Anderson, Richard C.

Eine vergleichende Felduntersuchung: Ein Beispiel vom Biologieunterricht in der Sekundarstufe

Wulf, Christoph [Hrsg.]: Evaluation. Beschreibung und Bewertung von Unterricht, Curricula und Schulversuchen. München : R. Piper & Co. Verlag 1972, S. 288-312. - (Erziehung in Wissenschaft und Praxis; 18)

Evaluation

Beschreibung und Bewertung von Unterricht,
Curricula und Schulversuchen

Texte
herausgegeben von Christoph Wulf

R. Piper & Co. Verlag
München

ISBN 3-492-01985-4

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

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

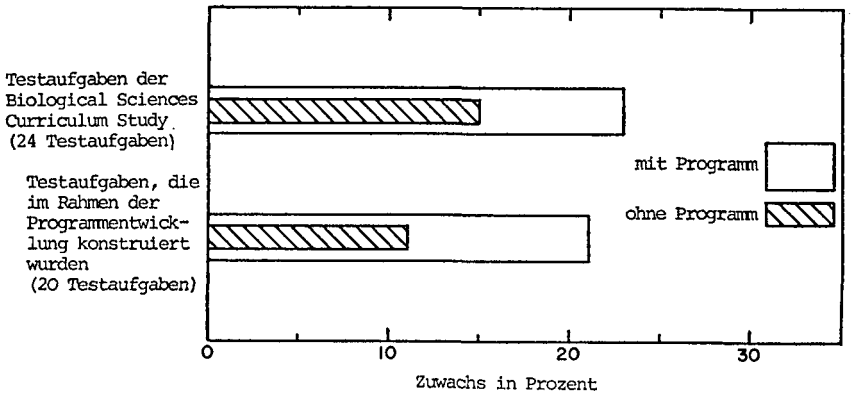
Während in der Felduntersuchung ein normenbezogener Test nach dem Auswahl-Antwort-Verfahren verwendet wurde, war der in den Voruntersuchungen des Programms verwendete Test kriteriumsbezogen; er war gekennzeichnet durch eigens konstruierte Testaufgaben. Für diese Änderung gab es zwei Gründe. Der erste 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



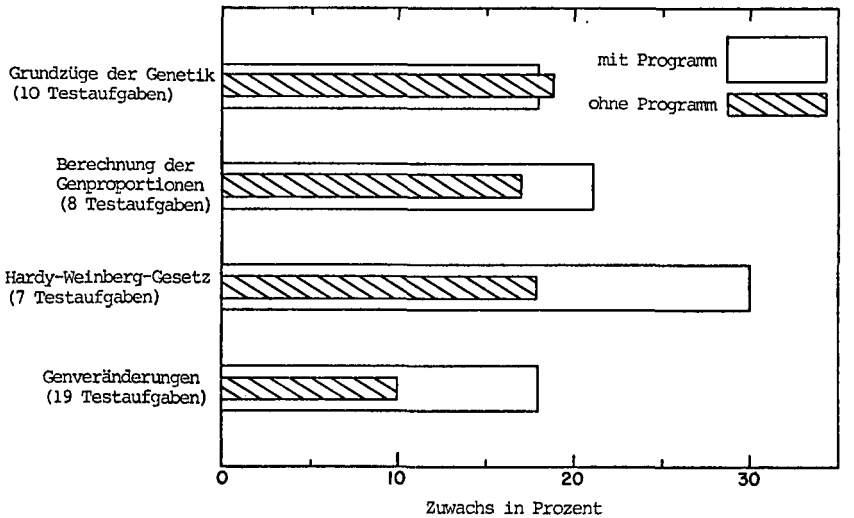
(1) die Berechnung der Genproportionen auf der Grundlage von ausgewählten Daten;
(2) die Logik des Hardy-Weinberg-Gesetzes;
(3) Faktoren, die eine Genveränderung bewirken (Mutation, Adaptation – Selektion, Wanderungssiebung, durch Zufall verursachter »genetic drift«, Paarungssiebung, Isolation).

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



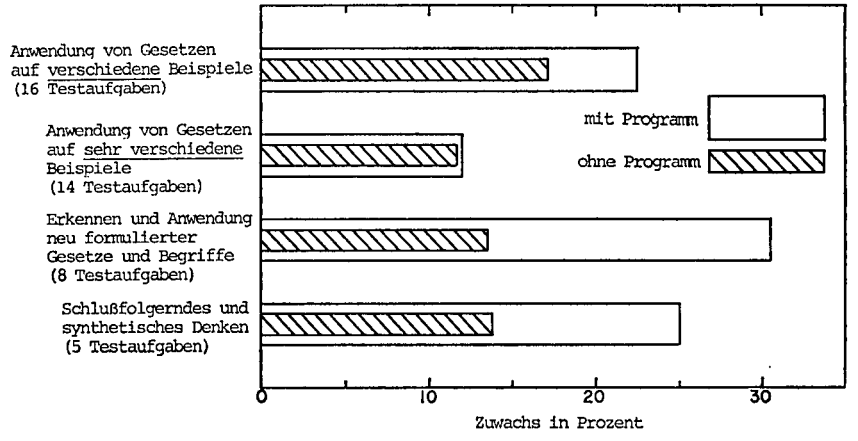
terrichtsverlaufs die Frage in einem Zusammenhang, in dem der Schüler die richtige Antwort finden muß. So muß sich auch bei einer schwierigen Testaufgabe das Ergebnis verbessern. Dessen ungeachtet, muß die Frage, was gelernt wurde, beantwortet werden. Es mag sehr wohl sein, daß die Schüler lediglich 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

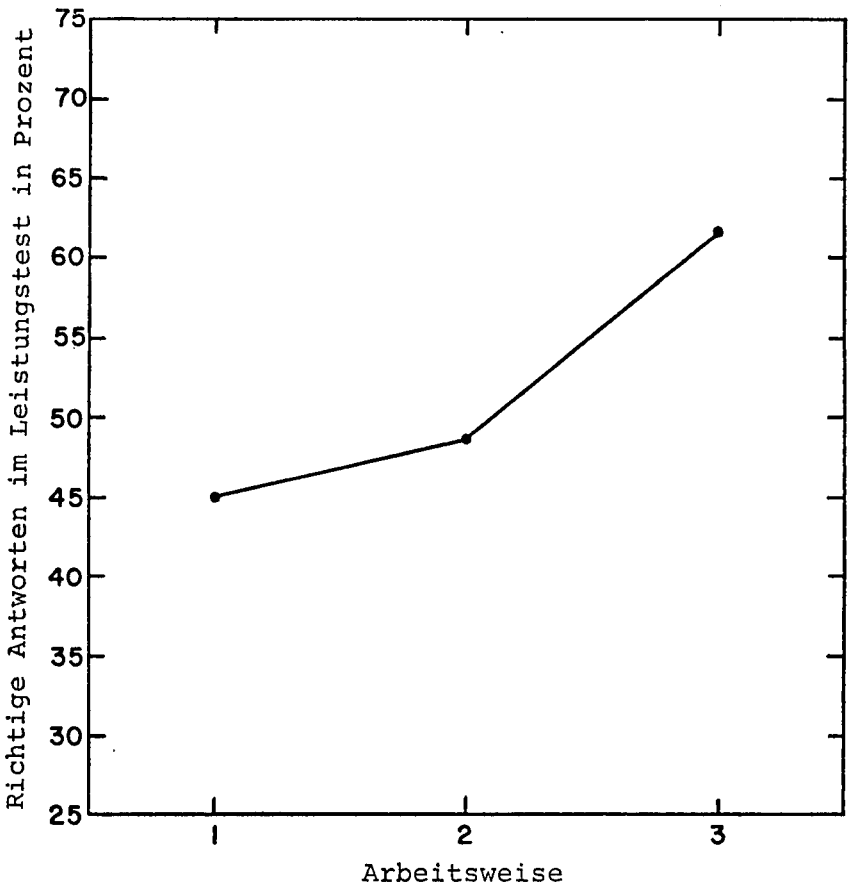
	Mit Programm		*Ohne Programm*	
	Schule A	Schule B	Schule A	Schule B
Durchschnittliche Unterrichtszeit zwischen Vor- und Nachtests	454	1031	454	1031
Zeit für Populationsgenetik				
mit Programm	151	0	0	0
andere Arbeit in Populationsgenetik . . .	52	451	228	451
insgesamt	203	451	228	451
Zeit für nicht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

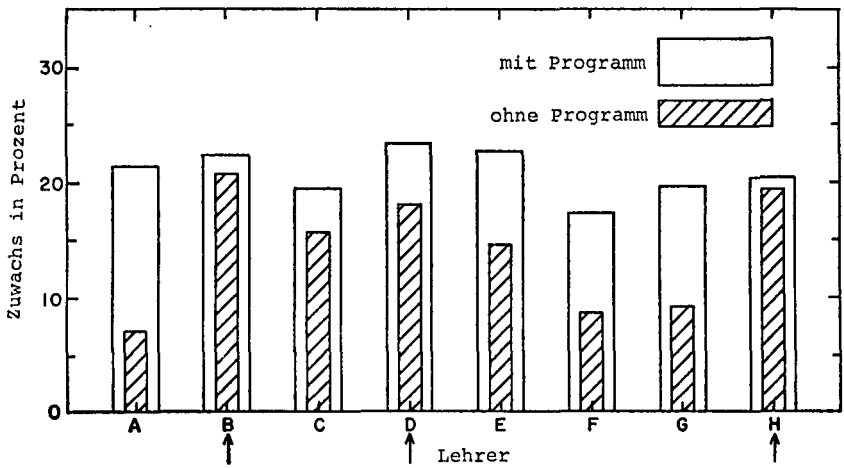
Die Lehrer wurden gefragt, ob das Programm ihren Unterricht ohne Programm beeinflußt hätte. Drei Lehrer (in Abbildung 5 mit einem Pfeil

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

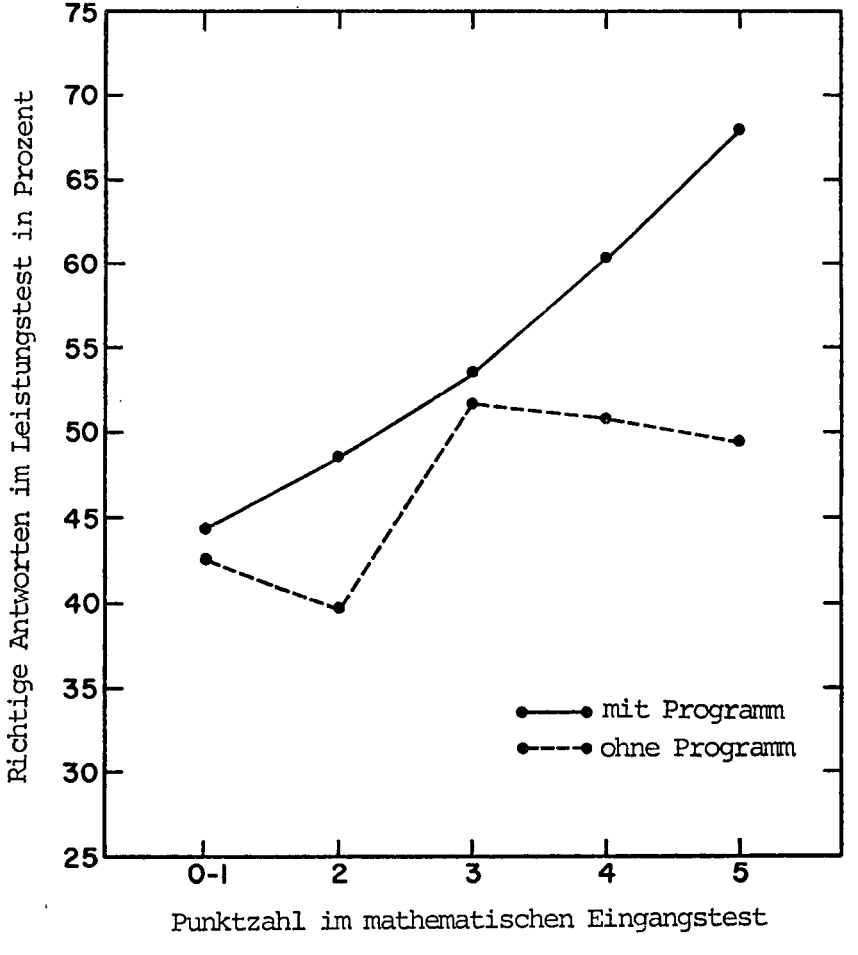
Jeder, der mit Unterrichtsplanung betraut ist, muß Voraussetzungen über den vorhandenen Kenntnisstand der Schüler machen, für die der Unterricht bestimmt ist. Das Programm in Populationsgenetik basiert auf der Voraussetzung, daß die Schüler, die damit arbeiten, mit Verhältniszahlen rechnen können und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

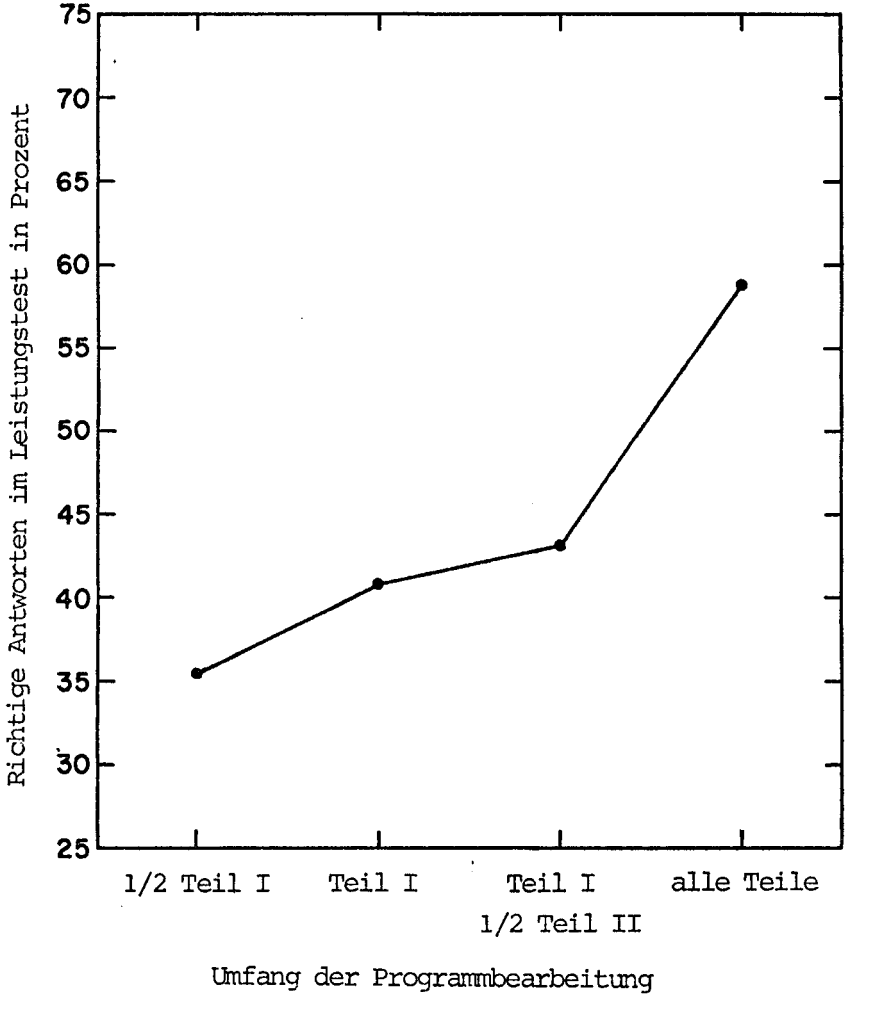
dung 6 zeigt die Leistung im Nachtest als eine Funktion der im Eingangstest festgestellten mathematischen Fertigkeiten. Das Programm war stets etwas effektiver, aber der Vorteil des Programms wirkte sich erheblich nur bei jenen Schülern aus, die in dem Eingangstest gute mathematische Fertigkeiten gezeigt 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



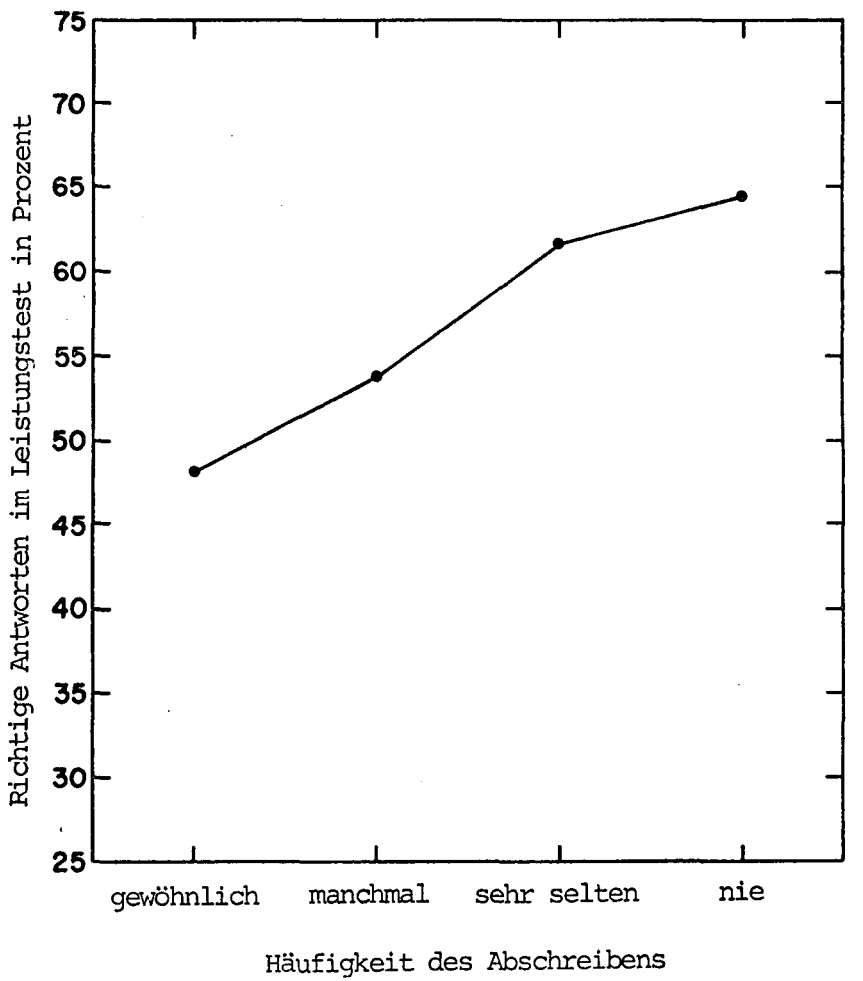
Im Fragebogen wurden die Schüler gebeten anzugeben, wie weit sie das Programm tatsächlich durchgearbeitet hatten. Etwa 75 % gaben an, das ganze Programm durchgearbeitet zu haben. Natürlich war die Leistung um so besser, je weiter das Programm durchgearbeitet worden war. Dieser Zusammenhang wird in Abbildung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

Dazu auch eine Kurzfassung: A summary of the major findings. In: The second year of Sesame Street: a continuing evaluation«.

RICHARD C. ANDERSON: Eine vergleichende Felduntersuchung
Ein Beispiel vom Biologieunterricht in der Sekundarstufe

Übersetzung von Otto Itzel (Dipl.-Soz.)

Originaltitel: A comparative field experiment: An illustration from high school biology, in: J. Th. Hastings (Ed.), Proceedings of the 1968 invitational conference on testing problems, Princeton, New Jersey: Educational Testing Service 1969.

1 Der Autor ist Gerald Faust, John Guthrie und Veronica Drantz, die bei der Entwicklung der Curriculumeinheit mitgeholfen haben, zu großem Dank verpflichtet; gleiches gilt für Gerald Faust, Marianne Roderick und Phillip Zediker, die ihm bei der Erhebung und Auswertung der Daten geholfen haben. Zu großem Dank ist er auch Robert Stake verpflichtet, der einen Entwurf dieses Beitrags kritisch begutachtete. Die hier dargestellte Untersuchung wurde teilweise von der National Science Foundation finanziell unterstützt.

2* Vergleiche dazu auch Block 1971 und Wulf 1971 b.

3 Der prozentuale Zuwachs ergibt sich aus dem tatsächlichen Zuwachs, dividiert durch die maximal erreichbare Punktzahl.

4 Zu beachten ist, daß ein Curriculum sich darauf beschränken kann, einen begrenzten Geltungsbereich eines Begriffs oder Gesetzes zu vermitteln.

5 Ein Lehrer blieb infolge eines Versehens bei der Verteilung des Leistungstests bei der Analyse der Ergebnisse unberücksichtigt.

6 Zuvor nicht erwähnt wurden drei Klassen von besonders leistungsstarken Schülern (für die das Programm eigentlich bestimmt war), die die BSCS »Blue Version« benutzten. Die zwei Klassen, die das Programm erhielten, erreichten im Nachtest einen Wert von 83,5 %, während die Klasse, die das Programm nicht erhielt, 72,9 % erreichte.