Cooley, William W.

Methoden der Evaluation von Schulinnovationen

Wulf, Christoph [Hrsg.]: Evaluation. Beschreibung und Bewertung von Unterricht, Curricula und Schulversuchen. München : R. Piper & Co. Verlag 1972, S. 313-329. - (Erziehung in Wissenschaft und Praxis; 18)

urn:nbn:de:0111-opus-14323

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

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

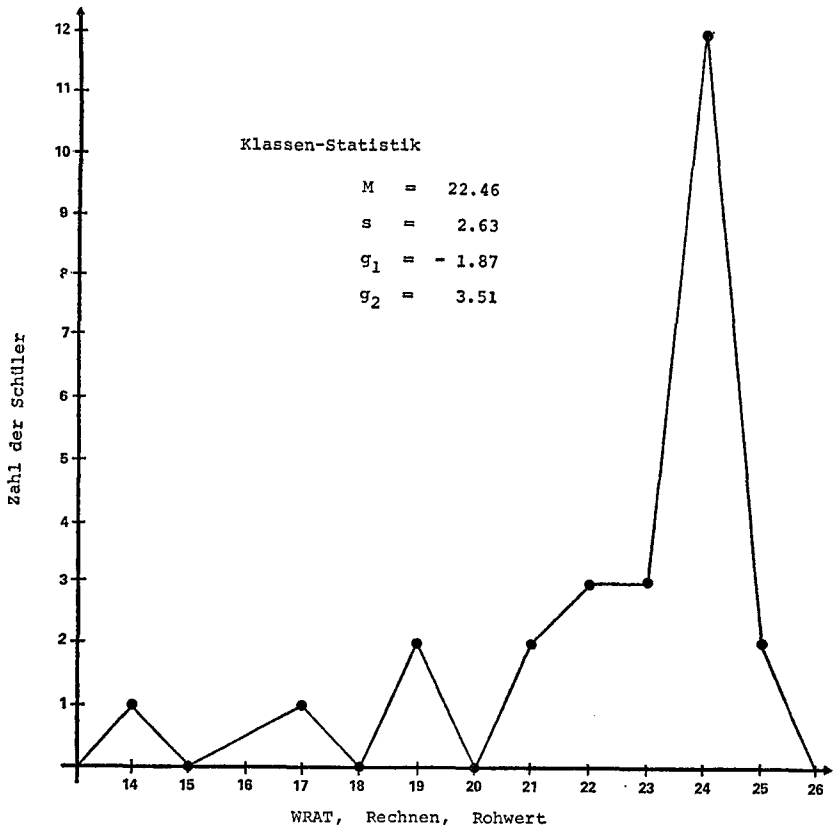

Abbildung 1
WRAT Rechnen, Verteilung für Klasse 1114
(N = 26)

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.

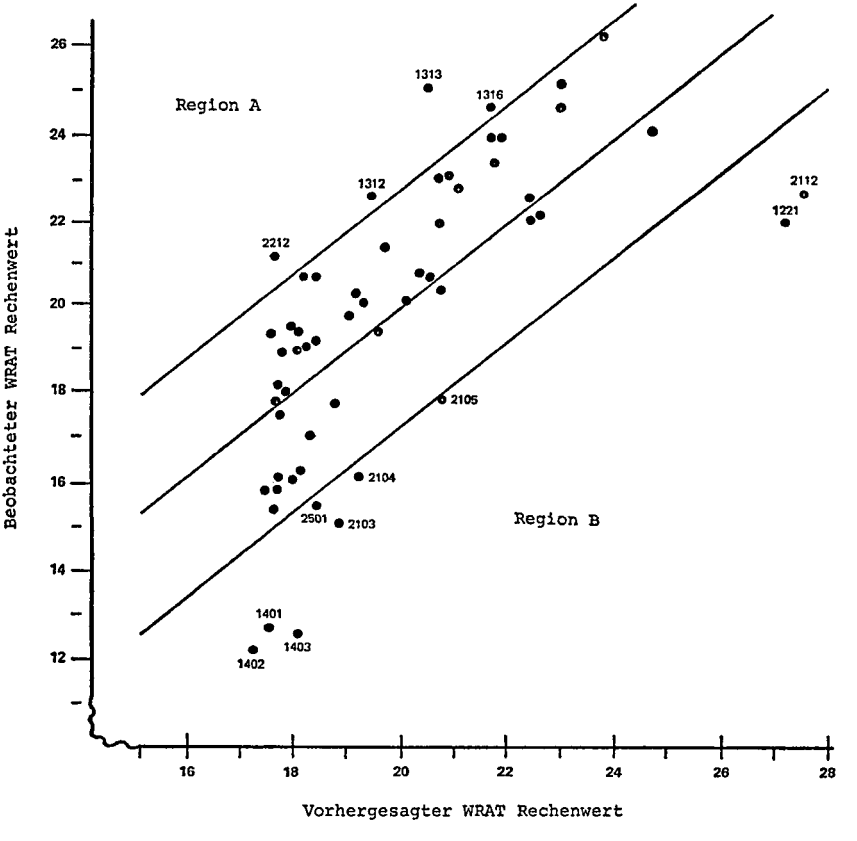

Abbildung 2

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.

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.

William W. Cooley: Methoden der Evaluation von Schulinnovationen

Übersetzung von Gudrun Eggert und dem Herausgeber.

Originaltitel: Methods of evaluating school innovations, Manuskript eines Vortrags auf der 79. Annual Convention of the American Psychological Association, Washington, D. C., 3. Sept. 1971.

1 Zur Diskussion dieses Redunanzkoeffizienten vgl. Cooley/Lohnes (1971).

2* »Mastery Learning« kann nach Auffassung von Carroll und Bloom durch die Beeinflussung der folgenden fünf Variablen für 90 % der Schüler erreicht werden:

Eignung für bestimmte Arten des Lernens,

Qualität des Unterrichts,

Fähigkeit, Unterricht zu verstehen,

Ausdauer,

zum Lernen gewährte Zeit.

Zu wichtigen Beiträgen und relevanten Forschungsergebnissen zu diesem The-

ma vgl. Blóck 1971; zur Kritik: Lee Cronbach, in: Eisner 1971, 69–75; vgl. auch Wulf 1971 b.

3* Vgl. dazu Popham/Husek 1969.

4 Als ich diesen Vortrag fertiggestellt hatte, erfuhr ich zu meiner Freude, daß meine Kollegen Glaser und Resnick (1972) gerade ihren Entwurf für einen kritischen Bericht über den Stand der Unterrichtspsychologie für die Annual Review 1972 fertiggestellt hatten, in dem sie die Untersuchungen diskutierten, bei denen »Eignung statt als Kontrollvariable als abhängige Variable behandelt wird und bei denen versucht wird, diese Variable durch unterrichtliche Maßnahmen zu beeinflussen.« Es dürfte für uns beim Learning Research and Development Center ein aufregendes Jahr werden, wenn wir diese Lücke zwischen dem, was psychometrisch sinnvoll ist, und dem, was wir über die Unterrichtspsychologie wissen, auszufüllen versuchen.