version unseres Programms für die erste Klasse. Die Evaluation des Programms bei den ersten Klassen der Frick-Schule, die 1969/70 erfolgte, führte zu den Modifikationen, die im Herbst 1970 vorgenommen wurden. Die Gegenüberstellung der Ergebnisse des ersten und des zweiten Jahres gibt uns nützliche Informationen für die Kontrolle der Programmentwicklung. Programmveränderungen können so lange nicht als Verbesserungen dargestellt werden, als ihre Auswirkungen nicht bekannt sind. Die hier erzielten signifikanten Verbesserungen bestärkten die Programmkonstrukteure in ihrer Überzeugung, daß sie auf dem richtigen Weg waren. Außer dem Leistungszuwachs von einem Versuchsjahr zum anderen, erreichen die Schüler der ersten Klasse jetzt genauso gute Leistungen wie die Schüler der zweiten Klasse vor Beginn des Programms (vgl. hierzu die Mittelwerte der zweiten Spalte von Tab. 3 mit den Mittelwerten der ersten Spalte von Tab. 1).

Für den Programmentwickler sind diese Ergebnisse ohne Zweifel ermu-

Tabelle 3
Schulleistungen im ersten Schuljahr nach Veränderungen im LRDC-Programm (Wide Range Achievement Test)

	Nach dem 1. Jahr (Frühjahr 1970) (N = 143)	Nach dem 2. Jahr (Frühjahr 1971) (N = 124)
Lesen		
Mittelwert (Rohwert)	34.27	41.37
Standardabweichung (Rohwert)	10.32	11.85
entsprechender Schuljahrswert	1;7	2;2
	$F = 27.41$; $df = 1$ und 265; $p < .001$	
Rechtschreibung		
Mittelwert (Rohwert)	20.64	25.53
Standardabweichung (Rohwert)	4.65	5.77
entsprechender Schuljahrswert	1;3	1;7
	$F = 58.89$; $df = 1$ und 265; $p < .001$	
Rechnen		
Mittelwert (Rohwert)	22.36	23.98
Standardabweichung (Rohwert)	3.24	2.58
entsprechender Schuljahrswert	2;1	2;4
	$F = 20.03$; $df = 1$ und 265; $p < .001$	

tigend. Doch können sie auch anderen, nicht an dem Programm beteiligten Personen helfen, den Wert unseres neuen Programms zu beurteilen? Innovationen führen nicht immer zu einer Verbesserung der Mittelwerte, obgleich man nur selten negative Ergebnisse in der Literatur findet. Können nun diese Ergebnisse jemanden davon überzeugen, daß dieses Programm in die Grundschule seiner Gemeinde gehört? Sicherlich nicht.

Viele Unzulänglichkeiten solcher Ergebnisse werden sofort deutlich:

1. Da die Ergebnisse nur aus einer Versuchsschule stammen, geben sie keine Auskunft darüber, wie das Programm sich in anderen Schulen bewähren würde.
2. Die Beschränkung des Leistungsvergleichs auf die Ergebnisse eines Leistungstests verringert bei skeptischen Adressaten ihre Aussagekraft.
3. Ein statistischer Beweis allein hat niemals jemanden von irgend etwas überzeugt.

Der Innovator hat die Aufgabe, nachzuweisen, wie gut das neue Programm sich bewährt. Es gibt keine sicheren Verfahren, jemanden von etwas zu überzeugen, und auch statistische Ergebnisse besitzen keinen sicheren Überzeugungswert. Die Auseinandersetzung um die Schädlichkeit des

Zigarettenrauchens ist dafür ein klassisches Beispiel. Die mit statistischen Verfahren ermittelte Tendenz, eine Verbindung zwischen dem Zigarettenrauchen und Krebs herzustellen, war seit langem vorhanden und bekannt. Solange man jedoch nicht zeigen konnte, *wie* Zigarettenrauchen Krebs erzeugt, haben nur wenige diese Ergebnisse ernst genommen. Dennoch war die anfängliche Tendenz wichtig, weil sie die entsprechende Forschung anregte.

Um die Unzulänglichkeit zu überwinden, die sich aus der Beschränkung der Evaluation auf eine Versuchsschule ergibt, können wir unser Netzwerk in die Evaluation miteinbeziehen. Wenn die neuen Programme aus der Versuchsschule auf andere Schulen übertragen werden, entstehen jedoch auch neue Probleme. Wie können wir Gewißheit erhalten, daß unser Modell wirklich im Unterricht realisiert wird? Sobald ein Lehrer mit den neuen Verfahren vertraut gemacht worden ist und die neuen Materialien in seiner Klasse sind, macht er seinen Unterricht, ohne daß man weiß, inwieweit er wirklich dabei nach den Intentionen des neuen Programms handelt. Man braucht Methoden, um festzustellen, in welchem Ausmaß das Unterrichtsmodell in jeder Klasse implementiert wird, und um die Daten über das Ausmaß der Implementation mit den Schülerleistungen in jeder Klasse in Verbindung zu setzen. Wählt man die Klasse als Analyse-Einheit, kann man dieses Problem vielleicht lösen und die grundlegenden Merkmale des Unterrichtsmodells besser verstehen.

Viele Evaluationsuntersuchungen neuer Curricula oder neuer Unterrichtsmodelle haben sich vor allem der Varianzanalyse als statistischen Hilfsmittels bedient. Neuere Bemühungen haben auch multivariate Modelle verwendet. Der allgemeine Versuchsplan ist dabei derselbe geblieben. Nach experimenteller oder statistischer Kontrolle der anfänglichen Unterschiede zwischen den Schülern werden zwei oder mehr grob definierte pädagogische Programme oder Programmvarianten anhand eines oder mehrerer Leistungskriterien verglichen. Weder die Programmentwickler noch der potentielle Adressat haben aus solchen Untersuchungen viel gelernt.

Da eine überzeugende Evaluation in zahlreichen unterschiedlichen Klassen stattfinden muß und da diese Klassen sich in dem Ausmaß unterscheiden, in dem die verschiedenen Aspekte des Unterrichtsmodells realisiert werden, müssen Dimensionen bestimmt werden, mit denen das Ausmaß der Implementation gemessen werden kann; außerdem muß die Klasse als Analyse-Einheit in einem Korrelationsmodell verwendet werden.

Drei Arten von Variablen müssen berücksichtigt werden:

1. Das Anfangsverhalten der Schüler (Input)
2. Die Dimensionen des Unterrichts (Prozeß)

3. Die Schülerleistungen am Ende des Jahres (Output).
Der Hauptgrund für die Verwendung der Klasse als Analyse-Einheit liegt darin, daß Prozeßwerte für die Klasse charakteristisch sind. Ein weiterer wichtiger Aspekt dieses Verfahrens liegt darin, daß man die Auswirkungen erfassen kann, die eine unterschiedliche Verteilung beim Input auf den Output hat. Außerdem kann man feststellen, inwieweit das Programm bzw. die Programmvarianten unterschiedliche Outputs zur Folge haben. Dies wird dadurch erreicht, daß alle Werte des (Schüler-) Inputs oder Outputs auf vier statistische Maßzahlen für jede Klasse reduziert werden: Mittelwert (M), Standardabweichung (s), Schiefe (g_1) und Exzeß (g_2). Abb. 1 zeigt die Häufigkeitsverteilung und die vier statistischen Maßzahlen für eine der Klassen des Frick-Programms. Die Informationen über negative Schiefe, Hyperexzeß der Verteilung, ihre Lokalisation und ihre Streuung werden in diesen vier Werten beschrieben. Wiley (Wittrock / Wiley 1970) hat die Brauchbarkeit dieses Ansatzes behauptet; Lohnes (1971) bietet in einer Reanalyse der Daten der Cooperative Reading Study eine gute Illustration dafür. Ich möchte die bisherigen Ausführungen mit Hilfe wirklicher Daten aus den Klassen der Frick-Schule und des Netzwerks veranschaulichen.

Eine Dimension des Schüler-Inputs ist der Einstufungs-Test in unserem Rechencurriculum (vgl. Resnick, Wang und Kaplan, 1970). Ein ähnlicher Wert des (Schüler-)Outputs ist der Rechenwert im WRAT. Die Werte dieser zwei Messungen von 1500 Schülern können in acht Werte von 57 Klassen umgewandelt werden. Die vier statistischen Maßzahlen jeder Klasse basieren auf der Einstufung im Rechencurriculum als Inputmaß und den vier WRAT-Maßen als Outputmaß.

Bevor wir die Meßwerte über die unterschiedliche unterrichtliche Realisation des Programms in den Klassen miteinbeziehen, sollte man die Beziehungen zwischen diesen 8 Input- und Output-Werten untersuchen. Anstatt auf eine Korrelationsmatrix von 64 Elementen zu starren, bietet die kanonische Korrelation eine gute Zusammenfassung davon, wie die Inputwerte auf den Output bezogen werden. Tabelle 4 faßt die Ergebnisse einer kanonischen Korrelationsanalyse zwischen den vier Inputwerten und den vier Outputwerten zusammen.

Nur eine der vier möglichen kanonischen Beziehungen war auf dem 5 %-Niveau signifikant. Die kanonische Struktur und die Koeffizienten für diese größte Beziehung zeigen, daß ein Faktor, der zur Zeit des Inputs auf den Mittelwerten und den Standardabweichungen positiv und auf der Schiefe negativ geladen ist, mit einem Faktor korrelierte, der primär durch die Mittelwerte zur Outputzeit definiert ist. So scheint also die Form und der Mittelwert der Verteilung der Schüler im Herbst die mittleren Leistungen

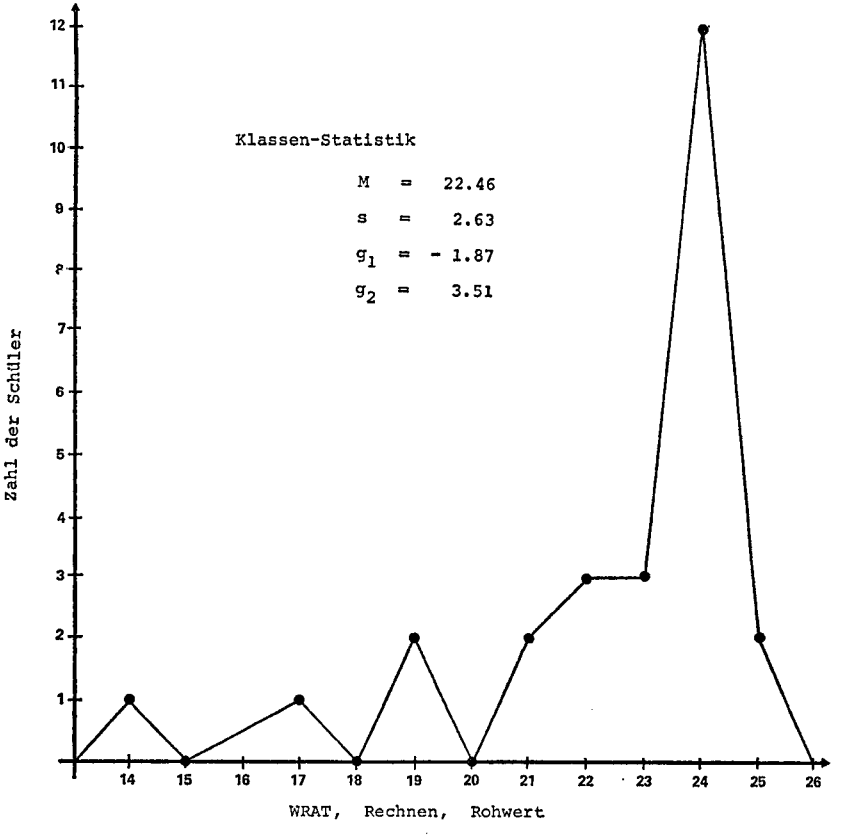


Abbildung 1
WRAT Rechnen, Verteilung für Klasse 1114
(N = 26)

der Klasse im Frühjahr zu beeinflussen. Jedoch besteht zwischen der Leistungsverteilung im Frühjahr und den Inputwerten im Herbst eine geringe Beziehung, d. h. das Ausmaß an Streuung, Schiefe und Exzeß im Frühjahr bezieht sich nur insoweit auf die Herbstwerte, als es von den Mittelwerten des Frühjahrs abhängt. Daher gibt es neben den Inputunterschieden noch andere Gründe für den Verlauf der Verteilungen im Frühjahr.

Der erste kanonische Faktor extrahiert ungefähr ein Drittel der Varianz jeder der zwei Gruppen der Variablen (.37 und .33). Die Varianz, die zusammen mit den kanonischen Relationen extrahiert wird, gestattet uns, die Redundanz des Output bei gegebenem Input einzuschätzen. Ein Redun-

danzkoeffizient von .18 zeigt an, daß 82 % der totalen Outputvarianz nicht durch diesen ersten Inputfaktor erklärt werden [1]. Daher muß ein Teil der Outputvarianz anders als durch die Inputvarianz erklärt werden.

Obwohl kanonische Analysen zwischen Input und Output interessant sein können, muß man die Prozeßdimensionen als eine dritte Art von Werten berücksichtigen und in die Analyse einbeziehen. Daher will ich zunächst beschreiben, wie die Prozesse gemessen werden, die wir auch als unterrichtliche Realisation oder Implementation bezeichnen.

Um den Prozeß der unterrichtlichen Implementation zu messen, müssen wir die Variablen identifizieren, die für das Unterrichtsmodell des Learning Research and Development Center besonders wichtig sind. Sieben Variablen scheinen für das Unterrichtsmodell relevant zu sein und lassen Unterschiede zwischen den Klassen erwarten:

1. Testverfahren

Tabelle 4
Kanonische Korrelationen zwischen den Werten im Herbst und den Werten im Frühjahr (N = 57 Klassen)

Klassenstatistik	Arithmet. Mittel	Standardabweichung	Kanonische Struktur	Kanonische Koeffizienten	
		INPUT			
Herbst-Quantifikation					
Mittelwert	7.12	8.12	.82	.92	erklärte
Standardabweichung	5.90	5.84	.53	−.29	Varianz = .37
Schiefe	1.11	1.11	−.66	−.85	Redun-
Exzeß	1.84	3.75	−.25	.66	danz = .20
		OUTPUT			
Frühjahrs-WRAT-Rechenwert					
Mittelwert	19.92	3.28	.99	.93	erklärte
Standardabweichung	3.17	1.01	−.57	−.12	Varianz = .33
Schiefe	−.49	.61	−.11	−.22	Redun-
Exzeß	.59	1.48	.09	−.16	danz = .18
		Kanonische Korrelation = .73			
		Chi-Quadrat = 50.12			
		df = 16			
		$p < .001$			

Andere mögliche kanonische Beziehungen sind nicht signifikant auf dem 5 %-Niveau

2. Unterrichtsanweisungen
3. Beweglichkeit des Lehrers (wie der Lehrer seinen Unterricht gestaltet und auf das Schülerverhalten angemessen reagiert)
4. Art des wirklich verwendeten Unterrichtsmaterials
5. Zeiteinhaltung
6. Ausnutzung des Klassenraums
7. Das curriculare Wissen des Lehrers und seine Kenntnis der ihm anvertrauten Kinder.

Um von diesen Bereichen zu meßbaren Dimensionen zu gelangen, bieten sich zwei Verfahren an. Im Bereich der Tests könnte man z. B. folgende Verfahren entwickeln, mit denen die Lehrer ihre Testpraktiken verbessern können:

1. Häufiges individuelles Testen der Schüler
2. Genaue Auswertung und Darstellung der Testergebnisse
3. Bestimmung eines festen Platzes, an dem im Klassenzimmer Tests bearbeitet werden
4. Verwendung des Mastery Level [2]
5. Testen aller Lernziele.

Ein Mitglied des Projektteams des Learning Research and Development Center (Champagne 1971) hat eine solche Liste entwickelt, die aus 108 Items für 7 Komponenten des Modells besteht, die alle von einem Unterrichtsbeobachter kontrolliert werden können. Ihre Erprobung im vergangenen Frühjahr zeigte, daß sie als ein Mittel für die Beurteilung der Effektivität des Fortbildungsprogramms für die im Netzwerk arbeitenden Lehrer geeignet war. In jedem Bereich müssen jedoch einige Haupt-Variablen identifiziert werden, wenn Datensammlung und -analyse im Rahmen der Evaluation durchführbar sein soll. Mehr als 150 Klassen könnten zur Evaluation herangezogen werden, jedoch müssen die Kosten für die Unterrichtsbeobachtung niedrig gehalten werden.

Reynolds (1971) hat ein gutes Beispiel für ein entsprechendes Verfahren gegeben. Seine Untersuchungen einiger Klassen der Oakleaf-Schule haben ergeben, daß die Korrelation zwischen der Einstufung des Schülers und den standardisierten Leistungswerten um so höher ist, je mehr die Einstufung und die Testverfahren mit dem Unterrichtsmodell übereinstimmen. Eine zentrale Voraussetzung unseres Unterrichtsmodells besagt, daß Lernen dann am wirksamsten ist, wenn ein Schüler in einem hierarchisch organisierten Curriculum an der Stelle arbeitet, die ein wenig über seinen bisherigen Leistungen liegt, jedoch unter dem, was er nicht mehr leisten kann. Die häufige Verwendung von Kriteriumstests [3] ist das Mittel, mit Hilfe dessen diese Einstufung fortwährend modifiziert werden kann. Wenn es jedoch nachlässig gehandhabt wird, verschwendet der Schüler seine Zeit

mit Aufgaben, die er bereits bewältigt hat oder für deren Bewältigung er keine Voraussetzungen hat.

Für eine bestimmte Klasse wird die Korrelation zwischen der Einstufung der Schüler im Curriculum und dem allgemeinen Leistungsniveau niedrig sein, wenn:

(1) die Schüler das ganze Curriculum durcharbeiten können oder sogar dazu ermuntert werden, ohne jede einzelne curriculare Einheit wirklich zu beherrschen;
(2) die Schüler im Curriculum unter ihrem Leistungsniveau arbeiten;
(3) Lehrer die Einstufung der Schüler dadurch beschränken, daß sie sie mehr oder weniger an der gleichen Stelle im Curriculum zusammenhalten.

Somit würde eine Korrelation innerhalb einer Klasse zwischen den im Herbst in standardisierten Tests erreichten Schülerleistungen und der Einstufung der Schüler im Herbst gute Aufschlüsse darüber erlauben, wie gut ein Lehrer Tests im Rahmen des Programms verwendet. Die anderen sechs Bereiche werden ähnlich behandelt, um festzustellen, welche Hauptvariablen man benutzen könnte, um den Grad der Implementation jedes Bereichs zu erfassen.

Nachdem nun die dritte Gruppe von Variablen behandelt worden ist, gilt es das Problem der Definition eines analytischen Schemas zu reflektieren, mit dessen Hilfe Prozeßwerte in Verbindung mit Input und Output untersucht werden können. Es gibt zahlreiche mögliche Ansätze, dieses Problem zu lösen. Vier davon sollen hier genannt werden:

1. Kanonische Korrelation zwischen Input und Output, um die Residuen der Outputfaktoren auf Prozeßwerte zu beziehen.
2. Multiple Korrelationen vom Input mit jedem Output, Berechnung der Residuen für jeden Outputwert und Verbindung dieser mit den Prozeßwerten.
3. Zunächst wurde eine Auspartialisierung des Inputs aus dem Output vorgenommen; sodann wurde eine kanonische Korrelation zwischen Output-Residuen und Prozeß berechnet.
4. Es erfolgte eine Auspartialisierung des Inputs aus dem Output und aus den Prozeßvariablen; sodann wurde eine kanonische Korrelation zwischen den Residuen des Outputs und der Prozeßvariablen bestimmt.

Ob die mit dem Input zusammenhängende Varianz vom Output und dem Prozeß oder nur vom Output getrennt werden soll, bedarf sorgfältiger Überlegung. Man kann zu Recht erwarten, daß die Inputwerte den Prozeß beeinflussen, d. h. die unterrichtliche Realisierung kann als eine Funktion der Lokalisation und der Form der Klassenverteilung im Input verschieden sein. Daher wäre es sicher nützlich, die Art solcher Beziehun-

gen zu kennen, obwohl wir vor allem wissen wollen, wie die wirklich verwendeten Unterrichtsverfahren die Varianz im Output, die nicht zum Input in Beziehung steht, erklären.

Um in dieser Frage einen ersten Schritt zur Lösung zu machen, wurde eine multiple Korrelation zwischen den vier Inputwerten im Herbst und den Mittelwerten im Frühjahr (*Tabelle 5*) bestimmt; darauf folgte eine Berechnung der Restwerte für die Mittelwerte des Frühjahrs, was eine Variation in den Mittelwerten des Klassen-Outputs erbrachte, die nicht durch die vier Inputwerte erklärt werden kann. Wegen der Dominanz der Mittelwerte des Frühjahrs bei der Definition des im Frühjahr ermittelten kanonischen Faktors in Tabelle 4 ist die multiple Korrelationsstruktur des Herbstes identisch mit der kanonischen Korrelationsstruktur des Herbstes, was die frühere Aussage über den Mangel an zusätzlicher Information bei den Verteilungen im Frühjahr bestätigt. Abb. 2 zeigt die Beziehung zwischen vorhergesagten und beobachteten Mittelwerten auch für die 57 Klassen. Die Restwerte sind die vertikalen Abstände jeder Klasse von der in der Mitte liegenden Regressionslinie.

Tabelle 5
Vorhersage der durchschnittlichen Rechenleistung vom Frühjahr aus den statistischen Maßzahlen im Herbst
(N = 57 Klassen)

Herbst Rechnen Prädiktor	Kriteriums-Korrelation	Standardisierte Partielle Regressions-Koeffizienten	Struktur
Mittelwert	.59	.64	.82
Standardabweichung	.39	−.19	.54
Schiefe	−.49	.63	−.68
Exzeß	−.20	.46	−.28
	Multiple Korrelation = .72		

Um von Mitarbeitern, die mit den Klassen vertraut waren, einige Vorschläge bezüglich der für die Klassenunterschiede wichtigen Dimensionen zu bekommen, entwickelte ich zwei Listen, von denen eine die Klassen mit hohen positiven Restwerten (Region A in Abb. 2), die andere die Klassen mit den hohen negativen Restwerten (Region B) enthält. Die beiden Listen wurden nicht als solche identifiziert. Anfangs ergaben sich Schwierigkeiten bei der Differenzierung der Unterschiede, weil Klassen, in denen der Lehrer sich offensichtlich bei der Dimension der Beweglichkeit und

den anderen Hauptdimensionen des Unterrichtsmodells richtig verhalten hatte, zusammen mit weniger wirksamen Klassen auf beiden Listen vertreten waren. Dennoch entstand eine störende Konsistenz. In Region A neigten die Lehrer dazu, den Einstufungstest vorzeitig abzubrechen, und unterbewerteten damit das allgemeine Niveau des Eingangsverhaltens ihrer Klasse. In Region B neigten sie dazu, die Einstufung der Schüler im Rechencurriculum des vergangenen Frühjahrs für die Plazierung im Herbst zu verwenden, und überbewerteten damit ihre Schüler, da sie die Sommerpause nicht berücksichtigten.

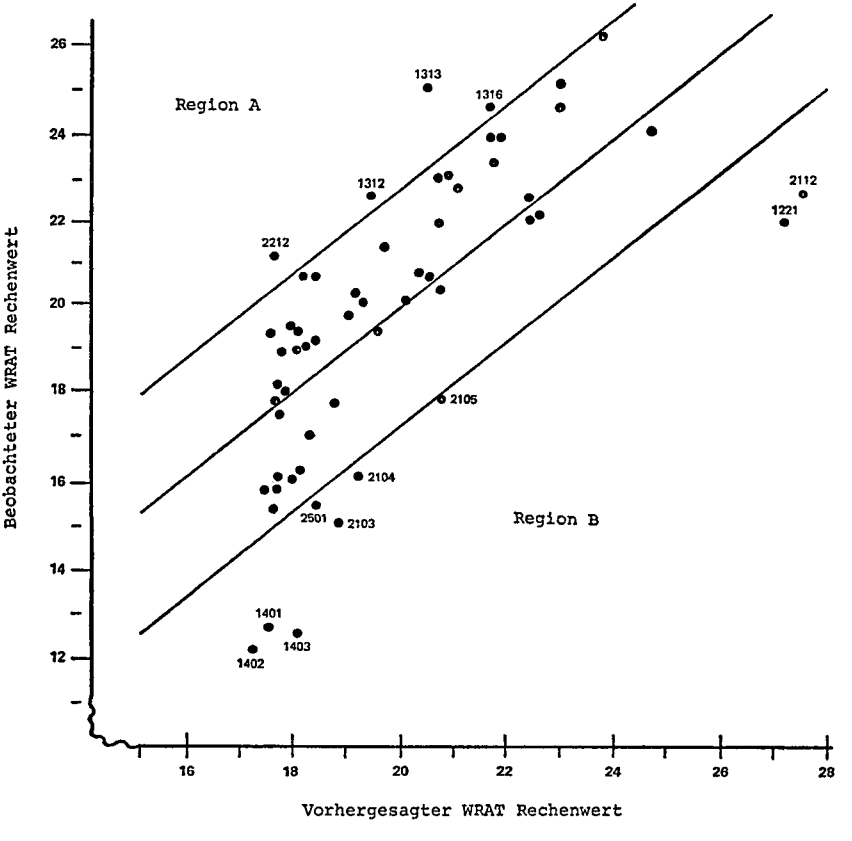


Abbildung 2

Stellung von 57 Klassen in einem zwei-dimensionalen Raum definiert durch eine lineare Funktion von vier Quantifikationswerten (Input) im Herbst und WRAT Rechenwerten (Output) im Frühjahr.

Dieser erste Schritt der Durchführung des Evaluationsplans teilte mir mehr darüber mit, wie sich die Klassen in bezug auf das Testen für die Einstufung der Schüler unterschieden, als über die Beziehungen zwischen den Unterrichtsverfahren und den Ergebnissen. Der Einstufungstest ist natürlich Teil des Unterrichtsmodells, und sein Einsatz steht unter der Kontrolle des Lehrers. Wenn aber erst Unterschiede bei der Realisierung dieses Aspekts des Modells entdeckt werden, kann aus ihrer Existenz bei diesem Regressionsansatz über das Unterrichtsmodell nichts mehr erfahren werden.

Wenn ein Forscher entdeckt, daß einer seiner Hauptwerte wie ein Gummiband ist, muß er bessere Untersuchungsverfahren entwickeln. Glücklicherweise kam zu dieser Zeit jemand auf ein besseres Verfahren. Lohnes (1971) überzeugte mich, daß eine Theorie der Input- und Outputmessungen notwendig ist, die den Forschungsprozeß stärker systematisch machen würde. Dies ist besonders wichtig, wenn man für jeden Versuch ein Schuljahr benötigt. Lohnes hat aber nicht nur deutlich gemacht, daß eine Theorie dieser Daten notwendig ist, er hat auch eine solche Theorie entwickelt. Um das zu verdeutlichen, muß ich auf einige Jahre zurückliegende Erfahrungen zurückgreifen. Lohnes und ich haben gemeinsam die Daten des Projekts TALENT bearbeitet, eine nationale Längsschnittuntersuchung, die mit über 400 000 Schülern der 9. bis 12. Klasse 1960 begann (Flanagan u. a. 1962). Eine Batterie von Tests und Fragebogen, deren Einsatz zwei Tage lang dauerte, wurde damals verwendet; ihre Daten wurden später durch Längsschnittwerte ergänzt, die an zentralen Stellen nach dem Sekundarabschluß erhoben wurden. Bei dieser Untersuchung überraschte uns die Vorhersagekraft einer kleinen Gruppe orthogonaler Faktoren, die Lohnes (1966) von der großen Batterie der TALENT-Prädiktoren abgeleitet hatte. Elf Faktoren für die Fähigkeiten und Motive schienen alle Informationen zu enthalten, die für die Vorhersage des von uns untersuchten Verhaltens nach dem Sekundarschulabschluß verfügbar waren. (Cooley/ Lohnes 1968).

Als ich Mitarbeiter am Learning Research and Development Center wurde, war ich über die mangelnde Berücksichtigung dieser grundlegenden allgemeinen Dimensionen individueller Unterschiede enttäuscht. Glaser (1968) und anderen Mitarbeitern gelang es, mich schließlich zu überzeugen, daß solche allgemeinen Einstellungen oder Motive nur wenig oder keine Relevanz für Unterrichtsentscheidungen haben. Die grundlegenden Dimensionen von TALENT, die sich als Prädiktoren für Erfolg und Befriedigung in unserer Gesellschaft so gut eigneten, sind nutzlos, um in der Praxis einen angemessenen Unterricht für einen Schüler zu entwickeln.

Lohnes überzeugte mich jedoch unlängst von der Notwendigkeit, die

TALENT-Dimensionen noch einmal nicht als *Prädiktoren* im Unterrichtsmodell, sondern als *Kriterien* für das Modell zu untersuchen. Nach seiner Auffassung müsse ein wertvolles Unterrichtsmodell auch dazu beitragen, die Wahrscheinlichkeit des Erfolgs und der Befriedigung eines Kindes nach seiner Schulzeit zu erhöhen. Aber auch wenn wir das Modell wiederholt definieren und modifizieren, können wir nicht zwanzig Jahre lang Längsschnittuntersuchungen durchführen, um festzustellen, welchen Fortschritt wir machen. Eine Möglichkeit bestand darin, diese Faktoren aus dem TALENT-Projekt, d. h. die Variablen in der Zeit vor der Sekundarschulerziehung und das Verhalten nach dieser Zeit, als Kriterien für die Wirksamkeit unseres Unterrichtsmodells zu verwenden. Die TALENT-Batterie selbst ist natürlich für Grundschulkinder nicht geeignet, aber die Primärfaktoren, die aus dieser Batterie hervorgingen, ließen sich auch in anderen Batterien finden.

Daher ist bei diesem Ansatz die Auswahl der Testbatterie für die Evaluation weit weniger willkürlich. Die Ergebnisse der Evaluation werden glaubwürdiger, wenn gezeigt werden kann, daß die Faktoren einen Übertragungswert auf das Erwachsenenleben haben. Es zeigt sich auch, wie man Grundschulpraktiken mit dem Prozeß der beruflichen Entwicklung in Beziehung setzen kann, woran kürzlich einige Beamte im Erziehungsministerium sehr interessiert waren.

Unter Evaluatoren ist die Frage umstritten, ob die Kriteriumsbatterie für die Evaluation aus standardisierten oder aus selbst angefertigten Tests bestehen soll, die sich auf die Items begrenzen, die eine Auswahl der Ziele des Curriculum repräsentieren, das evaluiert werden soll. Die Antwort darauf scheint mir jetzt klarer zu sein.

Unsere eigenen Tests sind wichtig, weil nur mit ihrer Hilfe die Frage beantwortet werden kann, ob unser Unterrichtsprogramm tatsächlich die Verhaltensweisen erreicht, die es erreichen soll. Eine umfassende Evaluation muß jedoch mehr leisten. Sie muß zeigen, wie Kinder durch dieses Programm befähigt werden, sich nach Abschluß der Schule im Leben zu bewähren. Wenn die Primärfaktoren für die Fähigkeiten und Motive gute Prädiktoren für den Erfolg und die Zufriedenheit junger Erwachsener sind, wenn sie eine Augenscheinvalidität (face validity) für die von ihnen vorausgesagten Kriterien besitzen und wenn solche Faktoren durch eine Verbindung zwischen Meßwerten aus der Verwendung des Unterrichtsmodells und standardisierten Tests gewonnen werden können, dann können und sollten diese Faktoren auch Kriterien für die Qualität unseres Programms sein.

Eine vollständige Beschreibung der Faktoren des TALENT-Projekts erfordert eine ganze Monographie (Lohnes 1966). Dennoch kann man we-

nigstens die Hauptfaktoren zusammenfassen, die in den Längsschnittsuntersuchungen Vorhersagekraft besaßen (Cooley/Lohnes 1968). Vier Kernfaktoren gingen aus 60 Eigenschaften des TALENT-Projekts hervor: Verbales Wissen, Englische Sprache, Mathematik und visuelles Erfassen. Der beste Prädiktor für die später erhobenen Kriterien und das wichtigste Konstrukt zur Erklärung der Interkorrelationen zwischen den 60 Eigenschaften des TALENT-Projekts ist der Faktor »verbales Wissen«. Lohnes (1966) sieht deutlich, daß dieser Faktor eine enge Approximation an die allgemeine Intelligenz darstellt. Er entschloß sich, ihn »verbales Wissen« zu nennen, weil »Intelligenz ein Begriff ist, der Mißverständnissen viel eher unterworfen ist als der Begriff Wissen«. Man sollte allmählich erkennen, daß ein Ergebnis des Unterrichts in der Maximierung der Punktwerte eines Schülers im allgemeinen Intelligenzfaktor[4] liegen kann und soll.

Von den 38 typischen Leistungswerten (Interessen und Bedürfnisse) leitete Lohnes 11 Motivfaktoren ab, von denen vier gute Prädiktoren dafür waren, wozu die Schüler nach dem Verlassen der Sekundarschule neigten. Drei dieser Faktoren waren sehr bekannte Interessendimensionen: Wirtschaft, Wissenschaft und Außenberufe. Der vierte Motivfaktor wurde mit »schulisches Interesse« bezeichnet. Lohnes (1966, 5–19) definiert diese Faktoren als »ein Motiv, das schulische Verhaltensweisen erklärt, denen die Gesellschaft zustimmt und die sie belohnt.«

Unsere evaluative Forschung in diesem Schuljahr wird von den Ergebnissen der Evaluation im vergangenen Jahr, der Lohnesschen Theorie über die Input- und Outputmessungen und dem Bedürfnis nach einer weiteren Erklärung des Ausmaßes der Implementation gesteuert. Im kommenden Herbst wissen wir über unser Unterrichtsmodell ein wenig mehr als in diesem Herbst. Evaluative Forschung kann und muß als integraler Teil der Curriculumentwicklung durchgeführt werden. Sie ist keine einmalige Handlung, die erst nach der Fertigstellung eines neuen Programms erfolgt. Evaluation kann nicht einfach in formative und summative Aktivitäten geteilt werden. Sie kann dem Programmentwickler Informationen liefern, während sie Informationen für potentielle Adressaten sucht. Sie ist Forschung. Sie wird durch Hypothesen gesteuert. Sie umfaßt eine Reihe von aufeinanderfolgenden Lösungsversuchen. Sie ist manchmal fehlerhaft, aber nie abgeschlossen.

Barry MacDonald

Informationen für Entscheidungsträger: Evaluation des Humanities Projects

Jede Evaluation zielt darauf ab, Informationen für Entscheidungsträger zu gewinnen; aber nicht alle Evaluatoren sind sich darüber einig, welches die wichtigsten Entscheidungsträger sind und welche Informationen sie brauchen. Eine These dieses Beitrags weist darauf hin, daß wenigstens in einigen Curriculumbereichen Evaluation sich stärker darum bemühen sollte, die Fragen der nicht unmittelbar an der Curriculumentwicklung beteiligten Entscheidungsträger zu entdecken und zu beantworten. Damit soll nicht der Wert formativer Evaluation in Frage gestellt werden. Zweifellos braucht man gute curriculare Materialien, aber ebenso empfiehlt es sich, wenn sie wirkungsvoll eingesetzt werden sollen, alle Einflüsse zu erforschen, die in den Schulen auf sie einwirken. Die folgende Darstellung des Humanities Projects und seiner Evaluation soll diese hier aufgestellte These unterstützen.

Wie soll eine demokratische Gesellschaft in ihren Schulen kontroverse Fragenkomplexe behandeln? Darin besteht kurz gesagt das Problem des Humanities Curriculum Projects, das die Nuffield Foundation und das Schools Council in Angriff nahmen. Nachdem drei Jahre an der Erforschung dieser Probleme gearbeitet worden war, wurden nunmehr einige Curriculummaterialien veröffentlicht und interessierte Schulen über die Entwicklung von Unterrichtsstrategien und die Behandlung der curricularen Probleme in diesem Bereich beraten.

Mit diesem Projekt wurde 1967 begonnen; es bildete einen Teil der Vorbereitungen für die Erhöhung der Pflichtschulzeit im Jahre 1972. Dem Projektteam wurde die Aufgabe gestellt, Materialien zu entwickeln und die Schulen zu beraten, wie man 14- bis 16jährigen Schülern mit durchschnittlichen oder unterdurchschnittlichen Fähigkeiten Probleme der Politischen Bildung (Humanities) vermitteln könne. Nach Auffassung des Teams besteht die entscheidende Aufgabe der Politischen Bildung in der Behandlung wichtiger humaner Problembereiche. Man entschloß sich daher, den Schwerpunkt der Bemühungen auf die Bereiche zu legen, in denen Wert-

konflikte auftreten. Die Teammitglieder waren der Überzeugung, daß sie so den Wunsch der Schulen nach einem für die Schüler relevanten Curriculum erfüllen könnten und dadurch den Schulen helfen würden, die kontroversen Problemkomplexe mit den Schülern offen und ehrlich zu erörtern.

Nach Auffassung des Teams bestand für die beteiligten Schulen das zentrale Problem darin, den Schülern Gelegenheit zu einer selbständigen Auseinandersetzung mit politischen Kontroversen zu geben, ohne dabei durch die politische Überzeugung der Lehrer gesteuert oder von anderen Schülern unter Druck gesetzt zu werden. Man wollte dieses Problem durch den Versuch lösen, ein Modell heuristischen Lehrens und einen entsprechenden Stil der Diskussionsführung im Unterricht zu entwickeln.

Als Grundlage für die Diskussion in den kontroversen Bereichen wurden Quellenmaterialien gesammelt. Lehrer übernahmen als Diskussionsleiter die Aufgabe, den Schülergruppen relevantes Material zur Untersuchung und Interpretation zur Verfügung zu stellen. Sie hielten ihre eigenen Anschauungen über die Probleme zurück und bemühten sich, die Schüler dazu zu bringen, widersprüchliche Ansichten zu formulieren. Obwohl der Gedanke der »Lehrerneutralität« in diesem Zusammenhang nicht neu ist, hat er während der Durchführung des Projekts viel Aufmerksamkeit auf sich gezogen. Manche sahen in dieser Neutralität das kennzeichnende Merkmal des Projekts. Dieses Urteil überbetont einen Aspekt in der Lehrerrolle, den das Projekt erforschen wollte. Das Team entwickelte zunächst Curriculummaterialien über Themen wie Krieg, Erziehung, Familie, Beziehungen zwischen den Geschlechtern, Mensch und Arbeit, Armut; (Materialien über das Rassenproblem, Recht und Ordnung, das Leben in den Städten sind in Vorbereitung.) Diese Materialien wurden in den Jahren 1968 bis 1970 von ungefähr 150 Lehrern in 36 Schulen in ganz England und Wales ausprobiert. Das Projektteam stellte Hypothesen über Lehrstrategien auf und legte es den Lehrern nahe, sie unter genauer Berücksichtigung der für sie charakteristischen Bedingungen zu prüfen. Zugleich bat man die Lehrer, ihre Urteile über die Brauchbarkeit des Materials abzugeben, alternative Verfahren oder Hypothesen vorzuschlagen und weitere für die Verbesserung der Diskussion notwendige heuristische Maßnahmen zu entwickeln.

Seit Ostern 1970 begann man damit, die Curriculummaterialien in den Handel zu bringen. Um den Anfragen einzelner Schulen und örtlicher Erziehungsbehörden nachzukommen, wurden überall in Großbritannien Kurse zur Einführung der Lehrer in die Benutzung des Materials eingerichtet. Während des laufenden Schuljahres haben etwa 600 Schulen das auf dem freien Markt erhältliche Material gekauft. Obwohl die meisten dieser Schulen die Absicht äußerten, die vom Projektteam entwickelten Lehrstrategien

zu übernehmen, sind bisher weniger als die Hälfte von ihnen bei den Fortbildungskursen vertreten gewesen.

Die Evaluation des Projekts begann 1968 mit meiner Einstellung. Meine Aufgabe war nicht genau bestimmt. Man wollte jedoch, daß ich die Entwicklung des Projekts untersuchte, um dem Projektteam über den Ablauf des Versuchs in den Schulen zu berichten und um ein geeignetes Evaluationsprogramm für die Implementation des Curriculum in den Jahren 1970 bis 1972 zu entwerfen. 1970 wurden drei weitere Kollegen für diese Aufgabe eingestellt; uns vier obliegt nun die Durchführung des Evaluationsprogramms.

Zu Beginn schien das Problem der Evaluation außerordentlich schwierig zu sein. Wenn ein Curriculum, wie in diesem Fall, in einem weitgehend unerforschten Bereich entwickelt wird, kann man nur auf geringe Erfahrungen zurückgreifen, um seine Auswirkungen oder etwa auftretende Probleme vorauszusagen. Erfahrungen aus früheren Innovationsprojekten lassen eine schlechte Prognose für dieses Projekt erwarten (vgl. Miles 1964). Auf den ersten Blick schien es durch die Mängel gekennzeichnet zu sein, die zu einem Fehlschlag früherer Innovationsprojekte geführt hatten. Es erforderte Einführungskurse für die Lehrer, war schwierig zu unterrichten und im Vergleich zu den in den Schulen verfügbaren Mitteln teuer. Außerdem stand es im Widerspruch zu vielen allgemein anerkannten Wertvorstellungen. Das Projekt zeigte zahlreiche Merkmale, die sein Scheitern durchaus möglich erscheinen ließen. Diese Ansicht beruhte auf der Voraussetzung, daß das Projekt danach beurteilt werden sollte, welche Lernergebnisse es in einer bestimmten Zeit bei den Schülern bewirkte. Das scheint mir heute ein unzureichendes Kriterium zu sein, um den Wert eines derartigen Curriculumprojekts zu beurteilen. Eine nähere Betrachtung einiger seiner wichtigen Merkmale soll dies deutlich machen und mir helfen, den Einfluß des Projektentwurfs auf die Entwicklung der Evaluation zu erklären. Dabei gilt es, drei Punkte zu erörtern:

Der erste Punkt betrifft den Entwurf des Projekts. Nach dem am meisten verbreiteten Modell der Curriculumentwicklung beginnt man damit, spezielle Lernziele zu formulieren, die das Endverhalten der Schüler angeben. Sodann werden die Inhalte und Lehrmethoden des Curriculum so lange modifiziert, bis das gewünschte Verhalten erreicht wird. Für den Evaluator stellen die curricularen Ziele die Erfolgskriterien dar. Seine Hauptaufgabe besteht darin, den Grad, bis zu dem sie erreicht werden, abzuschätzen.

Dieses Modell ist dann gut geeignet, wenn Lernziele einfach formuliert werden können, wenn bei ihrer Aufstellung leicht Konsens erreicht werden kann, wenn Nebeneffekte voraussichtlich unbedeutend und leicht feststell-

bar sind und wenn eine strenge Beachtung der Lernziele nicht dazu führt, die in ihnen nicht enthaltenen pädagogischen Werte zu verletzen. Das Projektteam war der Auffassung, daß die genannten Bedingungen auf das Humanities Curriculum nicht zutrafen, und hatte daher erhebliche Vorbehalte gegenüber diesem lernzielorientierten Modell. Man entschloß sich deshalb zu einem anderen Vorgehen (vgl. Stenhouse 1971), bei dem es folgende drei Fragen zu beantworten galt: Welche Inhalte sind von Bedeutung? Welches allgemeine Ziel ist für das Unterrichten dieser Inhalte angemessen? Welche Lernerfahrungen können dazu dienen, dieses Ziel zu erreichen? Zur Beantwortung dieser letzten Frage bedarf es eines umfassenden Unterrichtsversuchs. Durch die Verwendung von Hypothesen über die Auswirkungen des Curriculum an Stelle von Lernzielen hoffte das Team, eine gute Voraussetzung für die Entwicklung einer wirkungsvollen Lehrstrategie zu haben, die sich mit seinen besonderen Wertvorstellungen über die unterrichtliche Behandlung von Kontroversen deckte [1].

Dieser Auffassung lag die Überzeugung zugrunde, daß Lehrer, wenn sie sich an einem allgemeinen Ziel orientieren, eher wirkungsvolle Lehrstrategien entwickeln können. Daher sollte in diesem Modell der Versuch gemacht werden, das Ziel im Unterrichtsprozeß zu konkretisieren. Für den Lehrer bestand die Schwierigkeit darin, seine Unterrichtsführung mit dem Ziel in Übereinstimmung zu bringen und sie pädagogisch effektiv zu machen.

Bei diesem Modell, bei dem man von der Formulierung bestimmter Lernziele absieht, gibt es für den Evaluator kein eindeutiges Rezept. Er muß sorgfältig das Unterrichtsgeschehen beobachten und muß die verschiedenen Auswirkungen des Materials erforschen und die Beziehungen zwischen den Unterrichtsmodellen und ihren Auswirkungen aufdecken. Die *Ergebnisse* und der *Prozeß* bedürfen seiner Aufmerksamkeit. Ein besonderes Problem besteht darin, zu entscheiden, *welche* Auswirkungen untersucht werden sollen. In einem Evaluationsprogramm sollte man keine Antwort auf Fragen suchen, die keiner gestellt hat.

Der zweite Punkt betrifft die Zielsetzung des Projekts, Verständnis für gesellschaftliche Situationen, menschliche Handlungen und die sich daraus ergebenden Wertkonflikte zu entwickeln. Da für das Projektteam die wichtigste Aufgabe darin bestand, das Verständnis dieser Probleme zu vermitteln, lag es nahe, sich im Rahmen des Projekts eher um eine pädagogische Auseinandersetzung mit den kontroversen Fragen als um die Anpassung der Schüler an die bestehenden Verhältnisse und Normen zu bemühen. Die Folgen, die sich aus diesem Ziel für die Schüler- und Lehrerrollen und ihr Verhältnis zueinander ergaben, bestanden darin, daß in vielen Schulen Verhaltensweisen auftraten, die in Konflikt mit den allgemein ver-

breiteten Einstellungen und Gewohnheiten standen. Geht man davon aus, daß dieser Konflikt zweifellos Einfluß auf die Arbeit des Projekts hat und daß das Ausmaß des Konflikts von Schule zu Schule unterschiedlich ist, muß der Evaluator besonders den *Kontext* berücksichtigen, in dem das Curriculum realisiert wird.

Der dritte Punkt besteht in dem speziellen Einsatz, den das Team zur Lösung der Probleme der Curriculuminnovation gewählt hat. Während die meisten außerhalb der Lehrerschaft geplanten Ansätze zur Curriculumreform durch den Versuch charakterisiert sind, die Curricula von der Lehrfähigkeit der Lehrer unabhängig zu machen (teacher proof curricula), war das Projektteam davon überzeugt, daß es ohne eine sorgfältige Weiterbildung der Lehrer keine wirkungsvolle umfassende Curriculumentwicklung geben könne. Daher empfahl das Team den Lehrern, das Projekt eher als ein Mittel anzusehen, um die Probleme des Unterrichtens von Kontroversen besser selbständig handhaben zu können, als in ihm eine autoritäre, von Experten entwickelte Lösung zu erblicken. Für den Erfolg des Projekts war es wichtig, daß die Lehrer diese Auffassung verstanden und daß sie sich als Träger der Curriculumreform und nicht als Außenstehende fühlten.

Für die Evaluation folgte daraus, daß man die Kommunikation und die persönlichen Kontakte zwischen dem Projektteam und den Schulen untersuchen mußte, um Informationen über den Erfolg oder Mißerfolg dieser Bemühungen zu erhalten. Das heißt, man mußte in der Evaluation den *Input* des Projekts berücksichtigen.

Meine Aufgabe bestand darin, im Rahmen der Evaluation mit den verschiedenen Komponenten einer kreativen Curriculumentwicklung, zahlreichen möglichen Störfaktoren und einem neuen Curriculummodell fertig zu werden. In dieser Situation galt es für mich, die Entwicklung des Projekts so zu beschreiben, daß sie der Öffentlichkeit und dem professionellen Urteil zugänglich wurde. Angesichts der wahrscheinlich großen Bedeutung so vieler Aspekte des Projekts, fühlte ich mich anfänglich zu einer vollständigen Beschreibung seiner Auswirkungen und zur Beachtung aller relevanten Probleme verpflichtet. Evaluationsentwurf, -strategien und -taktiken würden sich, so hoffte ich, aus den Auswirkungen des Projekts auf die Struktur der Evaluationsprobleme ergeben.

Das Projekt in den Versuchsschulen (1968–1970)

Die 36 Schulen, die an dem Versuch im Herbst 1968 teilnahmen, wurden nicht mit den üblichen Stichprobentechniken ausgewählt. Statt dessen wurden sie von den zuständigen Verwaltungsbehörden empfohlen; die Verschiedenartigkeit der Schulen deutet auf interessante Unterschiede in den Beurteilungsmaßstäben und Prioritäten der örtlichen Schulbehörden hin. Nur in wenigen Fällen wurden die Kriterien der Nominierung explizit gemacht. In den meisten Fällen mußten sie erfragt oder erschlossen werden. Die Gründe für die Wahl der Local Education Authorities zu entdecken, war ein wichtiger Teil der Evaluation. Sie half mir, die Charakteristika der Schulstichprobe sowie die Politik und Strategie der örtlichen Schulbehörden in bezug auf die Curriculumentwicklung zu verstehen. Daraus ließen sich Schlüsse ziehen, wie gut die Local Education Authorities ihre Schulen kannten und auf Grund welcher Kriterien sie sie beurteilten. Ohne Zweifel bildeten zahlreiche unterschiedliche Kriterien die Grundlage für die Nominierung. In einigen Fällen wurden Schulen vorgeschlagen, die nach Meinung der Schulbehörden gut geeignet waren, an dem Versuch teilzunehmen; in anderen Fällen sollten bestimmte Schulen Anregungen zu neuen Ideen erhalten. Manchmal schien es, als habe man einer alten Schule die Teilnahme an dem Projekt als Entschädigung für die schlechten materiellen Verhältnisse, unter denen Lehrer und Schüler arbeiten mußten, zugedacht; dann wieder wurden vorbildliche Modellschulen empfohlen. Die Grundlagen für die Empfehlungen waren undurchsichtig. In einer Local Education Authority bestand ein wichtiges Kriterium für die Auswahl darin, ob die Schulleiter das Projekt dafür benutzen würden, größere finanzielle Anforderungen zu stellen oder nicht. Auf Initiative des Direktors und der Lehrerschaft hatten sich einige Schulen, die von dem Versuch gehört hatten, selbst beworben und ihr Anliegen bei der entsprechenden Schulbehörde vorgetragen. Im allgemeinen schienen die Local Education Authorities Schulen zu empfehlen, die ihrer Meinung nach für den Versuch geeignet waren. Aus einer näheren Kenntnis der Versuchsschulen wurde jedoch deutlich, daß einige Entscheidungsträger der Local Education Authorities ihre Schulen nicht genau kannten und folglich dazu neigten, ihr Urteil auf unzulängliche Kriterien zu gründen. Solche Beobachtungen über die Wahl der Local Education Authorities sollen nicht bedeuten, daß das Hauptkriterium für die meisten Empfehlungen nicht in der Geeignetheit der Schulen lag.

Die beteiligten Lehrer nahmen zusammen mit dem Projektteam an Regionalkonferenzen im Sommer 1968 teil. Auf ihnen wurde der Versuchsplan erläutert und die Aufgaben der Lehrer besprochen. Die meisten Leh-

rer verließen die Konferenz mit einigem Engagement für die ihnen gestellte Aufgabe.

Die Versuchsschulen waren über ganz England und Wales verteilt; sie lagen auf dem Lande, in Vororten, Städten und Großstädten. Die von allen Schulen ausgefüllten Fragebogen wiesen Unterschiede in der Art, Größe, Organisationsstruktur und in der formalen Beschreibung der Schüler auf. Aufgrund des Entschlusses, jeder Schule Entscheidungsbefugnisse zuzugestehen, vermehrten sich die bereits bestehenden Unterschiede noch durch die unterschiedlichen Entscheidungen über Einführung, Organisation und Verwirklichung des Versuchs. So standen z. B. in einer Schule vier Stunden, in einer anderen indessen 15 Stunden zur Verfügung. Die Zahl der in einer Schule beteiligten Lehrer schwankte zwischen 1 und 10. Einige Schulen ließen gute Schüler der 10. Klasse, andere die weniger guten Schulabgänger der 9. Klasse an dem Versuch teilnehmen. Die Situation wurde weiterhin durch Variablen wie Motivation, Verständnisfähigkeit und Erwartung der Versuchsteilnehmer erschwert, die erst im Laufe der Zeit deutlich in Erscheinung traten. Eine weitere Variable bestand im Ausmaß der Unterstützung, die eine Schule von ihrer Local Education Authority erhielt.

Die unmittelbaren Auswirkungen des Projekts waren im allgemeinen beunruhigend. Verwirrungen und Mißverständnisse waren so groß, daß ein Teil der Schulen nicht mehr den Empfehlungen des Projektteams angemessen nachkommen konnte. Es ergaben sich zahlreiche unerwartete Probleme und häufige Mißverständnisse über den Anspruch des Projekts wie z. B.:
(1) die Bedeutung der Schulleiter für die Realisierung von Innovationen wurde vom Projektteam unterschätzt, das anfangs das Ausmaß der Anforderungen, die an die wenig flexiblen Verwaltungen gestellt wurden, nicht richtig vorausgesehen hatte. Für die Schulen war es nicht einfach, die für den Versuch notwendigen Bedingungen zu schaffen; für die Lehrer war es nicht leicht, diese schwierigen und ungewohnten Aufgaben ohne Verständnis und Unterstützung durch den Schulleiter zu bewältigen.
(2) Die Lehrer waren sich nicht bewußt, daß die bisherigen Lernerfahrungen der Schüler die angestrebte Auseinandersetzung mit Kontroversen außerordentlich erschwerten und daß viele Schüler das Interesse für alle curricularen Inhalte verloren hatten. Auch vergegenwärtigte man sich nicht genügend, daß Lehrer und Schüler die traditionelle Rolle der Lehrerdominanz internalisiert hatten. Die Lehrer waren überrascht, daß die Schüler sich an den Diskussionen kaum beteiligten; sie waren darüber erstaunt, in welchem Ausmaß Schüler von der Initiative der Lehrer abhängig waren und mit welcher Zurückhaltung sie die Aufforderung, sich frei zu äußern, aufnahmen. Es schien, als seien fast alle Schulen und Lehrer autoritärer

eingestellt, als sie es selbst angenommen hatten. Die Auswirkung des Projekts auf die Autoritätsstruktur der Schule wurde immer deutlicher. Viele Lehrer kamen in schwierige Rollenkonflikte und versuchten vergeblich, das gestörte Vertrauensverhältnis zwischen sich und ihren Schülern zu überwinden. Das geht z. B. deutlich aus folgender Lehreräußerung hervor. »Ich bin sehr tolerant dem gegenüber, was die Gruppe in der Diskussion sagen möchte ... aber dann, wenn sie mich manchmal so ungezwungen oder direkt und unverfroren ansprechen, fühle ich doch, daß ich ein Erwachsener bin, und ich zeige ihnen meine Überlegenheit ... und meine Autorität, und gerade das gibt mir zu denken.«

(3) Offensichtlich hatte das Projektteam zunächst die Aufgaben des Projekts nicht klar und deutlich genug dargelegt. In den Augen der Lehrer hatte es eher einen moralischen als einen heuristischen Charakter. Die vorgeschlagenen Lehrstrategien wirkten eher so, als sollten sie nicht die Forschungshypothesen, sondern die Lehrfähigkeit der Lehrer prüfen. Beide Auffassungen führten dazu, daß die Lehrer aus ihren Erfahrungen wenig lernten und daher nur wenige Informationen von den Lehrern zum Projektteam gelangten.

Wenn die Bedingungen in den Schulen so einheitlich gewesen wären, wie die genannten drei Punkte darlegen, wäre die Evaluation anders verlaufen. Doch war das ganz und gar nicht so. Obwohl das Programm im allgemeinen sich als anspruchsvoll, schwierig und anregend herausstellte, gab es in manchen Fällen auch erhebliche Ausnahmen und Widersprüche. Während viele Schulen über ernsthafte Probleme wie den Schwierigkeitsgrad des Materials oder bestimmte Einstellungen der Schüler berichteten, waren andere über manche dieser Schwierigkeiten überrascht, die ihnen selbst überhaupt nicht begegnet waren. Daher können auch begrenzte Erklärungen des Erfolgs oder Mißerfolgs bei dem Versuch, die Schülerfähigkeiten, das Lehrerverhalten oder das Engagement eines Lehrerkollegiums zu verbessern, nicht ohne weiteres verallgemeinert werden. Es war nicht einfach, die Zahl der theoretisch zu berücksichtigenden Variablen aufgrund der Erfahrungen zu reduzieren. So legten z. B. Erfahrungen der Lehrer im Nordosten Englands die Vermutung nahe, daß das unterschiedliche Engagement von Jungen und Mädchen in kleinen Gruppendiskussionen nur durch starke Unterschiede zwischen den Geschlechtern erklärt werden könne, die in den Normen der Arbeiterklasse dieser Gegend ihren Ursprung hatten; während dagegen die Lehrer an einer wallisischen Schule halb scherzend behaupteten, sie kämen mit den Schülern in keine Diskussion, weil es in dieser Gegend von Wales keine kontroversen Fragenkomplexe gäbe. Selbst wenn solche regionalen Unterschiede außer Acht gelassen wurden, schienen die Probleme noch immer komplexer zu werden.

Während das Projektteam sich im ersten Jahr mit den Problemen der Schulen befaßte, um den Innovationsversuch funktionsfähig zu erhalten, konzentrierte ich mich darauf, die Vorgänge in den Schulen zu untersuchen und Informationen zu sammeln, die zur Erklärung unterschiedlicher Handlungs- und Reaktionsmuster beitragen konnten. Ich untersuchte die Aktivitäten der Teams, die Interaktionen zwischen seinen Mitgliedern, die Local Education Authorities und die Schulen. Ich sammelte Daten über die unterstützenden und hemmenden Einflüsse von außen, die bei der Implementation des Curriculum auftraten. Ich erarbeitete eine Liste mit empirisch abgesicherten und weniger abgesicherten Items, die ein institutionelles Profil für jede Schule ergaben. Mit Hilfe von Fragebogen versuchte ich abzuschätzen, wie weit die beteiligten Lehrer die Theorie des Projekts verstanden und wie sie dem Projekt gegenüber eingestellt waren. Ich organisierte mit Hilfe von Tonbandaufzeichnungen und ergänzenden schriftlichen Protokollen ein Feedback-System für die Lehrer und machte in mehreren Teilen des Landes Fernsehaufzeichnungen von Unterrichtsabläufen. Die Fragestellungen und das Erkenntnisinteresse des Projekts und des Evaluationsteams waren weitgehend identisch, so daß sich eine gute Basis für eine dauerhafte Zusammenarbeit ergab.

Ich begann, eine Reihe von Schulen zu besuchen, um sie unmittelbar zu untersuchen. Nachdem ich in etwa der Hälfte der Schulen gewesen war, gab ich diesen Plan zugunsten von Fallstudien einiger Schulen auf, weil ich die Gründe für das beobachtete Diskussionsverhalten der Gruppen nicht verstehen konnte. Warum waren in dieser Hinsicht die Unterschiede zwischen den Schulen größer als innerhalb der Schulen? Warum war eine Schülergruppe an den Problemen interessiert und eine andere mit ähnlichen Charakteristiken so desinteressiert und ablehnend? Weitere Fragen ergaben sich, als wir im Kontext der Schulen nach den Ursachen für die Unterschiede zu suchen begannen. Warum unterstützten einige Kollegen das Projekt, waren andere indifferent und wieder andere ablehnend? Warum reagierten Schulen auf ähnliche Probleme verschieden? Viele derartige Fragen stellten sich uns dabei, nachdem wir einen Überblick über die unterschiedlichen Reaktionen der Institutionen, Lehrer und Schüler gewonnen hatten.

Gegen Ende des ersten und im Verlauf des zweiten Jahres des Versuchs wurden in etwa sechs Schulen Felduntersuchungen durchgeführt. Der Auswahl der Schulen lagen zahlreiche unterschiedliche Kriterien zugrunde, die die verschiedenen Reaktionen der Schulen auf den Versuch berücksichtigten. Bei den Fallstudien wurden Unterrichtsbeobachtungen, Interviews mit dem Lehrerkollegium, den Schülern und den Eltern gemacht; man sammelte detaillierte Informationen über die verschiedenen innerhalb und

außerhalb der Schule entstehenden Einflüsse auf das Projekt. In diesem Zusammenhang kann jedoch über diese Untersuchungen nicht näher berichtet werden; ich will lediglich einige Einsichten erwähnen, die wir aus der Evaluation gewinnen konnten:

(1) Menschliches Verhalten in pädagogischen Situationen ist zahlreichen unterschiedlichen Einflüssen ausgesetzt. Dies ist zwar allgemein bekannt, wird jedoch manchmal bei der Curriculumevaluation übersehen, da man davon ausgeht, daß die Intentionen auch tatsächlich realisiert werden und daß sich die Unterrichtsereignisse nur geringfügig zwischen den Schulen unterscheiden.

(2) Die Bedeutung einer Innovation läßt sich nicht aus der Summe einzelner Wirkungen verstehen, sondern muß als ein System von Handlungen und Konsequenzen begriffen werden. Um eine einzelne Handlung zu verstehen, muß ihre Funktion innerhalb des Systems bestimmt werden. Daraus folgt, daß Innovationen viel mehr unerwartete Folgen haben, als man im allgemeinen bei der Innovations- und Evaluationsplanung annimmt.

(3) Nicht einmal zwei Schulen haben so ähnliche Bedingungen, daß Rezepte für Innovationshandlungen individuelle Entscheidungen ihrer Lehrerkollegien ersetzen könnten. Bereits aus den historisch verschiedenen Bedingungen der Schulen entstehen erhebliche Unterschiede im Innovationsverhalten, die beim Treffen von Entscheidungen berücksichtigt werden müssen.

(4) Ziele und Absichten der Curriculumentwickler entsprechen nicht immer denen der Adressaten des Curriculum. Wir stellten fest, daß das Projekt häufig in Machtkämpfen zwischen verschiedenen Gruppen im Lehrerkollegium als politisches Mittel eingesetzt wurde. Oder es wurde dazu benutzt, die Schüler besser kontrollieren zu können und das Ansehen der Institutionen zu verbessern, ohne jedoch die Unterrichtswirklichkeit innovierend zu verändern. Darin liegt ein Beispiel für eine Innovation ohne wirkliche Veränderung der Schulwirklichkeit.

Begründung und Bezugssystem des Evaluationsprogramms (1970–1972)

Um möglichen Mißverständnissen vorzubeugen, sollte ich darauf hinweisen, daß die Evaluation des Projekts nicht eine Aufgabe ist, die nur von besonderen Fachleuten ausgeführt werden kann. Alle Mitglieder des Projektteams haben viel Zeit für die Evaluation ihrer Arbeit aufgebracht, um sie besser zu verstehen und den Schulen besser behilflich sein zu können. Mit Hilfe des Projektteams haben viele Schulen Prüfungsprogramme und

Schülerbeurteilungsbogen erarbeitet. Ich bin lediglich für die Arbeit eines unabhängigen, in das Projekt integrierten Evaluationsteams verantwortlich, dessen Vorgehen ich jetzt beschreiben möchte.

Evaluation kann danach beurteilt werden, ob sie die richtigen Informationen den richtigen Leuten zur richtigen Zeit zur Verfügung stellt. Aber wer sind die richtigen Leute, was sind die richtigen Informationen und wann werden sie benötigt?

Da ich mit einem Projektteam zusammenarbeitete, das gegen den Gebrauch von Lernzielen war, mußte ich ein anderes geeignetes Konzept für die Evaluation entwickeln. Je stärker ich die Komplexität und Verschiedenartigkeit der Vorgänge in Versuchsschulen erkannte, desto skeptischer wurde ich dagegen, Evaluation auf die Messung des Erreichens von Lernzielen zu beschränken. Ich versuchte, meine Aufgabe im Hinblick auf die Adressaten meines Berichts zu bestimmen. Es erschien mir sinnvoll, die Evaluation an bestimmte Adressaten zu richten. Im Laufe der Zeit wurden sie als die Entscheidungsträger definiert. Vier Gruppen von Entscheidungsträgern ergaben sich: Geldgeber, örtliche Erziehungsbehörden (Local Education Authorities), Schulen und Prüfungsausschüsse. Als Aufgabe der Evaluation wurde die Beantwortung der Fragen der Entscheidungsträger angesehen. Bald erschien jedoch diese Aufgabendefinition als unbefriedigend, weil sie unterstellte, daß diese Gruppen im voraus wußten, welche Fragen wichtig waren. Solange über den Erziehungsprozeß so wenig bekannt ist, daß die Auswirkungen bestimmter Innovationen nicht voraussagbar sind, ist dies daher eine nicht gerechtfertigte Unterstellung.

Gegenwärtig sehen wir unsere Aufgabe darin, den Entscheidungsträgern behilflich zu sein, begründete Urteile zu fällen. Dazu liefern wir ihnen Informationen, die ihre Kenntnis der Faktoren, die auf curriculare Handlungen Einfluß haben, verbessern. Diese Aufgabenbestimmung enthält zwei wichtige Vorteile: Zunächst erhöht sich die Zahl der Personen, für die Evaluation wertvoll ist. Sodann berücksichtigt sie die oft geäußerten Einwände, daß die Daten der Evaluation zu spät verfügbar sind, um noch Entscheidungen über das Curriculum beeinflussen zu können. Solange Ergebnisse der Evaluation jedoch ausschließlich auf das Curriculum bezogen werden und nicht darüber hinaus generalisierbar sind, ist in vielen Fällen diese Kritik durchaus berechtigt. Unsere Ergebnisse sollten für die wiederholt auftretenden Probleme, die sich bei Entscheidungen über ein Curriculum ergeben, relevant sein, und sollten zu einem besseren Verständnis curricularer Innovationen beitragen.

Aufgrund dieser Überlegungen können wir die Ziele der Evaluation folgendermaßen bestimmen:

(1) Um sicher zu sein, welche Wirkungen das Projekt hat, müssen die Um-

stände, unter denen die Wirkungen auftreten, aufgezeichnet werden; sodann müssen die Informationen den Entscheidungsträgern so bearbeitet vorgelegt werden, daß sie ihnen helfen, die voraussichtlichen Folgen der Implementation des Curriculum zu beurteilen.
(2) Wir beabsichtigen, die gegenwärtige Situation und die Vorgänge in den Schulen so zu beschreiben, daß die Entscheidungsträger besser verstehen können, was sie zu verändern versuchen.
(3) Es gilt die Arbeit des Projektteams so zu beschreiben, daß es den Geldgebern und Bildungsplanern hilft, den Wert dieser Investition zu beurteilen und das geeignete Bezugssystem für die finanzielle Unterstützung, die Planung und Kontrolle genauer zu bestimmen.
(4) Um einen Beitrag zur Theorie der Evaluation zu leisten, müssen wir unsere Probleme klar formulieren, unsere Erfahrungen aufzeichnen und unsere Fehler öffentlich eingestehen.
(5) Es muß uns gelingen, zum Verständnis der allgemeinen Probleme einer innovativen Curriculumentwicklung beizutragen.

Nicht jeder würde der Wahl dieser fünf Punkte als Evaluationsziele eines Curriculumprojekts zustimmen. Jedoch sind Ziele nach meiner Ansicht zum Teil auch das Ergebnis bestimmter Situationen. Da Curriculumentwicklung immer mehr in den Aufgabenbereich neuer und relativ unerfahrener Institutionen fällt, sollten die Forscher, die Erfahrungen auf diesem Gebiet haben, sich bemühen, möglichst viel zum Verständnis der Probleme, die sich bei der Implementation von Innovationen ergeben, beizutragen.

Das Hauptproblem besteht noch immer darin, die *richtigen* Informationen auszuwählen und sie den Entscheidungsträgern zur Verfügung zu stellen. Die verschiedenen Gruppen der Entscheidungsträger unterscheiden sich danach, welche Daten sie benötigen. Lehrer sind hauptsächlich an der Erziehung der Schüler interessiert, Schulleiter an der Ausbildung der Lehrer, örtliche Erziehungsbehörden an der Verbesserung der Schulen, Curriculumplaner an Projektstrategien und Schulausschüsse an Prüfungen, die geeignet sind, die Leistungen der Schüler zu beurteilen. Die einzelnen Personen unterscheiden sich ferner darin, inwieweit sie den verschiedenen Daten vertrauen und wie hoch ihre Risikobereitschaft beim Handeln ist. Da wir unterschiedlichen Interessen gerecht werden müssen, versuchen wir eine umfassende Untersuchung des Projekts durchzuführen, in der wir für den Erwerb relevanter Informationen subjektive und objektive Verfahren verbinden (um eine einfache, wenn auch irreführende Dichotomie zu gebrauchen).

Unser Evaluationsplan enthält klinische, psychometrische und soziometrische Elemente. Wir versuchen, Informationen vor allem aus zwei sich

überschneidenden Schulstichproben, und zwar aus einer großen und einer kleinen Stichprobe zu gewinnen. Deshalb werden für eine bestimmte Zeit die Erfahrungen einer Zahl von Schulen genauer untersucht, während gleichzeitig aber auch genügend Informationen über die Unterrichtsabläufe einer größeren Zahl von Schulen gesammelt werden, damit Schlußfolgerungen von einer Stichprobe auf die andere gemacht werden können.

Der Entwurf sieht folgendermaßen aus:

a) *In der großen Stichprobe der Schulen (ca. 100):*
   (1) Mit Hilfe eines Fragebogens werden Daten über den Input, Kontext und die Implementation gesammelt.
   (2) Es werden Urteilsdaten von Lehrern und Schülern gesammelt.
   (3) Die Verhaltensänderungen der Lehrer und Schüler werden objektiv gemessen. (Wir haben zu Beginn dieses Jahres den Schülern Vortests gegeben, die nach dem gemeinsamen Urteil der Lehrer, der Schüler, des Projekt- und des Evaluationsteams die erwarteten Dimensionen der Änderung des Schülerverhaltens berücksichtigen. Das war ein ziemlich großer Aufwand, aber er ist gerechtfertigt, wenn er dazu beiträgt, die Auswirkungen des Curriculum auf die Schüler festzustellen und uns im nächsten Jahr die Verwendung einer kleinen, aber genauen Testbatterie ermöglicht.)
   (4) Die Veränderungen in der Lehrpraxis müssen durch den Gebrauch von speziell ausgearbeiteten Auswahl-Antwort-Aufgaben erfaßt werden, die nur einen geringen Zeitaufwand seitens des Lehrers erfordern und von den Schülern selbst gehandhabt werden können.
   (5) Die Wirkung auf die Schulen muß mit Hilfe von locker strukturierten Lehreraufzeichnungen festgehalten werden.

b) *In der kleinen Stichprobe der Schulen (ca. 20):*
   (1) Fallstudien über Art der Entscheidungsprozesse, über Kommunikationsprozesse, über Lehrerfortbildung und über das Ausmaß der Unterstützung in den örtlichen Schulbezirken.
   (2) Fallstudien an einzelnen Schulen innerhalb dieser Bezirke.
   (3) Erforschung der Dynamik einer Diskussion mit Hilfe eines Tonbands, Videorecorders und der Unterrichtsbeobachtung.

Wir müssen nun die Erfahrungen, die mehrere hundert Schulen mit dem Curriculummaterial des Humanities Projects gemacht haben, beschreiben und so darstellen, daß sie für die Entscheidungsträger nützlich sind. Unserer Meinung nach wurden die Probleme der Evaluation bisher zu sehr vereinfacht; oder aber man versuchte ausschließlich, sie mit Verfahren der empirischen Forschung zu lösen, wodurch die Funktion der Evaluation zu sehr eingeschränkt wurde. Vielleicht können bei unserem gegenwärtigen Verständnis komplexere Evaluationspläne mehr darüber Aufschluß geben,

was wir wirklich zu verändern versuchen und welche Mittel wir dazu brauchen. Deshalb haben wir einen so komplexen Evaluationsplan entwickelt. Mit dieser Evaluation wollen wir dazu beitragen, das Wechselspiel der Kräfte, die bei dieser Curriculuminnovation im Spiele sind, besser zu verstehen.