Schreiber, Walter

Methoden und Ergebnisse überregionaler Lernerfolgskontrollen in westlichen Industrieländern

Zeitschrift für Pädagogik 32 (1986) 1, S. 31-50

Zeitschrift für Pädagogik

Jahrgang 32 – Heft 1 – Februar 1986

I. Thema: Sinkende Schulleistungen – Mythos oder Realität?

II. Thema: Geschichte der Berufsbildung

III. Diskussion

IV. Besprechungen

V. Dokumentation

Contents

I. Topic: Downward Achievement Trends

II. Topic: History of Vocational Education

III. Discussion

IV. Book Reviews 129

V. Documentation

WALTER H. SCHREIBER

Methoden und Ergebnisse überregionaler Lernerfolgskontrollen in westlichen Industrieländern

Zusammenfassung

Überregionale Programme zu Messung von Schülerleistungen werden anhand von Beispielen aus den USA, Großbritannien und Australien kurz vorgestellt. Die angewandten Methoden (Item-Response-Theorie, Multiple-Matrix-Stichprobe) werden erläutert und exemplarische Resultate aufgeführt. Die Ergebnisse beziehen sich sowohl auf einzelne Querschnitte als auch auf Leistungsveränderungen, die sich in wiederholten Querschnitten gezeigt haben. In allen genannten Ländern sollten durch die Leistungsmessungen die Grundlagen dafür geschaffen werden, Defizite in Schülerleistungen festzustellen, um in der Curriculums- oder Bildungsplanung entsprechend korrigieren zu können. Die sorgfältige Planung der Programme und die Einbeziehung relevanter Gruppen in die Vorbereitung schienen ausschlaggebend für den Erfolg zu sein. Den Abschluß bildet ein Hinweis, wie nationale Leistungsmessungen in Schweden und den Niederlanden vorbereitet werden.

1. Einleitung

K. INGENKAMP hat in diesem Heft die wenigen Versuche einer Leistungsmessung in diesem Land beleuchtet und ist zu dem Ergebnis gekommen, daß aufgrund methodischer Mängel keine verläßlichen Aussagen über den Leistungsstand der deutschen Schüler gemacht werden können. Im gleichen Aufsatz wird aber deutlich gemacht, daß die Frage nach dem Erfolg des Schulunterrichts von pädagogischem, bildungspolitischem und allgemein öffentlichem Interesse ist. In anderen westlichen Industrienationen scheint im Bewußtsein der Bürger sehr viel stärker verankert zu sein, daß die Schule eine öffentliche Einrichtung ist und der Gesellschaft über ihre Erfolge Rechenschaft ablegen muß. Besonders die angelsächsischen Länder können auf eine lange Tradition von Längsschnitt- und wiederholten Querschnittsuntersuchungen zurückblicken (vgl. INGENKAMP 1985, S. 145 f.; INGENKAMP/SCHREIBER 1985, S. 223 f.).

In den USA ist in den letzten Jahren die Diskussion um die Qualität der Schulbildung wieder verstärkt geführt worden. Beispiel dafür war der Bericht *A Nation at Risk* der NATIONAL COMMISSION ON EXCELLENCE IN EDUCATION (1983). Die Aufzählung vieler Risikoindikatoren war die Grundlage einer pessimistischen Schilderung des gegenwärtigen Bildungswesens. Die Kommission sprach dann eine Reihe von Empfehlungen aus, die noch in der gleichen Woche Gegenstand einer Radioansprache von Präsident REAGAN wurden, der an die Eltern appellierte: „parents please demand these and other reforms in your local schools, and hold your local officials accountable" (*Radio Address* 1983).

Eine Reihe nationaler, einzelstaatlicher und örtlicher Beurteilungsprogramme, die über mehrere Jahre hinweg durchgeführt wurden, haben Aussagen über den Leistungsstand von Schülern erst ermöglicht. Im folgenden werden die bekanntesten Testprogramme vorgestellt, wobei ihr Aufbau und die wichtigsten Ergeb-

nisse besonders hervorgehoben werden sollen. Es handelt sich in den USA um das *National Assessment of Educational Progress* (NAEP) und das *California Achievement Program* (CAP). Eng damit verbunden sind die Anstrengungen, die der Staat Kalifornien zur Zeit unternimmt, um der Rechenschaftspflicht der Verantwortlichen im Bildungswesen gegenüber der Öffentlichkeit nachzukommen. Anschließend werden die nationalen Leistungserhebungen in England (*Assessment of Performance Unit* – APU) besprochen. Den Abschluß bilden Hinweise auf entsprechende Bemühungen in Australien, Schweden und den Niederlanden.

2. *National Assessment of Educational Progress (NAEP)*

2.1. Allgemeine Einführung

Seit 1969 führt das NAEP Leistungsuntersuchungen an Schülern in groß angelegten nationalen Stichproben durch. Damit wurde in die Tat umgesetzt, was dem US OFFICE OF EDUCATION schon 1867 bei der Gründung aufgetragen worden war (NAEP 1982a, S. xi). Die Bundesregierung verschaffte sich mit der Einrichtung und Finanzierung des NAEP die Möglichkeit, den Leistungsstand und die Leistungsveränderung zu erfassen, obwohl es kein Bundesbildungssystem gibt. Die Bildungshoheit liegt bei den Einzelstaaten; die Bundesregierung hat lediglich die Möglichkeit, auf Defizite hinzuweisen und im gegebenen Fall mit finanziellen Zuwendungen einen Ausgleich der Defizite anzustreben.

FORBES u.a. (1986) weisen in diesem Zusammenhang auf zwei Befürchtungen von Bildungspolitikern hin: eine nationale Lernerfolgsmessung könnte erstens zu einem nationalen Lehrplan führen, und die Ergebnisse der Untersuchungen könntens zweitens zu zwischenstaatlichen Vergleichen benutzt werden. Dem begegneten die Verantwortlichen des NAEP: Die Einzelstaaten bildeten keine Elemente im Stichprobenplan, und weniger der objektive Leistungsstand, vielmehr Leistungsveränderungen sollten Hauptgegenstand der Untersuchungen sein. Unabhängige Wissenschaftler, die 1981 vom NAEP beauftragt wurden, die bisherige Arbeit der Leistungsbeurteilung zu bewerten, bestätigten den ursprünglichen Beschluß. Es sollten weiterhin keine zwischenstaatlichen Vergleiche angestellt und keine entsprechenden Daten vorgelegt werden, um einer Bundeskontrolle des Bildungswesens vorzubeugen (LAPOINTE/KOFFLER 1982). Man regte allerdings an, daß solchen Staaten, die einen Vergleich mit den Bundesergebnissen wünschten, die Möglichkeit gegeben werden sollte, Items vom NAEP für eigene Leistungsbeurteilungen zu übernehmen. Bis 1983 hatten mindestens 12 Einzelstaaten NAEP repliziert, und 14 weitere haben das NAEP-Modell für eigene Zwecke adaptiert (POWER/WOOD 1984, S. 373).

Die Leistungsuntersuchungen werden periodisch in zehn Fächern vorgenommen: Lesen (LE), Schriftlicher Ausdruck (SA), Literatur (LI), Mathematik (MA), Naturwissenschaften (NW), Bildende Kunst (BI), Musik (MU), Sozialkunde (SK) *(social studies)*, Staatsbürgerkunde (ST) *(citizenship)* und Berufskunde (BK) *(occupational studies)*. *Tabelle 1* enthält eine Übersicht über die

Messungen, die in den einzelnen Jahren durchgeführt wurden (Quelle: NAEPIRS 1985).

Die Stichproben umfaßten 9-, 13- und 17jährige Schüler, teilweise auch 17jährige, die die Schule schon verlassen hatten, und junge Erwachsene im Alter von 26 bis 35 Jahren (NAEPIRS 1985). Nachdem die EDUCATION COMMISSION OF THE STATES (ECS) NAEP 15 Jahre lang durchgeführt hatte, ging aufgrund einer Ausschreibung die Verantwortung auf den EDUCATIONAL TESTING SERVICE (ETS) über. ETS erweiterte die Stichprobe um 30 Prozent durch die Einbeziehung von Schülern in den Klassen 3, 7 und 11, um den Vergleich der Ergebnisse mit anderen Leistungserhebungen zu ermöglichen (MESSICK 1984; GOODISON 1985).

Bis zum heutigen Tage wurden im Rahmen des NAEP über eine Million Schüler getestet (BEATON 1984, S. 1).

2.2. Methoden

Die Auswahl der Schüler erfolgt aufgrund einer dreistufigen Stichprobenziehung. Die Stichprobeneinheiten sind hierbei geographische Lage und Ortstyp, Schulen für jede Altersgruppe und Schüler innerhalb der Schulen (MESSICK u. a. 1983, S. 24). Die Größe der Stichprobe beträgt etwa 10000 Schüler in jeder Altersklasse (MESSICK u. a. 1983, S. 32).

Tab. 1: Beim NAEP durchgeführte Leistungsmessungen nach Jahr und curricularer Einheit.

	Curriculare Einheit									
Jahr	LE	SA	LI	MA	NW	BI	MU	SK	ST	BK
1969/70		×			×				×	
1970/71	×		×							
1971/72							×	×		
1972/73				×	×					
1973/74		×								×
1974/75	×					×				
1975/76								×	×	
1976/77					×					
1977/78				×						
1978/79		×				×	×			
1979/80	×		×							
1981/82				×				×	×	
1983/84	×	×								
1985/86	×			×	×					

Um einen möglichst großen Bereich innerhalb jedes Fachgebietes abdecken zu können, hat man beim NAEP schon früh die Methode der *Multiple Matrix Stichprobe* (MMS) angewandt. Diese Methode, Untermengen aller Items in einem Untersuchungsgebiet verschiedenen Schülern vorzulegen, geht auf theoretische Arbeiten von LORD (1955, 1962) und SHOEMAKER (1973; SHOEMAKER/SHOEMAKER 1981) zurück. Das MMS ist ein äußerst ökonomischer Ausweg aus dem Dilemma, bei Tests möglicht viele Aufgaben aus einer curricularen Einheit vorlegen zu müssen, um eine hohe *Validität* zu erreichen, während man auf der anderen Seite die Schüler nicht durch überlange Tests belasten will. Da die unterste Ebene, auf der Aussagen über Leistungen gemacht werden sollen, Klassenverbände oder ähnliche Aggregate sind, ist es ausreichend, jedem Schüler nur einen Teil der Aufgaben vorzulegen und durch Anwendung geeigneter Methoden den Leistungsstand für das Aggregat zu berechnen. LORD (LORD/NOVICK 1968, S. 256) hat gezeigt, daß der Gruppenmittelwert eines Itemclusters mit der höchsten *Reliabilität* geschätzt werden kann, wenn jedes Item von einer anderen Schülerstichprobe bearbeitet wird. ETS bemühte sich bei den von ihm durchgeführten Untersuchungen um eine Verbesserung der MMS und hat, gestützt auf Vorarbeiten von KNAPP (1968), die Methode des *balanced incomplete block spiralling* implementiert.

MESSICK u. a. (1983) erklären die Methode anhand eines Beispiels im Unterrichtsbereich Lesen bei 13jährigen. Hierfür sind 165 Aufgaben entwickelt worden, die in 15 Blöcke zu je 11 Aufgaben aufgeteilt werden. Je 3 dieser 15 Blöcke werden in einem Testbuch zusammengefaßt, dabei werden sie so permutiert, daß jedes mögliche Paar von Blöcken in mindestens einem Testbuch vorkommt. In dieser einfachen Blockordnung würden mindestens 35 verschiedene Testbücher benötigt. Bei der Testadministration innerhalb der Schule werden die Testbücher der Reihenfolge nach an die Schüler verteilt, die dann 33 anstelle von 165 Aufgaben bearbeiten. Dennoch wird das gesamte Aufgabenspektrum von den Schülern einer Schule bearbeitet, und der Leistungsstand auf Schulebene kann erfaßt werden. Eine erheblich geringere Belastung für den einzelnen Schüler, geringere Erfassungskosten und verringerte Meßfehlervarianz (vgl. hierzu FORBES u. a. 1986) sind die Vorteile dieses Verfahrens. Gegenüber dem konventionellen Matrix Sampling hat die BIB-Spiralling Methode den Vorteil, daß Korrelationen über Testbücher berechnet werden können (beim MMS sind diese nur innerhalb eines Testbuches möglich) (MESSICK 1984, S. 7); der Standardfehler verringert sich bei gleicher Stichprobengröße um 10–15 Prozent, oder die Stichprobe kann um ein Viertel verringert werden (MESSICK u. a. 1983, S. 33). Eine detaillierte Beschreibung der Anwendung kann bei GOODISON (1985) nachgelesen werden.

Da in den Leistungsmessungen des NAEP nicht der einzelne Schüler, sondern Schulen, Geschlecht der Schüler, Region und andere Faktoren bei der Berichterstattung im Vordergrund stehen sollen und die Methode des BIB-Spiralling eine Reduktion der Aufgaben je Schüler ermöglicht, werden besondere Ansprüche an die Auswertung der Daten gestellt. Die Aussagen über Gruppen stützen sich allerdings auf die Summe der gewichteten Ergebnisse der einzelnen Schüler in der Auswertung, die ihre Grundlage in der *Item Response Theory* (IRT) hat (MESSICK u. a. 1983, S. 46).

Auf der Grundlage der IRT kann eine Maßskala festgelegt werden, auf der sowohl die Fähigkeit eines Schülers als auch die Schwierigkeit eines Items abgebil-

det werden kann (TRAUB/WOLFE 1981, S. 378f.). Schülerleistungen können zwischen Gruppen verglichen werden, gleichgültig, ob die Tests zur gleichen Zeit abgelegt wurden oder ob mehrere Jahre dazwischenliegen und obwohl nicht alle Schüler alle Aufgaben in einem Fach bearbeitet haben (MESSICK u.a. 1983, S. 52).

IRT setzt unter anderem voraus, daß die Lösungen der Aufgaben den Kategorien *richtig, falsch* oder *nicht beantwortet* zuzuordnen sind, obwohl schon Vorarbeiten geleistet wurden, um Teillösungen von Aufgaben zu berücksichtigen (SAMEJIMA 1972, 1973, 1974). Die Methode ermöglicht es, die Wahrscheinlichkeit einer richtigen Antwort aufgrund von Personen und Aufgabenparametern zu ermitteln. Der Personenparameter wird in der Regel als *Fähigkeit* beschrieben; die Itemparameter im dreiparametrischen logistischen Modell, wie es bei NAEP angewandt wird, sind Schwierigkeitsgrad der Aufgaben, Aufgabendiskrimination und Pseudo-Rateparameter (LORD 1980; LORD/NOVICK 1968; HAMBLETON 1983). Die IRT ist in dieser Form eine Erweiterung des Rasch-Modells, mit dem Unterschied, daß bei diesem einparametrischen Modell der Pseudo-Rateparameter (die Wahrscheinlichkeit der richtigen Antwort eines Individuums ohne Kenntnisse hinsichtlich der Aufgabe) als Null angenommen wird und die Aufgabendiskrimination (das Ausmaß der Beziehung der Aufgabe zum erfaßten Gesamtgebiet) bei allen Aufgaben konstant ist. Eine ausführlichere Begründung dafür, warum diese Methode im NAEP angewandt wird, geben MESSICK u.a. (1983, S. 53f.).

Um weiterhin Leistungsveränderungen im Vergleich mit älteren NAEP-Erhebungen anstellen zu können, wurde es notwendig, den Einfluß der neuen Erhebungs- und Administrationsmethoden zu bestimmen. Dafür wurden zum Zeitpunkt T teilweise die gleichen Instrumente und Methoden eingesetzt wie im Zeitpunkt T-1. Da die Methoden identisch waren, konnten, bei bekanntem Standardfehler, auftretende Differenzen Leistungsveränderungen zugeschrieben werden. Parallel zur alten Methode wurden auch die neuen Instrumente zum Zeitpunkt T eingesetzt, und ein Vergleich mit den neuen Instrumenten führte zur Isolierung der Veränderung, die auf Instrumentenwechsel beruht (MESSICK u.a. 1983, S. 36f.).

2.3. Ergebnisse

Leistungsmessungen, wie sie vom NAEP durchgeführt werden, bringen eine solche Fülle von Ergebnissen mit sich, daß eine Auswahl schwerfällt. An einigen Beispielen sollen Ergebnisse der NAEP-Leistungsmessung aufgezeigt werden.

Anfang der siebziger Jahre wurden lediglich Beschreibungen des Leistungsstandes herausgegeben, die bei der Öffentlichkeit noch wenig Interesse fanden. Dies änderte sich, als nach 1975 die ersten Veröffentlichungen über Leistungs*veränderungen* erschienen. In der Erhebungsphase 1975 bis 1982 war in den meisten Unterrichtsbereichen die zweite oder dritte Untersuchung abgeschlossen, und Trendanalysen waren möglich. Zusätzlich zu Auswertungen nach demographischen Variablen wurden der Öffentlichkeit Ergebnisse von Leistungsveränderungen in vier Leistungsgruppen vorgestellt, die als Leistungsquartile definiert waren. Während der siebziger Jahre war das Gesamtergebnis in der Leseleistung anders als in Mathematik und Naturwissenschaft. Im allgemeinen verbesserte sich die Leseleistung der jüngeren Schüler, während Teenager das Niveau zu halten vermochten. Bei den Mathematikleistungen konnte während der siebziger Jahre ein Lei-

stungabfall bei älteren Schülern verzeichnet werden, während 9jährige von einem Untersuchungszeitraum zum nächsten die gleichen Leistungen erbrachten. In den Naturwissenschaften konnten 9- und 13jährige frühere Leistungen bestätigen, während sie bei 17jährigen etwas abfielen (vgl. NAEP 1982a, S. xi).

Wenn die Ergebnisse in den vier Leistungsquartilen nach Klassenstufen und Alter analysiert werden, wird etwas eindeutiger, wo Veränderungen eintraten und welche Schülergruppen Leistungsaufstieg oder -abfall verzeichneten. *Tabelle 2* zeigt eine Tabelle aus einer NAEP-Veröffentlichung (NAEP 1982c, S. 26).

Tab. 2: Nationale Veränderungen der Mittelwerte in Lesen, Naturwissenschaften und Mathematik in den niedrigsten (LG1) und höchsten (LG4) Leistungsgruppen zwischen zwei Leistungserhebungen bei 9-, 13- und 17jährigen.

	9jährige		13jährige		17jährige	
	LG1	LG4	LG1	LG4	LG1	LG4
Lesen	5,0%*	1,4%*	1,4%*	0,3%	–1,0%	–0,4%
Naturwissenschaften	1,0%	–2,5%*	1,5%*	–2,5%*	0,6%	–3,9%*
Mathematik	1,1%	–3,0%*	1,2%*	–3,4%*	–1,2%	–4,3%*

* Bedeutet signifikante Veränderungen zwischen zwei Leistungsmessungen.

Insgesamt waren die Verbesserungen für Schüler im untersten Leistungsquartil stärker als im höchsten Quartil. Die größten Verbesserungen im Lesen hatten die jüngeren Schüler, teilweise auch 13jährige.

Berücksichtigt man weitere Hintergrundvariablen, werden die Informationen noch detaillierter. Im untersten Quartil konnten sich 9-, 13- und 17jährige farbige Schüler, die sich in den ihrem Alter entsprechenden Klassen befanden, in Lesen und Mathematik verbessern. Schüler der 4. Klasse im höchsten Quartil steigerten sich ebenso in Lesen und Mathematik. Die gleichen Verbesserungen konnte man für farbige Schüler der 8. Klasse in den Leseleistungen im höchsten Quartil nachweisen. Die Leistungsanstiege der farbigen Schüler, die in Klassen waren, die ihrem Alter entsprachen, übertrafen die von weißen in der altersentsprechenden Klasse.

Sowohl farbige als auch weiße 17jährige in der 11. Klasse zeigten erhebliche Leistungsverluste im höchsten Leistungsquartil (NAEP 1982a, S. xi).

Weitere Ergebnisse finden sich in Publikationen des NAEP (1982a, 1982b, 1982c, 1983; NAEPIRS 1985), bei FORBES u. a. (1986), AHMANN (1983) sowie INGENKAMP/SCHREIBER (1985).

In einem Versuch, die Ursachen zu erklären, die zu einer Verbesserung der Schüler in der untersten Leistungskategorie führten, weisen FORBES u. a. (1986) darauf hin, daß dies mit Hilfsprogrammen von Bundesregierung und Einzelstaaten zusammenhängen könnte, die darauf abzielten, benachteiligten Jugendlichen mehr Förderung zukommen zu lassen. Im gleichen Abschnitt betonen sie, daß die jetzige Bundesregierung eben jene Programme drastisch kürzen wolle. An anderer Stelle (MESSICK u. a. 1983, S. 6) wird in diesem Zusammenhang auf den *Elementary and Secondary Education Act* von 1965 verwiesen, in dem eine ver-

stärkte Unterstützung benachteiligter Kinder gefordert wird (z. B. Behinderte, Kinder von Einwanderern und Indianern).

Die NAEP-Befunde zu den älteren Schülern bestätigen zugleich jene Beobachtung, die die aktuelle kritische Diskussion über die amerikanische High School wesentlich mit ausgelöst hat: Sowohl im Laufe der sechziger als auch der siebziger Jahre sind die Schulleistungen – wie sie sich zum Beispiel im *Scholastic Aptitude Test* spiegeln – bei den Absolventen, die zum College übergehen, kontinuierlich gesunken. Der Rückgang während der sechziger Jahre wird dabei primär auf die Änderung der getesteten Population zurückgeführt, nämlich auf die Tatsache, daß die High School einen wachsenden Anteil von Jugendlichen bis zum 12. Schuljahr und damit bis zum Übergang auf die Hochschule hält. Der Rückgang während der siebziger Jahre wird dann als eine Anpassung der Schulen an die sich in diesem Prozeß der Expansion verändernden Lernvoraussetzungen der Schüler verständlich: Der Anteil der traditionellen akademischen Kurse am Lehrplan der Oberstufe sinkt; Textbücher werden vereinfacht; die Anforderungen an häusliche Arbeiten werden reduziert; die Schulen bieten ein breites Spektrum von berufsbezogenen und sonstigen wahlfreien Kursen an, um die Lernmotivation derjenigen Schülergruppen zu stützen, die dem traditionellen akademischen Lehrplan wenig Interesse entgegenbringen. Das Jahr 1980 scheint einen Wendepunkt in dieser Entwicklung zu markieren: Daß die Durchschnittswerte im *Scholastic Aptitude Test* seither wieder steigen, kann wohl auch als Ergebnis einer selbstkritischen Reflexion der vorausgegangenen Entwicklung bewertet werden (vgl. CUSICK 1983; TURNBULL 1985).

Abschließend kann festgestellt werden, daß das NAEP sicherlich seiner Aufgabe, eine Bestandsaufnahme im Bildungsbereich vorzunehmen und in Verbindung damit Leistungsveränderungen festzustellen, nachgekommen ist. Die Leitung hat mit politischem Fingerspitzengefühl vermocht, drohende Finanzierungseinschränkungen abzuwehren und konnte aus Krisen gestärkt hervorkommen. Als nämlich die Bundesverwaltung versuchte, dem NAEP alle Mittel zu streichen, „schaltete sich der Kongreß ein und verabschiedete ein Gesetz, das an der bundesweiten Leistungsmessung festhielt und die Leitung allein in die Hände des ASSESSMENT POLICY COMMITTEE legte" (FORBES u. a. 1986).

Da Wissenschaftler, Lehrer und Laien eng in die Entwicklung der Aufgabenstellungen einbezogen wurden und auf der Basis völliger Übereinstimmung arbeiteten (POWER/WOOD 1984, S. 356), kam von dieser Seite kein Widerstand. Dies hat sicherlich zum Gelingen des NAEP beigetragen. Weiterhin wurde von seiten des NAEP immer wieder betont, daß es nicht seine Aufgabe sei, *Standards* festzulegen, sondern daß dies die Sache der Einzelstaaten sei. In diesem Zusammenhang wurde mehrfach auf die wichtige Aufgabe der örtlichen Schulbezirke hingewiesen (NAEP 1982b, S. 3).

3. California Achievement Program (CAP)

3.1. Allgemeine Einführung

1985 hatten 47 der 50 Staaten und der *District of Columbia* Testprogramme, darunter viele schon seit Jahrzehnten (FORBES u. a. 1986). Das am besten dokumentierte Schultestprogramm ist das *California Achievement Program* (CAP). Die Grundlage des CAP bildet ein Gesetz aus dem Jahr 1961, das Leistungstestprogramme in den öffentlichen Schulen des Staates forderte. Dieses Programm wurde 1969 als *California School Testing Act* revidiert. 1982 wurde durch ein weiteres Gesetz das CAP geschaffen (CAP 1982a, S. 6).

Anders als bei NAEP stehen hier einzelne Schulen und Schulbezirke im Mittelpunkt der Analysen. Da die untersten politischen Gliederungen in den USA den Schuletat bestimmen, fällt ihnen auch die Aufgabe zu, Eltern und Steuerzahlern Rechenschaft über den Leistungsstand der Schüler abzulegen. Die detaillierten Auswertungen des CAP sind als Hilfsmittel gedacht, dieser Pflicht nachzukommen.

Mittlerweile werden alle Schüler der Klassenstufen 3, 6 und 12 in öffentlichen Schulen des Staates Kalifornien jährlich in Lesen, schriftlichem Ausdruck und in Mathematik getestet. Allen Schulen gehen nach Abschluß jeder Erhebung individualisierte Auswertungshefte pro Klassenstufe zu, die einen Umfang von fast 40 Seiten haben. (Ausgesuchte Abbildungen finden sich bei INGENKAMP 1985, S. 151 f.) Im ersten Teil des Berichtes werden die Schulwerte des laufenden Jahres und der letzten beiden Jahre für jeden Untersuchungsbereich angegeben. Auf einem Vergleichsband ist dieser Wert der Schule graphisch ausgewiesen, so daß mit einem Blick der Wert der Schule mit statistisch vergleichbaren Schulen konfrontiert werden kann. Ebenso ist der Vergleich mit anderen Schulen des Schulbezirks möglich. Statistisch vergleichbare Schulen sind in diesem Zusammenhang Schulen mit vergleichbarem sozio-ökonomischem Index der Eltern, vergleichbarem Anteil von Wohlfahrtsempfängern und vergleichbaren Prozentsätzen von Schülern, die die englische Sprache nicht oder wenig beherrschen (CAP 1982b, S. 2; S. 26). Alle Werte werden auch für das betreffende Jahr und die beiden vorhergehenden berichtet.

Für den gleichen Zeitraum werden auch in jedem Fach die Prozentanteile derjenigen Schüler angegeben, die in den jeweiligen Quartilen der Gesamtverteilung im Staat pro Untersuchungsbereich anzusiedeln wären.

Im zweiten Teil des Berichtheftes werden die Ergebnisse der drei Untersuchungsbereiche sowie aller curricularen Unterbereiche aufgeführt. Im Bereich Mathematik der 6. Klasse werden zum Beispiel neben dem Gesamtwert 50 Werte für Unterbereiche (z. B. in der Mathematik: Addition/Subtraktion ganzer Zahlen; Addition/Subtraktion von Dezimalzahlen; Formen und Begriffe aus der Geometrie; Umfang, Fläche und Volumen) aufgeführt, wovon einige wiederum Durchschnittsergebnisse der kleinsten Ebene von Aufgabengruppen darstellen. Jedes Ergebnis in den Unterbereichen wird außerdem graphisch als ein Balken dargestellt (Mittelwert $\pm$ 1 Standardfehler). Wenn dieser Balken

den Schulmittelwert nicht einschließt, wird dieses Untergebiet als relative Schwäche bzw. relative Stärke ausgewiesen (für ein Beispiel im Bereich Mathematik s.INGENKAMP 1985, S. 153).

Teil 3 enthält die Ergebnisse für die Untergruppe Geschlecht, Klassenstufe der Aufnahme in die betreffende Schule, Berufsgruppe des Erziehungsberechtigten, Grad der Beherrschung der englischen Sprache und Förderung durch spezielle staatliche Hilfsprogramme. Dabei werden jeweils die Werte und Prozentanteile der Schüler in der betreffenden Schule den Werten des betreffenden Bezirks und des Staates gegenübergestellt.

Teil 4 enthält ausführliche Anleitungen zur Interpretation des Berichtsheftes, und Teil 5 enthält Tabellen zur Übertragung der Skalenwerte in auf Staatsebene bezogenen Perzentile.

3.2. Methoden

Die Methoden des MMS und der IRT wurden im CAP nach 1979/80 eingeführt. Zuvor ist für jedes Untersuchungsgebiet der Prozentsatz richtig beantworteter Fragen pro Schule und Schulbezirk ausgewiesen worden, ebenso das Perzentil der Schule oder des Bezirks innerhalb der Ergebnisse auf Staatsebene (BOCK/MISLEVY 1981, S. 66). Mit dieser Methode war allerdings die Vergleichbarkeit der Ergebnisse über die Jahre hinweg nicht gegeben. Die Berechnung von Skalenwerten auf der Grundlage der IRT erlaubt jetzt die Berechnung von Werten auf einer Skala, die zu jedem Zeitpunkt Gültigkeit hat.

Voraussetzung dafür, daß die Skalen ihre Gültigkeit behalten, ist die Unidimensionalität der sie bildenden Items. Auf diesen Punkt wurde bei der Entwicklung der Instrumente großer Wert gelegt, mit dem Resultat, daß für jedes Untersuchungsgebiet eine Vielzahl von Unterrichtsteilbereichen *(individual curricular elements)* gebildet werden. Die Itembank der dritten Klasse besteht zum Beispiel aus 62 dieser Unterrichtsteilbereiche, 17 im Lesen, 20 in Mathematik und 25 in schriftlichem Ausdruck. Jeder Teilbereich umfaßt etwa 16 Items (BOCK/MISLEVY 1981, S. 68). Die Überprüfung der Eindimensionalität dieser kleinsten Bereiche wurde faktorenanalytisch vorgenommen. „Das CAP führt ... periodische Validierungsuntersuchungen durch, um die Beständigkeit der Itemparameter nachzuprüfen und um größere Veränderungen im curricularen Vorgehen festzustellen, die die Meßgenauigkeit beeinträchtigen könnten" (FORBES u. a. 1986). Bei Bedarf werden die Aufgaben dem veränderten Curriculum angepaßt.

Hieraus wurden in unserem Beispiel 30 sich nicht überschneidende Testhefte erstellt, die je 34 Aufgaben enthalten (9 aus dem Bereich Lesen, 11 in Mathematik und 14 in schriftlichem Ausdruck). Die Testhefte sind so aufgebaut, daß möglichst jedes Item aus einem anderen Unterrichtsteilgebiet kommt und die Hefte die gleiche Anzahl leichter und schwerer Aufgaben haben. Die Testhefte wurden bei der Testadministration der Reihe nach ausgeteilt und wie konventionelle Tests bearbeitet. Die Testdauer beträgt etwa eine halbe Stunde (BOCK/MISLEVY 1981, S. 68; FORBES u. a. 1986).

Bei der Auswertung wird beim CAP das zweiparametrische logistische Modell angewandt. Da bei einer Vorstudie keine Rateeffekte gefunden wurden, verzich-

tete man auf diesen Parameter (BOCK 1979; BOCK/MISLEVY 1981, S. 75; FETLER 1982; PANDEY 1982). Für jeden Unterrichtsteilbereich wurden bei der ersten Kalibrierung Normskalen mit Mittelwert 250 und Standardabweichung 50 gewählt, um die Vergleichbarkeit zwischen Teilbereichen und Jahren zu gewährleisten. In späteren Jahren ausgetauschte Items wurden anhand der ursprünglichen Normskalen geeicht (FORBES u. a. 1986). Technische Einzelheiten der Normskalenentwicklung finden sich bei MISLEVY/BOCK (1980a, 1980b), BOCK/MISLEVY (1981) und MISLEVY u. a. (1981).

3.3. Ergebnisse

Im jährlichen Bericht über die Schülerleistungen in den Schulen Kaliforniens 1981/82 (CAP 1982a) werden u. a. folgende Ergebnisse in den einzelnen Unterrichtsbereichen besonders betont:

Lesen

- Die Testwerte beim Lesen verbesserten sich bei Schülern der 3. Klassen im fünfzehnten aufeinanderfolgenden Jahr; Verbesserungen zeigten sich in allen 27 Unterrichtsteilbereichen.
- 1981/82 wurde erstmals ein neuer Test für die 6. Klasse eingesetzt. Ergebnisse einer Vergleichsstudie zeigten, daß Verbesserungen im fünften aufeinanderfolgenden Jahr auftraten.
- Die Lesewerte der Schüler der 12. Klassen fielen während des letzten Jahres leicht ab, was einen anhaltenden Abfall seit 1975/76 bestätigt (S. 13).

Schriftlicher Ausdruck

- 1981/82 verbesserten sich die Werte der Schüler der 3. Klasse bei schriftlichem Ausdruck mit Verbesserungen in allen 34 Unterrichtsteilgebieten.
- 1981/82 wurde erstmals ein neuer Test für die 6. Klasse eingesetzt. Ergebnisse einer Vergleichsstudie zeigten, daß Verbesserungen im sechsten aufeinanderfolgenden Jahr auftraten.
- Der Gesamtwert im schriftlichen Ausdruck für Schüler der 12. Klassen stieg um 0,1 Prozent gegenüber dem Vorjahr mit Verbesserungen in 5 der 7 Unterrichtsteilgebiete.
- Ein Schüler am Median der Schüler der 12. Klasse in Kalifornien befindet sich jetzt am 35. Perzentil der nationalen Norm in schriftlichem Ausdruck, der der 6. Klasse am 57. Perzentil und ein Schüler am Median der 3. Klasse am 56. Perzentil der nationalen Norm (S. 49).

Mathematik

- Die Ergebnisse der Untersuchung werden auch als Prozentrichtigwerte ausgewiesen. In der 3. Klasse beantwortete ein „typischer" Schüler Kaliforniens 76 Prozent der Fragen richtig; das ist ein Anstieg von 1,3 Prozent gegenüber dem Vorjahr. Im Vergleich zu 1979/80 bis 1980/81 hat sich der Anstieg damit verdoppelt.
- In der 6. Klasse beantwortete ein durchschnittlicher Schüler 62,6 Prozent der Fragen richtig. Wegen der Revision der Untersuchung sind keine Langzeitdaten verfügbar. Da allerdings die alte Version einer Stichprobe der Schüler gegeben wurde, zeichnet sich ein Anstieg von 60,4 Prozent im Vorjahr auf 61,6 Prozent im Jahre 1981/82 ab.

- In der 12. Klasse ergab sich ein Anstieg um 1,2 Prozent von 1979/80 zum folgenden Jahr und ein Abfall um 0,3 Prozent während des nächsten Jahres (S. 89f.).

Die Ergebnisse sind natürlich viel ausführlicher dargestellt, wobei auch Analysen nach Untergruppen aufgezeigt wurden. In jedem Abschnitt des Berichtes sind außerdem noch Empfehlungen einer Expertenkommission aufgelistet, die den Schulbezirken als Hilfestellungen dienen sollten.

3.4. Maßnahmen zur Leistungsverbesserung

Die Ergebnisse des CAP liegen also auf mehreren Ebenen vor: auf der Schulebene, der Distriktebene und der Staatsebene. Kalifornien unternimmt derzeit einen Versuch, mit großem Aufwand die Ergebnisse zu verbessern. Das STATE DEPARTMENT OF EDUCATION hat in Zusammenarbeit mit Verwaltungsbeamten, Lehrern, Mitgliedern der Schulausschüsse und Geschäftsleuten einen Plan entwickelt, der auf die Verbesserung von Schülerleistungen abzielt (CALIFORNIA STATE DEPARTEMENT OF EDUCATION 1984a, S. iii).

Dabei werden „individuelle Leistungsberichte für jede Schule an die Schul- und Bezirksverwaltungen geschickt. Der Bericht faßt den Stand der Schule für jeden Qualitätsindikator (z. B.: Kursteilnahme, CAP-Ergebnisse, Anwesenheitsrate usw.) zusammen und beinhaltet Zustand und Ziel auf der Ebene des Staates. Jede Schule wird gebeten, lokale Ziele festzulegen, um damit zum Erreichen des Ziels innerhalb des Staates beizutragen. Es ist z. B. das Ziel im Staat, den Prozentsatz richtiger Antworten im Lesetest des CAP bis 1985/86 von 62,2% auf 62,7% richtige Antworten zu steigern. Jede Schule muß festlegen, was ihre lokalen Anstrengungen zum Erreichen dieses Zieles sein werden" (CALIFORNIA STATE DEPARTMENT OF EDUCATION 1984a, S. 2). Die geplanten Verbesserungsraten sind bewußt klein gehalten, um eine realistische Umsetzung zu ermöglichen.

Im individualisierten Teil des Bereichs sind dann z. B. verschiedene anspruchsvolle Kurse der High Schools aufgelistet, zusammen mit dem Anteil der Studenten, die staatsweit diese Kurse im Jahr 1983/84 belegten. Zielbelegungen sind dann bis in das Jahr 1990 aufgeführt. Die Schulen erhalten zusätzlich die Werte für andere Schulen, die ihnen im Hinblick auf die soziale Schichtung der Schüler usw. ähnlich sind, und sollen Schulziele einsetzen. Ähnlich werden die Ergebnisse des letzten CAP präsentiert sowie Daten über die Anwesenheitsrate der Schüler, *drop-out*-Rate, extracurriculare Aktivitäten und Umfang an Aufsätzen und Hausaufgaben (CALIFORNIA STATE DEPARTMENT OF EDUCATION 1984a, S. 37f).

Darüber hinaus werden Curricula und extracurriculare Aktivitäten der einzelnen Schulen von Begutachtungsgruppen untersucht, die anhand von detaillierten Anleitungen in Zusammenarbeit mit den Schulen Beurteilungen und Anregungen für die Steigerung der Unterrichtsqualität abfassen sollen (CALIFORNIA STATE DEPARTMENT OF EDUCATION 1984b, 1984c, FETLER 1984). Dabei werden u. a. die Ergebnisse des CAP evaluiert (1984b, S. 20; 1984c, S. 23). Die Begutachtung soll alle drei Jahre von einem Team von Lehrern, die nicht dem Lehrkörper der Schule angehören, auf kooperativer Basis durchgeführt werden.

Die Methoden sind Unterrichtsbesuche, Gespräche mit Lehrern, Schülern, Mitgliedern der Verwaltung und Eltern sowie Einsicht in sachdienliche Dokumente (GASTON/DARO 1985, S. 1).

Zwei weitere Initiativen begleiteten das Begutachtungsprogramm: die Entwicklung von modellhaften Anforderungen für den Abschluß der High School und die Entwicklung von Modell-Lehrplänen für sieben Unterrichtsfächer – Computerausbildung, Englisch, Fremdsprachen, Geschichte/Sozialkunde, Mathematik, Naturwissenschaft und visuelle/darstellende Kunst (ROST 1985, S. 8f.).

Bei den Anstrengungen, die Schülerleistungen zu verbessern, stieß man auch auf das Problem unterschiedlicher Wahlfachkombinationen innerhalb der High-Schools. In manchen Zügen *(tracks)* zeigten Schüler erheblich bessere Leistungen als Schüler mit anderen Kombinationen. Durch die Einführung eines verbindlichen Kerncurriculums glaubt man negativen Auswirkungen eines mehrzügigen Wahlfachsystems zu begegnen und *allen* Schülern intellektuell stimulierendere Kurse anzubieten (vgl. ROST 1985, S. 16). Darüber hinaus will man Schülern mit besonderen Bedürfnissen mehr als bisher entgegenkommen. Diejenigen, die sich besonders auszeichnen, sollen verstärkt die Möglichkeit haben, *advanced-placement-Kurse* zu belegen oder an Kursen teilnehmen zu können, die in Zusammenarbeit mit benachbarten Hochschulen angeboten werden. Schüler mit Defiziten will man kompensatorische Stützkurse anbieten oder das Angebot an bilingualen Kursen und Kursen für Schüler mit Englisch als zweiter Sprache verstärkt ausbauen (CALIFORNIA STATE DEPARTMENT OF EDUCATION 1984a, S. 50).

4. Großbritannien: Assessment of Performance Unit (APU)

4.1. Allgemeine Einführung

Die Assessment of Performance Unit (APU) ist eine Einheit des DEPARTMENT OF EDUCATION AND SCIENCE. Die APU ging 1975 aus der *Educational Disadvantage Unit* hervor, die nur ein Jahr früher mit dem Auftrag gegründet wurde, Defizite benachteiligter Gruppen zu beschreiben (POWER/WOOD 1984, S. 357). Dieser Aspekt kam jedoch in der neu gegründeten APU nicht zum Tragen. Leistungserhebungen werden fünf Jahre lang jährlich in Mathematik und Englisch bei 11- und 15jährigen Schülern durchgeführt, in den Naturwissenschaften an 11-, 13- und 15jährigen. Alle fünf Jahre sollen in diesen Bereichen Neuerhebungen in England, Nordirland und Wales (Schottland unternimmt eigene Untersuchungen) durchgeführt werden. 1983 gab es die erste Untersuchung der ersten Fremdsprache bei 13jährigen (Französisch, Deutsch oder Spanisch). Untersuchungen im Fach Französisch haben 1984 und 1985 stattgefunden (POWER/WOOD 1984, S. 358).

Die APU führte eine neue Generation von Leistungsuntersuchungen ein, die allerdings auf einer Reihe von groß angelegten Untersuchungen aufbaute. Unter anderem sind die *Scottish Surveys* zu nennen (vgl. INGENKAMP 1985, S. 146), die 40 Jahre lang durchgeführten Eingangstests 11 +, die bis in die sieb-

ziger Jahre in Mathematik, Englisch und Lesen gegeben wurden, sowie die Leseerhebungen der NATIONAL FOUNDATION FOR EDUCATIONAL RESEARCH (NFER) von 1948 bis 1964 (vgl. PEAKER 1966; START/WELLS 1972).

Trotz dieser Fülle an Daten hatte man beim DEPARTMENT OF EDUCATION AND SCIENCE den Eindruck, „zur Zeit gibt es wenige Fakten und Angaben darüber, auf die sich signifikante Aussagen über die Standards in unseren Schulen stützen könnten. Wir brauchen diese Information nicht nur, um den gegenwärtigen Stand zu beschreiben, sondern auch Veränderungen wahrzunehmen, wenn solche auftreten" (zit. nach GIPPS/GOLDSTEIN 1983, S. 10). Der Gebrauch des Begriffes *Standard* ist nicht zufällig. In der Mitte der siebziger Jahre gab es eine weite Debatte über diesen Begriff, der sowohl als Beschreibung des Ist- als auch des Soll-Zustandes gebraucht wurde (vgl. WOOD/POWER 1984; GIPPS/GOLDSTEIN 1983, S. 4–11).

4.2. Methoden

Die Schülerstichprobe basierte auf einer mehrstufigen stratifizierten Klumpenstichprobe mit Schultyp, Umfang der Zielgruppe innerhalb der Schulen, Region und Lage der Schulen als Stratifikatoren (SEXTON 1981, S. 2). Diese Auswahl war nicht unumstritten; Kritiker argumentierten, daß es sich hierbei mehr um eine Untersuchung als eine Leistungsmessung handele, denn Leistungsunterschiede zwischen Schultypen würden den jeweiligen Schultypen zugeschrieben und nicht z. B. unterschiedlichen Einzugsbereichen (GIPPS/GOLDSTEIN 1983, S. 70). Neben den Aufgaben im Bereich der Leistungsmessung wurden den Schülern in dem jeweiligen Fach noch Einstellungsfragen (z. B. Mathematik, Lesen) vorgelegt. Zusammenhänge zwischen Testleistungen und Einstellungen sind jedoch bisher noch nicht ausgewertet worden.

Die Größe der Stichproben wurde mit 10000 Schülern in England und 2500 in Nordirland und Wales festgelegt. Diese Entscheidung war weniger eine statistische als eine politische (vgl. GIPPS/GOLDSTEIN 1983, S. 65; S. 101).

Auch bei der APU wurde Matrix Sampling eingesetzt. Bei der dritten Leistungsmessung der 15jährigen in Mathematik bestand der Itempool aus 653 Aufgaben in 13 Unterrichtsteilgebieten. Jedes der 25 Testhefte enthielt Teilmengen aus drei dieser Teilgebiete, wobei jedes Item insgesamt in je zwei Teilmengen auftauchte.

In der zuvor erwähnten Mathematik-Untersuchung wurden in drei Teilgebieten mehr Items als bei den anderen eingesetzt, um nach der geplanten fünften Welle für jedes Teilgebiet detailliertere Aussagen machen zu können, wenn diese Rotation unterschiedlicher Gewichtung beibehalten wird (FOXMAN u. a. 1982a, S. 52).

Die Untermengen der Teilgebiete wurden so zusammengestellt, daß sie insgesamt einen vergleichbaren Schwierigkeitsgrad aufwiesen. Für jedes Teilgebiet wurde eine Rasch-Skalierung vorgenommen, deren Ergebnisse zu einer Einheitsskala in der Weise transformiert wurden, daß den durchschnittlichen Schwierigkeiten der Teilgebiete der gleiche Wert zugewiesen wurde (FOXMAN

u.a. 1982a, S. 54f.; SEXTON 1981, S. 15f.). Zur kritischen Diskussion über diese Entscheidung vgl. GIPPS/GOLDSTEIN (1983).

4.3. Ergebnisse

Nach drei Leistungsmessungen in *Mathematik* (1978, 1979, 1980) bei 11- und 15jährigen Schülern konnten folgende Ergebnisse zusammengefaßt werden:

Jungen schneiden im Durchschnitt besser ab als Mädchen: Bei den 15jährigen waren unter den besten 10 Prozent 62 Jungen und 38 Mädchen (FOXMAN u.a. 1982a, S. 145); bei den 11jährigen sind die Unterschiede jedoch noch gering (FOXMAN u.a. 1982b, S. 118).

Bezüglich der Regionen zeichnet sich bei den 15jährigen ein klares Bild ab. Die durchschnittlichen Leistungen lagen in Wales am niedrigsten, im Süden Englands am höchsten. Dies gilt für alle drei Messungen und nahezu alle Unterbereiche im Fach Mathematik. Bei den 11jährigen ergab sich ein anderes Bild: Schüler in Wales hatten im Durchschnitt die besten Ergebnisse in den leichtesten und schweren Aufgaben der Unterbereiche, die traditionelle Mathematik umfassen, während die neueren mathematischen Konzepte im Curriculum im Süden Englands mit mehr Enthusiasmus aufgenommen zu sein schienen und die Schüler dieser Region bessere durchschnittliche Ergebnisse zeigten.

Beim *schriftlichen Ausdruck* erbrachten unter den 11- und 15jährigen die Mädchen im Vergleich der ersten beiden Messungen (1979, 1980) signifikant bessere Durchschnittsleistungen als die Jungen; dieser Unterschied bestätigte sich auch bei der Leseleistung, mit zunehmender Tendenz im 2. Jahr. Im Lesen und Schreiben fielen die Durchschnittsleistungen, je mehr Schüler in der Schule zu kostenloser Schulspeisung berechtigt waren (GORMAN u.a. 1982, S. 153f.; GORMAN u.a. 1983, S. 124f.), ein Indikator für die soziale Zusammensetzung des Einzugsgebietes.

Bei 15jährigen Schülern zeigte sich auch, daß die höchsten Durchschnittswerte beim Lesen und schriftlichen Ausdruck bei den Schülern zu finden waren, die eine Kombination von Physik, Chemie und Biologie in den naturwissenschaftlichen Fächern gewählt hatten (APU 1983, S. 5).

1980 wurden auch die ersten Leistungsmessungen in *Naturwissenschaften* bei 11-, 13- und 15jährigen durchgeführt. Bei den 13jährigen konnten sich nur Jungen in einem Teilbereich signifikant verbessern, dagegen gab es erhebliche regionale Unterschiede. Schüler aus dem Süden schnitten deutlich besser ab als die im Norden (SCHOFIELD u.a. 1982, S. 191).

Da in 1982, 1983 und 1984 jeweils die ersten Wiederholungsuntersuchungen in Mathematik, Sprache und Naturwissenschaften abgeschlossen wurden, können jetzt zusammenfassende Abschlußberichte erstellt werden.

1982 gab es eine Reihe von Regionalkonferenzen zum APU-Unternehmen; die Ergebnisse der Messungen wurden insbesondere an Lehrer weitergeleitet, und man überlegte, ob nicht unabhängige Bewertungen der Berichte in Auftrag gegeben werden sollen (GIPPS/GOLDSTEIN 1983, S. 158).

GIPPS/GOLDSTEIN (1983), die vom SOCIAL SCIENCE RESEARCH COUNCIL mit der Begutachtung der APU beauftragt worden waren, betonen in ihrem Bericht, daß die Leistungsmessung allein wenig hilfreich ist, solange sie nicht von detaillierter Forschung nach der Ursache des unterschiedlichen Leistungsverlaufs be-

gleitet wird (S. 160). Die in den Leistungsmessungen gewonnenen Daten geben allerdings eine gute Basis, um die Ursachen der Leistungsunterschiede und Leistungsveränderungen in einem zweiten Schritt aufzuarbeiten.

5. Australien, Niederlande, Schweden

Neben den skizzierten Lernerfolgskontrollen in den USA und Großbritannien sind ähnliche Projekte in anderen Industrienationen zu nennen. Sie bewegen sich allerdings – wie in Schweden und in den Niederlanden – noch im Stadium der Planung oder haben – wie in Australien – nicht ein solches Ausmaß wie die Erhebungen des APU oder des NAEP erreicht, das dabei immer einen deutlichen Bezugspunkt bildet.

In den Niederlanden wurde 1985 von einem dazu eigens beauftragten Forschungsinstitut (S.C.O.) eine Pilotstudie abgeschlossen, die den Nutzen einer nationalen Leistungsmessung zu untersuchen hatte. Die Hauptarbeit lag dabei in der Entwicklung geeigneter Instrumente, die einer Stichprobe von Grundschülern vorgelegt wurden. Wenn der zuständige Parlamentsausschuß nach der angelaufenen Diskussion eine positive Entscheidung trifft, wird 1986 eine nationale Leistungsmessung in den Niederlanden beginnen, die über mehrere Jahre hinweg in den Grundschulen für verschiedene Fächer fortgesetzt werden soll.

Mit einer ähnlich terminierten Zielperspektive hat sich in Schweden seit längerem schon eine Forschungsgruppe um eine Erweiterung des NAEP-Modells bemüht, um eine Gruppe von Indikatoren zu bestimmen, die als zentrale Variablen zum Verständnis von festgestellten Leistungsunterschieden beitragen können (vgl. WEDMAN/WESTER 1981, S. 14; WEDMAN 1984, S. 12; WEDMAN u. a. 1984, S. 4). Die Verknüpfung einer Leistungsmessung mit Indikatoren ist für den Kenner des schwedischen Bildungssystems nicht überraschend, denn auf allen Ebenen war man stets bemüht, Defizite auszugleichen. Diese Entwicklung reicht von der frühen Umsetzung der Gesamtschulidee über Stützkurse für Einwandererkinder bis zur rollenden Reform: Auf Schulbezirksebene werden Innovationen ausgetestet, damit sie bei erfolgreichem Abschneiden für das gesamte Schulsystem verbindlich gemacht werden können (vgl. SCHREIBER 1979). Hinzu kommt der Auftrag aller Staatseinrichtungen und staatlich unterstützter Organisationen, in eigener Verantwortung die laufende Arbeit zu bewerten und aufgrund der Ergebnisse die weitere Arbeit zu bestimmen (vgl. WEDMAN u. a. 1984, S. 73). Da Schulen oder Schulbezirke diese Auswertungen aufgrund mangelnder Kapazität nicht eigenständig durchführen können, ist Hilfe „von oben" notwendig. Wegen der derzeit fortschreitenden Dezentralisierung des Bildungswesens (vgl. WEDMAN 1984, S. 12), die auch in den anderen Ländern einen nicht zu überschätzenden Impuls für entsprechende Maßnahmen gegeben hat, scheint dieser zusätzliche Bedarf an Informationen noch gestiegen zu sein.

Die schwedischen Bemühungen sind für uns nicht nur in theoretischer Hinsicht besonders interessant, sondern auch aufgrund gewisser Strukturaffinitäten zwischen deutschem und schwedischem Schulsystem (eine im Vergleich zu den angelsächsischen Traditionen stärkere Bedeutung der zentralen staatlichen Instan-

zen). Eine Implementierung einer überregionalen Leistungsmessung in Deutschland dürfte nur dann Erfolg haben, wenn das Indikatorenprinzip entsprechend berücksichtigt wird.

Das Beispiel der australischen Leistungsmessungen von 1975 und 1980 schließlich verdient deswegen Aufmerksamkeit, weil es eine wesentliche Bedingung für den Erfolg des gesamten Verfahrens verdeutlicht: Bei der zweiten Untersuchung war es wegen des Boykottaufrufes der australischen Lehrergewerkschaft zu erheblichen Ausfällen in der Stichprobe gekommen (vgl. POWER u. a. 1982, S. 191). Man hatte es aufgrund politischer Auseinandersetzungen unterlassen, die Lehrergewerkschaft förmlich an der Vorbereitung der zweiten Messung zu beteiligen. Gleichzeitig wurden aber, um Ärger zu vermeiden, Vergleiche zwischen Schulen, Schulsystemen, Staaten und ethnischen Gruppen bei der Auswertung der *Australian Studies in Student Performance* (ASSP) nicht zugelassen (vgl. POWER/WOOD 1984, S. 359).

Für die Gewerkschaft wurde daraufhin gerade das derart geschrumpfte Auswertungsprogramm zum Anlaß, ihre Unterstützung aufzukündigen, weil sie den Untersuchungen keine pädagogische Relevanz und keinen Nutzen für den Bildungsbereich zubilligte (vgl. HOGBEN u. a. 1982, S. 105).

6. Diskussion

Auch wenn die Beschreibung der Planung oder Durchführung von nationalen Leistungsmessungsprogrammen der verschiedenen Länder nur skizzenhaft war, können einige Beobachtungen abgeleitet werden. In allen Ländern stand die Sorge um die Qualität der Schulbildung im Vordergrund. Es ging immer darum, den Ist-Zustand festzustellen, denn nur dann ist es möglich, die Differenz zum Soll-Zustand zu definieren. Nirgendwo war es das Ziel, Schulen oder Lehrern Versäumnisse nachzuweisen. In einigen Ländern hat man sogar diesen Befürchtungen entgegengewirkt, indem Auswertungen auf dieser Ebene nicht vorgenommen wurden. Es ist weiterhin klargeworden, daß die Einbeziehung aller relevanten Gruppen im Bildungswesen auf breiter Basis im frühestmöglichen Stadium der Planung überaus wichtig war und als Garant für das Gelingen dieser großen Unternehmen bezeichnet werden kann. Dafür liefert insbesondere das NAEP anschauliche Belege. Das australische Beispiel hat hingegen gezeigt, daß das Programm ohne diese Vorkehrungen leicht scheitert.

Die Gewinner sind die Schüler. Alle Leistungsmessungsprogramme unternehmen den Versuch, Ansatzpunkte für pädagogische Hilfen zu finden – oft liegt die Betonung bei der Hilfe für unterprivilegierte Schülergruppen – oder den Erfolg pädagogischer Hilfen zu erfahren. Wiederholte, großangelegte Querschnittsuntersuchungen sind hierfür ein geeignetes Mittel. Damit gelangen Bildungsfragen auch wieder in die öffentliche Diskussion; ein solides Fundament für die Bildungsplanung wird geschaffen, und auftretende Fragen können Inhalte zusätzlicher pädagogischer Forschung werden.

Die Entwicklung neuer Methoden, insbesondere was Itemstichproben (MMS) und psychometrische Verfahren (IRT) angeht, erlaubt ökonomische Messun-

gen mit einem hohen Grad an Validität. Wenn die Diagnoseebene nicht der einzelne Schüler, sondern Gruppen von Schülern sind, dann kann auch ein breiteres Spektrum in Form von mehr Testaufgaben auf diese Gruppe verteilt werden, *ohne* irgendeine Einbuße in der Qualität der Testgütekriterien.

Wenn die Methoden, die bei den erfolgreichen Leistungsmessungen eingesetzt wurden, auch sehr anspruchsvoll sind, so kann dennoch am Beispiel des NAEP und CAP gezeigt werden, daß die Ergebnisse so dargestellt werden können, daß sie von jedem Interessierten ohne statistische Vorkenntnisse zu verstehen sind. Die Frage wird sich sicherlich stellen, welche Ergebnisse an welche Adressaten gegeben werden können. Die Zeit ist in der Bundesrepublik sicher noch nicht reif dafür, daß Ergebnisse auf Schulebene in Zeitungen veröffentlicht werden, wie es in Kalifornien und anderen Staaten der Fall ist. Es wäre aber sicherlich zu wünschen, daß sich die Mitglieder einer Lehrerkonferenz darüber Gedanken machen, warum manche Ergebnisse besser oder schlechter ausgefallen sind als erwartet.

Literatur

AHMANN, J.S.: The academic achievement of young Americans. Bloomington, In: Phi Delta Kappa Educational Foundation 1983.

APU: Assessment of Performance Unit. Summary report No. 13. Secondary language survey. London: Department of Education and Science 1983.

BEATON, A.E.: A new design for a new era. Princeton, NJ 1984.

BOCK, R.D.: A feasibility study of the one-, two-, and three-parameter logistic item-response models for the analysis and reporting of California Assessment data. Chicago: International Educational Services 1979.

BOCK, R.D./MISLEVY, R.J.: An item response curve model for matrix-sampling data: The California grade-three assessment. In: New Directions for Testing and Measurement 10 (1981), S. 65–89.

CALIFORNIA STATE DEPARTMENT OF EDUCATION: Performance Report for California schools. Indicators of quality. Sacramento, CA: California State Department of Education 1984. (a)

CALIFORNIA STATE DEPARTMENT OF EDUCATION: School program quality handbook. Preliminary. Sacramento, CA: California State Department of Education 1984. (b)

CALIFORNIA STATE DEPARTMENT OF EDUCATION: Concluding the program review. Preliminary. Sacramento, CA: California State Department of Education 1984. (c)

CAP: Student Achievement in California schools. 1981–82 annual report. Sacramento, CA: California State Department of Education 1982. (a)

CAP: Survey of Basic Skills: Grade 6–1982. School report for (Schulname). Sacramento, CA: California State Department of Education 1982. (b)

CUSICK, PH.A.: The Egalitarian Ideal and the American High School. Studies of Three Schools. New York 1983.

FORBES, R.H./PANDEY, T./CARLSON, D./MADLEY, C.: Überregionale Testprogramme im Bildungswesen der USA. In: INGENKAMP, K., u.a. (Hrsg.): Tests und Trends 5. Weinheim 1986 (in Herstellung).

FETLER, M.: Unidimensionality of the California Assessment Program third grade test. Vortrag AERA, New York 1982.

FETLER, M.: Accountability in California Public Schools: Local reactions to a statewide program. Sacramento, CA: California State Department of Education 1984.

FOXMAN, D.D./MARTINI, R.M./MITCHELL, P.: APU – Mathematical development. Secondary survey report No. 3. London: Her Majesty's Stationery Office 1982. (a)

FOXMAN, D.D./RUDDOCK, G.J./BADGER, M.E./MARTINI, R.M.: APU – Mathematical development. Primary survey report No. 1. London: Her Majesty's Stationery Office 1982. (b)

GASTON, M./DARO, P.: State of California elementary school quality program review document. Sacramento, CA: California State Department of Education 1985.

GIPPS, C./GOLDSTEIN, H.: Monitoring children. An evaluation of the Assessment of Performance Unit. London: Heinemann 1983.

GOODISON, J.: National Assessment of Educational Progress. An update of the data collection process. Princeton: ETS 1985.

GORMANN, T.P./WHITE, J./ORCHARD, L./TATE, A./SEXTON, B.: APU – Language performance in Schools. Primary survey report No. 2. London: Her Majesty's Stationery Office 1982.

GORMANN, T.P./WHITE, J./ORCHARD, L./TATE, A./SEXTON, B.: APU – Language performance in schools. Secondary survey report No. 2. London: Her Majesty's Stationery Office 1983.

HOGBEN, D./ELEY, M./POWER, C.: Reactions to the results of the testing. In: POWER, C./BAUMGART, N./ELEY, M./HEWITSON, M./HOGBEN, D./MCGRAW, B./THEOBALD, J.: National assessment in Australia: An evaluation of the Australia Studies in Student Performance project. ERDC report No. 35. Canberra: Australian Government Printing Office 1982, S. 95–115.

INGENKAMP, K.: Lehrbuch der pädagogischen Diagnostik. Weinheim 1985.

INGENKAMP, K./SCHREIBER, W.H.: Neue Methoden und Ergebnisse überregionaler Untersuchungen schulischer Lernerfolge in England und den USA. In: AURIN, K./SCHWARZ, B. (Hrsg.): Die Erforschung pädagogischer Wirkungsfelder. Freiburg: Universität 1985, S. 223–238.

KNAPP, T.R.: An application of balanced incomplete block designs to the estimation of test norms. In: Educational and Psychological Measurement 28 (1968), S. 265–272.

LAPOINTE, A.E./KOFFLER, S.L.: Your standards or mine? In: Educational Researcher 11 (1982), S. 4–11.

LORD, F.M.: Equating test scores – a maximum likelihood solution. In: Psychometrika 20 (1955), S. 193–200.

LORD, F.M.: Estimating norms by item sampling. In: Educational and Psychological Measurement 22 (1962), S. 259–267.

LORD, F.M.: Applications of item response theory to practical testing problems. Hillsdale, NJ: Erlbaum 1980.

LORD, F.M./NOVICK, M.R.: Statistical theories of mental test scores. Reading, MA: Addison Wesley 1968.

MESSICK, S.: Response to changing assessment needs: Redesign of the National Assessment of Educational Progress. Research Report. Princeton: ETS 1984.

MESSICK, S./BEATON, A./LORD, F.M.: National Assessment of Educational Progress reconsidered: A new design for a new era. Princeton: ETS 1983.

MISLEVY, R.J./BOCK, R.D.: Estimation of CAP scale scores. Technical report No. 100. Chicago: International Educational Services, 1980. (a)

MISLEVY, R.J./BOCK, R.D.: Introducing CAP scale scores. Technical report No. 101. Chicago: International Educational Services, 1980. (b)

MISLEVY, R.J./REISER, M.R./ZIMOWSKI, M.: Scale score reporting of national assessment data. Final report to the Education Commission of the States. Chicago: International Educational Services 1981.

NAEP: Reading, science and mathematics trends. A closer look. Report No. SY-RSM-50. Princeton: ETS 1982. (a)

NAEP: Standards and National Assessment: Synthesis of seven educator's responses to questions on the National Assessment's role relative to higher standards in education. No. AY-HS-50. Princeton: ETS 1982. (b)

NAEP: Technical report. Changes in student performance by achievement, class and modal grade: A different look at assessment data in reading, science and mathematics. Report No. SY-RSM-21. Princeton: ETS 1982. (c)

NAEP: The third national mathematics assessment: Results, trends and issues. Report No. 13-MA-01. Princeton: ETS 1983.

NAEPIRS: National Assessment of Educational Progress Information Retrieval System (Diskette). Ohne Ort: National Institute of Education 1985.

NATIONAL COMMISSION ON EXCELLENCE IN EDUCATION: A nation at risk: The imperative for educational reform. In: The Chronicle of Higher Education, May4 (1983), S. 11–16.

PANDEY, T.N.: Effects of violation of the assumption of unidimensionality in item response test scoring models. Vortrag AERA: New York 1982.

PEAKER, G.F.: Progress in Reading 1948–1964. Education Pamphlet No. 50. London: Her Majesty's Stationery Office 1966.

POWER, C./BAUMGART, N./ELEY, M./HEWITSON, M./HOGBEN, D./MCGRAW, B./THEOBALD, J.: National assessment in Australia: An evaluation of the Australia Studies in Student Performance project. ERDC report No. 35. Canberra: Australian Government Printing Office 1982.

POWER, C./WOOD, R.: National Assessment: A review of programs in Australia, the United Kingdom, and the United States. In: Comparative Education Review 28 (1984), S. 355–377.

Radio Address of the President to the nation on education. President Reagan discusses the findings of the National Commission on Excellence in Education. In: American Education 19, June 1983, S. 4–5.

ROST, J.C.: High School program review in the State of California. Vortrag AERA, Chicago 1985.

SAMEJIMA, F.: A general model for free-response data. Psychometrika Monograph Supplement No. 18, 1972.

SAMEJIMA, F.: Homogeneous case of the continuous response model. In: Psychometrika 38 (1973), S. 203–219.

SAMEJIMA, F.: Normal ogive model on the continuous response level in the multidimensional latent space. In: Psychometrika 39 (1974), S. 111–121.

SCHOFIELD B./MURPHY, P./JOHNSON, S./BLACK, P.: APU – Science in schools. Age 13: Report No. 1. London: Her Majesty's Stationery Office 1982.

SCHREIBER, P.: Pädagogische Entwicklungsblöcke als Ansatz zur Schulreform. Unveröffentlichte Diplomarbeit. Dortmund: Pädagogische Hochschule Ruhr 1979.

SEXTON, B.: A technical supplement on the analysis of APU monitoring in language. Ohne Ort: National Foundation for Educational Research 1981.

SHOEMAKER, D.M.: Principles and procedures of multiple matrix sampling. Cambridge, Mass.: Ballinger 1973.

SHOEMAKER, D.M./SHOEMAKER, J.P.: Applicability of multiple matrix sampling to estimating effectiveness of educational programs. In: Evaluation and Program Planning4 (1981), S. 151–161.

START, B./WELLS, K.: The trend of reading standards. London: National Foundation for Educational Research 1972.

TRAUB, R.E./WOLFE, R.G.: Latent trait theories and the assessment of educational achievement. In: Review of Research in Education 9 (1981), S. 377–425.

TURNBULL, W.W.: Student Change, Program Change: Why the SAT Scores Kept Falling. (Also College Board Report No. 85-2). Educational Testing Service, Princeton 1985.

WEDMAN, I.: Kejsarens nya kläder eller ... Ett diskussionsunderlag angående utvärdering genom utbildningsindikatorer. Umeå: Pedagogiska Institutionen 1984.
WEDMAN, I./GRANBERG, M./KARPBERG, E./SALOMONSSON, K.: Gymnasieskolan – kan de beskrivas? Om information och utbildningsindikatorer i anslutning till gymnasieskolan. Umeå: Pedagogiska Institutionen 1984.
WEDMAN, I./WESTER, A.: Kunskaper och färdigheter i ett indikatorperspektiv. Vortrag: Konferens om national assessment, Vindeln 1981.
WOOD, R./POWER, C.: Have national assessments made us any wiser about ‚standards'? In: Comparative Education 29 (1984), S. 307–321.

Abstract

On the Methods and Results of Large-Scale Assessment Programs of Student Achievement in Western Industrial Countries

Large-scale assessment programs of student achievement in the U.S., Great Britain and Australia are described. Special emphasis is put on the methods used (item response theory, multiple-matrix sampling) and a selection of results is presented. The results are based on single cross-sections as well as achievement trends from different cross-sections. In all countries mentioned, the assessment programs were to create a basis to locate deficits in student achievement in order to make appropriate corrections in curricula and in the educational system. Careful planning of the programs and the involvement of relevant groups seemed to be the guarantee for success. Finally, the planning stages for a national assessment in Sweden and the Netherlands are described.

Anschrift des Autors:
Dipl. päd. Walter H. Schreiber, M.A., Thomas-Nast-Str. 40, 6470 Landau