Sukzessiveichung von Prüfungen

Die mündliche Prüfung ist in viel zu geringem Maße objektivierbar, und die zu ihrer Verteidigung vorgebrachten Argumente sind eher gefühlsmäßiger Natur. Die Lehrer glauben an die Kraft ihrer Eingebung in der Beurteilung, die Eltern sehen in der mündlichen Prüfung eine zusätzliche Chance. Die Kommission ist aber der Ansicht, daß auf jeden Fall nach dem fünften Schuljahr eine solche weitere Chance besser auf andere Art gegeben würde, wenn sie überhaupt wünschbar ist. Die Kommission kann die Durchführung von mündlichen Prüfungen – mindestens nach nur fünf oder sechs Schuljahren – nicht befürworten. Die Berücksichtigung der Erfahrungsnote ist ein weit besseres Mittel zur Verbesserung der Auswahl als die Durchführung einer mündlichen Prüfung.

5. Ausblick

Die Kommission ist davon überzeugt, daß bei der Beachtung der unter 3 aufgestellten Forderungen die Zuverlässigkeit von Prüfungen dem erreichbaren Ideal schon näher kommt. Sie sieht es aber als wichtig an, dabei nicht stehenzubleiben, sondern beabsichtigt, gezielte Untersuchungen zur Abklärung weiterer Fragen und Details durchzuführen. Als wichtigste Fragen seien genannt:
– In welcher Form wird die Erfahrung der Vorstufe am besten einbezogen?
– Kann man die Anforderungen in den einzelnen Fächern und für die verschiedenen Stufen durch Umschreibung oder durch Beispielserien normieren?
– Welche Prüfungsarbeiten in den einzelnen Fächern versprechen guten Erfolg?
– Sind zur Vereinfachung und Verbesserung der Verfahren regionale Zusammenschlüsse denkbar?

April 1971

Für die Arbeitsgruppe
Der Leiter: E. TENGER

Sukzessiveichung von Prüfungen

August Flammer

Tests fangen an, auch in der Schule zum Alltag zu gehören. In der Psychologie wird häufig unterschieden zwischen Persönlichkeitstests und Leistungstests. Während erstere für eine verantwortbare Interpretation und oft auch für die richtige Durchführung eine umfassende fachpsychologische Bildung voraussetzen, werden die Schulleistungstests als Untergruppe der Leistungstests mit Vorteil auch durch Lehrer eingesetzt. Sie haben dabei ihre Erfahrungen aus dem Unterricht maßgebend mit ins Spiel zu bringen.
Was unterscheidet einen solchen Leistungstest von der typischen Klassenprüfung, wie sie der Lehrer etwa in den Klassen von der Mittelstufe bis zur Matura einsetzt? Primär ist es die *Eichung* oder *Normierung.* Ein Test ist typischerweise einigen tausend Schülern jener Klassenstufe(n), in der (denen) er später zur Anwendung kommen soll, durchgeführt worden; aus diesen Ergebnissen wird ein Leistungsmaßstab entwickelt, an dem der Lehrer hernach die Ergebnisse seiner Schüler messen kann. Dadurch wird die Schülerbeurteilung grundsätzlich unabhängig von der Einzelklasse; ein an sich überdurchschnittlicher Schüler einer ebenfalls guten Klasse wird nicht mehr unterschätzt, ebenso wenig wie der Star der schwachen Klasse nur von der Abhebung von seinen Klassenkameraden leben kann.
Es muß hier sogleich angefügt werden, daß ein Test diesen Namen allerdings nur verdient, wenn seine Konstruktion vor der Eichung sorgfältig durchdacht wurde. In einer wissenschatflichen Analyse wird er beispielsweise darauf hin untersucht, ob die Schwierigkeit jeder einzelnen Aufgabe dem anvisierten Schülerniveau entspreche, ob die Aufgaben eindeutig seien, ob die richtigen Lösungen tatsächlich und möglichst nur von Schülern erbracht werden, die auf oder über dem Niveau der Aufgabe stehen usw. Eine praktische Einführung für Lehrer in die Anwendung der einfacheren solcher Analysemethoden gibt WENDELER (1969). Der interessierte Leser, der mit der elemen-

taren Statistik vertraut ist, drüfte mit Gewinn HORST (1971) oder LIENERT (1969) lesen. Eine Übersicht ohne mathematische Ansprüche gibt FLAMMER (1967a; 1971a).

Der vorliegende Aufsatz enthält im ersten Teil eine Einführung in für Schultests gebräuchliche Eichmaßstäbe und im zweiten Teil praktische Vorschläge für jene Kollegen, denen die bis jetzt zur Verfügung stehenden Schulleistungstests zu wenig zahlreich oder/und für bestimmte Fragestellungen zu wenig präzis sind.

Eichmaßstäbe

Die Aussage, daß ein Schüler der 5. Primarklasse in einem Diktat sechs Fehler geschrieben hat, nachdem er noch vor zehn Wochen in einem vergleichbaren Diktat 14 Fehler gemacht hatte, sagt solange nichts aus, als wir nicht wissen, wie schwierig die beiden Diktate in Wirklichkeit sind. Auf welcher Basis läßt sich überhaupt «Schwierigkeit» erfassen? Man kann z. B. sagen: Das ist ein Diktat, das von (sozusagen) allen Sekundarschülern fehlerfrei geschrieben würde, oder: Am Ende der 5. Klasse kommen bei 90 % der Schüler nurmehr höchstens vier Fehler in einem «solchen» Diktat vor, oder: Dieses Diktat setzt insbesondere die Regeln der Zeichensetzung unter Ausschluß des Semikolons voraus, oder: Ein Schüler, der dem eben folgenden Unterricht gewachsen ist, schreibt in diesem Diktat höchstens fünf Fehler, oder: Von denen, die hier acht Fehler machen, erreichen nur noch 20 % das folgende Orthographie-Halbjahresziel. Die letzteren Aussagen sind die wertvolleren, aber sehr schwer zu belegen. Unterrichtsziele sind grundsätzlich nicht aus statistischen Durchschnittszahlen definitiv ableitbar, sondern sind Setzungen, Entscheidungsprodukte. In diesem Entscheidungsprozeß geht aber auch die Berücksichtigung von empirischen Befunden ein. Und das im Speziellen, wenn es darum geht, eine allgemeine Zielsetzung mit Hilfe konkreter Aufgaben zu «operationalisieren». In einer Jahresabschlußprüfung könnte ein Mathematiklehrer z. B. festlegen: Je 2 (ganz) falsche Aufgaben eine Note Abzug. Gelegentlich sind dann 40 % der Schüler unter der Note 4. Schwache Klasse? Schlechter Unterricht? Zu schwierige Prüfung? Ziel unrealistisch gesetzt? Hier setzt die Bedeutung des Eichmaßstabes ein.

Der Eichmaßstab bringt die einzelnen Leistungen in den Zusammenhang mit vielen Schülerleistungen, und das auf einer Skala, die auch Vergleiche zwischen verschiedenen Prüfungen zuläßt. Von allen geläufigen Skalen oder Maßstäben ist die *Prozentrangskala* vielleicht die zentralste. Wenn von 2000 Schülern 500 oder 25 % in einem Test 8 oder weniger Punkte erzielt haben, sagen wir von jenen Schülern, die genau 8 Punkte erreicht haben, sie ständen im 25. Prozentrang. *Der Prozentrang gibt an, wieviele Prozent von Schülern eine bestimmte Punktzahl nicht überschritten haben.* Die Spalten 1 bis 5 der Tabelle 1 sollen das verdeutlichen. Zu jedem Rohwert (z. B. Anzahl richtig gelöster Aufgaben) ist die Zahl der darauf entfallenden Schüler eingetragen. Diese wurde von den schwächsten bis zu den besten aufaddiert und in Prozente umgerechnet.

Tabelle 1: Prozentränge und Dezile

(1) RW	(2) f	(3) f_c	(4) f_c %	(5) PR	(6) D
14	15	2000	100.0	100	10
13	60	1985	99.2	99	10
12	312	1925	96.2	96	10
11	429	1613	80.6	81	8
10	303	1184	59.2	59	6
9	361	881	44.0	44	4
8	212	500	25.0	25	3
7	108	308	15.4	15	2
6	65	200	10.0	10	1
5	60	135	6.7	7	1
4	49	75	3.7	4	0
3	13	26	1.3	1	0
2	11	13	0.7	1	0
1	2	2	0.1	0	0
0	0	0	0.0	0	0

RW = Rohwert = Punktzahl

f = Frequenz = Häufigkeit = Anzahl Schüler, die diesen RW erreicht haben.

f_c = kumulierte Frequenz = von unten aufgezählte Häufigkeiten

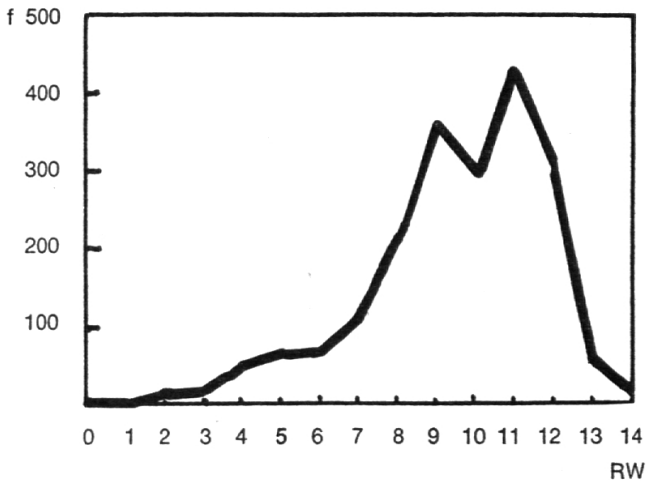
f_c % = auf 100 relativierte f_c = f_c in %

PR = Prozentrang = auf ganze Zahlen gerundete f_c %

D = Dezil = auf ganze Zehner gerundete PR

Einige Ablesebeispiele:
a) Wieviele von diesen 2000 Schülern haben genau 10 Aufgaben (irgendwelche 10!) richtig gelöst?
b) Wieviele Schüler haben 7 oder weniger Aufgaben gelöst?
c) Wieviele % der Schüler haben 9 oder mehr Aufgaben gelöst?
d) In welchem Prozentrang befindet sich ein Schüler, der 4 Aufgaben gelöst hat?
e) gehört ein Schüler, der 9 von den 14 Aufgaben gelöst hat, zur besseren oder schwächeren Hälfte der Klasse?
Antworten: a) 303; b) 308; c) 100,0 % – 25 % = 75 %; d) 4 (3,7); e) zur schwächeren, obwohl er mehr als die Hälfte der Aufgaben gelöst hat.
Die Kommastellen der % dürfen für die PR ruhig aufgegeben werden, weil sie Überexaktheit vortäuschen. Das nicht, weil 2000 Schüler zu wenig wären für eine saubere Eichung, sondern vor allem, weil die Prüfung mit ihren 14 Aufgaben, die – hier so angenommen – nur richtig oder falsch sein können, zum vorneherein sehr wenige Abstufungen zuläßt. Zwischen 59,2 % und 80,6 Prozent gibt es hier z. B. keine Zwischenstufen. Am untern Ende der Skala sind die Unterschiede allerdings kleiner. Der Test läßt offensichtlich feinere Unterscheidungen zu zwischen mehr und weniger schwachen Leistungen als z. B. zwischen mittleren Graden. Das hängt mit den Aufgabenschwierigkeiten zusammen. Die Häufung von leichteren Aufgaben in einem Test läßt feinere Unterscheidungen zwischen verschieden schwachen Schülern zu; größeres Gewicht auf die schweren Aufgaben gibt den Besten Gelegenheit, sich noch über die sehr Guten zu erheben. Wenn also der Lehrer z. B. einen Test einsetzen möchte, um weitere Unterlagen für die Gymnasialempfehlungen zu erhalten, tut er gut daran, durch Inspektion der Eichwerte oder der gelegentlich mitveröffentlichten Rohwertverteilung (vgl. Fig. 1) einen Test zu wählen, der «oben» deutlicher «differenziert». Natürlich sind die Resultate auch dann mit gewissen Vorbehalten aufzunehmen. In unserem Beispiel (Tab. 1) hat vielleicht ein Schüler zwei Aufgaben gelöst und die dritte beinahe. An einem seiner «besseren» Tage wären ihm vielleicht eine dritte oder gar eine vierte geglückt. Die Erfahrung der psychologischen Forschung hat ergeben, daß wir auch bei größter Sorgfalt mit kleinen «Störungen» rechnen müssen. (Stichwort in der test-statistischen Literatur: «Reliabilität» oder «Zuverlässigkeit») Wenn wir z. B. die sinnvolle Annahme machen, daß bei Ausschaltung aller Unregelmäßigkeiten der RW eines Schülers mit größter Wahrscheinlichkeit in einem Bereich von ± 1 um seinen effektiv erreichten Rohwert ausfallen würde (Näheres hierzu z. B. in WENDELER 1969), kann das im Fall eines effektiven RW = 10 bedeuten, daß wir statt PR=44 einen PR-Bereich von 25 bis 59 anzunehmen hätten. Nicht nur die Dezimalstellen, auch die Einerposition fängt bei so kurzen Tests (unser Beispiel ist der Handlichkeit halber besonders kurz gewählt) an, unsicher zu werden. Um den Unvorsichtigen vor falscher Sicherheit zu bewahren, werden darum gelegentlich statt Prozentränge sogar nur *Dezile* (D) angegeben, d. h. *auf Zehner aufgerundete Prozentränge* (s. Spalte 6 der Tabelle 1). In unserem Beispiel hat diese Skalenvergröberung mindestens ab RW=6 nichts geschadet. Die schwächsten Leistungen sind jedoch möglicherweise unnötig «unterdifferenziert», zu Unrecht «in den gleichen Topf geworfen».
Ebenfalls aus Gründen der Absicherung vor Überdifferenzierung hat die IMK sogar einen noch mehr vergröberten Prozentrangmaßstab gewählt, sog. *Quartile* (Q) mit zusätzlicher Kennzeichnung des 90. Prozentrangs. Der Quartilmaßstab enthält nur vier Werte, von denen jeder einem Viertel der ganzen Häufigkeitsverteilung entspricht. Bei dieser

Fig. 1: Rohwertverteilung



massiven Erweiterung der Skalenintervalle wird allerdings eine neue Gefahr der «Überinterpretation» offenkundig: Zwei Schüler, von denen einer den PR = 24 und der andere einen solchen von 26 erreicht, sind ohne weitere Informationen als etwa gleich leistungsfähig zu bezeichnen, obwohl sie auf dem Quartilsmaßstab in verschiedene Quartile fallen (Q = 1 und Q = 2); Quartilsdifferenzen werden aber wegen der weitgefaßten Kategorien als gewichtig empfunden. Man müßte wohl Grenzfälle als solche bezeichnen oder – vorteilhafter – bei den PR bleiben und darauf Intervalle (wie oben) angeben [1].

Daß die IMK-Leistungsstäbe zudem die PR-Grenze von 90 angeben, hat seinen Grund darin, daß die Tests auch als Unterlagen für die Mittelschulempfehlungen dienen können sollten. Aus den bisherigen Ausführungen ist klar geworden, daß eine solche Differenzierung im «obern» Bereich auch relativ viele schwere Aufgaben, resp. auch einen langen «rechten» Ausläufer der RW-Verteilung voraussetzt. Der Stab a der Fig. 2 trägt die 90-PR-Marke wohl zu Recht, der (erfundene!) Stab c jedoch kaum. Es wird in diesem Zusammenhang auch klar, daß ein gleicher Test für mehrere Klassenstufen wegen des (hoffentlich) zu erwartenden allgemeinen Jahresfortschrittes recht viele schwierige Aufgaben enthalten muß, um auch noch in der obern Klasse die guten Schüler einigermaßen zu differenzieren. Aus Gründen der Unterschiede zwischen den kantonalen Schulsystemen ist es aber z. B. wünschbar, daß Mittelstufentests (solange es so wenige gibt) die Klassen 4 bis 6 umfassen. Test-statistisch wünschbar wäre

Fig. 2: IMK-Leistungsstäbe

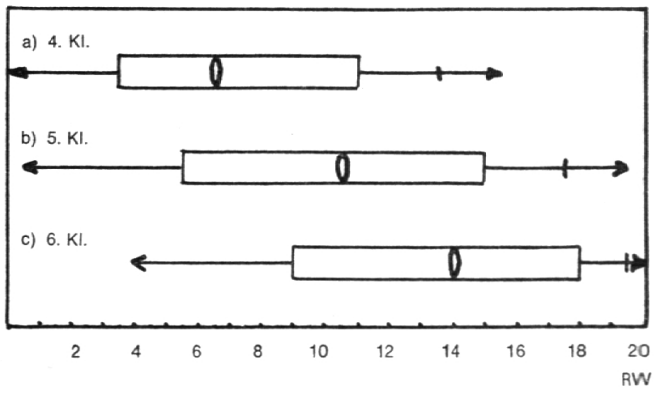


Tabelle 2: Prozentränge und Dezile von Intervallmitten

(1) RW	(2) f	(3) f_c	(4) $^1/_2$f	(5) f_c'	(6) f_c' %	(7) PR	(8) D
		2000					
14	15		7.5	1992.5	99.6	100	10
		1985					
13	60		30.0	1955.0	97.8	98	10
		1925					
12	312		156.0	1769.0	88.5	89	9
		1613					
11	429		214.5	1398.5	69.9	70	7
		1184					
10	303		151.5	1032.5	51.7	52	5
		881					
9	361		180.5	680.5	34.2	34	3
		500					
8	212		106.0	414.0	20.7	21	2
		308					
7	108		54.0	254.0	12.7	13	1
		200					
6	65		32.5	167.5	8.4	8	1
		135					
5	60		30.0	105.0	5.2	5	1
		75					
4	49		24.5	50.5	2.5	3	0
		26					
3	13		6.5	19.5	1.0	1	0
		13					
2	11		5.5	7.5	0.4	0	0
		2					
1	2		1.0	1.0	0.0	0	0
		0					
0	0		0.0	0.0	0.0	0	0

Symbole: wie in Tabelle 1, überdies:

$^1/_2$f = halbiertes f = halbe Frequenzzahl des betreffenden Rohwertes, stellt Intervallmitte dar.

f_c' = unter Berücksichtigung der Intervallmitten kumulierte Häufigkeiten, zum Beispiel
für RW = 9 : f_c' (9) = f_c (8) + $^1/_2$f (9) = 500 + 180.5 = 680.5

f_c'% = kumulierte Häufigkeitsprozente unter Berücksichtigung der Intervallmitten.

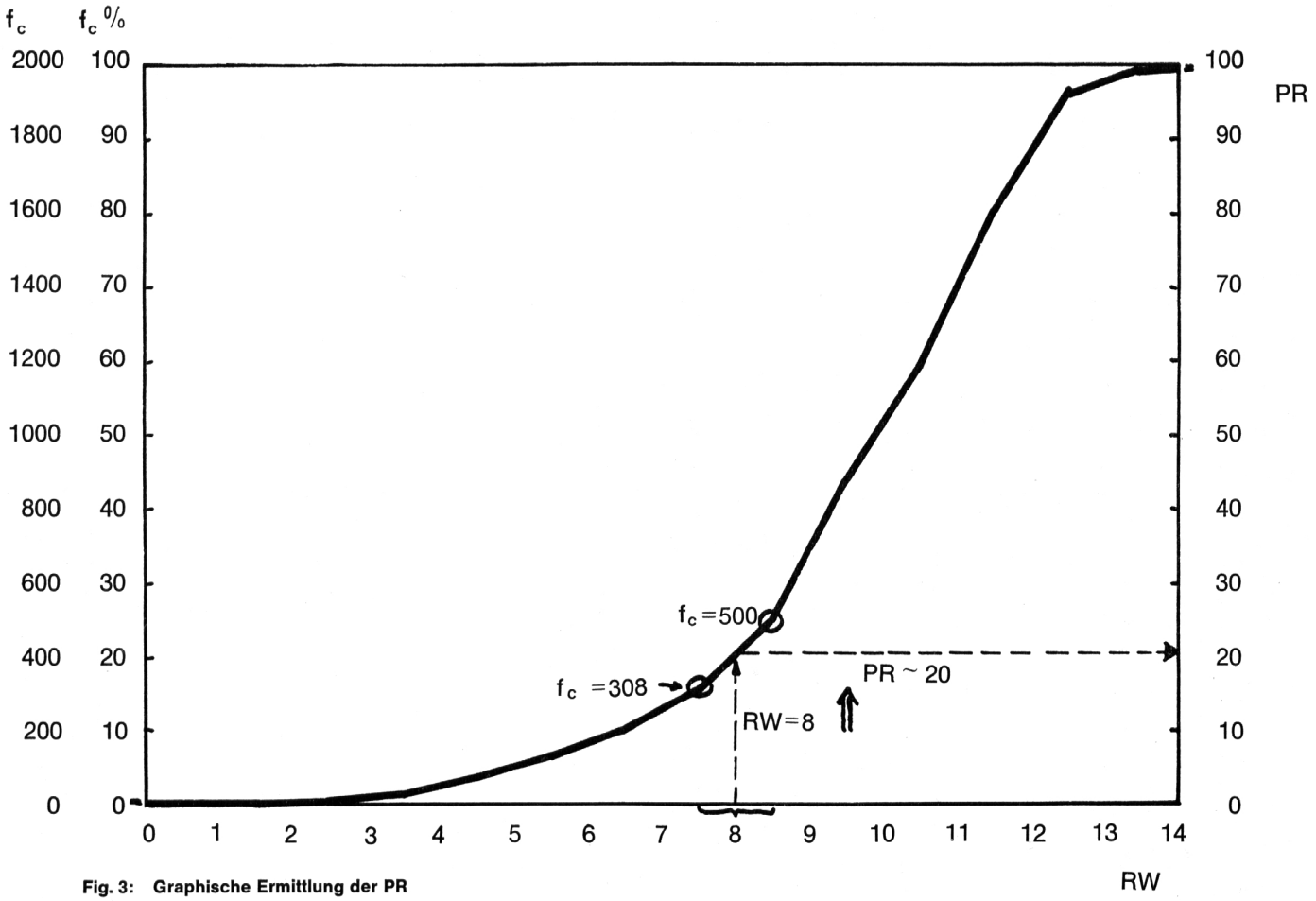


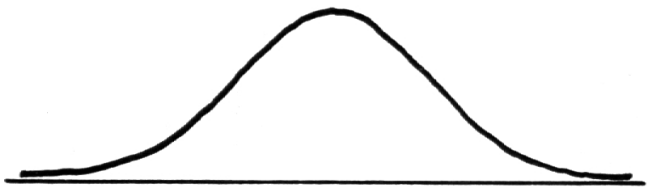
Fig. 3: Graphische Ermittlung der PR

dann aber, daß mit einem solchen Test in der 4. Klasse Gymnasial- und in der 6. Klasse Hilfsschulkandidaten zu ermitteln wären ... (Aus diesem Grund mußte z. B. der deutsche AzN 4 zum schweizerischen AzN 4–6 fast vollständig neu aufgezogen werden.) [2]

Dieser Abschnitt kann bei einer ersten Orientierungslektüre übergangen werden, ist jedoch wichtig für den Fall der praktischen Anwendung, wie sie der zweite Teil des Aufsatzes vorschlägt. – Der leichteren ersten Verständlichkeit halber ist in Tabelle 1 der PR immer so gewählt worden, daß er der oberen Grenze der jeweiligen RW-Häufigkeit entspricht. Das ist theoretisch nicht ganz befriedigend und in extremen Fällen auch praktisch irreführend. Angenommen, ein Test enthalte nur leichte Aufgaben, und 50 Prozent der Schüler hätten alle Aufgaben richtig gelöst, dann bekämen auch die Hälfte der Schüler den PR von 100. 10 von 20 Schülern könnten sich je als klassenbeste präsentieren, der Rangdurchschnitt würde 62,5. Um die größten solcher Fehler in schlecht differenzierenden Testbereichen kleinstmöglich zu halten, wird darum normalerweise die Rangmitte zum PR erhoben, im genannten Fall PR = 75. Da jeder Test nur eine endliche Anzahl von Aufgaben enthält, ist jeder PR definiert als die vorausgehende Prozentgrenze der kumulierten Häufigkeiten plus den halben Prozentanteil der gerade in Frage stehenden Klasse. In der Tabelle 2 sind die Werte entsprechend korrigiert, wobei, um Kumulation von Rundungsfehlern möglichst zu vermeiden, die Häufigkeitshalbierung bereits bei den f ansetzt und nicht erst bei den Prozentzahlen. Eine zweite, sehr elegante Möglichkeit besteht natürlich darin, die f_c% der Tabelle 1 graphisch darzustellen und darauf die Intervallmitten abzulesen (Fig. 3).

Manchem Leser ist vielleicht aufgefallen, daß die Fig. 1 bei weitem nicht der berühmten sogenannten Normalverteilung oder Gauss'schen Glockenkurve entspricht [3] (Figur 4). Psychologisch bedeutet die Annahme der Normalverteilung, daß auf einem Merkmal (z. B. Lesetempo, Ängstlichkeit, Stilempfinden, mathematische Begabung) sehr viele Menschen durchschnittlich sind und desto seltener werden, je größer die Abweichung nach oben oder nach unten

Fig. 4: Normalverteilung



wird. Überdies wäre die Abweichung nach unten und nach oben etwa gleich häufig zu erwarten (Symmetrie der Verteilung). Vielen scheinen diese Aussagen sinnvoll, einige sehen darin sogar ein (statistisch unbeweisbares) psychologisches Naturgesetz. Warum entspricht aber die Fig.1 nicht der Normalverteilung? Liegt der Beweis für eine Ausnahme vor? Wir haben oben gesehen, daß ein Übergewicht von leichten Aufgaben einen langen linken Ausläufer der Rohwertverteilung verursacht. Sehr wenige mittelschwere Aufgaben ergäben z. B. einen sehr hohen schmalen Gipfel usw. (s. FLAMMER 1967b). Die Auswahl von Aufgaben für einen Test ist aber willkürlich, und darum ist es die Verteilungsform auch. Wenn wir die Normalverteilung in einem konkreten Fall wirklich allen andern vorziehen, dann können wir entweder solange verschiedene Aufgabenkombinationen ausprobieren, bis wir unser Ziel mehr oder weniger erreicht haben, oder wir können die Skala transformieren. Mit vorbereiteten statistischen Tafeln ist das einfach.

Tafel 1 gibt für die sogenannte *T-Skala,* auf der die Häufigkeiten normalverteilt sind, die zugehörigen Prozentränge. Die T-Skala mit dem Mittelwert 50 und ihrer typischen Breite [4] wird aus Handlichkeitsgründen in der Teststatistik sehr häufig gebraucht, obwohl sich die Normalverteilungsform natürlich über jeder linearen Transformation der Skala auch ergeben würde. Orientierungshalber sind in Tafel 1 auch einige Punkte der IQ-Skala angegeben, obwohl wir hier von Schulleistungstests sprechen [5]. (Psychologen vermeiden es heute in zunehmendem Maß, von IQ zu sprechen, weil es die Intelligenz als einheitliche Fähigkeit nach den bisherigen empirischen Untersuchungen nicht gibt; man spricht eher von Intelligenzfaktoren oder von – sprachlichen, schlußfolgernden usw. – Seiten der Intelligenz.)

Der Hauptvorteil der T-Skala besteht darin,

Tafel 1: Tafel zur Umwandlung von Prozentrangplätzen in T-Werte

Prozentrang	T	IQ	Prozentrang	T	IQ
1	27		50	50	100
2	30		51	50	
3	31		52	51	
4	33		53	51	
5	34		54	51	
6	35		55	51	
7	35		56	52	
8	36		57	52	
9	37		58	52	
10	37		59	52	
11	38		60	53	
12	38		61	53	
13	39		62	53	
14	39		63	53	
15	40		64	54	
16	40	85	65	54	
17	41		66	54	
18	41		67	54	
19	41		68	55	
20	42		69	55	
21	42		70	55	
22	42		71	56	
23	43		72	56	
24	43		73	56	
25	43		74	56	
26	44		75	57	
27	44		76	57	
28	44		77	57	
29	45		78	58	
30	45		79	58	
31	45		80	58	
32	45		81	59	
33	46		82	59	
34	46		83	60	
35	46		84	60	115
36	47		85	60	
37	47		86	61	
38	47		87	61	
39	47		88	62	
40	48		89	62	
41	48		90	63	
42	48		91	63	
43	48		92	64	
44	49		93	65	
45	49		94	66	
46	49		95	66	
47	49		96	68	
48	50		97	69	
49	50		98	71	
50	50	100	99	73	

daß sie im Mittelbereich viele Schüler in gleichen Intervallen faßt und gegen die Enden mehr Abstände zwischen die Leistungen bringt, während die PR-Skala (resp. auch D und Q) die Schülerleistungen darstellt, als ob in der Durchschnittsklasse der Unterschied z. B. zwischen dem Zweit- und dem Drittbesten gleich groß wäre wie der zwischen dem 11. und dem 12. Schüler (sogenannte Rechteckverteilung über der PR-Skala).

An dieser Stelle ist wenigstens ein kurzes Wort zur *Notenskala* fällig. Welcher der eben genannten Skalenkategorien ist sie zuzuordnen? In den üblichen Zeugnissen ist sie nicht über eine Verteilung der Häufigkeiten definiert, sondern durch Adjektive wie «sehr gut», «gut» usw. Da man sich beliebig lang darüber streiten kann, wann ein Schüler «genügend bis gut» ist, verwundert es nicht, daß die Skala auch von verschiedenen Lehrern sehr unterschiedlich gehandhabt wird. Von Normalverteilung kann bei der gegenwärtigen Handhabung ebenfalls nicht die Rede sein. Ausgehend von empirischen Belegen zu diesen Aussagen gelangt darum der Autor an einem andern Ort (1971b) zur Forderung, die Notenskala über eine Häufigkeitsverteilung zu definieren, und zwar über diejenige (allenfalls unregelmäßige), von der die gegenwärtigen, jedoch von Lehrer zu Lehrer und von Schule zu Schule schwankenden Praktiken am wenigsten abweichen. Dadurch würde die Skala nicht nur einheitlicher, es ließe sich auch eine exakte Zuordnung zwischen Testnormen und Notenskala errechnen (und z. B. in Tafel 1 anfügen). Dabei darf nicht übersehen werden, daß eine allfällige Norm-Häufigkeitsverteilung nur für die gesamte Schulpopulation gelten würde, sozusagen als allgemeine Vorstellung für den Lehrer, nicht aber für die konkrete Einzelklasse. Sonst müßte zwischen den Klassenkameraden ein unerbittlicher Wettkampf darum entstehen, wer das nächste Jahr repetieren muß . . . Es wäre wohl überhaupt angemessen, die Steigungslimiten anhand von Minimallernzielen zu definieren (und die bisher so mehrdeutige Notenskala mit ihren vielen psychologischen Nachteilen für die schwächeren Schüler am Ende gar fallen zu lassen). Doch das ist nicht der Gegenstand dieses Aufsatzes.

Sukzessive Eichung von Prüfungen durch den Lehrer

Die Eichung der Tests ermöglicht den Vergleich der Leistung einzelner Schüler sowohl mit der ihrer Kameraden außerhalb der Klasse als auch mit ihren eigenen Leistungen in andern Tests, z. B. nach Ablauf eines Jahres. Das sind Eigenschaften, die mancher Lehrer auch für seine selbsterstellten Prüfungen wünschen würde. In einem etwas eingeschränkten Ausmaß ist das auch möglich, und zwar ohne daß über die ganze Schweiz sorgfältig verteilt etwa hundert Kollegen angeschrieben und gebeten werden müssen, eine bestimmte Prüfung durchzuführen und einer zentralen Stelle zur Verrechnung einzuschicken. (vgl. die enorme Arbeitsleistung der IMK im letzten Jahrzehnt!)

Der einfachste Fall einer Annäherung an die Eichung besteht darin, daß der Lehrer von seinen wichtigsten Prüfungen, zwischen denen er Vergleiche anstellen möchte, die Schülerergebnisse in Prozenträngе umrechnet. Durch unterschiedliche Kombination der Aufgabenschwierigkeiten verursachte Unregelmäßigkeiten in der RW-Verteilung werden dadurch bereits ausgemerzt. Der Maßstab ist aber dennoch nur klassenintern: die Schüler einer überdurchschnittlichen Klasse erhalten dann zum Beispiel im allgemeinen zu schlechte Werte. Wenn aber der Lehrer seine Prüfung mitsamt den Ergebnissen seiner diesjährigen Klasse aufbewahrt, kann er im nächsten Jahr die Prozenträngе bereits auf den Rohwerten von zwei Klassen basieren. Tut er das über mehrere Jahre hinweg, wächst nicht nur einfach die Eichstichprobe, sondern auch die Wahrscheinlichkeit, daß Verzerrungen durch nach unten oder oben abweichende Klassen (oder auch durch breit oder schmalstreuende Klassen) gegenseitig ausgewogen werden. Eine dritte sehr wirkungsvolle Maßnahme besteht darin, daß verschiedene Kollegen die gleichen Prüfungen verwenden und so die Eichstichprobe rasch anwachsen lassen. Auch wenn solche Eichstichproben nicht rasch auf einige hundert ansteigen, stünde doch auch ein solches vereinfachtes Verfahren einer Aufnahmeprüfung besser an als A priori-Festsetzungen wie «x Fehler eine Note Abzug, Noten aller Fächer mit Gewichten gemittelt, und wer unter 3,8 ist, fällt durch», was dann ja nachträglich auch gelegentlich revidiert werden muß . . .

Soviel zur Idee. Im nächsten kürzeren Teil sei ein (erfundenes) Beispiel zur Sicherstellung des Verständnisses der minimalen Technik durchgeführt; hernach seien einige Punkte genannt, die bei einem solchen Unternehmen besondere Beachtung verdienen.

Beispiel: Die 28 Schüler der ersten Sekundarklasse des Lehrers A hätten am Ende des Schuljahres in einer zusammenfassenden Französischprüfung folgende Zahl von Fehlern geschrieben: 4, 16, 13, 8, 9, 17, 4, 15, 14, 9, 10, 13, 8, 19, 13, 15, 10, 7, 17, 8, 12, 6, 12, 13, 10, 15, 12, 9. Der Leser ist eingeladen, die Tabelle 3 erst einzusehen, wenn er die PR selber ebenfalls errechnet hat.

Als Resultat der ersten Stufe dieses Beispiels können die Spalten 1 bis 7 der Tabelle 3 gelten. Daß im Gegensatz zu den ersten beiden Tabellen hier die hohen RW unten stehen, hat seinen Grund darin, daß hier Fehler statt Gutpunkte angegeben wurden. Es ist eine Konvention, die PR immer so zu rechnen, daß schwache Schüler niedere und gute Schüler hohe PR erhalten.

Zur Fortsetzung des Beispiels nehmen wir an, es hätte im nächsten Jahr wieder eine 1. Sekundarklasse, diesmal 22 Schüler, zur gleichen Prüfung folgende Fehlerzahlen gegeben: 9, 17, 3, 20, 16, 10, 14, 7, 18, 11, 8, 14, 11, 17, 0, 15, 14, 9, 19, 12, 6, 11. (Der Leser ist wieder eingeladen mitzurechnen). – Die Spalten 8 bis 14 der Tabelle 3 geben die Zwischenergebnisse und die neuen Prozenträngе.

Offensichtlich hat diese neue Klasse eine größere Zahl schlechter Resultate beigesteuert. Verifizieren Sie zum Beispiel die Veränderung des Prozentranges für 16 Fehler. Während letztes Jahr ein Schüler mit dieser Fehlerzahl nur auf den Prozentrang 9 kam, kommt er dieses Jahr durch die relative Ausbalancierung der beiden Klassen auf PR = 18. (Ohne Mitverrechnung der letztjährigen Klasse würde sich aber im zweiten Jahr für RW = 16 ein sehr viel höherer PR ergeben, d. h. Überschätzung schwächerer Schüler in schwachen Klassen.) [6]

Wir verfolgen unser fiktives Beispiel bis ins 3. Jahr, d. h. bis zum Augenblick, da die Ergebnisse von 3 ersten Sekundarklassen zur kombinierten Maßstaberstellung zur Verfü-

Tabelle 3: Sukzessive Eichung

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)	(20)	(21)
RW	L_1*	f_1	f_{c1}	$^1/_2f_1$	f_{c1}'	PR_1	L_2*	f_2	$f_{T,2}$ **	$f_{c\,T,2}$	$^1/_2f_{T,2}$	$f_{c\,T,2}'$	$PR_{T,2}$	L_3*	f_3	$f_{T,3}$ ***	$f_{c\,T,3}$	$^1/_2f_{T,3}$	$f_{c\,T,2}'$	$PR_{T,3}$
0		0	28	0	28.0	100	I	1	1	50	0.5	49.5	99		0	1	75	0.5	74.5	100
1		0	28	0	28.0	100		0	0	49	0	49.0	98		0	0	74	0	74.0	99
2		0	28	0	28.0	100		0	0	49	0	49.0	98		0	0	74	0	74.0	99
3		0	28	0	28.0	100	I	1	1	49	0.5	48.5	97		0	1	74	0.5	73.5	98
4	II	2	28	1.0	27.0	97		0	2	48	1.0	47.0	94		0	2	73	1.0	72.0	96
5		0	26	0	26.0	93		0	0	46	0	46.0	92	I	1	1	71	0.5	70.5	94
6	I	1	26	0.5	25.5	91	I	1	2	46	1.0	45.0	90		0	2	70	1.0	69.0	92
7	I	1	25	0.5	24.5	89	I	1	2	44	1.0	43.0	86	I	1	3	68	1.5	66.5	90
8	III	3	24	1.5	22.5	81	I	1	4	42	2.0	40.0	80	II	2	6	65	3.0	62.0	83
9	III	3	21	1.5	19.5	69	II	2	5	38	2.5	35.5	71	II	2	7	59	3.5	55.5	73
10	III	3	18	1.5	16.5	59	I	1	4	33	2.0	31.0	62	III	3	7	52	3.5	48.5	65
11		0	15	0	15.0	53	III	3	3	29	1.5	27.5	55	IIII	4	7	45	3.5	41.5	55
12	III	3	15	1.5	13.5	48	I	1	4	26	2.0	24.0	48	IIIII	5	9	38	4.5	33.5	45
13	IIII	4	12	2.0	10.0	36		0	4	22	2.0	20.0	40	III	3	7	29	3.5	25.5	34
14	II	2	8	1.0	7.0	25	II	2	4	18	2.0	16.0	32	I	1	5	22	2.5	19.5	26
15	III	3	6	1.5	4.5	16	I	1	4	14	2.0	12.0	24	I	1	5	17	2.5	14.5	18
16	I	1	3	0.5	2.5	9	I	1	2	10	1.0	9.0	18	II	2	4	12	2.0	10.0	13
17	I	1	2	0.5	1.5	5	III	3	4	8	2.0	6.0	12		0	4	8	2.0	4.0	5
18		0	1	0	1.0	3	I	1	1	4	0.5	3.5	7		0	1	4	0.5	3.5	5
19	I	1	1	0.5	0.5	2	I	1	2	3	1.0	2.0	4		0	2	3	1.0	2.0	3
20		0	0	0	0	0	I	1	1	1	0.5	0.5	1		0	1	1	0.5	0.5	1

gung stehen. Wir nehmen an, die Klasse hätte vor allem Durchschnittsschüler, d. h. wenige Abweichungen nach unten und nach oben. Die Werte seien: 10, 7, 12, 15, 11, 9, 12, 16, 9, 14, 11, 13, 8, 10, 12, 5, 13, 8, 10, 11, 16, 12, 13, 11, 12. – Ergebnis in den Spalten 15 bis 21. Je größer die bereits vorliegende Stichprobe ist, desto weniger wirken sich dazukommende Klassen aus; so verursacht diese durchschnittsbetonte Klasse nur noch relativ geringe Vergrößerungen der Abstände im Mittelbereich.

Es scheint unumgänglich, abschließend einige an sich selbstverständliche Bedingungen sicherheitshalber zu diskutieren:

1. Die Prüfung darf nie verändert werden, resp. die sukzessive Eichung hat neu zu beginnen, wenn sich Änderungen aufgedrängt haben. Es lohnt sich darum, die Prüfung sorgfältig aufzubauen, und notwendige Änderungen gleich anzubringen und mit dem Start der sukzessiven Eichung bis zur nächsten Klasse zu warten.
2. Es muß klar sein, ob den Schülern soviel Zeit zum Bearbeiten der Prüfung gelassen wird, wie sie wünschen. Wenn ja, dann sollen sie auch jedesmal nicht zum Abschließen gedrängt werden. Da dies häufig zu umständlich ist, empfiehlt sich eine großzügige Zeitbegrenzung. Diese ist aber exakt festzulegen und jedesmal genau einzuhalten. Am besten wird sie z. B. auf dem vervielfältigten Prüfungsblatt von Anfang an notiert. Versehentliche Abweichungen in der Durchführung müssen den Ausschluß der Resultate von der sukzessiven Eichung zur Folge haben.
3. Die Durchführung der Prüfung muß immer unter vernünftigerweise gleichen Bedingungen erfolgen. Ausnahmen (z. B. Unterbrechung der Klasse durch hohen Besuch) machen die Werte für die Eichung unbrauchbar. Betrifft die Ausnahme nur einzelne Schüler (z. B. Nasenbluten), sind nur diese Einzelresultate auszuschalten.
4. Die Auszählung der Fehler oder Gutpunkte muß immer nach den gleichen und eindeutigen Prinzipien erfolgen. Zweifelsfälle sind zu entscheiden und in einer Auswertungsanleitung schriftlich festzuhalten. Besonders im Fall der Teilnahme mehrerer Kollegen an dieser Eichung sind solche ausdrücklichen Festlegungen von größter Wichtigkeit.
5. Offensichtlich schlechte oder gute Klassen dürfen nicht ausgeschlossen werden. Sie gehören auch zum normalen Bild einer Schulpopulation. Solange die Eichstichprobe noch klein ist, soll der Lehrer bei der interpretation allenfalls daran denken, daß die Eichwerte vorläufig noch etwas einseitig liegen könnten.
6. Repetenten gehören natürlich zur Klasse und auch zur Eichstichprobe, da der Maßstab eines Schulleistungstests klassentypisch und nicht z. B. alterstypisch sein soll.
7. Stammen alle Resultate z. B. aus einem gleichen Quartierschulhaus, ist der Maßstab evtl. nur für das Quartier typisch. Tauscht der Lehrer die Prüfungen nicht aus und glaubt er, ein bestimmtes Fach überdurchschnittlich gut zu erteilen, ist der Maßstab nicht einmal für das Schulhaus repräsentativ, sondern nur für die Kombination «Schüler dieses Quartiers bei diesem Lehrer». Oft ist diese theoretisch selbstverständliche Einschränkung nicht von praktischem Belang, sie gewinnt aber z. B. rasch an Bedeutung, wenn die Prüfung nur einen engen Lernausschnitt betrifft, der ja leicht von verschiedenen Kollegen unterschiedlich betont werden kann. In solchen Fällen hat auch der Lehrer selbst gut darauf zu achten, daß er nicht Werte aus eigenen Klassen dazunimmt, die einen außerordentlichen Unterricht genossen haben, z. B. Prüfung über das amerikanische Präsidentschaftswahlsystem in Schaltjahren (=Wahljahren).
8. Es ist selbstverständlich, daß solche Prüfungen nicht beim Schüler bleiben dürfen, sonst könnten sich bei geeignetem Informationssystem der Schüler im Laufe der Jahre die «Bedingungen» ändern. Es sind allenfalls schon bald (separat zu eichende!) Parallelprüfungen bereitzustellen.

Erläuterung zu Tabelle 3:

* L_1; L_2; L_3 = Strichlisten der 3 Prüfungen
** $f_{T,2} = f_1 + f_2$
*** $f_{T,3} = f_{T,2} + f_3$

Anmerkungen:

[1] Es gibt natürlich statistische Methoden zur Berechnung der Wahrscheinlichkeit, mit der der «wahre» Wert innerhalb bestimmter Grenzen liegt (LIENERT 1969; HORST 1971). Als Faust-

regel könnte gelten: Mit etwa 95prozentiger Wahrscheinlichkeit ist der «wahre» Wert zu erwarten innerhalb der Grenzen RW ± 0.15 n, wobei n die Anzahl Testaufgaben ist und die Testzeit nicht unter 15 Minuten liegen soll.

[2] Diese ist eine mathematisch definierte Funktion und macht Annahmen, deren Erfüllung in psychologischen Variablen zwar nicht beweisbar, deren hinreichende Annäherung jedoch kaum widerlegbar, oft jedoch plausibel ist: Gegen unendlich strebende Zahl von additiv verbundenen Zufallsvariabeln. (s. Zentraler Grenzwertsatz)

[3] Standardabweichung = 10

[4] Früher wurde der IQ als Quotient aus Intelligenz- und Lebensalter definiert. Seine Definition als Abweichungs-IQ ist normalerweise: IQ = (T—50) (1.5) + 100 = 1.5 T + 25

[5] Die Wahl der Intervallmitten als Prozentränge bedingt, daß, wenn mehr als 1 % der Ergebnisse das höchstmögliche Resultat darstellen, der Prozentrang 100 gar nicht mehr vergeben wird.

Literaturnachweis:

Aufgaben zum Nachdenken AzN 4–6. Begabungstest für den Übergang auf weiterführende Schulen. Schweizer Fassung 1971 von A. Flammer, E. Broch, J. Bründler und J. Imfeld. Basel, Beltz-Verlag.

FLAMMER, August (1967a): Psychologische Tests in der Schule. In: *«schweizer schule». 54,* Seiten 174—178, 183, 432—435.

– (1967b): Der Schwierigkeitsindex in der Endform von Niveau-Tests. In: *Menschenbild und Menschenführung; Festschrift Montalta.* S. 521–535. Fribourg: Universitätsverlag.

– (1971a): *Leistungsmessung in der Schule.* Der innere Aufbau und der Einsatz von Leistungstests in der Schule. Hitzkirch: Comenius. 53 S.

– (1971b): Zur Definition der Notenskala. In: *Schweizerische Zeitschrift für Psychologie und ihre Anwendungen.* 30, S. 204—218

HORST, Paul (1971): *Messung und Vorhersage.* Eine Einführung in die psychologische Testtheorie. Basel: Beltz. 539 S.

IMK (1968): *Handbuch zur IMK-Reihe.* Winterthur: Schubiger.

LIENERT, Gustav A. (1969): Testaufbau und Testanalyse. 3. Auflage. Basel: Beltz

WENDELER, Jürgen (1969): *Standardarbeiten; Verfahren zur Objektivierung der Notengebung.* Basel: Beltz

Aktuelle Kurzmeldungen der «schweizer schule»

CH: Hat der Nationalrat die Zauberformel für die Neufassung von Art. 27 der BV gefunden?

Nationalrat Dr. Alfons Müller-Marzohl (CVP, Luzern) präsentierte am 22. Juni 1972 einen Vorschlag zur Neufassung des Artikels 27, der, nach langer Debatte, die Zustimmung der Volkskammer fand: «Die Ausbildung vor und während der obligatorischen Schulzeit fällt in die Zuständigkeit der Kantone. Die Kantone sorgen für die Koordination in diesem Bereich. Der Bund fördert die entsprechenden Bestrebungen: er kann Vorschriften über die Koordination erlassen.»

Entschieden setzte sich Müller für die Koordination auf dem Konkordatsweg und gegen eine zentralistische Lösung des Schulproblems ein: «Wir wollen die Kantone und unsere Schulen davor bewahren, daß in Bern ein Schulamt entsteht, welches nach französischem Vorbild alles reglementiert und inspiziert. Wir wollen ein koordiniertes, aber kein gleichgeschaltetes Schulwesen. Nichts wäre fortschrittsfeindlicher als eine schweizerische Einheitsschule mit Einheitslehrmitteln und reizlosen Eintopfgerichten.»

Die Jugendfraktion der BGB hat sich bereit erklärt, ihre Initiative zurückzuziehen, sofern Müllers Kompromißvorschlag auch die Zustimmung des Ständerats findet.

CH: Rechtschreibereform

Das Eidgenössische Departement des Innern hat einen vorberatenden Ausschuß für Fragen der Rechtschreibereform eingesetzt. Dieser ist beauftragt, zusammen mit den in der Bundesrepublik Deutschland und in Österreich zuständigen Organen die gegenwärtige Lage und die Absichten hinsichtlich der Rechtschreibebestrebungen in diesen Ländern abzuklären und dem Departement hierüber Bericht zu erstatten:

Zu Mitgliedern des Auschusses wurden ernannt: Landammann und Ständerat Dr. Fridolin Stucki, Vorsteher der Erziehungsdirektion des Kantons Glarus, Vorsitzender des Ausschusses, Nationalrat Dr. Alfons Müller, Luzern, Professor Dr. Stefan Sonderegger, Uetikon a. S., und Professor Dr. Louis Wiesmann, Basel. Nach Eingang des Be-