DOCUMENT RESUME

ED 173 434 TM 009 571

TITLE Proceedings of the Invitational Conference on Testing Problems (New York, New York, October 31, 1953).

INSTITUTION Educational Testing Service, Princeton, N.J.

PUB DATE 31 Oct 53

NOTE 188p.

EDRS PRICE MF01/PC08 Plus Postage.

DESCRIPTORS Course Content; *Decision Making; Educational Benefits; Educational Improvement; *Educational Testing; Elementary Secondary Education; Higher Education; Interviews; Measurement; Officer Personnel; Personality Assessment; *Profile Evaluation; Student Testing; *Testing Problems; *Test Results; *Test Scoring Machines

ABSTRACT

Seven major topics were included in the conference proceedings: (1) Improving Evaluation of Educational Outcomes at the College-Level; (2) Individual versus Group Decision Making; (3) Problems and Procedures in Profile Analysis; (4) Making Test Results Meaningful; (5) The Teaching of Educational Measurement; (6) The Interview as an Evaluation Technique; and (7) Impact of Machines and Devices on Developments in Testing and Related Fields. Most sessions concluded with a summary of the discussion. The first topic emphasized the benefits of testing. Topic three dealt with Rorschach data on Marine Corps officer candidates and a geometric model for profile problems. Topic four was concerned with testing programs in local schools, locally produced tests, and the usefulness of the test manual. Topic five dealt with introducing measurement courses; extension courses for inservice training; and psychological research training. Section six included papers on evaluation of interviews for selection purposes; interpersonal aspects; and personality appraisal. Section seven was concerned with topics such as the IBM scoring machines, electronic computers, and new developments in test scoring machines. (MH)

INVITATIONAL CONFERENCE
ON
TESTING PROBLEMS

OCTOBER 31, 1953

WALTER N. DUROST, *Chairman*

Improving Evaluation of Educational Outcomes at the College Level

Individual Versus Group Decision Making

Problems and Procedures in Profile Analysis

Making Test Results Meaningful

The Teaching of Educational Measurement

The Interview as an Evaluation Technique

Impact of Machines and Devices on Developments in Testing and Related Fields

EDUCATIONAL TESTING SERVICE

PRINCETON, NEW JERSEY LOS ANGELES, CALIFORNIA

FOR[illegible]D

It is most encouraging to note the [illegible]g growth and success of the Invitational Conference on Testing Problems, particularly when we recall that the first conference, held in 1936, attracted a total of seven testing leaders! This year the Conference brought together almost 500 persons from all over the United States and from several foreign countries.

The expanded program, judging from the responses of those attending the Conference, appears to have been very well received, due in very large part to the competence of the speakers and the excellence of their papers. ETS is particularly appreciative of their participation in the Conference and their contribution to its success.

We are indeed deeply indebted to Walter Dunost for his expert chairmanship of the 1953 Conference and for the imagination and energy with which he planned and conducted the program. To him goes the full credit for the enthusiastic response to the Conference and with it the sincere thanks of ETS for a most successful meeting.

HENRY CHAUNCEY
President

PREFACE

When it came time to plan the 1953 Invitational Conference on Testing Problems, one fact stood out as being of paramount importance, namely, that this was no longer a small gathering of individuals with a concentrated community of interest. The invitation list had reached nearly 1,000 names and the fields of specialization represented, while having measurement overtones, were nevertheless very diversified. For this reason, the plan was hit upon of dividing up the second hour into five sectional meetings, planned especially to cater to more specialized interests. Counting the general sessions and the luncheon meeting a total of 23 papers were presented. In all almost 500 persons attended the conference and a substantial majority of those questioned, approved of this program diversification.

The substantially larger number of papers and hence the greater cost of producing the Proceedings posed another problem, which has been met this year by making a modest charge for the volume. This has the added advantage of making the Proceedings more generally available and in view of the excellence of the papers presented and their permanent value, this is especially desirable. The luncheon address by Dr. Truman L. Kelley was devoted to the description of some new test developments in the field of interest testing. Dr. Kelley spoke extemporaneously and presented in an informal manner some very tentative data describing his Activity Preference Inventory. Because of the informal nature of his report, as well as the tentative nature of the findings described, Dr. Kelley has requested that his address be omitted from the Proceedings. To this the Chairman and Educational Testing Service have reluctantly agreed but only because we have been assured that further publication describing his work in this field is planned for the near future.

The Chairman would like to take this opportunity to express once again his deep appreciation to the speakers, the chairmen, and to Educational Testing Service for the 100 per cent cooperation received on every point in the development and implementation of this Conference.

Walter N. Durost, *Chairman*
1953 Conference

CONTENTS

Improving Evaluation of Educational Outcomes at the College Level

C. ROBERT PACE

REMARKS OF THE CHAIRMAN

ONE TIME a few years ago, on the spur of the moment, I defined an evaluator as an expert in measurement with a social conscience. Our three speakers this morning have demonstrated that there may really be a kernel of truth in this definition.

Evaluation is always concerned with three things.

It is concerned first with philosophy. What are the objectives? What are the goals? Toward what ends are education and evaluation directed?

It is concerned, second, with the science of measurement—with such problems as reliability, validity, relevance, comprehensiveness, discrimination, and sampling.

It is concerned, third, with the psychology of human relations, learning, and personality. What modes of action are most likely to produce changes in behavior? What interpersonal and intergroup relationships must be considered? What is the evaluation dividend for the professor and his teaching, for the student and his learning, and for the institution and the effectiveness of its program?

Perhaps, unfortunately, over the past two decades, much that is pertinent for evaluation has tended to be produced separately in education, psychology, and measurement, and with what Dr. Dressel judges to be a widening gap in understanding.

Many college professors and educational philosophers have been criticizing the failures of higher education in sweeping generalizations unsupported by systematic objective evidence, and finding refuge in what seems to me the curiously anti-intellectual notion that anything which can be measured, or even dealt with empirically, cannot be very important.

Many of the major advances in measurement theory and technique in recent years have come out of statistical laboratories, national testing agencies and research bureaus. And many of the good people in these environments appear to be divorced from the on-going responsibilities of teaching and learning, and from direct and daily contact with the processes of higher education.

Many of the recent major research efforts in psychology concerned with human relations, group behavior and decision making, personality, and attitude change have been conducted in the context of industry, government, the armed forces, and voluntary community groups—again appearing to be divorced from the framework of higher education and its evaluation. Indeed, I suspect that the experience gained in cooperative educational projects—such

as the Eight-Year Study, the Cooperative College Study of General Education, the Commission on Teacher Education, or the recent Cooperative Study of Evaluation in General Education—is largely unknown by many of the psychologists who are now wrestling with similar problems on a different level of theory and research, and in a different context.

If we are to make a major improvement in the evaluation of educational outcomes at the college level, we need to bring these related groups and these related problems into better focus.

The concept of systems research could be productive at this point if it forced our attention on the necessity of integration.

When evaluation is seen as contributing to the improvement of learning, of instruction, and of the efforts of the institution as a whole to attain its objectives; and when it is seen as a necessary integration of philosophy, science of measurement, and psychology of human relations, it may be well on the road toward a significant development in behavioral science and a major advance in education.

Improving Evaluation of Educational Outcomes at the College Level

ROBERT L. EBEL

HOW AN EXAMINATION SERVICE HELPS COLLEGE TEACHERS TO GIVE BETTER TESTS

I. Descriptive Background

May I begin by giving a brief description of the Examinations Service of the State University of Iowa. It was established in April, 1943 under the leadership of Professor E. F. Lindquist. The first director of the service, Professor Paul Blommers, was largely responsible for organizing and equipping it, and for setting up its procedures.

Currently the Examinations Service occupies several small rooms in the principal office building of the university. The staff consists of seven clerical employees, a graduate assistant, and the director. Multilith duplicators, IBM scoring machines and typewriters, Monroe calculators, and the usual desks, filing cabinets, and storage cupboards constitute the principal items of equipment. The service operates on an annual budget for salaries and supplies of approximately $30,000.

One of the principal functions of the Examinations Service is to relieve university instructors of the mechanical and clerical burdens involved in duplicating, scoring, and analyzing course examinations. Each year a few examinations are constructed as special projects, but this is not one of the main functions of the service. Its principal responsibility is to provide, upon request, assistance and advice on technical problems involved in the construction, administration, and interpretation of examinations of all types. Another area of responsibility is the administration of special testing programs, such as entrance examinations and various achievement testing programs. Special individual testing for the purposes of guidance and counseling is the function of a separate office.

II. Procedural Strategy

The long-range goal of the Examinations Service is to contribute to the effectiveness of the university's educational programs by improving the quality of the testing done. The strategy for achieving this objective involves three main considerations. The first is to make it as convenient as possible for college instructors to give good tests. No charge is made for any of the services offered. The offices are centrally located on campus. Service is rendered as promptly as possible and pains are taken to prevent errors.

The second consideration is to avoid any suggestion of interference with the instructor's independence with respect to his own examination procedures. The service is operated on a voluntary, permissive basis. If an instructor chooses to give a speed test when a power test seems more appropriate, or to call for raw scores reported as 100 less the number of errors, rather than as the

number of correct answers, or to weight each multiple-choice item 4 times as much as each true-false item in his test, we may inquire the reason for his procedure, but it is his preference, not ours, which determines how the job will be done.

It may seem somewhat inconsistent that a service established to improve examination practices should cooperate in carrying out unimproved practices. Yet it appears that helpful service offers the best opportunity for gradual improvement. College professors are human and cherish their feelings of adequacy and independence. However much their methods need mending, they do not like to be told so directly. Further they feel some distrust for those they call "educationists." Many of us who work in the field of educational measurements have backgrounds of public school teaching or hold degrees from colleges of education. We could not succeed, even if we tried, in *imposing*, upon them, our ideas of what is right and proper in educational measurement.

The third consideration is to maintain free and friendly communication with staff members. They receive notices regarding the extent and availability of the service. They are sent bulletins dealing with such matters as item writing, test administration, and the interpretation of item analysis data. Opportunities to consider examination problems with departmental faculties and with individual staff members are welcomed.

III. Basic Problems

Having given this brief description of our organization and service, may I turn next to consideration of some of the basic problems we have encountered in our efforts to help college teachers give better tests. It is emphatically asserted by wounded students and harassed professors and generally agreed by impartial observers, that college testing practices need improvement. In a recent survey of student opinions of teaching in our College of Liberal Arts, the item on which instructors were ranked lowest was "Quality of Examinations." This low rating may be explained in part as defensive rationalization by students who received lower grades than they expected or desired. But, I am convinced, it also indicates that there is very great room for improvement in many of our classroom tests. What has stood in the way of more rapid improvement? It seems to me that certain common attitudes and misconceptions on the part of college teachers constitute the chief barriers.

To begin with there are some professors who do not take the responsibility of evaluating student achievement seriously. They regard testing and grading as non-essential administrative red tape, largely divorced from the essential process of education. Since, in their view, the whole process is unimportant, it is unnecessary for them to expend much time or effort to achieve high validity or precision in the process. Such professors often solve the problem of grading according to the principal of least annoyance. "Give few low grades and you will have few complaints from the students, and hence there will be few occasions for outsiders to question your grading procedure."

It is interesting to note that the same professors who shun the onerous tasks of accurately evaluating student achievement in their own classes are likely, quite inconsistently, to object very strenuously to the consequences of similar behavior by other teachers. "My students can't read or write," they cry in anguish. "How did they ever pass freshman English?" Sometimes a professor finds himself in this unhappy predicament. He is serving on the doctoral examination committee of a very weak candidate. Ready to object stren-

uously, and possibly to cast a negative vote, he suddenly discovers that, in the recent past, he has himself rewarded the student's mediocre talents with a grade of "A."

The second obstacle to the improvement of college tests is the refusal of many professors to acknowledge the importance of objectivity in the evaluation of achievement. Note that we *did not say*, "the importance of objective tests." Such tests are valuable but they do not provide the only pathway to objective measurement. Evaluation is objective to the degree that equally competent observers can agree in their evaluation of a particular achievement. Our thesis here is that no measurement or evaluation, however obtained, is worth anything if it is not objective in some degree. To put it another way, only that component of a measurement or evaluation which can be verified by independent observation can serve any useful purpose.

Strange as it may seem, there exists a considerable group among college professors who reject the idea that objectivity is an essential ingredient in good evaluation. Two subspecies of this group may be recognized. The first is composed of those who worship the mystical goddess of intangibility. The second consists of those who are certain that any judgment which differs from their own must be wrong.

Here is a professor who bases his evaluation on a student's extended discussion of a rather vague topic. He prefers this procedure because, as he says, it yields occasional flashes of insight into how the student's mind is working. If these flashes please him, he is likely to give the student an A. It bothers him not at all that an equally competent professor, failing to receive these same flashes, or valuing them less, might give the same paper a C. He may know that his grade of A will be placed on record together with the A's assigned by a great many other instructors, and that when such grades are combined to give an overall indication of achievement, each "A" is assumed to have essentially the same meaning as any other "A." But the obvious inference from these facts that some sort of common standards (objectivity) in grading are required seems to have escaped him. Our Examinations Service does very little in the way of analysis of essay tests, not because such analyses would be impossible to perform, but rather because many users of essay tests are subjectivists who fail to see any significance in such analyses.

A third major obstacle to the improvement of college tests lies in the emphasis upon testing for recall of course content and the neglect of testing the attainment of course objectives. Some of the poorest examinations we handle come from the older academic fields where students are simply asked to recall this fact or that statement which was presented during the course. Some of the best tests come from professors of medicine, and law. Concrete case situations are described and the examinee is required to make decisions concerning treatment or procedures.

We hold it to be a fundamental principle of good educational achievement testing that a test should measure as directly as possible as many as possible of the ultimate objectives of instruction in the course. We have little patience with those who assert that many of the important outcomes of instruction are intangible. Does this mean that no one can observe them? If so they cannot possibly be of any importance, except in the internal life of the particular individual. And none of us can possibly be concerned about the internal life of another individual except when it manifests itself in overt behavior. I strongly suspect that many of those who insist upon the importance of intangible outcomes of education are

simply using it as a shield for their reluctance, or inability, to describe specifically what a given course of instruction ought to accomplish.

Many college instructors have seriously limited and distorted notions concerning what can be measured in well constructed tests of educational achievement. Having had limited experience with objective type questions they assume that only the recall of factual details, or, worse still, the recognitions of names and verbal symbols, can be tested with such a device. You frequently hear it asserted that "we do not yet know how to measure certain important outcomes of education." This may be true, but I am convinced that most of the difficulty is due to inability to define precisely what is to be measured. I learned recently of a college instructor who was seeking help in measuring leadership potential. He felt the measurement specialists were being evasive and uncooperative when they persisted in asking him precisely what he meant by "leadership potential."

A fourth obstacle to the improvement of college tests arises from the reluctance of many professors to recognize the inherently relative character of most measurements of educational achievement. Many of them feel quite strongly that the standards they should employ are set by the subject matter itself, and are best interpreted and applied by one who possesses expert knowledge of that subject matter. They would deny that the best basis for judging whether a given student's achievement is superior or inferior is a comparison of his achievement with that of other students in the same class, course, or grade group. The latter type of relative evaluation they refer to, somewhat disdainfully, as "grading on the curve."

This preference for presumably absolute external standards seems to be based on several misconceptions. One is that standards of achievement are inherent in the subject matter. If they are, they are extremely elusive. Repeated attempts to set minimum standards of achievement in various subject matter areas seem to have demonstrated quite convincingly that such standards are arbitrary, depending largely on the preferences and judgments of the individuals concerned with establishing them.

A second misconception is that standards set by an expert instructor are likely to represent more stable standards than those based on group performance. Again, experience has demonstrated quite clearly that equally competent instructors in the same subject matter have quite different standards both qualitative and quantitative. Further, any given instructor's standards are likely to shift markedly as time passes.

It is easy for an instructor to delude himself about the absoluteness of his standards when he relies on subjective processes of evaluation. So long as essay tests prevailed an instructor could claim absoluteness and still see to it that not too many students failed. His standard of 70% correctness for a passing grade *seems* to be independent of group performance, but actually may not be. If too many students seem to be getting scores below 70, the instructor can quietly shift his basis for calculation without letting any one else know about it.

With objective tests, however, such a readjustment is much more obvious. The decline of percentage grading (and the 70% pass mark) following the growth of objective testing was no accident. It came about, in part at least, because of the great difficulty of building an objective test which a consistent and reasonably large number of students could pass when the minimum was set at 70% correctness.

Another misconception is that of rela-

tive standards based on student performance tend to be too low. The use of relative standards does not, however, imply fixed percentages of A's, or F's, or any other grade. These proportions can be set independently of the group performance. The point is that consideration should be given to that performance so the standard will not be too high or too low. Others have feared that relative grading will encourage academic "slow-downs." Students might say to one another, "If none of us performs at capacity it will be easier for all of us to pass." So far as I know there is no empirical evidence to support this argument, and even the logic of it appears shaky when one considers the strength and prevalence of individual aspirations toward academic achievement.

Four obstacles to more rapid improvement of classroom tests at the college level have been mentioned. These are (1) failure to recognize the importance of measurement, (2) failure to see the need for objectivity, (3) emphasis on content details and (4) preference for allegedly "absolute" standards of achievement. These are essentially problems of attitude and orientation. But there is another serious obstacle or problem of a different type. It is lack of knowledge of appropriate techniques of measurement and lack of skill in their application. May we consider briefly what an examinations service can do to help solve this problem.

IV. Test Analysis

Detailed test analysis has proved to be one of the most useful avenues for progressive improvement of the tests constructed by classroom instructors. It possesses the two very important characteristics of objectivity and impersonality. Following test analysis it is not necessary for the test specialist to make critical comments concerning the test. Such criticisms are obvious and implicit in the data presented. Instead of placing the test specialist in the role of judge and critic, the use of test analysis places him in the role of consultant and adviser, working with the instructor to solve a common problem.

You have been issued sample copies of some of the materials we use in the test analysis.* May I direct your attention first to the test analysis report form. You will note that two major test characteristics, relevance and discrimination, are the subject of analysis. These are the most important qualities which an educational achievement test can possess. Most of the suggestions for improved test construction relate to these two basic qualities either directly or indirectly. In order to improve the relevance of their tests, instructors are urged to write items based on course objectives rather than on content details, to emphasize useful and important information rather than trivial or esoteric details, and to test the students understanding and ability to apply rather than his recognition or recall of details. When item writers are urged to choose questions of moderate difficulty, to express them clearly, to make sure that there is one best response, but that each of the alternative answers has some basis for plausibility, the purpose is to improve the discriminating power of the individual items and hence the test as a whole.

Evaluation of the relevance of a test is a difficult matter which is largely a responsibility of the subject matter specialist rather than of the test specialist. However, it is possible to roughly classify the items with respect to the type of achievement they call for, on the basis of very little subject matter competence. This is what we attempt to do in our analysis of relevance. Six

* Materials distributed at the Conference are reprinted at the end of this address.

categories of relevance are recognized: content details, vocabulary, facts, generalizations, understanding, and applications. These constitute an ascending scale of values. That is, we regard items dealing with the understanding and application as being much more valuable than those dealing in general with content details or vocabulary. While it is often difficult to be sure where a particular item ought to be classified, the usual effect of the classification process is to show quite clearly where the emphasis of the test as a whole lies. One of the main purposes of this analysis for relevance is to make each instructor directly aware of the desirability of writing more items which deal with generalizations, understanding, and applications.

The so-called ideals listed in the second column of the test analysis report are practical ideals rather than "ideals" ideals. Their purpose is to indicate the desired emphasis without giving any instructor the impression that his test is hopelessly inadequate. The second sheet in the handout is a relevance worksheet filled in with data from a specific test. A copy of this classification of items is returned to the instructor along with the test analysis report giving the over-all distribution of emphasis.

The lower half of the test analysis report sheet is concerned with the discrimination of individual items and of the test scores as a whole. The index of item discrimination used in this report is the U-L index suggested by Johnson. It is based upon the number of correct responses to an item in upper and lower 27% criterion groups, and is defined as the difference between the number of successful responses in upper and lower groups divided by the maximum possible difference. It is an index which favors items of near 50% difficulty, but for most tests of educational achievement this is an advantage rather than a handicap. Separate indices of item difficulty are reported to the instructor, to aid him in analyzing the causes of low discriminating power among certain items, but they are not used as a basis for selecting the best items or of indicating the over-all quality of the items. This is done on the basis of the U-L index alone. One should remember that the frame of reference for this analysis is classroom testing. In wide-scale achievement testing programs, where tests are designed to cover a range of grades, more attention would certainly need to be given to the distribution of item difficulty values.

If all of the items in the test possess high relevance to the objectives of instruction, then the only other necessary quality for the test scores as a whole is reliability. It is well known that the standard deviation (or variance) of test scores is an important factor in the reliability of those scores, and that the general level of scores, as indicated by the mean, has some bearing on the variance obtainable. Data on all of these matters are reported to the instructor. A copy of the score analysis worksheet is included in the handout. The reliability coefficient it calculates is based on Angoff's simplification of Kuder-Richardson formula 8. At least one competent statistician has questioned the superiority of this formula over the more familiar Kuder-Richardson formula 20. On the other hand, this formula gives results identical to those obtained from K.R. 20 when all items are equally difficult. When the items differ in difficulty this formula has the advantage of an upper limit of 1.00, while that of K.R. 20 is less than 1.00. At the moment I am prepared to argue that Angoff's formula provides the best reasonably convenient estimate of test score reliability that can be derived from a single test administration. The issue between the two may be of considerable theoretical importance. Practically, the

coefficients yielded are not widely different.

We have developed a table to simplify the computation of item variances and the so-called "true item variances" which are needed in this reliability formula. The table is entered with response counts obtained in the usual upper-lower 27% item analysis procedure. You will note that a blank row was left at the top of this table and at the left of it. This permits a clerk to enter directly the *number* of correct responses corresponding to a *percentage* of correct responses for any given size of upper and lower criterion groups. When this has been done it is possible to obtain the values of total item variance and "true" item variance working directly from the response counts which the scoring machines produce. Since we would be performing an item analysis in any case, we have found this approach to calculation of test reliability much simpler and more convenient than the split-halves technique.

The probable error of measurement of each test is calculated primarily to call the instructor's attention to the magnitude of such errors, even with well constructed tests. Instructors are cautioned, however, against regarding this probable error as a direct index of the quality of the test. Some highly unreliable test scores, which also show low variability, turn up with very low probable errors of measurement.

V. Conclusion

It has been our purpose, in this brief presentation, to describe some of the facilities available in the Examinations Service at the University of Iowa, to outline some of the procedures used in improving examinations, to discuss some of the obstacles to more rapid improvement, and to describe specifically our test analysis procedures. In these efforts to help college teachers give better tests, we have gained a little ground, but there is ample room for earnest and imaginative efforts on our campus, and on others as well, for many years to come.

University Examinations Service — State University of Iowa

TEST ANALYSIS REPORT

Test Title Security Transactions k = 100 Job Number 6314

Group Tested Class students N = 52 Date of Test 2-2-53

Time Limit 2 hr. Calculator Dorothy Checker Ruth

Characteristic	Ideals	Observed	Rating
I. Relevance			
A. Content details	0%	2%	OK
B. Vocabulary	less than 20%	6%	good
C. Facts	less than 20%	16%	OK
D. Generalizations	more than 10%	4%	OK
E. Understanding	more than 10%	5%	Low
F. Applications	more than 10%	68%	good
II. Discrimination			
A. Item			
1. High (.41 and up)	more than 25%	22%	OK
2. Moderate (.21 to .40)	more than 25%	41%	good
3. Low (.01 to .20)	less than 15%	30%	High
4. Zero or Negative	less than 5%	7%	OK
B. Score			
1. Mean	(a) 62.5	64.60	good
2. Standard Deviation	(b) 12.5	9.05	Low
3. Reliability	more than .70	.80	good
4. Probable Error		2.71	
III. Speededness			
A. Percent of Complete Papers	more than 90%	100%	good

(a) Midpoint of range between highest possible and expected chance score.
(b) One-sixth of range between highest possible and expected chance score.

University Examinations Service State University of Iowa

RELEVANCE WORKSHEET

Test Title Security Transactions k = 100 Job Number 6314

Group Tested Class students N = 57 Date of Test 2-2-59

Time Limit 2 hr. Classifier Eckel Checker

A. Content Details 52, 99

k = 2

% = 2

B. Vocabulary 1, 2, 5, 29, 70, 81

k = 6

% = 6

C. Facts 3, 9, 46, 27, 28, 46, 47, 48, 49, 50, 57, 64, 73, 74, 78, 100

k = 16

% = 16

D. Generalizations 6, 10, 19, 24, 44, 45, 54, 56, 82, 91, 82

k = 11

% = 11

E. Understanding 15, 17, 32, 58, 79,

k = 5

% = 5

F. Applications 4, 7, 8, 11, 12, 13, 14, 18, 20, 21, 22, 24, 25, 26, 29, 30, 31, 33, 35, 36, 37, 38, 39, 40, 41, 42, 43, 55, 57, 58, 59, 60, 61, 62, 63, 65, 66, 67, 68, 69, 71, 72, 75, 76, 77, 80, 83, 84, 85, 86, 87, 88, 89, 90, 93, 94, 95, 96, 97, 98

k = 60

% = 60

University Examinations Service · State University of Iowa

DISCRIMINATION WORKSHEET

Test Title [illegible] 100 Job Number 6814

Group Tested Class stude[illegible] 57 Date of Test 2-2-53

Time Limit 2 hr. Calculator Dorothy Checker Ruth

Highly Discriminating Items Number 22 Percent 22

.85	.90	.95
.70	.75	.80
.55	.60 100,	.65 10, 52,
.46 2, 13, 14, 16, 21, 36, 62, 65, 85	.48	.50 88, 95,
.40 3, 11, 27, 35, 43, 64, 77, 90	.42	.44

Moderately Discriminating Items Number 41 Percent 41

.34	.36	.38 36, 72, 73, 75, 94, 97
.29	.30 93,	.32 12, 22, 23, 32, 39, 42, 50, 53
.26	.27 1, 5, 6, 9, 20, 40, 43, 51, 59, 78	.28
.23 67, 69, 70, 79, 80, 83, 91, 92,	.24	.25
.20 19, 25, 26, 41, 45, 54, 58, 64,	.21	.22

Poorly Discriminating Items Number 30 Percent 30

.17 87, 89, 96, 99,	.18	.19
.13 4, 15, 18, 38, 42, 55, 60, 64, 82,	.15	.16
.07 29, 30, 48, 52, 74, 81, 86,	.09	.11
.01	.03	.05
.00 2, 12, 31, 32, 44, 49, 71, 74, 84,	.00 28,	.00

Negatively Discriminating Items Number 7 Percent 7

-.01	-.03	-.05
-.07 24,	-.09	-.11 8, 28, 33, 34, 68,
-.15	-.20	-.25
-.30	-.35	-.40 46,
-.50	-.60	-.80

University Examinations Service State University of Iowa

SCORE ANALYSIS WORKSHEET

Test Title Security Transaction k = 100 Job Number 63/4

Group Tested Class students N = 57 Date of Test 2-2-59

Time Limit 2 hr Calculator Dorothy Checker Ruth

I. Basic Data

N	ΣX	ΣX^2	k	Σpq	$\Sigma r_{it}^2 pq$
20	7703	612837	15	2.92	.49
20	1314	86797	20	3.78	.47
17	1265	94538	20	3.95	.38
			20	3.84	.60
			25	4.31	.68
57	3682	242512	100	18.20	2.62

II. Statistics

Mean (M) = $\frac{\Sigma X}{N}$ = $\frac{3682}{57}$ = 64.5965

Variance (σ_t^2) = $\frac{\Sigma X^2}{N} - M^2$ = $\frac{242512}{57} - (64.5965)^2$ = 81.8886

Standard Deviation (σ_t) = $\sqrt{81.8886}$ = 9.0492

Reliability (r_{tt}) = $\frac{\sigma_t^2 - \Sigma pq}{\sigma_t^2 - \Sigma r_{it}^2 pq}$ = $\frac{81.89 - 18.20}{81.89 - 2.62}$ = .8035

Probable Error (P.E.$_{meas}$) = .6745 $\sigma_t\sqrt{1-r_{tt}}$ = .6745 x 9.0492 x .4433 = 2.7057

TABLE OF TOTAL ITEM VARIANCE (pq) AND "TRUE" VARIANCE (r_{it}^2 pq)

Prepared by Robert L. Ebel, University Examinations Service, State University of Iowa

$$= \frac{\sigma_t^2 - \sum pq}{\sigma_t^2 - \sum r_{it}^2 pq}$$

Example: If an item is answered correctly on 75% of the upper group papers and on 30% of the lower group papers, its total variance is .25 and its "true" variance is .05. These values, summed over all k items in the test, may be used with the variance of the test scores σ_t^2 to find the reliability coefficient, α, or α_t, of the scores.

Up. \ P	00	03	06	09	12	15	18	21	24	27	30	33	36	39	42	45	48	51	54	57	60	63	66	69	72	75
00	00	01	03	04	06	07	07	08	10	10	11	12	12	13	14	15	16	16	17	18	18	19	20	20	21	21
	00	01	02	02	05	05	06	07	08	09	10	11	11	12	13	14	14	15	16	16	17	18	19	19	20	20
03	01	03	04	06	07	07	08	10	11	11	12	13	13	15	15	16	17	17	18	19	19	20	21	21	22	22
	01	00	00	00	01	01	01	02	02	02	03	03	04	05	05	06	06	07	08	08	09	10	11	12	12	13
06	03	04	06	07	08	09	10	11	12	13	14	15	15	16	17	18	18	19	20	20	21	21	22	22	23	23
	02	00	00	00	00	01	01	01	01	02	02	03	03	04	04	05	05	06	07	07	08	09	09	10	11	12
09	04	06	07	08	09	11	11	12	13	14	15	16	17	17	18	19	20	20	21	21	22	22	23	23	24	24
	03	00	00	00	00	00	00	00	01	01	02	02	02	03	03	04	05	05	06	06	07	08	08	09	10	11
12	06	07	08	09	11	11	13	13	15	15	16	17	18	18	19	20	21	21	21	22	22	23	23	24	24	24
	05	01	00	00	00	00	00	00	01	01	01	01	02	02	03	03	04	04	05	06	06	07	08	08	09	10
15	07	07	09	11	11	13	13	15	15	17	17	18	19	19	20	21	21	22	22	23	23	24	24	24	24	25
	05	01	01	00	00	00	00	00	01	01	01	01	01	02	02	03	03	04	04	05	05	06	07	07	08	09
18	07	08	10	11	13	13	15	15	17	17	18	19	20	20	21	21	22	22	23	23	24	24	24	25	25	25
	06	01	01	00	00	00	00	00	00	00	00	01	01	01	01	02	03	03	03	04	05	05	06	07	07	08
21	08	10	11	12	13	15	15	17	17	18	19	20	20	21	21	22	22	23	23	24	24	24	25	25	25	25
	07	02	01	00	00	00	00	00	00	00	00	00	01	01	01	02	02	02	03	03	04	04	05	06	06	07
24	10	11	12	13	15	15	17	17	18	19	20	20	21	21	22	22	23	23	24	24	24	25	25	25	25	25
	08	02	01	01	01	01	00	00	00	00	00	00	00	01	01	01	02	02	02	03	03	04	05	05	06	06
27	10	11	13	14	15	17	17	18	19	20	20	21	21	22	22	23	23	24	24	24	25	25	25	25	25	25
	09	02	02	01	01	01	00	00	00	00	00	00	00	00	01	01	01	01	02	02	03	03	04	04	05	06
30	11	12	14	15	16	17	18	19	20	20	21	21	22	22	23	23	24	24	24	25	25	25	25	25	25	25
	10	03	02	02	01	01	00	00	00	00	00	00	00	00	00	01	01	01	02	02	02	03	03	04	04	05
33	12	13	15	16	17	18	19	20	20	21	21	22	22	23	23	24	24	24	25	25	25	25	25	25	25	25
	11	03	03	02	01	01	01	00	00	00	00	00	00	00	00	00	01	01	01	02	02	02	03	03	04	05
36	12	13	15	17	18	19	20	20	21	21	22	22	23	23	24	24	24	25	25	25	25	25	25	25	25	25
	11	04	03	02	02	01	01	01	00	00	00	00	00	00	00	00	00	01	01	01	01	02	02	03	03	04
39	13	15	16	17	18	19	20	21	21	22	22	23	23	24	24	24	25	25	25	25	25	25	25	25	25	24
	12	05	04	03	02	02	01	01	01	00	00	00	00	00	00	00	00	00	01	01	01	01	02	02	03	03

42	14	15	17	18	19	20	21	21	22	22	23	23	24	24	24	25	25	25	25	25	25	25	25	25	25	24	24	24	23	22	22	21	20	17
	13	05	04	03	03	02	01	01	01	01	00	00	00	00	00	00	00	00	00	01	01	01	02	02	02	03	03	04	05	05	06	07	08	10
45	15	16	18	19	20	21	21	22	22	23	23	24	24	24	25	25	25	25	25	25	25	25	25	25	24	24	24	23	23	22	21	20	19	17
	14	06	05	04	03	03	02	02	01	01	01	00	00	00	00	00	00	00	00	00	01	01	01	02	02	02	03	04	04	05	05	07	08	09
48	16	17	18	20	21	21	22	22	23	23	24	24	24	25	25	25	25	25	25	25	25	25	25	24	24	24	23	23	22	21	21	20	18	16
	14	06	05	05	04	03	03	02	02	01	01	01	00	00	00	00	00	00	00	00	00	01	01	01	02	02	02	03	04	04	05	06	07	09
51	16	17	19	20	21	22	22	23	23	24	24	24	25	25	25	25	25	25	25	25	25	25	24	24	24	23	23	22	21	21	20	19	18	15
	15	07	06	05	04	04	03	02	02	01	01	01	01	00	00	00	00	00	00	00	00	00	01	01	01	02	02	02	03	04	04	05	06	08
54	17	18	20	21	21	22	23	23	24	24	24	25	25	25	25	25	25	25	25	25	25	25	24	24	24	23	23	22	22	21	20	19	18	17	15
	16	08	07	06	05	04	03	03	02	02	02	01	01	01	00	00	00	00	00	00	00	00	00	01	01	01	02	02	03	03	04	05	06	07	
57	18	19	20	21	22	23	23	24	24	24	25	25	25	25	25	25	25	25	25	25	25	24	24	24	23	23	22	22	21	21	20	19	18	16	14
	16	08	07	06	06	05	04	03	03	02	02	02	01	01	01	00	00	00	00	00	00	00	00	01	01	01	02	02	03	03	04	05	07		
60	18	19	21	22	22	23	24	24	24	25	25	25	25	25	25	25	25	25	25	25	24	24	24	23	23	22	22	21	21	20	19	18	17	15	13
	17	09	08	07	06	05	05	04	03	03	02	02	01	01	01	01	00	00	00	00	00	00	00	00	01	01	01	02	02	03	04	04	06		
63	19	20	21	22	23	24	24	24	25	25	25	25	25	25	25	25	25	25	25	24	24	24	23	23	22	22	21	21	20	19	18	17	16	15	12
	18	10	09	08	07	06	05	04	04	03	03	02	02	01	01	01	00	00	00	00	00	00	00	00	00	01	01	01	02	02	03	04	05		
66	20	21	22	23	23	24	24	25	25	25	25	25	25	25	25	25	25	25	24	24	24	23	23	22	22	21	21	20	19	19	18	17	15	14	11
	19	11	09	08	08	07	06	05	05	04	03	03	02	02	02	01	01	01	00	00	00	00	00	00	00	00	01	01	01	02	03	03	05		
69	20	21	22	23	24	24	25	25	25	25	25	25	25	25	25	25	25	24	24	24	23	23	22	22	21	21	20	19	19	18	17	16	15	13	10
	19	12	10	09	08	07	07	06	05	04	04	03	03	02	02	02	01	01	01	00	00	00	00	00	00	00	00	00	01	01	02	02	03	04	
72	21	22	23	24	24	24	25	25	25	25	25	25	25	25	25	25	24	24	24	23	23	22	22	21	21	20	19	19	18	17	16	15	13	12	10
	20	12	11	10	09	08	07	06	06	05	04	04	03	03	02	02	02	01	01	01	00	00	00	00	00	00	00	00	01	01	02	02	04		
75	21	22	23	24	24	25	25	25	25	25	25	25	25	24	24	24	24	23	23	22	22	21	21	20	19	19	18	17	16	15	14	13	11	09	
	20	13	12	11	10	09	08	07	06	06	05	05	04	03	03	02	02	02	01	01	01	00	00	00	00	00	00	00	00	01	01	02	03		
78	22	23	24	24	25	25	25	25	25	25	25	25	24	24	24	24	23	23	22	22	21	21	20	19	19	18	18	16	15	14	13	12	11	08	
	21	14	13	12	10	10	09	08	07	06	06	05	05	04	03	03	02	02	02	01	01	01	00	00	00	00	00	00	00	01	01	02	02		
81	22	23	24	25	25	25	25	25	25	25	25	24	24	24	24	23	23	22	22	21	21	20	19	19	18	17	16	15	14	13	12	11	09	07	
	21	15	14	12	11	11	10	09	08	07	06	06	05	05	04	04	03	02	02	02	01	01	01	00	00	00	00	00	00	00	01	01	02		
84	23	24	24	25	25	25	25	25	25	25	24	24	24	23	23	23	22	21	21	21	20	19	19	18	17	16	15	14	13	12	11	10	08	07	
	22	16	14	13	12	11	11	10	09	08	07	06	06	05	05	04	04	03	03	02	02	01	01	00	00	00	00	00	00	00	00	01	02		
87	24	24	25	25	25	25	25	25	25	24	24	24	23	23	22	22	21	21	20	20	19	18	18	17	16	15	14	13	12	11	10	09	07	06	
	22	16	15	14	13	12	11	10	10	09	08	07	07	06	05	05	04	04	03	03	02	02	01	01	01	00	00	00	00	00	00	00	00	01	
90	24	25	25	25	25	25	25	25	24	24	24	23	23	22	22	21	21	20	19	19	18	17	17	16	15	14	13	12	11	10	09	07	07	05	
	23	17	16	15	14	13	12	11	11	10	09	08	08	07	06	05	05	04	04	03	03	02	02	02	01	01	01	00	00	00	00	00	00	01	
93	25	25	25	25	25	25	24	24	24	23	23	22	22	21	21	20	20	19	18	18	17	16	15	15	13	13	12	11	10	09	07	07	05	04	
	23	18	17	16	15	14	13	12	12	11	10	09	09	08	07	07	06	05	05	04	04	03	03	02	02	01	01	01	00	00	00	00	00	01	
96	25	25	25	25	25	24	24	23	23	22	22	21	21	20	20	19	18	18	17	16	15	15	14	13	12	11	11	09	08	07	07	05	04	02	
	23	19	18	17	16	15	14	14	13	12	11	10	10	09	08	08	07	06	06	05	04	04	03	03	02	02	02	01	01	00	00	00	00	00	
99	25	25	24	24	23	23	22	22	21	21	20	19	19	18	17	17	16	15	15	14	13	12	11	10	10	09	08	07	07	06	05	04	02	01	
	24	21	20	19	17	17	16	16	14	14	13	12	11	11	10	09	09	08	07	07	06	05	05	04	04	03	02	02	02	01	01	00	00	00	

ITEM ILLUSTRATING CATEGORIES OF RELEVANCE

A. Content Detail

". . . 'title' is a formal word for a purely conceptual notion; I do not know what it means and I question whether anybody does, except perhaps legal historians." Statement of

(1) Charles Clark
(2) Felix Frankfurter
(3) Harry Chase
*(4) Learned Hand

B. Vocabulary

A security interest in a chattel, created by a bailment for the purpose of securing the payment of a debt, is properly called

(1) equitable chattel mortgage
(2) deposit of title bonds
*(3) pledge
(4) equitable conditional sale
(5) conditional sale

C. Fact

The title of the mortgaged personal property is held in Iowa by

*(1) the mortgagor
(2) the mortgagee

D. Generalization

Probably the outstanding recent development in the area of the conditional sales contract is

*(1) its gradual coalescence with the mortgage security devise
(2) the development of the right to bar the equity of redemption
(3) the inequitable treatment meted out to it by the courts of equity
(4) its total replacement of the chattel mortgage

E. Understanding

If a creditor ever got your advise on loan arrangements, you might recommend the taking of a deed absolute in form rather than a mortgage as security for a loan because (most persuasive reason)

*(1) the debtor will have certain procedural hurdles to overcome if he comes in seeking to get the deed declared a mortgage
(2) the creditor can move on the property on default
(3) the creditor can sell, after default, to a third party free and clear and get the market value
(4) by taking a deed absolute in form, the creditor can obviate the necessity of foreclosure and thus eliminate the equity of redemption

F. Application

A married to B. A alone mortgaged certain property. On A's death, B asserted her right to her statutory share in the property. She claimed a one third interest in the realty. She can redeem by

*(1) paying off the entire mortgage
(2) paying off her pro rata share
(3) having the court divide the property

Improving Evaluation of Educational Outcomes at the College Level

LILY DETCHEN

THE EVALUATION DIVIDEND FOR THE INDIVIDUAL STUDENT

FORTUNATELY, it won't be necessary to decide here whether a comprehensive program of evaluation best serves the purposes of the administration, the faculty or the students, for we all know that these are truly interrelated. But in a sense it is the student who is the real customer in most evaluation situations, and if the program is to function for him in a useful and realistic way, we must assure him dividends in it. It is therefore periodically desirable to consider quite closely what the benefits are for the individual. This exercise can even be, if you will, a sort of examination of conscience for the evaluator. Furthermore, because the evaluation process takes a heavy toll of student time and energy, it is only practical to insure that its outcomes be tangible to him, so that on the one hand he will accept the work burden more cheerfully and on the other so that his motivation represent a maximum personal effort. For, after all, this is an important aspect of obtaining reliability.

In addition to these considerations, there is that of multiplying student dividends by extending the evaluation resources so that students may also make use of them in that planning for which they are usually personally responsible, as in their so-called "student activities."

Evaluation, as we utilize it at Pennsylvania College for Women, ranges beyond the academic situation. We try to help students to determine the objectives of their co-curricular activities, for instance, and to study the degree to which they are fulfilled by their various groups. We work with them to check unfounded rumors with facts and to determine what is the will and reaction of the majority in given situations. Evaluation thus serves an educational role in helping to promote a democratic spirit of inquiry and action. To this end, the questionnaire type of inquiry has been utilized extensively. Structured as a process rather than as a mere measuring device, it has had learning and other psychological effects of significance in improving the College program for the individual and her attitude towards it. In using the questionnaire, we make the overall assumption that student participation in planning pays off in good dividends of rapport and learning. We have found that the questionnaire method may be used to reveal to the student group, from time to time, the whys and wherefores of some of our educational devices and the reasons for planning in a certain way, so that students may gain a greater sympathy for over-all objectives and some appreciation of the difficulties that attend their achievement.

We have certain long range objectives aiming at the improvement of course examinations and thereby indirectly at the improvement of instruction. Our office is also a "research" bureau in that we attempt within the

limitations of a modest staff to assemble data to answer or to clarify miscellaneous types of problems which arise from time to time. It is perhaps necessary to note that ours is a small college of but several hundred students and that we do have an excellent faculty-student ratio. I am not so sure that a larger institution could or should proceed in just the same way. But certainly there would be a way. All of our activities of evaluation are covered at a cost of approximately $25.00 per student per year. This is the direct cost for personnel and supplies. It does not cover such over-all items as maintenance, which never get charged against the budget of any office, nor does it include the services of the faculty and staff who may co-operate from time to time on given projects or who co-operate continuously in the general instructional and advisory programs of the College.

In a situation where it is intended to give a fair quota of time to follow-up of routine evaluation, there first has to be a resistance to a temptation to administer too many tests, questionnaires, or whatever, or to work with too many groups simultaneously, for a lot of busy work can ensue that seems always to get in the way of accomplishing the goals of any one program. I have referred to this as a temptation because while the preparation of the evaluation instrument is certainly not an easy task, the administration of evaluation instruments with their attendant scoring, listing, collating and first crude interpretations is by far the easiest task of all for a person who is trained in those skills. This is even more true when published tests are utilized. But after the first excitement and interest of the more apparent results have been communicated to faculty and administration, interest can ebb dismally. The hardest task for the evaluator is to keep that interest alive long enough to realize all the objectives of the activity and to proceed rather like the thrifty housewife who continues to secure healthful sustenance for her family even from the bones of the chicken which she was able to serve so much more ostentatiously on the first day! An important portion of this sustenance lies in the opportunity the evaluation program has to work with students to develop in them attitudes of self-appraisal and self-acceptance and to plan multiple applications that will operate for student benefit.

It may be of interest to note here first the nature of the provisions made to acquaint students with our services. Of course our students are conscious of our existence on campus from the day they enter, since like other colleges, we test them for several days upon arrival. But at that time we make an effort to tell them why. A period of time is set aside for such orientation on the first morning just preceding the testing. We discuss each test and why it has a place in the program, and specifically how it is to be used later for the benefit of the student. We also talk about the limitations of interpretation and the dangers of over-stressing test results. Within the first semester there is another session with this group but in small group units. At that time we endeavor to acquaint the students with the more general objectives of the office, to help them establish some acceptance of them, and to enlist their interest. We also discuss course examinations, the problems that their instructors face in grading them fairly, the reasons for examinations. At a later time we discuss more specifically preparation for examinations. From then on as the necessity for any general examination program arises to fulfill administrative or other needs, students are carefully instructed in its purpose. They also learn in advance that each will receive a report of the over-all results and a special report of her personal results which is usually issued in some profile form.

Also, explanations are attached to the profiles that usually satisfy the great majority of students; others may drop in at the office to talk about these results or discuss them with one of the deans or other faculty friend.

So much for our generalized student relations techniques. Let us now consider some cases of individual students and what they receive in the way of benefits from the fact they attend a college with an evaluation program. I am here glancing over the generally accepted benefits of better curriculum through improved objectives, better examinations, better testing conditions and greater efficiency in handling and reporting their results, for Dr. Ebel will have covered this aspect. My assignment here is to describe some of the more tangible student benefits.

Our first case is just an ordinary student named Nancy who is not particularly outstanding as a good or poor student or problem case. Nancy was graduated from high school as a member of the National Honor Society, but considering the fact that two-thirds of the freshman class that year had graduated in at least the upper fifth of their classes, this was not too significant. During the freshman week tests, Nancy distinguished herself in no outstanding way. With one or two exceptions not particularly noted at that time, her scores indicated that at least twenty-five per cent of the class would surpass her later and that she would certainly in most instances always be better than another twenty-five per cent. Nancy knew that she could come in to see me, or her faculty adviser, or her freshman dean about her test results, but she didn't feel any urge to do so until after her third month in college. The impression was distinctly getting through to her by that time that something was amiss. To make a long story short, we had to get Nancy reoriented to the idea that the competitive picture was entirely different; it was easier to do this by discussing her test profile against a backdrop of data for all other students. This was done with an intention to urge her to put forth with a greater effort, not to accept a mediocre rating. However, it had to be done in a way to help Nancy become more selective about her efforts; and to give her a new concept of the grade of C, which is a quite respectable grade in our College. Nancy's case is representative of a common one among freshmen who for a while suffer the rigors of increased competition and who could be hopelessly disturbed if they are not assisted to some better acceptance of the change. It is probably the most universal "common complaint" suffered in our freshman class, and all advisers are alerted to it.

Sometimes a student benefits because a faculty member benefits first. I am thinking of an instance where a student fell into difficulties in a certain required course in the social sciences. Esther had made one of the lowest grades in the class on any quiz the intructor saw fit to give; she had turned in two poorly written reports and failed to turn in a third; she was beginning to cut frequently. The instructor was a sensitive guy who lost sleep over the occasional necessity of giving a failing grade. He stopped in to talk about Esther. Her other grades were not remarkably good, but at least she wasn't reported as having undue difficulty in other classes. The instructor suggested the possibility of moving her to some other section because he thought there might be something personal in the situation that was causing this block. So we looked up Esther's test records. Anywhere that she had encountered a test in the social sciences she had managed to flub it. At entrance she had made one of the lowest grades in a test of reading and interpretation in the social sciences; this was noteworthy because results for all of her other tests, altogether about a

dozen, were quite good. At the end of the freshman year, she had done just as poorly in a contemporary affairs test —in fact, had turned the paper in without completing it. Her vocational interest test showed an extremely low interest score in the social services area. It wasn't the instructor that Esther rejected, just his subject! With this reassurance, the instructor was better able to cope with Esther. I can't report that she chose his field as a major, but she managed to pass the course respectably. The test records here had given some inkling of a problem this student had that might have ended sadly for her, if someone had not bothered to try to work it out. In this instance it is not Esther herself who is as typical as the instructor. Students' problems are frequently instructors' problems, and our office can frequently help.

Questions similar to these arise daily in the office, in the corridor, across the lunch table.

All of our test reports go out to faculty in list as well as profile form. Many use them to spot their problem situations in advance so that they may provide for them. What these provisions may be differ with the faculty. One faculty person sets up tutoring arrangements for slower students; an English teacher may make a somewhat different assignment of term paper topics to individuals; another may expect a great deal more of some than of others and ride them when they don't produce; academic loads are adjusted; the Committee on Academic Standing is more apt to take a sterner view of a failing student of good potential than of a failing student of lower potential, but on the other hand take more drastic final severance action in the latter instance than in the former.

Jeannine's case is typical of several other students who have recently made a similar type of adjustment. Jeannine has an A.B. degree With Honors from PCW, but Jeannine never graduated from high school. We accepted her after her junior year. She was unusually mature for her age with a high seriousness of purpose directed towards a career in music as a flutist. She was admitted on that premise, her great desire to get on with her college education, her family's willingness to enter into the plan, chiefly because of the expense of her music education, and her tested background of excellent achievement in high school academic subject matter. She became a Mrs. Jeannine in her senior year marrying a young dental student. And she is now playing in a local symphony orchestra. Everything that Jeannine did was characterized by a singleness and earnestness of purpose. I talked with her frequently during her four years with us. She never regretted the move she had made, and her gratitude toward the program that had made it possible for her was sincere. We have admitted a number of students on this plan, with no casualties.

Many students benefit in the exemption examination program. The exact operation of our program of screening tests and exemption examinations is complex and has been described elsewhere.[1] Our students after first being introduced to its opportunities are screened in various initial tests, and then selected ones are given exemption examinations in our so-called basic courses. If they successfully qualify they are excused from that course and enter immediately into the more advanced work for which the course ordinarily serves as a prerequisite. For example, a student who qualified in freshman English enters the advanced writing course, or in modern society one of the more specialized social sciences

[1] Lily Detchen, "A Program of Required Exemption Examinations," *Journal of Higher Education*, (May, 1953), pp. 249-254.

courses. An interesting feature of the program is that the student is strongly advised to follow this recommendation of accelerated work, but not actually required to do so. Most do follow our advice in the matter. The uniqueness of our program is that our students are required to take the preliminary screening examinations. We believe that this program cuts through a lot of waste for these students, stimulates their academic efforts, and gives a supporting recognition which has resulted in fewer dropouts in this particular group.

From the outset we have emphasized in our test construction work the use of the examination as a teaching and learning device rather than as a grading instrument alone. I do not wish to say too much about this phase of our program because it is the subject of Dr. Dressel's paper. But we have been examining and trying out various procedures to achieve these ends and will continue to do so. Some of the devices which have been employed have been study sheets which reiterate the objectives of the course and give the student sample questions for measuring her own progress, lists of broad study questions around which students can plot a study plan for their final examinations with the knowledge that two or three of the ten or so questions which they are given to prepare will be chosen as the actual final examination essay questions; open-book examinations taken in the classroom, which it is hoped will encourage students to take and organize better notes; open-book examinations which the student is allowed to do outside of the classroom; "illustrated" examinations in science where the student is given a situation both in "words" and in a "demonstration" and it is expected that some additional learnings accrue as a result of the demonstrations; discussions in the classroom of those concepts which a majority of students fail to achieve, these weak spots having been located behind-the-scenes through the use of the IBM graphic item counter.

At PCW a number of instructional situations are built directly around tests or questionnaires. In the Human Development and Behavior course, a freshman course, the administration of a vocational interest test serves as a focal point for the consideration of careers and the features of temperament that must be taken into consideration in choosing a job. The larger matter of values in selecting careers is also explored. This class unit and the activities of a Vocational Guidance Week which brings representative career counselors to campus are combined into something more meaningful and helpful for the student, the classroom instructor and the visiting counselors. In a senior group a Life Goals Inventory is used at the outset of the course in Philosophy of Life as a sort of curtain raiser for the course. This activity leads directly into a consideration of purposes and values. Some students request copies to administer to their fiances, and these we score for them also.

As I indicated previously evaluation services are made available to students who wish such help with some immediate problem of student government or other student activity. In such instances, the initial stimulation generally comes from some faculty adviser who works with the group and who may make the suggestion to them. One or two such projects have been undertaken annually. These enterprises have been related to such problems as the elimination of faculty domination in interest groups, the clearer defining of the roles of students as leaders and as followers, the determination of the kind of religious program that might best serve PCW students, a survey of interest in the type of assembly program offered in a particular year, a survey of the attitudes of students toward

some specific aspects of student government, analysis of senior opinion on the one hand and faculty opinion on the other on the value of the senior tutorial.

The ultimate aim in these projects is a structuring that will result in better understandings for those concerned through the exchange of ideas and the assumption of responsibility for leadership in planning. It is easiest to describe the process with an example. In one situation the student government decided to evaluate the "clubs" (interest groups), class organizations, and the student-government association. The student planners of the evaluation developed the following major strategy: (a) the utilization of a questionnaire technique to provide the college community, both leaders and rank and file, with criteria by which to judge the effectiveness of student organizations and (b) the provision of an opportunity for the student to consider in some systematic, way her contribution, as an individual, to student co-operative enterprises. Because the questionnaire that the group helped develop emphasized that the success of an endeavor rests as much with the general membership as with its leaders, the answering of the questionnaire was expected to alert more students to their responsibilities. That this questionnaire situation was formulated as a process and not a mere data-collecting device is attested by the other activities which stemmed from it:

1. In the planning stages several thoughtful sessions were conducted by a number of student leaders, who, in the course of analyzing the criteria by which the clubs should be judged, acquired some needed, relevant understandings themselves.
2. The task of answering the questionnaire actually was a process of self-analysis by each student of her attitude toward community endeavor and of her role as a constructive college citizen.
3. Besides using the results for the immediate purposes of making adjustments as suggested by student opinion, a new group of student leaders, entirely different from those who worked on the original planning of the questionnaire, was inducted into its philosophy and use. This occurred at a leadership conference several months later, when the officers of organizations for the new year received their general training in officership. The questionnaire results for each organization furnished specific points for special review by the new officers of the organization.

The use of the questionnaire achieves its maximum value when those to whom the results are of interest can be given some share in its planning and when multiple interpretational uses are made of the data. Incidentally, when such questionnaires are given, usually the entire student body is polled. At such times a student committee prepares the materials, doing all of their own mimeographing and stapling, administers the questionnaire, and tabulates the results.[2]

In summary may I say that an evaluation program in a college situation must certainly include within its objectives that of scrutinizing its *own* contribution to the progress of the institution and that unless this can be done fairly concretely in terms of benefits to the individual student, there is no real assurance that its activities have much meaning.

[2] Lily Detchen, "Instructional Values Associated with the Use of Questionnaires," *School Review*, (November 1953), pp. 481-486.

Improving Evaluation of Educational Outcomes at the College Level

PAUL L. DRESSEL

EVALUATION AS INSTRUCTION

As a reformed (or possibly renegade, depending on the training of the observer) mathematician and statistician, I retain sufficient respect for the concept of integration to wince at the loose and indelicate usage of the term which has become current in the field of education. Yet, experiences of the past ten years as a counselor, organizer and director of a counseling center, as a test constructor, organizer and head of a staff of examiners, researchers and evaluators, as a director of an evaluation project in general education, and as a speaker and consultant on general education have brought me to a state of introspection and to an attempt to organize my varied experiences and reactions to them which can perhaps best be conveyed by the word "integration." I should like to share with you an initial attempt at integrating or organizing my thoughts in measurement and evaluation as they relate to the educational process. In so doing I make no claim to originality nor even to a complete and coherent collating of the ideas of others on this subject.

Increasingly I have felt, and I know that others have shared with me the feeling, that a wide and widening gulf separates the foremost thinking in measurement from the reality of measurement and evaluation as carried on in the classroom. I fully recognize the need to push ahead on the technical and theoretical front but concerned as I am with the implications of measurement for the improvement of the educational process, I would suggest that there is an equally urgent need for relating some of our thinking about measurement to these matters. It is for this reason that I have chosen the topic Evaluation as Instruction.

Perhaps some slight residue of my mathematical training prompts me to start with a set of assumptions. After numerous attempts at producing such a set I have arrived at ten assumptions paired in such a way as to suggest what seem to me to be parallel elements of evaluation and instruction. Indeed, there is more than a suggestion that good instruction is simply continual evaluation. The assumptions, which are in part also definitions and principles of learning, are:

1. Instruction is effective as it leads to desired changes in students.
2. Evaluation is effective as it provides evidence of the extent of the changes in students.
3. New behavior patterns are best learned by students when the inadequacy of present behavior is understood and the significance of the new behavior patterns thereby made clear.
4. Evaluation is most conducive to learning when it provides for and encourages self-evaluation.
5. New behavior patterns can be more efficiently developed by

teachers who know the existing behavior patterns of individual students and the reasons for them.

6. Evaluation is conducive to good instruction when it reveals major types of inadequate behavior and the contributory causes.
7. Learning is encouraged by problems and activities which require thought and/or action by each individual student.
8. Evaluation is most significant in learning when it permits and encourages the exercise of individual initiative.
9. Activities which provide the basis for the teaching and learning of specified behavior are also the most suitable activities for evoking and evaluating the adequacy of that behavior.
10. Activities or exercises developed for the purposes of evaluating specified behavior are also useful for the teaching and learning of that behavior.

As one immediate result of these assumptions we conclude that:

Evaluation does not differ from instruction in purposes, in methods, or in materials and can be differentiated from instruction only when the primary purpose is that of passing judgment on the achievement of a student at the close of a period of instruction.

I should like now to examine with you what I believe to be some of the major implications of these assumptions.

First, in regard to teaching it appears to me that the following statements may be readily inferred from the preceding assumptions:

1. Classroom practices which are restricted to textual or teacher presentation of knowledge and the testing of the extent of recall of this knowedge are unworthy of the name instruction.

   Such practices, regardless of stated objectives, make knowledge the sole objective of instruction. Learning is unrealistic in that what is learned is divorced from reality. The teacher covers content but does not instruct students. The majority of students remain completely passive, and work only to memorize what the teacher emphasizes.
2. Good instruction will be concerned less with what the teacher is doing or wants to do and more with what the student is doing.

   It is readily accepted that the supervisor who devotes most of his time to planning his own work will not last long; he must accept as his major responsibility the planning of the work of each of his workers and the assisting of them when difficulties arise. The analogy is not amiss in considering the task of the instructor. Learning is an individual phenomenon and results largely from the efforts and activities of the individual. Good instruction involves careful planning of specific tasks with definite purposes which can be undertaken and completed by each student. It is not that which the instructor does that counts; it is what he stimulates the student himself to do that yields the most significant educational results.
3. As new materials and skills are presented, the instructor will simultaneously assign tasks which require applications of these by the student, thereby providing both to the teacher and the student concrete evidence of the utility and of the mastery of those materials and skills.

   This conclusion is simply an extension of the preceding one suggesting that the assigned tasks not be repetitive or copywork in nature but require relation of new

materials and skills to old ones and the application of both to the achievement of a deeper insight into old or new problems.

4. For motivation of learning and for continual evaluation of the effectiveness of instruction the instructor will check almost daily for evidence of changes in students and will seek to make these changes as evident and understandable to them as to himself.

   A common complaint of students is that they receive less than perfect grades and have no knowledge of the imperfection. In composition and speech work a similar complaint by students is that they see no improvement over the period of time in which they take the course. I am convinced that there is real ground for both complaints in that instructors themselves are not commonly aware of changes in individual students and certainly make little attempt to provide concrete evidence of change to the individual himself. The irritation evoked in many teachers by students who want to know precisely what is wrong with their work is perhaps less a resentment that a judgment is being questioned than it is in the realization that that judgment is highly fallible and hardly defensible.

5. To encourage integration and transfer, the instructor will seek for tasks from or impinging on other fields of knowledge which do or can utilize materials and skills presently in focus in the course.

   The association of ideas, concepts, and principles from one field with those in another, or with current problems is one mark of a well-educated person but it is very likely to be frowned upon in the classroom in fear that the associations made will be superficial. The instructor himself, trained in one field, hesitates to make reference to others, and the student who attempts it courts ridicule or reprimand. Yet life is not departmentalized like colleges and the development of individuals of the ability to integrate knowledge and experience and to transfer it requires practice in it.

I am aware that many teachers presently regarded by colleagues as highly competent and even outstanding would hotly debate or as coldly ignore some of these points either on the grounds of irrationality or impracticability, but I am much more troubled by the fact that many teachers who assent to these views depart so far from them in their own practice. Yet, since our primary concern here today is with measurement, I must ignore the many issues raised by these remarks about teaching in order to deal with other more immediately relevant matters.

In regard to evaluation I find my assumptions above leading to the following conclusions:

1. Testing for grading should be relegated to a minor role in instruction and possibly even assigned to an independent evaluation agency.
2. Testing for knowledge should be supplemented and even in part replaced by broad, pervasive, and continuing evaluation or assessment which becomes the major part of instruction and therefore indistinguishable from it.
3. Testing practices which involve comparison with status norms for superficially similar groups should be replaced by practices which emphasize comparison with maximum gains made by students of similar background and ability in regard to the same objectives.

4. Concern with reliability and validity as statistical concepts characterizing evaluation instruments or procedures should be replaced by concern with the permanency and relevancy of learning as concepts characterizing the quality of instruction.
5. Psychometrics—at least as practiced by those concerned with instruction and with learning—should become somewhat less concerned with rather exoteric mathematical investigations of measurement theory and more concerned with the proposing of models and hypotheses directly useful and testable in classroom and life situations.

These conclusions indicate the need for some changes in our evaluation and measurement practices. Each of them is sufficiently significant to call for some discussion.

I have suggested that testing for grading should be relegated to a minor role in instruction. During a period of visiting classes in many different colleges last winter and spring I found practically no cases in which there was any indication that testing or evaluation was carried on for any purpose other than grading. Frequently, testing sessions are reduced to one or two per quarter in order that there be more time for covering the ground. Half-jokingly, but none-the-less seriously, instructors still hold the threat of presence in an examination as a reason for paying attention and learning of a particular point. Tests and examinations are still catastrophic events which are greeted by the students with dismay, for testing for grading is inevitably a threat to students and not a procedure which is either conducive to motivation or to learning. With rare exceptions I doubt that an instructor can have really close association with his students and get a true picture of what they know or what they are thinking about while he brandishes the weapon of a grade. If you doubt this, may I raise the analogous question as to how frequently an instructor gives vent to his actual feelings and reactions about an idea proposed to him or to the staff by his dean. Authority in the form of the power to reward or to punish inevitably modifies the overt thinking, expression, and action of those to whom that power is applicable. I must hasten to add that experience in working with a staff in which all grading was taken from their hands indicates that such divorcement cannot be successful unless the staff itself accepts the principle evoked in these remarks. I hope, however, and I believe that it is not circular reasoning on my part to say that those teachers who have given most evidence of fully understanding general education objectives and have evidenced most interest in the needs of individual students have been those who have rather readily accepted the principle of divorcing grading from instruction.

So far my remarks have applied perhaps more to teaching than they have to evaluation but they have implications for evaluators. So long as teachers are primarily interested in the use of tests which enable them to assign a grade on content materials, teachers and test agencies will of necessity spend most of their time on the making of achievement tests which emphasize knowledge of specific facts. There is more than an indication, however, in the light of my experience in the Cooperative Study of Evaluation in General Education that many teachers only await the availability of instruments and techniques which will provide evidence of the development in students with regard to other types of objectives to make extensive use of such evaluation materials.

To illustrate the insidious effects of the grading practices, I want to remark

briefly upon some of our experiences with instruments dealing with affective objectives. One instrument developed in our Cooperative Evaluation Study was an Inventory of Beliefs, which seeks to get at certain attitudinal elements which are related to general education. The statements in the inventory are stereotypes or prejudices commonly and freely expressed by individuals. When students are asked to react to this Inventory, they must be assured very definitely that the results will be in no way used for grading. There have been cases in which students have indicated that they have not given an honest answer to the statements because they feared that an individual instructor might look at the results and be likely to recall and hold them against him at a later date. Similarly an inventory of attitudes toward the humanities and the people working in that area was made up of statements actually extracted from students in informal situations or from essays in which they were writing on other topics. Placed, however, in the context of other similar items of this type, we found that the humanities attitude inventory gave very little in the way of a range of scores, simply because students realizing now what the inventory had been developed to do, gave what they readily perceived to be the approved response to the items. Their inner feelings on the matter were concealed partly, at least, because they did not trust the instructor to deal fairly with them if he knew how they really reacted.

It has been said that the grade determines what the students work for. I think it may equally be said that the possibility of grading a student determines to a considerable extent what the instructor considers important. Certainly, teachers are inclined to ignore objectives for which they have no ready or defensible means of evaluation and grading. I also feel quite certain, and I do not know whether it is fortunate or unfortunate, that students generally speaking are less concerned about grades than are their teachers. I feel that there is only a limited set of objectives upon which we dare to grade individuals and that so long as a grade holds the most prominent position in our thinking in the classroom it will restrict both our instruction and evaluation to this limited set of objectives.

My second conclusion regarding evaluation was that testing for knowledge should be supplemented and in part replaced by a broader evaluation. As this is done, evaluation becomes a major part of instruction and ultimately indistinguishable from it. I should like to illustrate by a number of concrete examples what I have in mind by saying that instruction and evaluation can become one and the same thing. The Angell-Troyer Self-Scorer, sold by Science Research Associates, provides a means whereby a student taking a test continues to select answers until he obtains the right answer. If the answer that he first chooses is incorrect, in punching the answer sheet he finds only a white space revealed. He continues punching until a red mark shows up under the hole punched out and then he is assured that the correct answer has been obtained. I have recommended to numerous teachers that this Self-Scorer be used. While somewhat inconvenient to handle, everyone who has worked with it has found that very favorable student reaction is created. A technique frequently used has been to have a small number of questions, perhaps ten or a dozen, covering the major points of an assignment to be discussed for the day. The students answer these questions on the Self-Scorer. Those who really know the answers are assured that they are correct. The larger number who may have been, when they came to class, reasonably sure they knew something about

the assignment and who find that they must make two or more selections before finding the right answer are motivated to find out why. Those individuals who had to punch all four answers before getting the right one are concerned about it. There is little difficulty in getting a warm discussion when the test items have been completed. An interesting point in connection with the Self-Scorer is that since the student has punched a hole in the answer sheet he may retain the answer sheet in front of him throughout a discussion with no concern on the part of the teacher that any modification will be made in the results. There has been at least one investigation which would indicate that the use of the Angell-Troyer Self-Scorer results in greater retention on the part of the student and the suggestion in the same investigation that the Self-Scorer followed by a discussion of the results would result in still greater retention. One of the major values of the device is that each student is forced to go on record with regard to each item. The usual tendency in most classes is for either the instructor or a few students to take the lead in answering questions and discussing points. The reticent individual easily arrives at the decision that the answer which is finally indicated as right is the one he felt was right all the time. He just does not realize that if he had had to put himself on record at the beginning of the discussion he would have marked an incorrect answer and he has not consciously analyzed the reasons for his incorrect position. There is real value in having individuals realize just how much or how little they know and thereby give them some incentive to do something about it.

As a second example of the coincidence of evaluation and instruction I should like to refer to the Theme Analysis Handbook developed in our Cooperative Evaluation Study. This Handbook originally developed out of a concern on the part of our Communications teachers that there be some improvement in theme reading. A large number of themes were collected from all of the colleges involved in the project and teachers undertook to agree on a grading for them. They soon found that they were reading themes from different viewpoints and with different sets of values in mind. It became clear that they had to formulate some kind of a statement of intent or purpose before they could move to any agreement. Gradually they found that many of their differences arose out of the fact that the reading was rather superficial in many cases, that one individual was emphasizing one aspect of a theme and another individual another aspect. Too frequently all readers were missing the point of examining what the individual student was really trying to communicate—a viewpoint which requires constructive suggestions as to how the intent of the student might better have been accomplished. As this became the guiding purpose in theme reading it was found that much more agreement could be reached, although the task of reading a theme in this fashion became a very arduous one. The reader must determine just what it is the student is trying to accomplish in the first place and having reached this judgment he must then re-examine all phases of the theme in such a way as to be able to make concrete suggestions to this student as to how the theme might be improved. This becomes less a task of assigning a grade to a student and much more a task of helping the student to evaluate his own effort. Finally, a rather large number of themes were thus evaluated and put together in a theme analysis handbook. The first purpose which the committee had in mind in using this was that it would be a device whereby teachers of Communications could work together

on a set of themes and reach some commonality in their thinking. A second usage envisaged was that of giving it to new instructors on a staff as a basis for helping them develop the skills of theme reading. Some teachers, however, found that to use these themes with an opaque projector provided an interesting and instructive experience for the students. The students could be asked to make the same kind of criticism of the theme as the teachers had engaged in. After this criticism had been rendered it was then possible to throw on the screen the analysis made by the Communications teachers and the students could then, point by point, compare their own reactions with those of the committee. In a sense, this is only a substitute for making such an analysis for each individual. It is quite obvious, however, that a teacher cannot take the time to make the detailed sort of analysis that was involved for every theme of every student. There is some indication in a subjective way that a student who sees a theme analyzed in detail after having undertaken himself to make such an analysis obtains a better idea of the problems of communication than he had before. Thus, a device originally intended as an evaluation device has been found to have tremendous instructional value. The Theme Analysis Handbook may be regarded as the placing of the evaluation of themes on a concrete objective level whereby the evaluation has obvious significance to students in indicating the qualities of good writing.

The Test of Critical Analysis and Judgment in the Humanities is still another example of the same coincidence of instruction and evaluation. Teachers of the humanities indicated a concern that students faced with a work of art be able to arrive at some judgment of it based on an application of facts, principles, and personal reactions. They found that when students were simply asked to react to a picture or to a piece of music, that practically nothing was obtained. An attempt to evaluate the students' ability in critical analysis and judgment failed simply because most students in most classes had been given no opportunity to develop such facilities and therefore were unable to exhibit them. In most cases it was obvious that the task of analyzing and judging a work of art was something rather new to the student and that he was even bewildered by the request. This being true it might seem that we should concentrate on teaching. On the other hand, asking just what kind of an analysis and judgment or just how a judgment might be arrived at for a particular work of art holds promise of clarifying ways of instructing students in order to develop these abilities. As agreement was reached on a number of points and a series of suggestions set up for students as to things to which they might react in analyzing the work prior to making an integrated judgment. When this was presented to students, it was found that something more tangible in the way of a response was obtained from students. However, the vast majority of students still betrayed a lack of familiarity with this objective. The structured essay format adopted for the critical analysis and judgment in the humanities is probably most valuable in dealing with students who have had little experience in this type of work. In other words, the instrument can be used in making assignments or it can be used as a guide in class reactions to a work of art. Ultimately, however, when a student has had enough practice it would no longer be necessary to put before him the specific points in the structure of the critical analysis and judgment test. He would already have arrived at that structure or a similar one based upon his own experience to which he could resort when faced with such a task. The de-

vice is still a test and can be used as such, but if anything its value as a teaching device exceeds its value as an evaluation device.

I should like to use one more example to illustrate the combination of instruction and evaluation. The science committee of our project became interested in the use of current science materials because of a feeling that the further contact of many students taking only one general education science course would be largely with popular and semi-popular science articles. If students develop an interest in reading this material and some skill in applying their knowledge of science and their understanding of the scientific attitude toward these materials, it might be reasonably expected that:

1. They would continue some contact with science.
2. They would continue to increase their knowledge of the area by this constant reading.

The task of evaluation resulted in an extensive search of science articles to see what kinds of materials might be appropriate for evaluating students' ability in science reading. It soon became evident that some articles were much more suitable than others. It also became clear that having an article was not the same thing as knowing exactly what one would do who read it critically and understandingly. It became necessary, therefore, to build up a list of behaviors which might characterize the critical and understanding reader of science materials. These in turn were formulated into both essay and objective tests. Use of these in the classrooms indicated that students did not perform very well on such tasks and this was indicative of the fact that in most classrooms it was the first time that students had been asked to make such an analysis of any reading material. As a result both objective test exercises and essay questions raising issues about current articles and extracts from articles, were found by numerous teachers to be a challenging type of activity for the classroom. Such materials evoked much interest and discussion on the part of students, and, often for the first time, caused each individual student to attempt for himself to do something with the material in front of him. Textbooks and teachers who present both problems and solutions to students rarely challenge them because it is an exceptional student who can undertake to disagree with a solution proposed by a teacher. Problems, with solutions not given, involve a much less threatening situation in which each student can propose his own solution and argue for it up to the point where he is convinced that there are better solutions than his. Such use of current science materials in the classroom results in a coincidence of evaluation and instructional practice. Even more significant, the type of behavior envisioned by a group of science teachers for their students at a period later in life—that is the reading of science materials becomes a part of daily classroom experience. Surely the problem of transfer must be much less complicated for the student who has been asked and even forced to read current science materials. It is far more likely that he will continue that behavior than it is that a student with no such experience will initiate it just because he has learned a little more science.

The third conclusion which I have suggested as applying to evaluation and arising out of my earlier assumptions is that there should be a de-emphasis of national and regional norms. To illustrate what I have in mind let me discuss for just a moment the nature of some of the gains which we found on the Test of Critical Thinking in Social Science. In some schools we found over a year's time a gain of one or two

points on the average, a gain scarcely more than the practice effect determined for that test. In other schools we found a mean gain of as high as ten or twelve points. In almost all cases we found that the students who were lower at the beginning made much larger gains than the students who had a high score on the test at the beginning. Particularly it was true that a student whose score on the test of critical thinking was lower relatively than his intelligence level tended to make large gains. Yet there were places where the most able students made larger gains than the least able students in other schools. There were places in which the difference in the gains between the most able and the most inept students was much less than was the case in other schools. Such variations are undoubtedly associated with variations in the educational programs although it was not always possible to identify them. In general, the large gains were made where more overt attention was given to the objective. In accord with the prevailing trends we threw together the data from three major classifications of schools based on the ability of the students admitted and made up what might be called norms. The general implication of these is that a school in which the students make a gain of roughly four points over a year's time on the test of Critical Thinking in Social Science is doing an average job. The difficulty with this is that an average job is—speaking frankly—a darn poor showing. Most Social Science classes in general education are still heavily content centered, most students are given no opportunity to practice or develop critical thinking, and only in a few schools with very outstanding gains can one find in the instructional practice anything which is indicative of real attention to the objectives. For schools and for teachers to adopt as their standard the mean gain found in a set of norms as presently derived is simply to give blessing to mediocrity or worse. The real standard, which should be sought is the largest gain made for students of the same ability level and general background. It would then be necessary to find out just what were the characteristics of the program which achieved those gains and to adopt as many characteristics as are appropriate or even to improve upon them in order to gain still more. Another point which I have noted is that most norms commonly are only given in terms of a classification of types of schools and occasionally in terms of intelligence level. Our finding that low ability students, with low pre-test status on the Social Science Critical Thinking Test, are unlikely to make very large gains is only common sense. The fact that students with average or high scores on the Critical Thinking Test in Social Science, who have low academic ability, make almost no gain which is also very reasonable. Yet with the usual norms provided, a teacher who is aware of some of these situations would find it very difficult or impossible to obtain any information on what would be a reasonably expected improvement on the part of a given student. Perhaps we should take a leaf from the practice of photographers, both amateur and professional. Those of you who are interested in the field know that it is customary when displaying a photograph to provide a detailed description of the conditions under which it was taken, the kind of exposure, lens, etc., used. Likewise, it might be appropriate in educational circles, when tests are given and gains are recorded, to describe in some detail the essential characteristics of the educational experience which was provided. In closing my remarks on norms I think I should make it clear that I do not feel that norms are necessarily evil in themselves. There are some limited values served by them.

On the other hand, the prevalence of national and regional norms emphasizing mediocrity often leads teachers to regard them as the only kind of standard which is significant in judging test results. I suggest that it is a major obligation of everyone in the measurement and evaluation field to see that the almost exclusive use of status norms for superficially similar groups is supplemented or even replaced by practices emphasizing comparisons of the gain made by students with the maximum gain made by students of similar background and ability in regard to the same objectives.

I have suggested that there is a need for de-emphasis of the statistical concept of reliability and validity as characteristics of tests. Among professional measurement people there is a full realization that a coefficient of reliability may mean many different things depending upon how it is computed. It is also a highly variable quantity depending upon the group upon which it is computed. Similarly validity is the characteristic not only of the test but of the situation in which the test is used and of the group with which it is used. Let me revert again to some experiences in the Cooperative Evaluation Study. Dealing with objectives which are not overtly emphasized in many classrooms and even in many schools, it is perhaps to be expected that the reliability of tests would vary a great deal. I recall one test which yielded a coefficient of reliability in the .6 to .7 range in most schools but which resulted in a coefficient of .9 when used in another school. Just why this was true I never ascertained. In other cases we found that certain schools in which there had been obvious concern with critical thinking and numerous things in the program directed at it, yielded higher reliability coefficients for these tests than other schools in which there was little attention to these matters. In regard to critical thinking, we did not find any really satisfactory criterion of critical thinking, but we did attempt to compare the results of the tests with a variety of judgments of teachers and others. Correspondence of teachers' judgments in a particular school with the test results on a Critical Thinking Test varied extremely within a school, but there was even more variation in the kind of judgments made from one school to another. As one might expect, in some schools where the objective was regarded with indifference, the very reactions of teachers to the request for making judgments about Critical Thinking ability almost insured that the ratings would be of little value. Rather consistently we found that teachers who had worked directly with us in the committees made judgments which corresponded closely with the test results. One may of course argue that this is true simply because they knew what was in the test, but it seems to me equally plausible to argue that as a result of working on the test they had a better concept of what critical thinking is and therefore were able to render more reliable and valid judgments of its presence or absence in students. Clearly, reliability and validity are not simply characterizations of instruments or evaluation procedures.

Primary concern in education must not be with the evaluation procedures but with the learning to which the evaluation procedures provide some index. Instead of reliability of a test we should be concerned with the permanency of learning. Instead of the validity of a test we should be concerned with the utility or the relevancy of learning. Viewed in this way, these characteristics are related to the tests but they are also concepts which characterize qualities of instruction and predetermine test performance.

Finally, I have suggested that psychometrics has become concerned with

theoretical activity so far removed from the classroom as to not only have little applicability but actually to alienate the classroom teachers. I am well aware that no hard and fast line can be drawn between pure and applied research. I should be the last to wish to draw one. On the other hand, I do become a little impatient with what seems to be an increasing tendency to pursue fine points of testing theory, which go far beyond any utility which I can foresee for many years in actual instructional and learning situations. I should like to see some of our best minds in measurement confront the reality of educational problems and attempt to propose some hypotheses or models which are directly relevant to and testable in the classroom and in life situations. Let me give you a very simple example of what I have in mind. I referred, at the beginning of this paper, to integration as one of the bywords of the educator. Certainly it is one of the concepts which has been given a great deal of emphasis in general education. There are many different concepts of integration, but I prefer to regard integration as a characteristic of an education person, one which implies that he is able to interrelate everything that he knows, his skills, his abilities, his beliefs and his values in such a way as to deal in a more effective way with situations in which he finds himself day by day. In short, the integrated person is a well-organized individual who is able to make the most effective use of all of his resources in dealing with the problems that confront him. He is the antithesis of our tendency in education to break our objectives into subject matter groups and even to regard critical thinking in science, critical thinking in social science, etc., as distinct and independent abilities. We find some evidence that the intercorrelation between tests or other evidence collected on these various abilities is rather low and this reinforces our feeling that they are more or less independent factors. I am not sure that this must be so. May it not be an artifact of culture and of our present educational policies?

Some time ago I proposed to myself and some of my colleagues the hypothesis that a really good general education program would be characterized by the fact that over a period of a year or two the intercorrelations between tests, giving evidence of various general education objectives, would show a definite increase. In other words, if a person is really interrelating what he has learned there should be an increasing tendency to coordinate all resources on a particular task. It has been of interest in the Study to find that in one school where marked gains are made in the separate abilities but where the courses are quite distinct and where the general education experience is almost restricted to these courses, that the intercorrelations between various tests actually decreased over a period of one or two years. In certain other schools, where there are no clear cut distinctions between a variety of living experiences and classroom experiences on the campus, where instructors in one course are quite cognizant of what is going on in other courses and make every effort to interrelate them, we find that over a similar period of time the correlations among a number of different tests increased in size. There are other possibilities which might explain this and we have explored a number of them. At the present moment I have not been able to explain this data on any basis other than the original hypothesis, but I should still like to regard it as a tentative one and have a great deal more study of it before accepting it as a simple way of ascertaining the extent of integration in an educational program.

I have presented these remarks with some fear that I may be misunderstood

and that some people in the measurement field may, in some way or other, feel that I am trying to demean the importance of what has been accomplished. I assure you that this is not the case, but it is true that in working closely with classroom teachers over the period of the past few years I have become increasingly aware of the deep and increasingly deeper suspicion with which many teachers regard our whole field of activity. Yet in every case where I have successfully made contact with an individual and interested him in the problems of evaluation and the interrelationship of these to the problems of instruction, I have found the attitude to change completely. In every case there comes an awareness that these things are not distinct fields of operation and that evaluation has a great deal to contribute to the improvement of instruction—that, indeed, as I have already said repeatedly — evaluation and good instruction are indistinguishable. To those of us interested in evaluation this seems to be only just recognition of the worth of our activity. However, we need also to realize that those working in the field of instruction can contribute to evaluation a great deal more than we have solicited from them.

Improving Evaluation of Educational Outcomes at the College Level

PAUL DIEDERICH

SUMMARY OF DISCUSSION

Dr. North raised the question as to whether instructors avoid re-using effective items for fear that students may learn to expect these items and make special preparation for them. Dr. Ebel replied that item files are kept and that he has seen examinations in which as many as two-thirds of the items had previously been used. He was asked whether under these circumstances students anticipate these items and prepare for them. Dr. Ebel felt that this was not the case and cited the example of a medical college instructor who had exactly the same items in one examination for five years and had found no upward drift in scores. Dr. Gulliksen asked whether this result was regarded as favorable or unfavorable and Dr. Ebel stated that at least it showed that students are not as adept at taking advantage of the repetition of items as is sometimes supposed.

Dr. Carroll asked whether there is any evidence whether the different categories of items listed under "relevance" actually test different things. He cited a factor analysis of a similar classification of items at Harvard which indicated clusters in other ways than logical analysis had supposed. Dr. Ebel stated that no claim was made for the factorial purity of the categories. He said that it would be impossible to carry out factor analyses on all his examinations. The analysis was essentially that of content, he said, of what the examination emphasizes. Dr. Ebel remarked that a few mistakes in classifying individual items do not seem to affect seriously the picture of the examination that this type of analysis discloses.

Individual Versus Group Decision Making

IRVING LORGE

INDIVIDUAL VERSUS GROUP DECISION MAKING

THE PROVERBS and maxims of any people do summarize their experiences and wisdom over the generations. In respect to individual and group behavior, of course, there are many such proverbs. Most of you have learned, and, at times, acted upon such aphorisms as "Two heads are better than one" and "In the multitude of counsellors, there is safety." By contrast, however, there are adages that suggest "Truth is lost with too much debating" and "If you want something done, do it yourself." Indeed, formal group thinking has been disparaged by an unknown phrase-maker's definition of a *committee* as "a group of men that keep minutes and waste hours," and, informal group procedures have been reviled in Ambrose Bierce's definition of *discussion* as "a method of confirming others in their errors."

Though these succinct summary statements are in contradiction, each may express a true and useful generalization. The apparent disagreement may be limited for certain kinds of tasks, or for specified sorts of groups, or some classes of conditions. Variations in the tasks to be done, or the organization of the groups to solve them, or the circumstances under which the action is to be taken, may produce different outcomes or results. Today's social scientist, however, is not inclined to search the sayings of the fathers for generalizations about the differences in group and individual performances. He tends to go through a process of making a *review of the literature*. Such a reading of the research literature should, indeed, reveal the range of generalizations as well as the means and evidence upon which they were formulated. When experimental studies are approached (in the aggressive mood) the variables of organization, task and criterion are found to be related to the so-called generalizations about the successes of groups and of individuals.

In the restricted range of studies that distinguish between the products produced by groups and by individuals, it soon becomes apparent that the conclusions in the textbooks are broader than the evidence warrants. Very quickly comes the realization that in considering group versus individual *decision making*, "a group is not a group is not a group." In social psychology and in sociology, the concept of group implies the sense of two or more persons interacting among themselves to accomplish some objective. In fact, the essential differentiation between group as "an assemblage or aggregate" and group as "a social organism" inheres in the notion of interpersonal interaction. Groupness implies interaction.

Yet, some of vaunted superiority of the group is based upon a kind of group whose individual members never met together, and, who, as a matter of experiment, never knew that other persons were working at the same time at the same task. For instance, Kate Gordon asked individual college students to arrange a set of weights in the

correct order. Each student did the task independently of all other students. From the resulting *protocols*, so-called "groups" were formed by averaging the rankings of five (or of ten, or of fifty) individuals selected at random from the full supply. By the criterion of correlation with the actual physical order of the weights, she concluded that "the results of a group are distinctly superior to the results of the average member" —a conclusion which cannot come as a complete surprise to those conversant with the elementary facts of tests and measurements.

Or again, the superiority of the group is founded upon a kind of group whose individual members meet in the same social setting although each one works independently upon the same task without any personal interaction. A typical instance of such experimentation would seek the answer to the question, "Do individuals make better test scores working in complete isolation than working in a group climate?"

More recently, the superiority of the group has been demonstrated on the basis of *ad hoc* groups whose individual members were designated to work together to accomplish some externally imposed task. For example, in the widely quoted Shaw study, a college class was divided in half with one of the halves of the class arbitrarily formed into groups of four to work on problems proposed by the experimenter.

These three illustrations suggest that in actual experimental procedure, groups are of different kinds. At the upper extreme the theorists think of genuine interacting face-to-face groups who have a tradition of working together with the responsibility for the accomplishment of a broad task. Solutions that such groups make have not been studied by psychologists. At the lower extreme is the concocted group made by some external authority who adds together the independent products of several different isolated individuals, each of whom worked alone. It is the solutions that such "groups" produce that have been a principal source for textbook generalizations.

Moreover, experimentally, there is even wider variation in the nature of the tasks groups and individuals are required to perform. Groups have been contrasted with individuals on the basis of such tasks as estimating the number of beans in a bottle, as judging the aesthetic value of music, as predicting the date for the end of World War II, as selecting a bread for nutritional purposes, as planning a course of action for accepting a bequest, as learning a maze, as solving a difficult mathematical problem.

Each different kind of task has been subsumed under the broad term "decision-making." The thesaurus recognizes as related to the verb "to decide" the verbs, judge, conclude, ascertain, determine, deduce, infer, estimate, appreciate, value, assess, consider, settle, and choose. All strongly suggest that "thinking" or "reasoning" is basic to the accomplishment of the task. Indeed, one can only wonder why the jargonistic "decision-making" was substituted for the readily available "deciding."

Deciding, or (not to lose face with my fellow workers) decision-making, involves such ideas as the task, its acceptance, a process of active search for ways and means to get it done, with an ultimate selection of a plan and consequent action. The history of the study of thinking, indeed, is replete with formulations like this. As a matter of fact, the early tasks or "Aufgabe" set by the Wurzburg School, to a degree, set the nature of subsequent experimentation. Basically, the Wurzburgers were studying what was happening in thinking and willing. So great was the attention on the process that they tended to underemphasize the quality of the product or solution. For instance,

Ach asked for a free or controlled association to a given word, or Bühler asked, "Was the theorem of Pythagoras known in the Middle Ages?" or "Can you get from Vienna to Berlin in seven hours?"

Solutions for such tasks may be assessed readily as "right" or "wrong." The standard appraisal for such simple task problems is truth from the viewpoint of fact or actual measurement. The answers to the estimation of weights, the number of peas in a bottle, the prediction of a specific date in the future, each had an objective criterion of correctness. In most studies of the *process* of thinking, some readily available abjective criterion for correctness was available. Yet most real life problems, that must be pondered, do lack an objective external criterion of correctness or adequacy. Many problems of interpersonal relations are so complex that policy, or plan, or course of action cannot be evaluated as just right or wrong. Such problems range from those whose solutions are clearly right or wrong to those that can be appraised only by consideration and evaluation of many different consequences.

For instance, Marjorie Shaw's problems in the first part of her experiment consisted of three different but closely related mathematical puzzles: each was a transport problem like that involving the three jealous husbands and the three beautiful wives who had to get across a deep river in a rowboat for three under the constraints that only husbands can row and that the wives cannot be trusted in the presence of another man unless the husband is also present. Twenty-one individuals and five groups attempted the three problems. In terms of just the number right, there were five solutions for the individuals over the three problems in contrast with eight solutions for the groups. Disregarding the possibility of differential transfer-of-training in groups and in individuals, her conclusion about group superiority was based on comparing 7.9 percent for individuals versus 53 percent for groups. Incidentally, no test of significance was made.

In the second part of her experiment, the problems involved the rearrangement of the scrambled words of a final sentence of a prose passage, the completion of three and a half lines of a sonnet, and the location of a school building under the constraint of minimizing school bus mileage. No group or individual solved the final two problems, but on the sentence rearrangement four of the five groups and three of the seventeen individuals solved the problem without error. An additional group and seven individuals made just one error which involved a word reversal or the placement of the word "there." Although the errors did not meet the criterion of perfection, they did not affect either the smoothness or the sense of the sentence. Solutions, in other words, could vary in their closeness to the original as well as in their adequacy. As the problem allows for more and more different adequate solutions, the criterion of absolute right or wrong seems more and more arbitrary. The review of the literature suggests that the differences between groups and individuals in decision-making may be related not only to the nature of the task but also to the method for evaluating the quality of the solution.

In general, the more complex the problem for decision, the more difficult it is to find adequate means for the appraisal of its solutions. It was, indeed, fortunate that the Air Force was concerned with the appraisal of its staff work. For under two different contracts from the Human Resources Research Institute with cooperation of the Air University, it gave the Institute of Psychological Research an opportunity to study the differentiation between the

goodness of solutions of groups and of individuals for two very different kinds of problems. At one extreme, were used problems for which there were a known or knowable solution. These are designated as "eureka" problems. At the other extreme, were used problems so complex that it would be impossible to ascertain the complete efficacy of the solutions. Such complex problems are designated as "rational decision" problems.

An example of the rational decision type was Wilco Air Force Base for which a plan had to be devised to raise the morale of airmen stationed at an isolated Air Force Base in desert country. The base facilities were inadequate to care for wives and families of the airmen; the AWOL and VD rates were unusually high; the social and recreational facilities of the nearest town, some forty miles away, seemed limited to gambling, drinking and prostitution. The elected officials and the town's leading citizens were greatly disturbed.

The usual procedure was to administer this problem to individuals and to groups before instruction at the Air University. The groups were *ad hoc* leaderless groups of six to eight officers, usually majors and lieutenant-colonels. All solutions were to be written in fifty minutes. The results for such individuals and *ad hoc* groups, relatively naive in group interaction and problem solving techniques, demonstrated that the average quality of the individual decisions was superior to the quality of the average group decision.

The "eureka" problems were adapted from those used by the Office of Strategic Services in its assessment program. For example, one problem required the formulation of a plan of action for getting a cadre of five men across a road mined with [illegible] sensitive enemy mines that could neither be neutralized nor dug up. Scattered about was potentially [illegible]is, including beams, ropes, discarded truck tires, etc. The groups were *ad hoc* groups of four or five Air Force officers in training, usually junior and senior college undergraduates. The solutions had to be accomplished in an hour. The results were in sharp contrast for individuals and for *ad hoc* groups who were also relatively naive in group interaction and in problem-solving techniques. They showed that the average quality of individual plans was markedly *inferior* to that of the average group.

The "eureka" problems, of course, could be assessed in terms of an absolute criterion of "Do the men get across?" In fact, each different plan of action was tried out in the real field situation to estimate its work-ability on a pass-fail basis. In comparison, however, the goodness of the solutions for the "rational decision" on Wilco Air Force Base could only be estimated by experts judging their adequacy in terms of forseeable consequences.

The "rational decision" solutions written for the Wilco Air Force's problem showed an unusual range. Experts could recognize that many solutions tended to treat just a specific symptom of the problem; on the other hand, experts appreciated the nicety of the perceptions in some of the solutions. These, indeed, indicated not only the diagnosis of the many factors in the problem but also an anticipation of the consequences of the several aspects of the action plan. The extraordinary diversity among the solutions suggested that it would be possible to identify the factors diagnosed in the plans for alleviation of the situation. The technique, essentially, requires an analysis of the decisions by broad areas and factors and by specific courses of action.

The steps involved: first, making a representative sampling from all decisions collected; second, developing a content analysis of each positive point or idea regardless of its location in each

written decision; third, arranging in a master scheme, the broad areas in which each specific point could be located; and fourth, assigning numerical credits to each broad area and within it to each specific point.

For instance, for Wilco Air Force problem, 300 solutions were selected at random. These were content-analyzed into broad groups such as:

Regulations and discipline
Leadership and training
Morale and military customs
Problem solving procedures
Housing and facilities
Civilian-military relations
Passes, leave and work schedules
Recreation

Each separate point under each broad category was identified separately by a code number, e.g. "communicate disciplinary policy to personnel" as point 122, or "organize meetings and/or committees of airmen to work on problems to formulate policy" as point 411.

The content analysis based on the reading of three hundred decisions encompassed more than 95% of the points subsequently found in a second random sample of two hundred more decisions. The summary of the different points and ideas found in the five hundred decisions constituted the basis for the subsequent quantification of the quality score.

Four expert judges, after reading the problem, assigned numerical credits among the categories and to each point within a category under the constraint that the total sum for all credits would equal 100. The score for each point was the average of the values by the four independent judges.

Each decision, then, could be coded according to the content analysis scheme. The sum of the point values for any decision could be considered as the total quality score. Such a coding and scoring procedure, of course, allows, in addition, a frequency count of each specific point made by individuals or by groups as well as a separate count of the number of broad areas in which such points are made. It cannot be expected that any one individual or group will make *all* the points in the complete master key. As a matter of fact, the average score before instruction is about 20 with a standard deviation of about 7.

There is no practical way for validating the quality score for the solutions to a problem as complex as Wilco Air Force Base. Try-out of each suggested decision is neither possible nor feasible. At second best, one can estimate whether the quality score does correlate with the judgments of independent experts about the over-all goodness of the decisions.

To test this, the relation between the rank-order of fifty decisions for over-all goodness and the quality scores was estimated. Six competent judges, independent of those used in establishing the rating system, arranged the decisions in order from best to worst, a process that took each judge about four hours. The average inter-correlation among the six judges was .74 so that the rank corresponding to the sums of judges' ranks have an estimated reliability of .94. The correlation between the rank based on the sums and the quality score was .82. For other problems, the results are equally satisfactory.

In passing, it may be stated that the reliability of coding the solutions is extraordinarily high. Two independent content analysts analyzed the fifty decisions for points. The correlation between the two analysts for the quality score was .97 without significant difference between the mean quality scores or the standard deviations.

The technique of content analysis can be done objectively, reliably, and

quickly. The scores are sufficiently sensitive to allow the measurement of the effects of instruction in problem-solving and in group dynamics, and of the differences between the decisions of individuals and groups.

The results were so illuminating that content analyses and scoring systems were also established for the "eureka" problems. The advantages of the content analysis is that it allows not only the designation of the kinds of points made by individuals and by groups but also serves to suggest to teachers the kinds of points that are often, and the kinds that are rarely, proposed. In a sense, a content analysis of several hundred decisions is a summation of the best of many minds. It tends to codify all the possible positive points for handling a problem situation.

Against the panorama of such ideas, every specific decision, whether of an individual or of a group, can be appraised. Not only does the content-analysis scheme allow for evaluation of each such decision, but also it gives the teacher a basis for making a diagnosis of faults and errors in problem solving. For instance, it allows the instructor to recognize that the solver placed most of his attention on a single factor and its alleviation; or again, that some solvers attack a symptom without considering its cause; or again, that a solver made a plan without anticipating the new difficulties that would ensue.

For example, the content analysis of the "eureka" type solutions, when utilized for contrasting those of individuals and groups, indicated that the individuals apparently do not check their ideas for getting across the road, for example, as well as do groups. In groups, each member tries to consider the workability of each suggestion with the consequence that poor and inadequate ideas are rejected. Groups, by comparison, produce not only a greater number but also more easily workable solutions for "eureka" problems.

Groups before writing their solutions, tend to exercise a greater amount of criticism of the ideas suggested by the various members. As a group, therefore, they reject the apparently inadequate and the clearly wrong. Individuals, however, are not quite so autocritical. For they do suggest fewer good ideas and more poor and even inconsistent suggestions.

In the "rational decision" solutions, the behavior of groups and of individuals as behavior is not very much different, i.e. the groups are more critical of the ideas and suggestions of their fellow-members. The result, however, with this kind of problem tends to lead to a reduction in the number of good ideas actually written down. This does not mean that good ideas were not introduced into the group deliberation. As a matter of fact, in process observation and in controlled studies, the group fails to specify in its report at least two-thirds of the good ideas produced within the group. The individual, however, in his less critical behavior does record a variety of suggestions related to different aspects of the complex problem, i.e. the individual tends to record ideas over the full range of the factors in the problem. In complex probems with wide potential range of ideas and actions the tendency within the group toward critical review has the effect of constraining the area in which ideas get recorded, salvaging just about a third of the ideas contributed by the group. It should be stated, however, that these ideas, in general, are the more important when appraised in terms of the average value of the recorded points, i.e., dividing the total Q.P.S. by the number of different points recorded.

The Quality Point Score with its prerequisite content analysis gives promise for the appraisal of high level executive

decisions, those by the individual executive, or by staffs, or by an interaction of an executive with staffs. A complex problem really extends a good executive: he must appraise the situation, discriminate and identify the essential factors, and then innovate plans and strategy for handling the situation. The Q.P.S. provides a basis for the *appraisal* of the thinking of individuals and of groups. As such it emphasizes the product more than the process. Inference from the quality of the product, however, leads to useful hypotheses about group and individual problem-solving and decision-making.

As a method, it is different from one that asks the individual or the group to select from a limited number of alternatives, a single best plan or action. The difference inheres in giving credit for the actual number of ideas recorded. The Q.P.S., therefore, for evaluation of the so-called "higher mental processes" is a method of greater scope and greater significance. Of course, objective tests have been, and could be, produced based on the selection from among alternatives; and, obviously the results of such tests will correlate quite highly with a Q.P.S. appraisal. Yet, in a sense, the objective tests do not estimate adequately the ability of an individual or of a group to produce ideas, to innovate plans, and to evidence originality in policy or concept.

Desirable as machine scoring may be, it should not obscure the fact that the quality of high level decision-making lies not only in the selection from among alternatives but also in the capacity for developing the alternatives to be considered.

Individual Versus Group Decision Making

WILLIAM G. MOLLENKOPF

SUMMARY OF DISCUSSION

Dr. Langmuir raised several questions: one with regard to the make-up of the groups, a second with regard to the variability of responses in the two situations, and a third as to whether the results which indicated a superiority of the individual over the group might possibly have been an artifact of the scoring system. Dr. Lorge indicated that officers had been assigned to groups systematically from a roster; that there had been a greater variability of response for individuals, with the best individual solution better than the best group solution but with the poorest individual solution being distinctly inferior to the poorest group solution; and that he did not believe the results were an artifact of the scoring procedure. Dr. Lorge further stated that the group recorder screened out some of the ideas of members of the group, and kept them from getting into the record.

Dr. Andrews pointed out that the recorder's actions might have led to an apparent smaller variability of response for the group, and wondered whether tape recordings might indicate to what extent ideas given in the group situation had not appeared in the record kept by the group leader. Dr. Lorge then reported that for the several group sessions which had been tape recorded, some two-thirds of the ideas observable in the discussions had not appeared in the group's record. In part, this apparent shrinkage in number of ideas arose out of the group's ability to reject obviously wrong or distinctly poor ideas.

Dr. Spencer referred to some work of the Office of Strategic Services indicating that the more forceful a person, the more successful he was in introducing his ideas into a group discussion.

Dr. Burke wondered whether the group had been handicapped by the shortness of the time, and made the point that rejection of poor ideas might be as important as the acceptance of good ones. Dr. Lorge replied that when he had tried longer time limits more ideas were obtained both from individuals and from groups, but that the difference between groups and individuals nevertheless tended to persist.

Problems and Procedures in Profile Analysis

F. LORD

REMARKS OF THE CHAIRMAN

I WOULD LIKE to discuss a definition of terms. What do we mean by profile analysis? A profile is commonly considered to be a sort of collection of mountains and valleys arranged in a row. A typical problem in profile analysis involves the comparison of two such mountain ranges with a view to obtaining some measurement of their degree of similarity.

This is not an unreasonable way to look at the problem; however, I think it will be better to approach profile analysis from another point of view. If we have the profile on 30 tests of each of 300 individuals, these profiles represent 300 sets of thirty measurements each. Such a set of data is precisely the subject matter of a branch of statistics called multivariate analysis. I suggest, then, that the problems of profile analysis are no more nor less than the problems of modern multivariate analysis.

Let me mention very briefly a few of the problems and techniques of multivariate analysis. Suppose I walk into my office one morning and find upon my desk the scores of 300 individuals on each of 30 tests. What procedures are available for dealing with these data?

One important function of statistics is to reduce a large, confusing collection of measurements to a smaller, more manageable set of numbers. I might apply the principle component method of factor analysis and derive perhaps five, or ten measures for each individual which would convey virtually all the information present in the original set of thirty measures. A different but analogous method of approach might be to use an appropriate method of inverse factor analysis to show that for certain purposes the 300 individuals tested may be usefully grouped into 10 distinctive types so that for certain purposes it is no longer necessary to consider each individual separately.

So far, I have mentioned two techniques of multivariate analysis which may be applied when there is no hypothesis and no further information available beyond the bare set of data.

Suppose, next, that some of the examinees may be identified as clerks, and some of them as mechanics. First of all, it is likely to be appropriate to use Hotelling's T Test to determine whether or not the two groups can be shown to be significantly different from each other on the basis of the test scores available. This statistical test is simply the multivariate analog of the usual test of the difference between two means. Hotelling's T Test may also be used to determine whether or not it is plausible to assume that a certain individual is a member of a certain group. A related but more complicated problem is that of determining to which of two or more groups a certain individual most plausibly belongs. This is the main problem discussed today by Dr. Tiedemann.

A somewhat different problem is the

following: Although, we may be able to show, using all 30 available scores, that successful Army officers can be discriminated from unsuccessful Army officers, we may be unwilling to consider each of these 30 scores individually. We would like some index for summarizing the information contained in these 30 scores insofar as it relates to discriminating between a successful and an unsuccessful officer. This is one of the problems discussed today by Dr. Anderhalter.

Dr. Tiedeman has given us a very clear description of an important and practical procedure in profile analysis. I should like to mention briefly certain further problems. Dr. Tiedeman can tell us which of two or more groups a certain individual most resembles on the basis of the data at hand. Such a statement, he will be the first to admit, does not necessarily exactly parallel the best possible recommendations for a vocational choice or for a job assignment. Even from a purely statistical point of view, a person making a vocational choice should consider not only the degree of similarity between himself and other members of each vocational group, but he should also consider the following information, if it is available: 1) the number of job opportunities or openings in each field, 2) the probability of success in each vocation, and 3) the rewards of success in each vocation. Given sufficient data on these considerations, the whole problem is amenable to treatment by modern statistical decision theory.

Without discussing recent statistical theory, we are all aware of the fact that if we have available some measure of the success or the degree of success of each of Dr. Tiedeman's mechanics and clerk typists, we can, by multiple correlation methods, obtain some statement predicting the probable success or degree of success of a given examinee if he becomes a mechanic or a clerk typist.

In many situations, one would prefer the information given by such prediction to the information given by a centour score; of course, it would undoubtedly be still better to have both kinds of information. Good measures of the success or the degree of success of large numbers of people working on the job are in practice very difficult if not impossible to obtain, so that Dr. Tiedeman's procedures will doubtless be of widespread practical use. In case criterion data are available, however, it should not surprise us if the conclusions reached making use of this additional criterion information are somewhat different than the conclusions reached without this information. Additional relevant information is, by definition, always capable of yielding improved conclusions.

Suppose that in the foregoing situation we are given still further information. Suppose, in fact, that we are given the number of vacancies in each job area, and that we are required to select individuals to fill the specified vacancies. Here we have what is often called the multiple selection problem. Various solutions to this problem are available, tending to maximize in some sense the net success or the net product of the totality of assignments.

I should point out that the problems of profile analysis are not all easily solved by currently available statistical techniques. If the data at hand do not have a multivariate normal distribution, and cannot be transformed so as to have such a distribution, serious difficulties arise.

It is frequently asserted, epecially among personality psychologists, that it is not enough to work with an average of the test scores of the examinees. On the contrary, one must pay attention to the pattern of scores, considering each score individually in relation to

each of the others. The meaning of a given score on the first variable changes according to the score observed on the second variable. This point of view is clearly represented, for example, by the insistence upon the use of multiple cutting scores for the selection of examinees, rather than the use of a composite or average score. In the field of profile analysis, this point of view would require in the limit that each shape or pattern of profile should be considered separately from every other shape or pattern of profile.

If the psychologist is adamant in his insistence that each measure in a profile has a different meaning depending on the value of every other measure in the profile, it is then clear that we are not dealing with a normal multivariate distribution. If the psychologist can set up specific hypotheses predicting the behavior of the data in such a situation, these hypotheses can probably be tested fairly readily. If, however, he must rely on the data to suggest hypotheses to him, and if the number of variables in the profile is more than two or three, then any thoroughgoing investigation will require huge numbers of cases and an amount of labor that will often surpass even the potential capacities of electronic computing equipment.

Dr. Anderhalter's report is relevant to this problem. His group made a careful, intensive, and persistent effort to deal with the individual patterns of their profiles. With his particular data, however, the most satisfactory results were obtained by the use of the linear discriminant function. The linear discriminant function is, of course, nothing more nor less than an optimally weighted average of the test scores.

Problems and Procedures in Profile Analysis

O. F. ANDERHALTER

AN APPLICATION OF PROFILE SIMILARITY TECHNIQUES TO RORSCHACH DATA ON 2161 MARINE CORPS OFFICER CANDIDATES

IN THIS paper I shall present the results obtained when applying several methods of profile analysis to Rorschach data collected in a Marine Corps officer selection situation. The emphasis intended here is not so much upon the results and conclusions concerning this particular measuring instrument applied in this selected situation, but rather, upon the comparative effectiveness—as empirically demonstrated—of analyzing this type of data in terms of profiles of scores instead of as isolated parts, along with a comparative evaluation of the several methods of profile analysis applied.

To make the results obtained perhaps a little more interesting I will review briefly the method of Rorschach administration and scoring used, as well as its setting in the Marine Corps screening program. The test is administered through a group free association method. Cards are placed on a projector in only one position for three minutes. Between 15 to 18 individuals at a time write out free responses to the cards. Following the administration, individual inquiries are made, these averaging about 20 minutes for each individual. The psychologists making the inquiries total all responses and do the main portion of all scoring, although the computing of ratios, listing of scores, and similar tasks are done by clerks under supervision. The scoring system used is essentially that of Klopfer, with Beck's "populars." Some modifications, primarily in terms of the ratios used, were made by the psychologists involved. In all, a total of 40 separate scores are recorded for each individual.

The candidates included in the screening program are all selected from the enlisted ranks of the Marine Corps, some from combat, some from training programs, and others from recruit depots. Selection for the screening program is made on the basis of various combinations of GCT scores and college training, plus the recommendation of the Commanding Officer. The lowest allowable GCT is 110, with a college degree. This group makes up a very small portion of the total involved. The minimum GCT score required without any college training is 120, along with successful completion of a college equivalence test.

The screening program itself runs three weeks, with the major emphasis being placed upon field type problems, group discussions, and a wide variety of stress situations. Peer ratings, sociometric ratings, and many standardized tests are assigned to each section of about 15 candidates each week. At the end of the screening program, the three assessors who have observed each

group vote for acceptance or rejection of each candidate, as does the commanding officer. Psychological results are utilized primarily in situations in which the 4 voters disagree, and then only a composite final psychological ranking is used. In no instance does the assessor see Rorschach protocols.

Those passing the screening program are recommended for commissioning. Following commissioning, all officers attend approximately 20 weeks of Basic School, after which they are graded on both leadership and academic standing. All grading is done by observers.

Since estimates of leadership are weighed heavily in the screening program, selection of the criterion groups take this into account. Dichotomous criterion groups were established, with one group consisting of those individuals rejected after the screening program, and the other group consisting of individuals receiving a leadership grade of at least 90 following basic training. The manner in which rejections were made in the screening course and in which grading was done at the end of basic school, permitted limited contamination in-so-far as the Rorschach is concerned. So far only a small sample of cases is available upon which to check the suitability of the immediate criterion used, but the result is favorable. A correlation of .33 (which is significant) has been obtained between overall screening results and Korean combat ratings.

It should be pointed out that the Rorschach—as used in the situation—is being called upon to discriminate within a very select group. A vast majority of the individuals involved have GCT scores above 120, and all were originally recommended by their commanding officers as good officer potential. The criterion groups are, therefore, the extremes of a group already selected in terms of ability, leadership, and other characteristics ordinarily considered to be associated with a successful officer.

Before attempting to analyze the Rorschach data as profiles, tests of significance were applied to each of the 40 scores and ratios separately. Of the 40 measures only one differentiated between the criterion groups, and this barely at the 5% level of confidence. In view of the fact that one would expect more than twice this number of scores to discriminate at this level through the operation of chance, it appears safe to conclude that none of the 40 Rorschach scores independently differentiates between leaders and non-leaders as defined in this situation.

The first difficulty encountered in analyzing the data as profiles stemmed from the lack of equivalence of scales. In a strict sense, the Rorschach raw scores did not constitute a profile, since with the raw scores, it was impossible to say whether an individual scored "higher" on the D scale than on the M scale, etc. To overcome this difficulty all 40 scores and ratios had to be converted to some common scale. In this process Rorschach data for 2161 individuals were utilized. These individuals had all been in the Marine Corps screening program in the past three years. The distributions included those accepted and commissioned as well as those who were rejected for any reason. The raw scores were converted to normalized standards scores with a mean of 100 and a standard deviation of 20. Since these standard scores were based upon individuals going through the screening program over a period of years, and since the standards for admitting candidates to the program have not changed, these scores were considered representative of individuals who might later enter the screening program.

It should be pointed out that this procedure places no restrictions upon

the profiles of individuals—such as equalizing means or variances. It was still possible for any one individual to have all low scores, all high scores, or any scattering of scores.

While the primary purpose of this paper is to illustrate the application of profile-similarity techniques, it might be well to refer briefly to a number of other analyses that were made possible through the conversion to standard scores. First, raw scores for each individual were converted to standard scores, and the median score was computed for each individual in the criterion groups. Tests of significance, when applied to the two resulting distributions, showed no significant difference between the groups. This might be expected since "elevation" of profile has been associated with intelligence, and all individuals in the two groups have GCT scores near, or above, 120.

Another study involved the scattering of scores in individual profiles—or the flatness of profiles. For each individual, the range from the 5th through the 35th score in the profile was determined—i.e., the 5th highest, and the 5th lowest scores. Distributions of these ranges were formed separately for the two criterion groups. As in the case of the previous study, no significant difference was observed between the two groups in this respect.

A third study involved the individual Rorschach scores. The 40 scores for each individual were ranked from 1 to 40 on the basis of the standard scores. Distributions of ranks were then formed separately for each of the 40 scores, and for both criterion groups. Tests of significance were then applied to determine whether any score or scores differed significantly in position in a profile for the two groups. Significance was obtained for one-fourth of all scores. In view of the non-significance of individual scores when tested using raw scores, this evidences the fact that similar raw scores for members in the criterion groups might often differentiate if considered in relation to other scores in the profile.

The first attempt to analyze the data as a profile—intended as a screening device—utilized all 40 Rorschach measures. Essentially it consisted of first sorting the 40 scores for each individual into a limited number of categories on the basis of the standard scores; next attempting to locate a sort which would be typical of candidates in the high criterion group; then finding the Q-correlation between each individual sort and this typical sort; and finally, noting any difference which might occur between such Q-correlations for members in the two criterion groups. It was recognized that the use of the Q-correlation results in equalizing both means and variances in the patterns of all individuals, which—when profiles being compared are flat—would tend to emphasize differences as well as any errors of measurement. However, the previous analysis involving the range of profiles indicated that the scatter of individual patterns varied only slightly over all individuals. Based upon the 40 scores in a profile, the variability of the flattest profile did not differ significantly from the variability of the profile with the widest range of scores. The effect of equalizing means and variances consequently would have an only negligible effect upon the outcomes of the analysis.

The sort itself consisted of 9 categories, with the highest score in the first category, the three next highest scores in the scond category, and the next highest 5, 7, 8, 7, 5, 3, and 1 scores in the next 7 categories respectively. The use of a normal sort was used primarily to facilitate interpretation of the resulting correlations.

The typical sort of the high group was determined by computing the mean sort of each score for the top

criterion group, and then sorting the scores on the basis of these means. This method did not allow for the possible occurrence of several typical sorts; however, the fact that the distribution of each of the 40 scores for the top criterion group was unimodal gave some support to the assumption of a unique sort.

The distributions of the Q-correlations obtained from the two groups were irregular, however a cut-off point was readily observable at the point +.35. The 2 x 2 table in the upper left hand corner of the sheet previously passed out illustrates the extent of the discrimination obtained. If a tetrachoric coefficient of correlation were computed for this table, its value would be +.60.

It was noted, however, that most of the discrimination occurred in the middle range of scores. Below a Q-correlation of .19 exactly the same number of candidates in each group could be found, while for a Q-correlation above .50 the two groups differed by only 1 frequency.

When the same procedure was applied to independent criterion groups, using the same typical sort, the results were inconsistent with the original analysis. The Q-correlations classified in the same 2 x 2 table as previously used yielded the results given in the upper right hand corner of the sheet in your hands. Here, a tetrachoric r, if computed, would actually yield a slight minus value. It was felt that the disappointing results of this analysis might be due in major part to the fact that many of the 40 scores used were individually unreliable. For further analyses it was decided to limit the number of variables.

The exact basis for selection of scores among the 40 available, presented a real problem. It has already been demonstrated that scores which do not discriminate individually might be of value. Rather than use individual indices of discrimination, the basis for selection was made so as to best meet the assumptions involved in the analyses that were to follow. The 40 scores were first screened on the basis of reliability. While no conventional reliability coefficients were available for the Rorschach as used in this situation, a previous study had been completed wherein 100 records were independently re-inquired. Coefficients resulting from this analysis might be called "rater" or "scorer" reliabilities. After retaining only those scoring categories with coefficients of .80 or greater, 17 categories which were at least potentially reliable remained. I should emphasize the fact that scores of known unreliability were rejected, rather than the different procedure of accepting those of known high reliability.

Next, all intercorrelations among the 17 remaining scores were computed. Through a trial and error method, 11 scores were selected as being relatively independent of each other. The average correlations, based upon the ten correlations of each variable with each of the other variables, ranged from −.045 to +.140, with 9 of the eleven groups being below +.100. Only 3 individual intercorrelations making up the averages were above .40, and only 7 of the intercorrelations were above .30. Eleven Rorschach scoring categories were thus available which had been screened for reliability and which were relatively independent. Two separate analyses were then performed using these 11 scores.

In the first of these, the profile for each individual was represented as a point in 11 dimensional space—following Cronbach and Gleser's geometric model. Orthogonal axis were justified in view of the low intercorrelations of variables. The procedure planned involved locating a point in this space which might be typical of high leader-

ship, and then determining the distance of each point for the two criterion groups from this point. The distance measure used was Cronbach and Gleser's $\overline{D}$.

The centroid of the high leadership criterion group was selected as the typical leadership profile. As in the case of the preceding analysis, this did not allow for the possibility of several typical leadership profiles. Again, however, the unimodal distribution of each of the 11 scores gave some assurance to the assumption of a unique typical profile.

When $\overline{D}$ was obtained for all members in each group, the conventional test of significance was applied. The non-leaders tended to be farther from this typical profile than did the leaders, but the difference was not statistically significant. An examination of a number of individual profiles suggested that a good many profiles for which the $\overline{D}$'s were almost identical, differed considerably in the patterning of scores. Thus, one might have a standard score of 120 on the first variable and 80 on the second while a second profile might have the scores in reverse. Should the typical coordinates be 100 for each of these two variables in question, the distance measure would be the same. What would differ would be the direction of the profile from the typical profile. In the case of ordinary correlation between variables this direction would be an artifact of the order in which the individuals names were listed, but here a real difference might be meaningful. Consequently, it was decided to use some direction indicator along with the distance measure as a possible means of obtaining a more discriminative measure.

While directional cosines would have provided an exacting measure, they were ruled out as impractical for the number of variables involved. Instead, a series of signed values were utilized. The distribution of each score was divided at plus and minus 1 probable deviation. Scores above a plus 1 probable deviation were given a plus sign; those below a minus 1 probable deviation were given a minus sign; and those in between were scored zero. The directional measure was further limited by the decision to use only three dimensions rather than the 11. While these decisions resulted in a coarser measure of direction than directional cosines, they were dictated by practical necessity.

Since it was possible that a combination of variables found to discriminate in terms of direction might not discriminate as well when considered with a distance measure: or again, one which did not discriminate alone, might discriminate when considered with distance, the two measures were considered simultaneously. For this purpose, the distance measure was categorized into one of three intervals on the same basis as were the individual scores. In all, the use of three scores, each divided into 3 categories, and the distance measure divided into three groups, resulted in 81 breakdowns. The breakdowns were made independently for each possible combination of 3 variables among the 11 variables. For each of these breakdowns, frequencies were noted for the two criterion groups.

As a means of selecting the one combination of variables that best differentiated between the criterion groups, the statistic used was:

$$\frac{D}{\text{sum sq. of } D}$$

where D is the difference between frequencies in comparable cells in the tables for the two criterion groups. While not an exact test of significance —and not intended as one since its purpose was to select the best among

those available—it would result in the selection of the combination for which the ratio of difference to variance of difference is the greatest. The statistic is somewhat comparable to the test of significance between correlated variables, except that the absolute difference is used.

Because the data were put in 81 categories, the combination selected had many categories with small frequencies. The method used to combine these categories consisted in combining the directional breakdowns systematically until at least 5% of the sample fell in each remaining category. This process was applied only to the top criterion group, the grouping arrangement from the upper group being applied to data from the lower group. A total of 12 categories resulted, involving the three signed variables and the distance measure. When Chi-square was applied to the frequencies for the criterion groups, significance beyond the .001 level of confidence was obtained.

Probabilities were computed for each category, in each case in terms of the probability of a candidate being rejected. These probabilities ranged from .23 to .67 for the various categories.

Independent criterion groups were selected and frequencies for the same 12 categories tabulated separately for the two groups. Again probabilities were computed for each of the categories. The comparison of probabilities obtained are shown on the front side of the materials previously handed out. The distributions given are those of probabilities from original and cross validation groups for the 12 categories. The product moment correlation coefficient computed between those probabilities is +.442, pointing out considerable consistency of results.

The final analysis made involved the use of the linear discriminant function to the same 11 variables used in the preceding analysis. The raw scores on these variables were used rather than standard scores to facilitate later use. I might mention that the IBM equipment in New York was very useful in the analysis. The simultaneous solution of the eleven equations involved was completed in about 6 minutes at a total cost of $35.

The scale of measurement resulting when the final equation was applied to individual profiles ranged from approximately −800 to +1100. The difference between the means for the two criterion groups was significant beyond the .001 level of confidence.

Scores were grouped, this time in terms of regular intervals of 100, and probabilities of rejection were obtained for each interval. In view of the more regular trend noted in these probabilities, they were plotted, and a smooth curve fitted to the points. On the second sheet of the materials the heavy smooth curve is the resultant curve, yielding probabilities of rejection corresponding to scores obtained from the discriminant equation.

A new sample of upper and lower groups (N = 100 in each group) was again selected, and the discriminant equation applied to each member of the criterion groups. Scores were classified in the same categories as those used previously, and probabilities were determined for each category. The irregular line plotted on the same area as the smooth curve in the diagram on the second sheet presents the results of this cross validation.

The distributions below the graph are the probabilities from the original analysis and the cross validation. The first column of probabilities were read from the smoothed graph, and the second column of probabilities represent actual probabilities from the cross validation. The product moment coefficient of correlation between actual probabili-

ties (i.e. not utilizing the smoothed curve, but the irregular points from both analyses) was +.820.

It should be borne in mind, that while this correlation is the result of cross-validation, it primarily represents consistency of results—hence reliability rather than validity. Validity more properly should be in terms of prediction efficiency in this setting. As a means of estimating the validity for this purpose the smoothed curve obtained in the original analysis was used to predict success or failure. The probability of predicting success or failure without the use of the Rorschach was 50% since there were an equal number of candidates in each group. All individuals with a probability of failure above 50% were predicted failures, and all individuals with a probability of failure below 50% were predicted successes. With this procedure, 121, or 60.5% were correct predictions for the cross validation group. Hence the improvement in prediction was 10.5/50, or 21%. In a more conventional situation involving product-moment correlation, a coefficient of about .61 would provide the same amount of improvement in prediction. This figure is perhaps a better estimate of the validity of the results than the previous correlation between probabilities.

From the preceding results, we have seen that the use of a group of variables, none of which individually discriminates between the criterion groups, can differentiate when considered as a profile. Of the profile measures used the linear discriminant function appeared to be most satisfactory for the present situation. The use of Cronbach and Gleser's $\overline{D}$ when associated with some directional measure, provided significant discrimination; however, it is likely that the use of a more exacting directional measure would have improved the discrimination effected by this measure. It is likely that the more sensitive weighting system involved in the linear discriminant function—which reflects direction as well as distance—is a primary determinant of its superiority to the other methods used here.

Problems and Procedures in Profile Analysis

DAVID V. TIEDEMAN

A MODEL FOR THE PROFILE PROBLEM*

NOT LESS THAN five papers (2, 5, 8, 10, 12) published during the past five years are introduced by a statement that psychologists are increasingly becoming interested in the problem of studying the similarity of the psychological profile of an individual to some reference profile. This independent collaboration of interest in the problem is reassuring to a person like myself who has been working on this problem for the past year. This work is being conducted by the Educational Research Corporation under a contract between the Corporation and the United States Government, represented by Dr. Lloyd G. Humphreys, Director of Research, Personnel Research Laboratory, Human Resources Research Center, Lackland Air Force Base. Professor Phillip J. Rulon of Harvard University is the principal investigator for this project. I am indebted to Professor Rulon for many of the ideas discussed in this paper. However, responsibility for these remarks is completely mine.

Cronbach and Gleser (5), and Kogan (10) indicate that psychological profiles are studied for several reasons. However, the lists in both of these papers have the comparison of an individual with a group as a common purpose. In the short time available, I shall deal with just this problem. I do this not only because a model for a set of observations on individuals grouped together by some common *known* characteristic or characteristics is useful for vocational guidance, psychological diagnosis and prognosis, and anthropology, but also because any treatment of data designed to isolate types when the classification is *unknown* should be consistent with the model appropriate when the classification is *known*.

In order to reason from a concrete example, let us presume that the Air Force has administered experimentally an *Activity Preference Inventory* to all airmen inducted during a given month. The *Activity Preference Inventory* has a scale consisting of thirty pairs of activities, one member of the pair being an indoor activity and the other member of the pair being an outdoor activity. Each airman must indicate preference for one activity of each pair. The score is the number of preferences for outdoor activities. The *Inventory* also has a second set of 35 pairs of activities, one member of the pair being a solitary activity, and the other member of the pair being a convivial activity. Airmen must indicate preference for one or the other activity in each pair and the score is the number of

* This research was supported in whole or in part by the United States Air Force under Contract No. AF 18(600)-361 monitored by Director of Research, Personnel Research Laboratory, Lackland Air Force Base, San Antonio, Texas. Permission is granted for reproduction, translation, publication, use and disposal in whole or in part by or for the United States Government.

preferences for convivial activities. Thus, each airman has two scores, X_1 and X_2.

The scores on the *Activity Preference Inventory* are not made available to career counselors during the career counseling of the airmen. Each airman is counseled and assigned to an Air Force specialty in the ordinary manner. After a time sufficient for airmen to assume duties of their specialties and for the Air Force to judge its satisfaction with the performance of the airmen in their specialties, airmen who were satisfied with and were performing satisfactorily in a specialty were classified according to that specialty. G groups or specialties resulted from this classification. Provided the bivariate distributions for each specialty are not all coincidental, the Air Force now has data for inferring the regions of an outdoor and convivial activity preference reference plane from which later satisfactory and satisfied airmen in each of these G specialties arise. From information such as this, we wish to determine the similarity of the outdoor and convivial activity preferences of a new airman to the outdoor and convivial activity preferences of airmen later satisfied with, and satisfactorily performing in, *each* of the G specialties.

Now consider the case of Tom Basic, who when tested at induction indicated preference for 18 outdoor and 19 convivial activities. If we define a test plane by constructing Cartesian reference axes such as those in Figure 1, we may indicate all the information concerning the outdoor and convivial activity preferences of Tom Basic by placing a point in the test plane at the intersection of a line parallel to the convivial axis through the point representing 18 outdoor activity preferences and a line parallel to the outdoor axis through the point representing 19 convivial activity preferences. The information that Tom Basic prefers 18 outdoor and 19 convivial activities is then called the *coordinates* of the point for Tom Basic. Thus, the scores X_i where i takes the values 1 and 2 define a point in the test plane. Cronbach has referred to this model in at least three of his papers (3, 4, 5). Cattell (1) refers to the model in his book on personality, Osgood and Suci (12) found the model useful in resolving some of the problems concerned in the analysis of semantic data, and Gaier and Lee (8) mention the model in a recent review. This form of representation of the set of scores is familiar to psychologists, since they have been making scatter diagrams for some time. However, it seems to be only recently that consistent attention has been given to this conceptualization of the profile of an airman.

We may, of course, indicate the two coordinates for Tom Basic on parallel reference axes as in Figure 2. If we do this, we have a *profile*.

Let us suppose that among the airmen to whom the Air Force administered the *Activity Preference Inventory* experimentally, 85 of them later became satisfactory and satisfied Clerk-Typists. The records of outdoor and convivial activity preferences of these 85 airmen are sorted from the records for all the airmen tested in the experiment and recorded in a roster such as that of Table 1. If the outdoor and convivial activity preferences of these later satisfactory and satisfied Clerk-Typists are taken from Table 1 and enclosed in two sets of parallel lines as they are in Figure 3, and if the labels for rows and columns are deleted, the result is called a *matrix*. The score matrix which has been designated X_1 in Figure 3 conveys exactly the same information as the roster for Clerk-Typists reported in Table 1. The matrix, however, has an advantage over the roster in Table 1; it immediately implies that the pairs of numbers like (10, 22), (14, 17), and

so on, are coordinates of 85 points in the test plane. Cartesian representation of the matrix is given in Figure 4.

Figure 4 represents all information available to the Air Force concerning the meaning of expression of preference for outdoor and convivial activity for later presence in, satisfaction with, and satisfactory performance in, the Clerk-Typist specialty. The figure suggests: (1) that later satisfactory and satisfied Clerk-Typists tend to be found in the upper left portion of the test plane; (2) that the joint preferences for outdoor and convivial activities of later satisfactory and satisfied Clerk-Typists are more dense in the region of the test plane near the *centroid* of the Clerk-Typist group; (3) that the density of outdoor and convivial activity preferences of Clerk-Typists becomes less as one moves in all directions from the centroid; and (4) that dispersion along the new reference axis labelled x_{112} is more widespread than dispersion along the new reference axis x_{122}.

We noted earlier that a profile resulted from representation of the coordinates of the outdoor and convivial activity preferences of Tom Basic on parallel reference axes. In a similar way we may form the profiles of outdoor and convivial activity preferences for each of the 85 Clerk-Typists as in Figure 5. However, in Figure 5 we have not connected the outdoor activity coordinate of each airman with his convivial activity coordinate. Therefore, it is impossible to tell in Figure 5 which outdoor activity preferences are associated with which convivial activity preferences. Hence, we have *Profile Problem 1:* How may the joint preferences of each airman be represented? To represent the joint preference of each airman on a profile results in such a mess between the two profile stalks that the profile for Tom Basic when plotted with respect to the profiles for Clerk-Typists becomes obscure.

Because of this fact, we frequently indicate only two profiles when we wish to compare the profile for Tom Basic with those for Clerk-Typists. The two profiles are the profile for Tom Basic and the profile for the centroid of the Clerk-Typist group as indicated in Figure 6. Figure 6 suggests immediately *Profile Problem 2:* How may the similarity of an airman to airmen in a particular specialty be expressed? In order to answer this question let us note first that all of the 85 airmen whose outdoor and convivial activity preferences are represented in the test plane in Figure 4 and on profile axes in Figure 5 have the *same* label. Each one of these 85 airmen is a later satisfactory and satisfied Clerk-Typist. And yet, Figures 4 and 5 indicate that these 85 airmen expressed outdoor and convivial activity preferences at the time of induction that were neither coincidental nor collinear. This condition has occurred so frequently in my experience and I am sure in yours as well that I suggest we accept it, summarize it, and even give it the formal name, "axiom." Let's have the axiom read as follows:

> *Axiom 1.*—Points representing pairs of psychological observations on airmen grouped according to a common designation will ordinarily be dispersed about the centroid in a space of dimensionality two.

This axiom does no more than formalize for bivariate data the principles of individual differences and reliability that most of us accept.

Application of this axiom to the example under discussion results in a feeling of relaxation concerning the bivariate dispersion of outdoor and convivial activity preferences of the 85 airmen later classified as satisfactory and satisfied Clerk-Typists as represented in Figure 4. All of these airmen *are* of the same kind. However, the outdoor and convivial activity preferences of all of these airmen are *not* of the same kind.

Our task may then be stated this way: How likely is it that a random point (X_1, X_2) whose coordinates are the two test scores will have a Clerk-Typist label associated with it?

An answer to this question necessitates a definition of similarity. Obviously, our definition of similarity cannot be so rigorous that points must be coincidental. If such a definition of similarity were employed it would result in missing almost all of the later satisfactory and satisfied Clerk-Typists. Cronbach and Gleser's index of eccentricity (5, p. 5) and Cattell's coefficient of pattern similarity (2) treat as similar all points lying on a circle centered at the centroid of the group. I have already indicated that wider dispersion along the x_{112} axis in Figure 4 than along the x_{122} axis is to be expected in the outdoor and convivial activity preferences of later satisfactory and satisfied Clerk-Typists. Consequently, if I were to define as similar points lying on a circle centered at the centroid of the Clerk-Typist specialty, I would be treating as similar points that occur with different relative frequencies in the profiles for later satisfactory and satisfied Clerk-Typists. In order to overcome this difficulty, I define similarity in terms of equal probability of occurrence of a point within a specialty. This definition is a second fundamental tenet of my remarks and consequently I also state it as an axiom.

Axiom II.—Similar profiles within a specialty are those that occur with the same probability.

In order to apply this axiom to a comparison of the outdoor and convivial activity preferences of Tom Basic with those of later satisfactory and satisfied Clerk-Typists, it is necessary to specify the regions of equal probability in the test plane of Figure 4. Obviously, it is impossible to do this from the empirical data of Figure 4 itself, since there is no consistent location of points that occur with a frequency of 3, a frequency of 2, or a frequency of 1. Therefore, we turn to a theoretical distribution.

Figure 5 indicates that it is reasonable to assume that both the outdoor and convivial activity preferences of a large sample of later satisfactory and satisfied Clerk-Typists are distributed normally. Figure 4 indicates that the regression is approximately linear. When both conditions pertain, the bivariate distribution is distributed in a bivariate normal manner. Consequently, *we will assume that outdoor and convivial activity preferences of a large number of later satisfactory and satisfied Clerk-Typists will be distributed normally with means of that of this sample and with dispersion matrix of that of this sample.*

The dispersion matrix may be computed directly from the raw score matrix X_1 given as Figure 3 by defining both a matrix of means as illustrated in Equation 1 and a matrix of number of cases as illustrated in Equation 4, and performing the matrix operations indicated by Equations 2, 3, and 5. For the Clerk-Typist data, the operations result in the dispersion matrix given as Equation 6. The matrix indicates that dispersion along the outdoor activity preference reference axis is somewhat greater than dispersion along the convivial activity preference reference axis, and that outdoor and convivial activity preferences are related positively to a small degree. Consequently, variation along the reference axis x_{112} in Figure 4 is greater than variation along the reference axis x_{122} in that figure. The dispersion matrix is, of course, a direct function of the raw score matrix since it was computed from Equation 7, the expanded way of writing Equation 5. Thus, when treated in the right way, the score matrix X_1 reports all the information concerning outdoor and convivial preferences that is of psychological significance for inferring later duty

as a Clerk-Typist. I say this with complete confidence because the indoor-outdoor and solitary-convivial scales of the *Activity Preference Inventory* are a figment of Professor Rulon's imagination as influenced by Professor Kelley's work on activity preferences and data for Clerk-Typists were obtained by throwing dice to approximate specification of the matrix of means and the dispersion matrix.

In a bivariate normal distribution the probability, $P(X_1, X_2)$, that a point, (X_1, X_2), drawn at random from the bivariate distribution will be from a small area surrounding the point is given by Equation 8, where the symbols μ, σ, and ρ have their usual meaning of population mean, standard deviation, and correlation respectively. For fixed values of these parameters the probability depends only upon the quantity in the wiggly brackets of Equation 8. This quantity is copied in Equation 9 and is called χ^2, following Pearson (13).

Since we do not know the values of the parameters in Equation 9, we choose their maximum likelihood estimates, sample means, standard deviations, and correlation. Computation of the correlation coefficient itself is not necessary, because if the matrix variable x_1 is defined as it is in Equation 10, χ^2 may be computed directly from Equation 11. I have written the subscript 1 after χ^2 in Equation 11 because this is the χ^2 for the deviation scores with respect to the means of the Clerk-Typist, or first, specialty. Equation 11 indicates that there are a number of values of the matrix variable x_1 that yield the same value of χ^2. The locus of points with equal values of χ^2 is that of an ellipse with center at the centroid of the Clerk-Typist group. The ellipse is symmetric about the Kelley (9) principle components of the Clerk-Typist data. Two of the set of iso-frequency ellipses are shown in Figure 7.

I have used the symbol χ^2 to denote the value obtained from the triple matrix product in Equation 11 because the distribution of this χ^2 is the same as the distribution of χ^2 used ordinarily in tests of independence or in tests of goodness of fit. The number of degrees of freedom associated with this χ^2 is the number of variates, in our illustration, two. Consequently, a value of χ^2 computed from Equation 11 may be evaluated in terms of probability from any of the tables of χ^2 commonly available. If greater accuracy for the relative frequency of the χ^2's which exceed a given χ^2 is desired, use Table IX of Pearson's *Tables for Statisticians and Biometricians, Part I* (13, pp. 22-23).

Thus, $\sqrt{\chi^2}$ is the distance in the Mahalanobis (11) sense of a point from the centroid for a specialty which is consistent with Axioms I and II, and contains all of the psychological meaning for later presence in, satisfaction with, and satisfactory performance in, the Clerk-Typist specialty inherent in the matrix X_1. If χ^2 is computed for every point in the bivariate Gaussian distribution, and if a frequency distribution of these χ^2's is determined, it is possible to compute the percentile rank of each χ^2 value. If computation is done such that a χ^2 of zero corresponds to the percentile rank 100, the percentile rank will indicate the per *cent* of points in the bivariate distribution beyond the cont*our* ellipse upon which a point with a given χ^2 value falls. For this reason, Dr. Rulon, Dr. Bryan and I (15) have called this percentile rank a centour score. The centour score is not only a percentile index of distance but also an estimate of the relative frequency with which the outdoor and convivial activity preferences of Clerk-Typists will exceed the χ^2 of the point representing the outdoor and convivial activity preferences of a given airman, provided outdoor and convivial activity prefer-

ences of Clerk-Typists reasonably approximate a bivariate Gaussian distribution. Thus, the centour score indicates the per cent of actual Clerk-Typists that will be assigned to some specialty other than the Clerk-Typist specialty if assignment to other specialties is made for values of the Clerk-Typist χ^2 equal to or greater than the value associated with a given point.

If we now evaluate Equation 11 for the 18 outdoor and 19 convivial activity preferences of Tom Basic, we get 3.8423 as the value of χ^2 for the comparison of the profile for Tom Basic with those for later satisfactory and satisfied Clerk-Typists. The χ^2's for approximately 15 per cent of the later satisfactory and satisfied Clerk-Typists exceed this value of χ^2. Thus, Tom Basic expressed preference for outdoor and convivial activities of a type whose discrepancy from the centroid of the Clerk-Typist specialty was exceeded by the discrepancies of 15 per cent of the later satisfactory and satisfied Clerk-Typists. If all airmen with a discrepancy this large or larger were excluded from the Clerk-Typist specialty, approximately 15 per cent of those who would normally become satisfactory and satisfied Clerk-Typists later would be excluded from that specialty.

The centour score permits interpretations of this nature for all points in the test plane. Centour scores may be summarized in a table such as Table 2. Table 2 contains the percentile equivalents of the distances of a sample of points in the test plane from the centroid for the Clerk-Typist specialty. Such information is quite useful in the career counseling of airmen.

This model is developed in terms of a single specialty. When the outdoor and convivial activity preferences of airmen who later became satisfactory and satisfied Aircraft & Engine Mechanics are assembled, they may be written as a new matrix of scores as in Figure 8. The matrix X_2 is written without influence on the matrix X_1. The scores indicated in the rows of the matrix in Figure 8 represent the coordinates of a point in a test plane such as Figure 9 for each of the later satisfactory and satisfied A & E Mechanics. Figure 9 is constructed by adding red points to the previous Figure 4. Figure 9 indicates that the location of the point representing the outdoor and convivial activity preferences of an airman in the test plane is related to later identification as a Clerk-Typist or an A & E Mechanic. Figure 9 indicates also that the meaning of outdoor and convivial activity preferences is different psychologically among later satisfactory and satisfied A & E Mechanics than it was among later satisfactory and satisfied Clerk-Typists. Among later satisfied and satisfactory A & E Mechanics dispersion of outdoor and convivial activity preferences is greater along the axis x_{212} than it is along the axis x_{222}. This information is implied in the matrix X_2, but is not explicit.

In order to make the information of X_2 explicit, it is necessary to form both a matrix of means and a matrix of number of A & E Mechanics as has been done in Equations 12 and 14 respectively. The dispersion matrix for the A & E Mechanic group may then be computed from Equations 13 and 15. If the matrix variable x_2 is formed for every pair of scores X_1 and X_2 according to Equation 16, equal frequency points in the test space for the A & E Mechanic specialty lie on one of the set of homothetic ellipses given by Equation 17, provided the bivariate distribution of outdoor and convivial activity preferences of A & E Mechanics is reasonably normal. Addition of two of this new set of homothetic ellipses to Figure 7 results in Figure 10.

For Tom Basic, evaluation of Equation 17 gives 0.2629 as the value of χ^2 for the comparison of his point with

those for A & E Mechanics. This χ^2 is exceeded by the χ^2's of approximately 88 per cent of the A & E Mechanics. 88 is the centour score for Tom Basic's point compared with the A & E Mechanic distribution. It indicates that the χ^2 of Tom Basic's point is exceeded quite frequently by the χ^2's of the points of later satisfactory and satisfied A & E Mechanics.

We now know that the centour score of the point for Tom Basic is 15, when compared to the later satisfactory and satisfied Clerk-Typists, and 88, when compared with the later satisfactory and satisfied A & E Mechanics. Tom's preferences for outdoor and convivial activities are more typical of those of later satisfactory and satisfied A & E Mechanics than they are of later satisfactory and satisfied Clerk-Typists.

Centour scores for comparison with A & E Mechanics may be added to Table 2 as they have been in Table 3. Table 3 contains all information from the outdoor and convivial activity preferences expressed by an airman at induction for inferring later presence in *either* the Clerk-Typist or the A & E Mechanic specialty.

In this brief consideration of a second specialty I have indicated sufficient steps for extension of the logic to 3, 4, and so on to the last or G group by means of induction. Computations specified by Equations 12-17 are simply repeated for each new group added. Table 3 may be augmented by a new row in each of the row blocks associated with a particular convivial activity preference score for each group.

The centour score model for inferring group membership provides implicit answers for Profile Problems 3 and 4. *Profile Problem 3* is: On what type of scales should profiles be represented? Centour scores are based upon the assumption that the bivariate distribution of each group is normal. Thus, the scale on which profiles are represented and the scale on which Cartesian representation of profile information is plotted should be such that an approximate bivariate normal distribution results in *each* group. When the distribution of each variate is normal in the group, and when regression is linear in the group, a bivariate normal distribution results for the group. Therefore, the basic requirements for the scale on which each variate is represented are that it produce an approximately normal distribution in *each* of the groups with which the profile for an airman is to be compared, and that it be related linearly to other variates in *each* group. Raw scores themselves may fulfill these properties. However, if the raw scores do not fufill these properties, some transformation similar to the transformation used by Flanagan (7) in the construction of his Scaled Scores for the Cooperative Tests may provide the desired conditions. The transformation is necessary for only those variates that do not fulfill both conditions in *each* group.

Profile Problem 4 is: On which group should profiles be standardized? In the comparison of a point with specialty 1 the comparison should be in terms of scores standardized on specialty 1. In comparison of a point with scores for specialty 2, the standardization should be in terms of scores for specialty 2, and so on, to the last, or G^{th}, specialty.

The centour score model has a further advantage of considerable importance. Equation 17 does not depend on the number of variates on which each individual is observed. Equations 12-16 need only be augmented appropriately for every variate added. Pearson (13, p. xxiv) has indicated that all points on the ellipsoid defined by the generalized form of Equation 17 are equally probable in a multivariate normal distribution. The value of the centour score can then be determined from a table of χ^2

with degrees of freedom equal to the number of variates, or from Pearson's (13) tables.

The volumes necessary to bind the tables of centour scores that would result from even as few as four or five variates and three or four groups will undoubtedly lead to immediate efforts to reduce the number of variates. In my opinion, these efforts will be most fruitful if they start from this model and evaluate the efficiency of the reduced number of variates with this model as the standard of efficiency. The Rao-Tukey-Bryan multiple discriminant function should not be overlooked in efforts at reduction.

The centour score conversion of χ^2 is, I believe, the index most useful in guidance work, especially of a vocational guidance nature. In work of this kind, I believe that the counselor should be aware of the realities of the ratios of individuals in various jobs and that these realities should influence the interpretation of centour scores. However, I do not believe that the ratios given by numbers currently in jobs should have the influence on the interpretation of centour scores that they would have if the ratios were to be incorporated in this model and if they were to affect classification regions in a test space such as that depicted in Figure 10. The introduction of these ratios into the interpretation of centour scores is, of course, necessary for the classification or assignment of men instead of their guidance. The introduction of these ratios will increase the efficiency of classification of men. However, boundaries of these classification regions are still functions of Equation 17. Therefore, this equation is one that must be determined whether the data are to be used for guidance purposes or for classification purposes. If the data are to be used for classification purposes, the centour scores need to be modified by the proportion of men in the specialty before classification is done. If the centour scores are to be used for guidance purposes, it is my feeling that each counselor should exercise his own judgment in dealing with the question of the ratio of men in each specialty.

In the development of the model I have stated two axioms. One axiom expressed the faith that the observations of individuals sharing a common classification will contain multivariate dispersion of the observations about the centroid. The second axiom defined as similar those points within the common classification that occur with the same probability. With these two tenets in mind, I suggest that you trace the locus of points in a test plane resulting from efforts to describe psychological types by person to person comparison, mean scatter, vocabulary scatter, the selection of similar profiles, Cronbach and Gleser's (5) index of distance, Cattell's (2) coefficient of pattern similarity, and DuMas' (6) coefficient of profile similarity. If you do this, you will find the loci of points inconsistent with either or both of the axioms I have stated. In this situation, you are forced to abandon either the techniques as a means of isolating types or the axioms as a description of what you would expect the bivariate distributions for a type to be. Personally, I have questioned the techniques.

I indicated earlier that I considered it desirable to have an adequate model for inferring a known classification from a knowledge of psychological characteristics because I felt that efforts to develop an unknown classification system on the basis of multivariate observations should be consistent with the other model.

I wish that I could offer a solution to the problem of resolving a multivariate set of N points into the multivariate distributions of G types which is consistent with my two axioms. Someday

I hope that someone will. If anyone is interested in thinking about this problem, I can at least share a lead with him. On pages 300-306 of Rao's recent book (14) you will find that Rao treats the probem of resolving a mixed series into two Gaussian components. Rao indicates that a solution of this problem for the case of a single variate and two groups in terms of the method of moments was discussed by Karl Pearson as early as 1894. The solution of this problem requires estimation of the two means, the two standard deviations, and the proportion of mixture from the data on the single variate of the mixed series. Rao gives an adaptation of Pearson's method. Rao further discusses the problem of sexing osteometric material on the basis of multiple measurements. These solutions should, I think, receive the attention of more psychologists.

Before closing, I acknowledge Dr. Lord's kindness in directing my attention to an error in my original statement of the sufficient test for bivariate normality and in permitting me to correct the error before reading my paper.

Problems and Procedures in Profile Analysis

DAVID V. TIEDEMAN

FIGURES, TABLES, EQUATIONS, AND REFERENCES FOR A MODEL FOR THE PROFILE PROBLEM

Figure 1

Cartesian Representation:
Outdoor and Convivial Activity Preferences—
Tom Basic

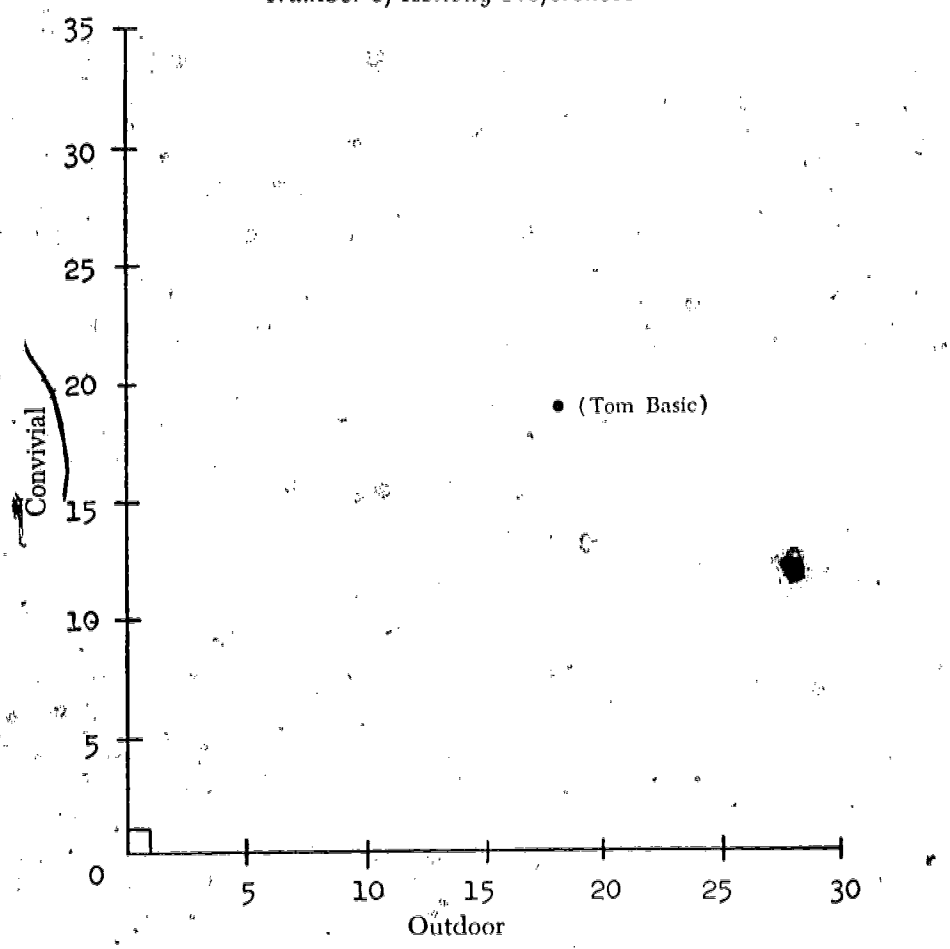


Figure 2

Profile Representation:
Outdoor and Convivial Activity Preferences—
Tom Basic

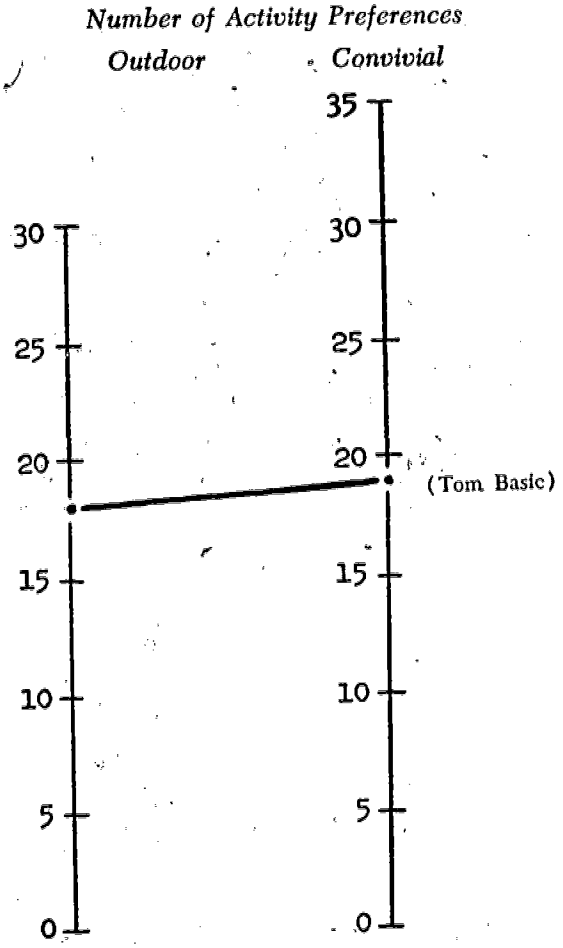


Table 1

Outdoor and Convivial Activity Preferences of Later Satisfactory and Satisfied Clerk-Typists

Airman No.	*No. of Activity Preferences*	
	Outdoor	*Convivial*
1	10	22
2	14	17
3	19	33
⋮	⋮	⋮
85	16	24

Figure 3

Matrix Representation:
Outdoor and Convivial Activity Preferences—
Clerk-Typists

$$x_1 = \left\| \begin{matrix} 10 & 22 \\ 14 & 17 \\ 19 & 33 \\ \vdots & \vdots \\ 16 & 24 \end{matrix} \right\| \quad (85 \text{ rows})$$

Figure 4

Cartesian Representation:
Outdoor and Convivial Activity Preferences—
Clerk-Typists

Number of Activity Preferences

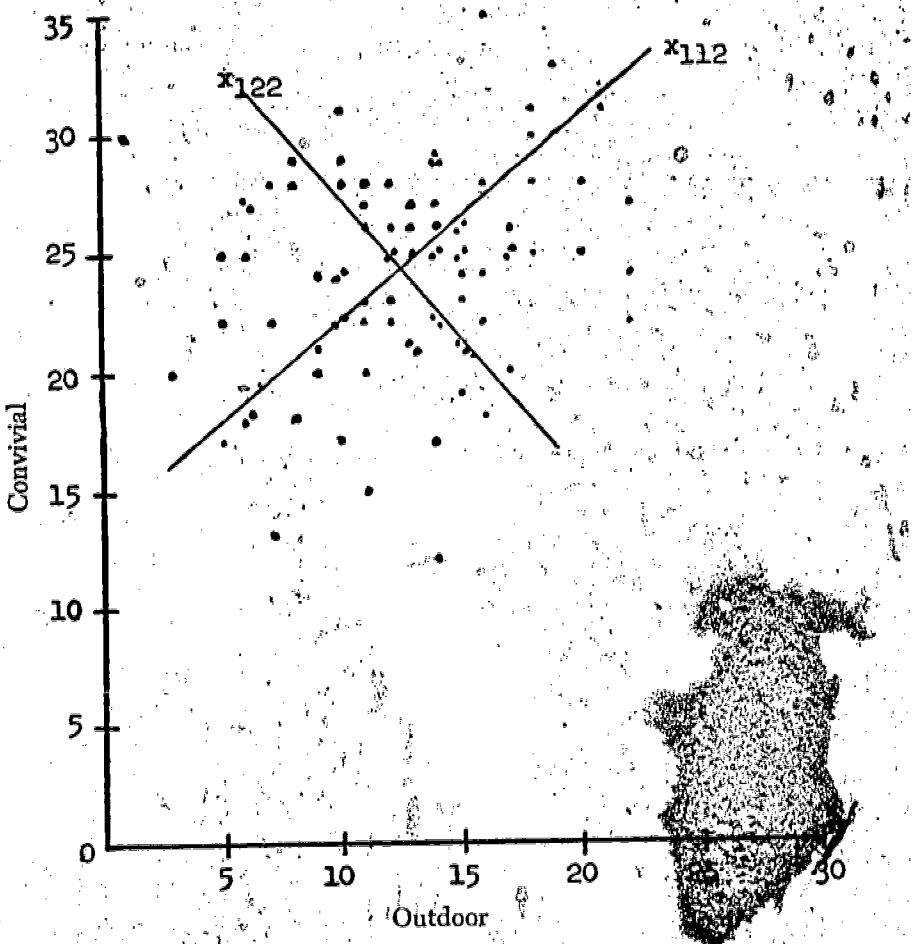


Figure 5

Profile Representation:
Outdoor and Convivial Activity Preferences—
Clerk-Typists

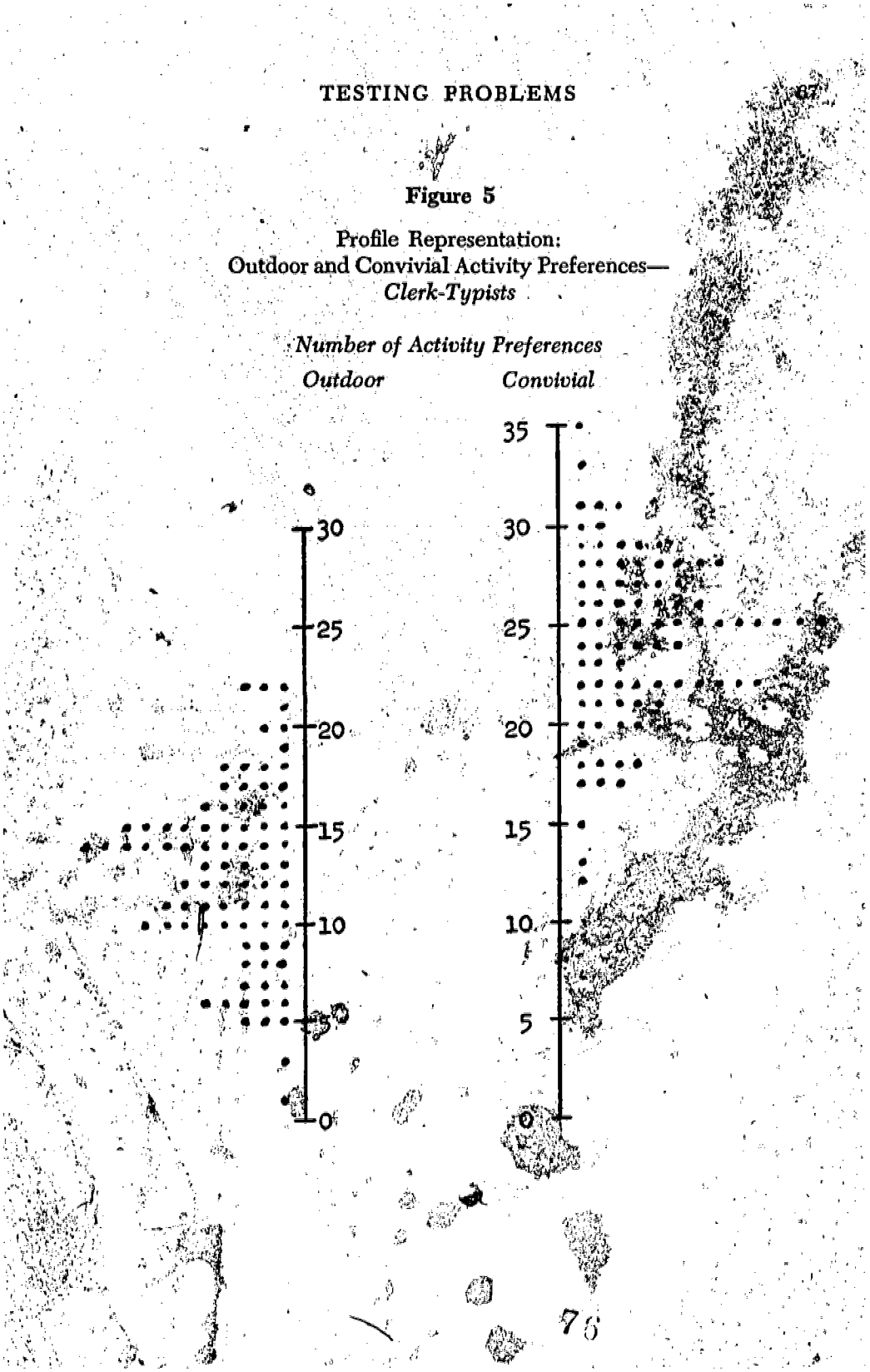
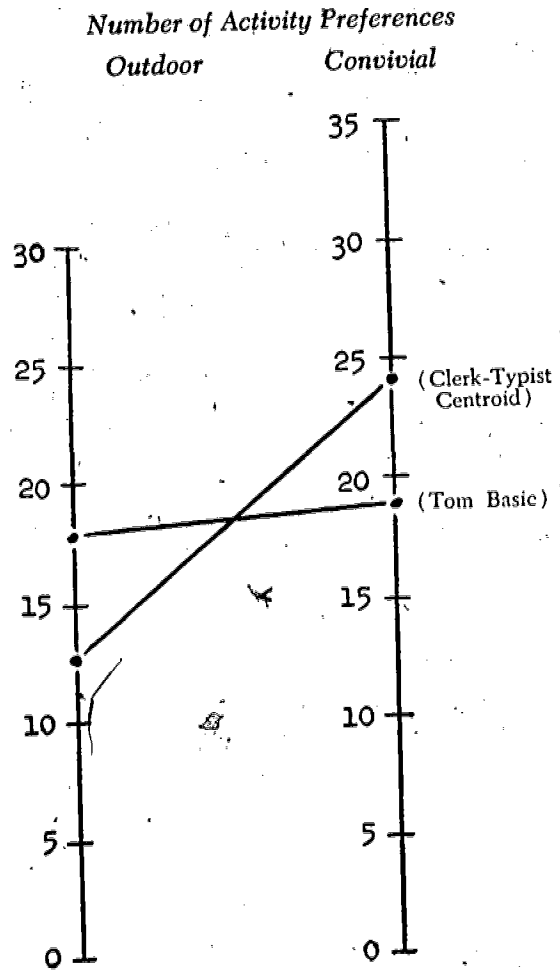


Figure 6

Profile Representation:
Outdoor and Convivial Activity Preferences—
Tom Basic and Clerk-Typist Centroid



$$\overline{X}_1 = \left\| \begin{matrix} 12.5882 & 24.2235 \\ 12.5882 & 24.2235 \\ 12.5882 & 24.2235 \\ \vdots & \vdots \\ 12.5882 & 24.2235 \end{matrix} \right\} \text{(85 rows)} \quad \text{[Eq. 1]}$$

$$x'_1x_1 = X'_1X_1 - \overline{X}'_1\overline{X}_1 = \Sigma_1 \quad \text{[Eq. 2]}$$

$$\Sigma_1 = \left\| \begin{matrix} \sum_{p=1}^{n_1} x^2_{p1} & \sum_{p=1}^{n_1} x_{p1}x_{p2} \\ \sum_{p=1}^{n_1} x_{p2}x_{p1} & \sum_{p=1}^{n_1} x^2_{p2} \end{matrix} \right\| \quad \text{[Eq. 3]}$$

$$N_1 = \left\| \begin{matrix} \sqrt{85} & 0 \\ 0 & \sqrt{85} \end{matrix} \right\| \quad \text{[Eq. 4]}$$

$$D_1 = N_1^{-1}\Sigma_1N_1^{-1} \quad \text{[Eq. 5]}$$

$$D_1 = \left\| \begin{matrix} 20.006920 & 4.562629 \\ 4.562629 & 18.573564 \end{matrix} \right\| \quad \text{[Eq. 6]}$$

$$D_1 = N_1^{-1}(X'_1X_1 - \overline{X}'_1\overline{X}_1)N_1^{-1} \quad \text{[Eq. 7]}$$

$$P(X_1,X_2) = \frac{1}{2\pi\sigma_1\sigma_2(1-\rho^2)^{\frac{1}{2}}} e^{-\frac{1}{2}\left\{\frac{1}{1-\rho^2}\left[\left(\frac{X_1-\mu_1}{\sigma_1}\right)^2 - 2\rho\left(\frac{X_1-\mu_1}{\sigma_1}\right)\left(\frac{X_2-\mu_2}{\sigma_2}\right) + \left(\frac{X_2-\mu_2}{\sigma_1}\right)^2\right]\right\}} dX_1dX_2 \quad \text{[Eq. 8]}$$

$$\frac{1}{1-\rho^2}\left[\left(\frac{X_1-\mu_1}{\sigma_1}\right)^2 - 2\rho\left(\frac{X_1-\mu_1}{\sigma_1}\right)\left(\frac{X_2-\mu_2}{\sigma_2}\right) + \left(\frac{X_2-\mu_2}{\sigma_2}\right)^2\right] = \chi^2 \quad \text{[Eq. 9]}$$

$$x_1 = \| \ (X_1 - \overline{X}_1) \qquad (X_2 - \overline{X}_2) \ \| \quad \text{[Eq. 10]}$$

$$x_1D_1^{-1}x'_1 = \chi^2 \quad \text{[Eq. 11]}$$

Figure 7

Cartesian Representation:
Outdoor and Convivial Activity Preferences—
Clerk-Typist Centour Specialty Psychographs

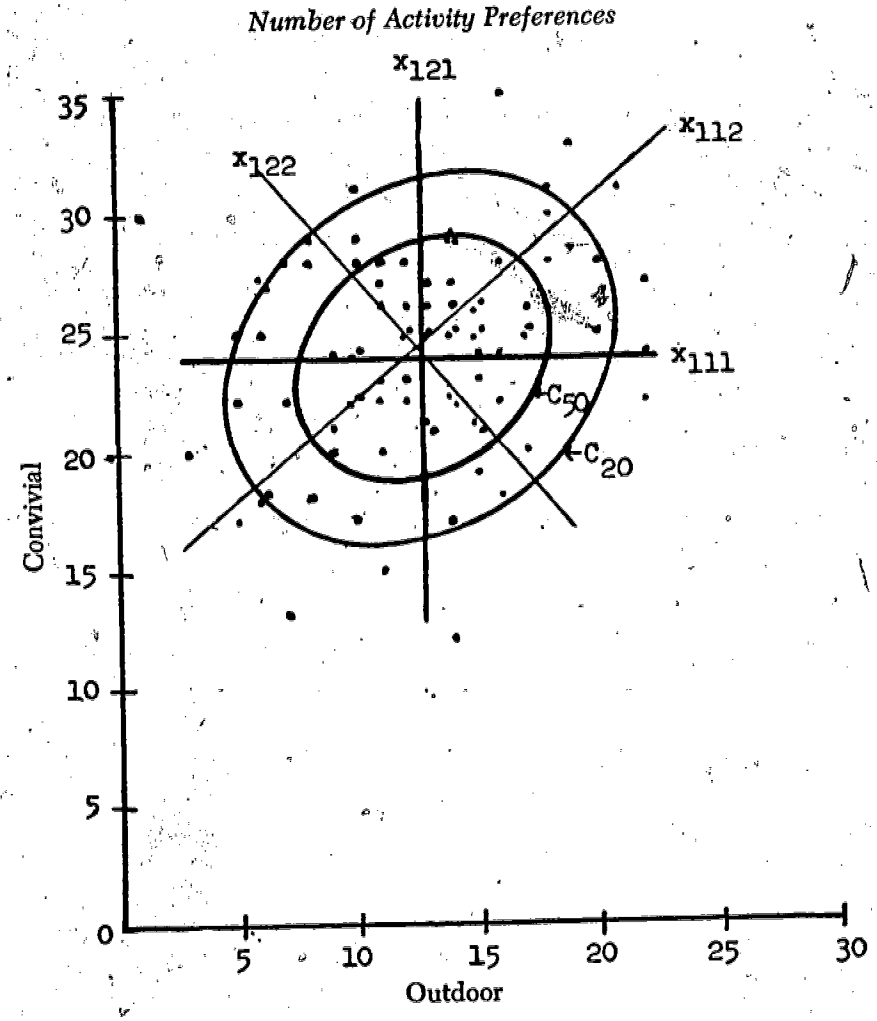


Table 2

Centour Scores:
Outdoor and Convivial Activity Preferences—
Clerk-Typists

Convivial	*Number of Activity Preferences Outdoor*						
	0	5	10	15	20	25	30
35	*	*	02	04	02	*	*
30	*	05	27	40	16	02	*
25	01	20	80	86	25	02	*
20	02	20	58	45	09	01	*
15	01	05	10	06	01	*	*
10	*	*	*	*	*	*	*
5	*	*	*	*	*	*	*
0	*	*	*	*	*	*	*

* Centour ≤ 0.5

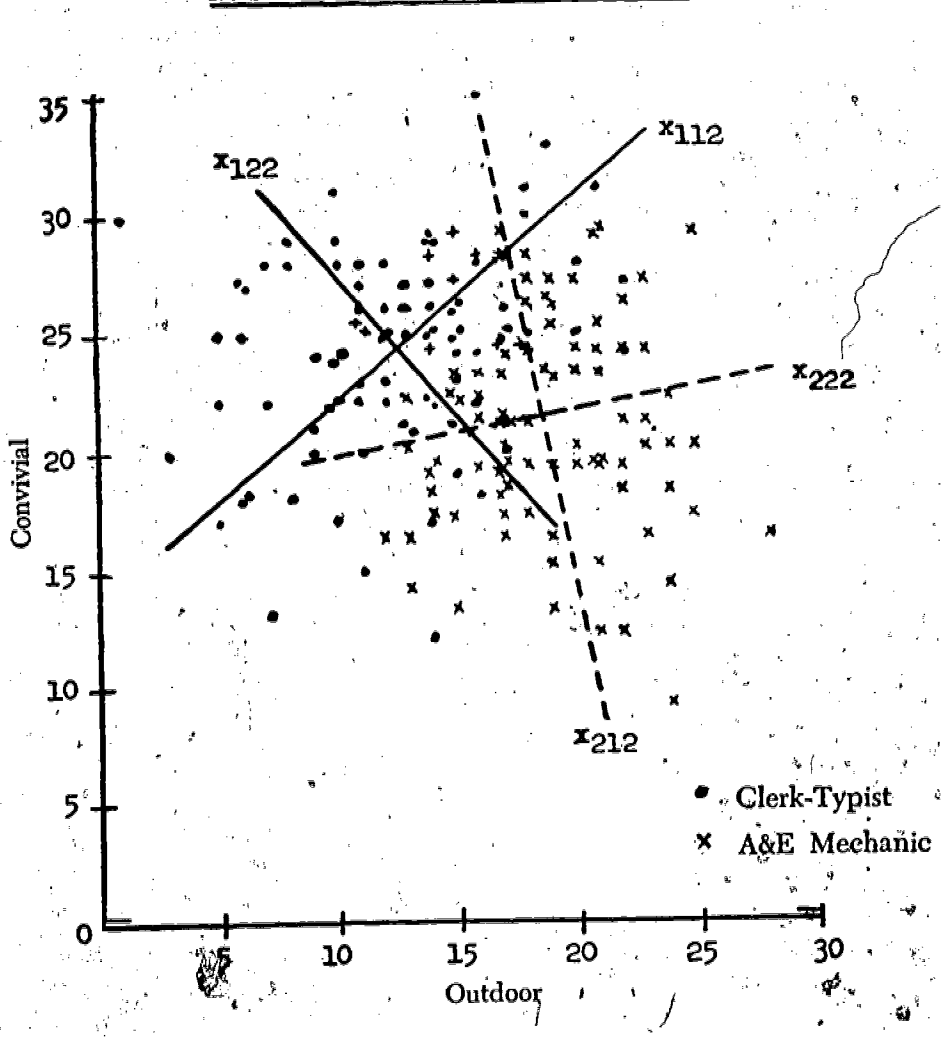
Figure 8

Matrix Representation:
Outdoor and Convivial Activity Preferences—
A&E Mechanics

$$X_2 = \begin{Vmatrix} 20 & 27 \\ 21 & 15 \\ 15 & 27 \\ \vdots & \vdots \\ 19 & 16 \end{Vmatrix} \text{ (93 rows)}$$

Figure 9

Cartesian Representation:
Outdoor and Convivial Activity Preferences—
Clerk-Typists and A&E Mechanics



$$\bar{X}_2 = \begin{Vmatrix} 18.5376 & 21.1398 \\ 18.5376 & 21.1398 \\ 18.5376 & 21.1398 \\ \vdots & \vdots \\ 18.5376 & 21.1398 \end{Vmatrix} \quad (93 \text{ rows}) \quad [\text{Eq. 12}]$$

$$X'_2X_2 - \bar{X}'_2\bar{X}_2 = \Sigma_2 \quad [\text{Eq. 13}]$$

$$N_2 = \begin{Vmatrix} \sqrt{93} & 0 \\ 0 & \sqrt{93} \end{Vmatrix} \quad [\text{Eq. 14}]$$

$$D_2 = N_2^{-1}\Sigma_2 N_2^{-1} \quad [\text{Eq. 15}]$$

$$x_2 = \| \ (X_1 - \bar{X}_2) \qquad (X_2 - \bar{X}_2) \ \| \quad [\text{Eq. 16}]$$

$$x_2 D_2^{-1} x'_2 = \chi_2^2 \quad [\text{Eq. 17}]$$

Table 3

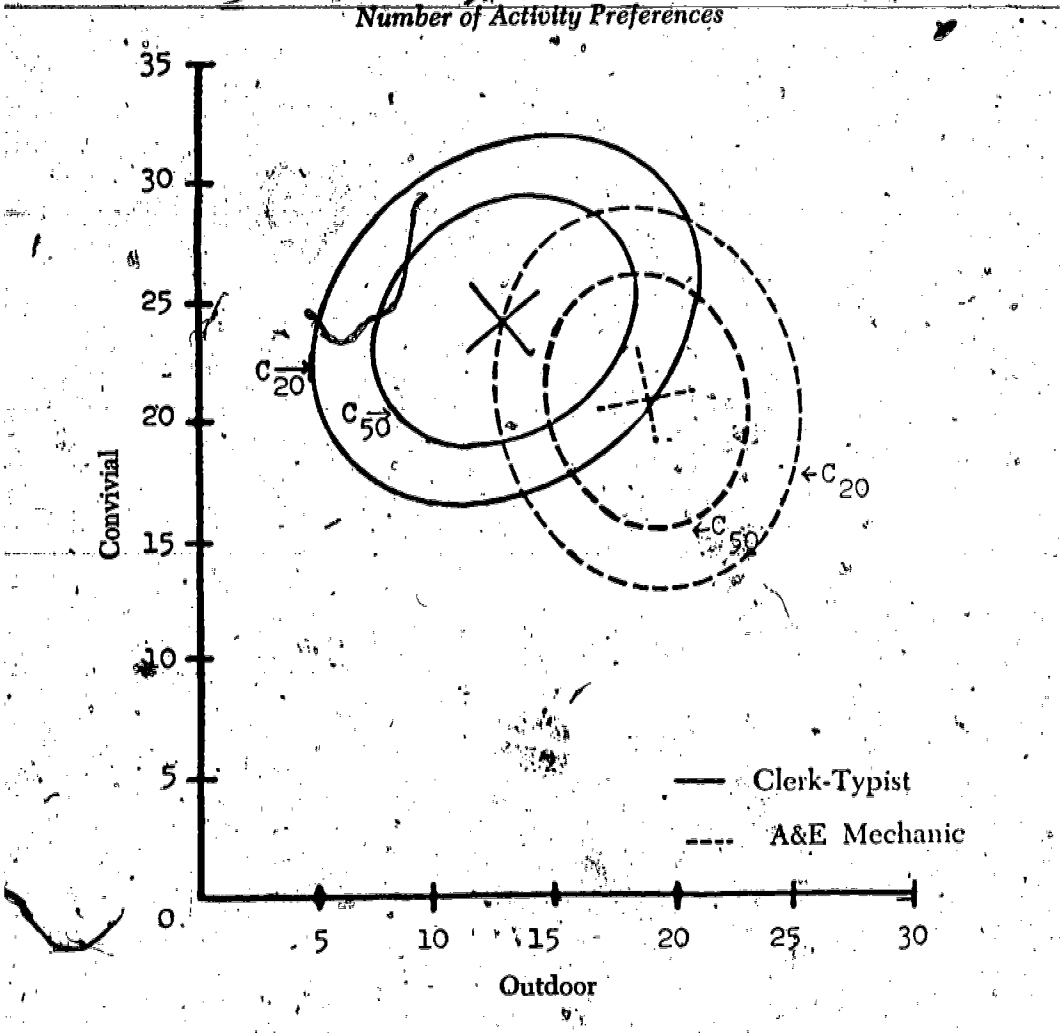
Centour Scores:
Outdoor and Convivial Activity Preferences—
Clerk-Typists and A&E Mechanics

Convivial	*Speciality*	*Number of Activity Preferences Outdoor*						
		0	5	10	15	20	25	30
35	C-T	*	*	02	04	02	*	*
	A & E	*	*	*	01	01	*	*
30	C-T	*	05	27	40	16	02	*
	A & E	*	*	01	11	12	02	*
25	C-T	01	20	80	86	25	02	*
	A & E	*	*	05	46	61	11	*
20	C-T	02	20	58	45	09	01	*
	A & E	*	*	05	57	90	19	01
15	C-T	01	05	10	06	01	*	*
	A & E	*	*	02	21	38	09	*
10	C-T	*	*	*	*	*	*	*
	A & E	*	*	*	02	05	01	*
5	C-T	*	*	*	*	*	*	*
	A & E	*	*	*	*	*	*	*
0	C-T	*	*	*	*	*	*	*
	A & E	*	*	*	*	*	*	*

* Centour ≤ 0.5

Figure 10

Cartesian Representation:
Centour Specialty Psychographs—
Clerk-Typist and A&E Mechanic



REFERENCES

1. CATTELL, RAYMOND B. *Personality, A Systematic Theoretical and Factual Study.* New York: McGraw-Hill Book Co., Inc., 1950. Pp. 689 + xii.
2. CATTELL, RAYMOND B. "r, and Other Coefficients of Pattern Similarity," *Psychometrika,* XIV (1949), 279-298.
3. CRONBACH, LEE J. "'Pattern Tabulation': A Statistical Method for Analysis of Limited Patterns of Scores, with Particular Reference to the Rorschach Test," *Educational and Psychological Measurement,* IX (1949), 149-171.
4. CRONBACH, LEE J. "Statistical Methods for Multi-Score Tests," *Journal of Clinical Psychology,* VI (1950), 21-26.
5. CRONBACH, LEE J. and GLESER, GOLDINE C. *Similarity Between Persons and Related Problems of Profile Analysis.* Urbana, Ill.: Bureau of Research and Service, College of Education, University of Illinois, Technical Report No. 2, April, 1952. Pp. 53.
6. DUMAS, FRANK M. "The Coefficient of Profile Similarity," *Journal of Clinical Psychology,* V (1949), 123-131.
7. FLANAGAN, JOHN C. *Scaled Scores.* New York: Cooperative Test Service, December, 1939. Pp. 41 + iii.
8. GAIER, EUGENE L. and LEE, MARILYN C. "Pattern Analysis: The Configural Approach to Predictive Measurement," *Psychological Bulletin,* L (1953), 140-148.
9. KELLEY, TRUMAN L. *Essential Traits of Mental Life.* Cambridge, Mass.: Harvard University Press, 1935. Pp. 145.
10. KOGAN, LEONARD S. "Statistical Methods," *Progress in Clinical Psychology,* Vol. 1, Section 2, (1953), 519-535.
11. MAHALANOBIS, P. C. "On the Generalized Distance in Statistics," *Proceedings of The National Institute of Science, India,* XII (1936), 49-55.
12. OSGOOD, CHARLES E., and SUCI, GEORGE J. "A Measure of Relation Determined by Both Mean Difference and Profile Information," *Psychological Bulletin,* XLIX (1952), 251-262.
13. PEARSON, KARL, *Tables for Statisticians and Biometricians, Part I.* University College, London: The Biometric Laboratory, Second Edition, 1924. Pp. 143 + xxxiii.
14. RAO, C. RODHAKRISHNA. *Advanced Statistical Methods in Biometric Research.* New York: John Wiley and Sons, Inc., 1952. Pp. 390 + xvii.
15. TIEDEMAN, DAVID V., BRYAN, JOSEPH G., and RULON, PHILLIP J. *The Utility of the Airman Classification Battery for Assignment of Airmen to Eight Air Force Specialties.* Cambridge, Mass.: Educational Research Corporation, 1951. Pp. 328. (Reprinted by Educational Research Corporation, March, 1953.)

Problems and Procedures in Profile Analysis

DAVID R. SAUNDERS

SUMMARY OF DISCUSSION

I. *Comments on Dr. Anderhalter's paper.*

It was pointed out that in Dr. Anderhalter's case a multiple correlation analysis using the same final eleven scores would have given equivalent results, *i.e.*, the result is obtained by combining a number of effects that are significant separately at only the 10-15% level.

Dr. Thorndike pointed out that since wide-spread upper and lower criterion groups had been used to obtain cross-validities in the multiple discriminant analysis, the obtained figure of 0.61, *et al.*, would be expected to be lower for a total population. However, this would not affect statistical significance.

Dr. Alman wondered about the effect of skewness in Rorschach scores upon the discriminant analysis. Dr. Anderhalter indicated that the most highly skewed variables had been eliminated because of their unreliability, and that none were left among the final eleven used.

II. *Comments on Dr. Tiedeman's paper.*

Dr. Solomon suggested that the multivariate normal distribution is, in many cases, quite adequate for psychological data, from the statistical point of view. He asked Dr. Cronbach whether any conventional test of significance of the distribution is used in the Cronbach-Gleser technique.

Dr. Cronbach indicated that he saw no basic disagreement between the Cronbach-Gleser $\overline{D}$ and the Mahalanobis D, except as to purposes. There are really three distinct types of problem —(a) determining the similarity of two individuals to each other, (b) determining the resemblance of individuals to a single group (e.g., Anderhalter's study), and (c) assessing the relative resemblance of one individual to two or more groups (e.g., Tiedeman's study). D is better for *a* and *b*, while $\overline{D}$ is better for *c*. The point is that, within a single population, it is not always reasonable to regard persons with equal centour scores—"Isofreaks" —as effectively equivalent. This can be seen by considering a situation in which two discriminant functions are highly correlated within a particular criterion grouping. A report is due to appear in the Psychological Bulletin (1953, 50, 456-473).

Dr. Cronbach also suggested that we may do ourselves harm by considering people who have modal discriminant scores as best by definition; the best person may actually be one who deviates from the mode in an advantageous fashion. For example, the best veterinary doctor may have certain qualities that are more like regular physicians than like the typical veterinary, and receive a lower centour score as a veterinary. The best psychometricians and clinicians may be more like one another than the separations of the modes of

their two distributions, based on present selection procedures, would indicate. Some technique is needed for integrating this sort of consideration into our thinking about profiles.

Dr. Tiedeman indicated that he and Dr. Rulon have a report pending that will consider these and other matters. He suggested that the audience should experiment themselves with the different indices proposed, as a means of discovering the inconsistencies which appear as to which people are classified together.

The discussion had to be stopped before the issues raised could be further explored.

Making Test Results Meaningful

RALPH F. BERDIE

BRINGING NATIONAL AND REGIONAL TESTING PROGRAMS INTO LOCAL SCHOOLS

MOST TESTS have multiple uses. When test scores are used for college admission purposes and not used for purposes of counseling, important information is wasted. When objective test data are used for purposes of evaluating instruction but not used for purposes of providing academic guidance to students, again incomplete use is made of the tests. The concept of "percentage efficiency of use" can be used in testing to indicate the degree to which maximum use is made of test results. The purposes of this paper are to discuss problems arising in increasing the efficient use of tests administered in large-scale programs and to discuss means of increasing the efficiency of use.

National and regional testing programs have developed in response to a variety of educational needs. Usually these programs have been conceived as means for solving certain broad educational problems. Occasionally these problems have been rather specific in nature and when this was the case, the testing program has had the potential for making an educational contribution somewhat greater than the solution of the posed problem. In every case, however, we must face the questions of how these testing programs fit in to the broader educational picture and how the data obtained through the means of these programs can be integrated with instruction and counseling.

The extent and coverage of national and regional testing programs have expanded since the first World War until now they constitute a significant activity in both education and psychology. A number of national programs have developed in response to needs expressed by professional schools and professional associations. National testing programs related to professional training are found now in Medicine, Dentistry, Accounting, Law, Engineering, Nursing, Journalism, Architecture, Pharmacy, and Education. Even now, attempts are being made in other professions to develop similar programs. Thousands of students each year are tested in these programs.

The early history of psychometrics suggested that national and regional testing programs would develop in response to demands for vocational guidance aids. Only one large scale national program, that sponsored by the United States Employment Service, has developed in that direction but this development has resulted in the promising General Aptitude Test Battery that eventually may be used with thousands of our young people.

Other testing programs have developed in response to needs for the appraisal of educational efforts. The

Sophomore Culture testing program, the Graduate Records Examination, and the testing programs in the State of Iowa all have been developed as a means to help educators evaluate the results of their labors.

Somewhat related to the programs coming from the professional schools are the programs developed to help colleges identify and admit qualified students. The programs of the College Entrance Examination Board, including the college transfer test, and the programs of the Association of Minnesota Colleges have had unique developments. More recently another type of national testing program has appeared in response to problems raised through the contiguity of the National Defense Program with our institutions of higher learning. The NROTC and the Selective Service Examinations provide important coordinating links between colleges and the defense department.

Another variety of a regional testing program has appeared, a variety that has purposes overlapping in large part with the purposes of many of the programs just mentioned. These are the state-wide testing programs now found in about forty states, programs sponsored by colleges and universities, educational associations, state departments of education, and combinations of these agencies. These several programs involve the testing of hundreds of thousands of pupils each year, from the elementary level to the college level, and they have resulted in a flood of objective test information pouring into our schools.

Obviously, national and regional programs serve a large number of purposes. Some programs aim to provide information to the schools to be used primarily for institutional purposes, sometimes for the purpose of individualizing instruction and sometimes for the purpose of evaluating instruction.

Some of these regional and national programs are utilized for what might be frankly called "recruiting" purposes. In these cases the purposes of the program involve the identification of qualified students with the expectation that suitable action can then be taken to increase the probability of the student's entering the appropriate institution of learning.

Closely related to this purpose are purposes involving the selection and admission of students, both in colleges and professional schools. Programs and curricula in most schools are designed for a relatively homogeneous group of individuals and schools frequently assume the responsibility for determining that all students admitted can have a reasonable expectation of doing successful work.

Finally, many of these programs have as their purpose the providing of objective information to be used in counseling individual students. In schools at all levels, the importance of considering psychological individual differences is being recognized and this recognition has resulted in an expansion of counseling and guidance programs.

Large-scale testing programs provide certain advantages that programs of a more limited or local nature cannot provide. Let us first recognize some of the advantages. In a program involving the testing of thousands or perhaps hundreds of thousands of individuals, many resources, both personnel and financial, are available to be used in the development of better technical instruments. No one who has gone through the shop of the Educational Testing Service can fail to be convinced of this. In constructing a good test, although it may not be necessary, it certainly helps to have available experienced item writers, trained editors, batteries of IBM machines, and high powered statisticians.

Another advantage resulting from

large scale programs comes from the samples available for test standardization purposes and for the thousands of persons who can provide normative data.

A third advantage derives from the extensive experience possible to accumulate with a test used in many situations. A test used in hundreds of schools may be subjected to research by dozens of individuals able to develop hypotheses, test these hypotheses, and attempt to verify the experiences reported by others. Not only is the range of research extended, but also the experiences obtained through the application of testing programs are different and varied. Finally, the whole activity of psychological testing may gain through the prestige accrued from these large scale programs.

Now let us look at possible disadvantages of national and regional programs as compared to local programs. In a way, every one of the advantages just listed also can be a disadvantage. In the past, it has been all too easy for those responsible for national or regional programs to ignore specific needs and goals of local test users and this, in some cases, has resulted in a minimizing of contributions that could be made by persons acquainted with local situations. For instance, a medical school that decides to make use of the tests available nationally may fail to give support to its local testing people who might be in a favorable position to develop superior tests for the selection of medical students.

The development of large scale programs also has apparently discouraged the use of local norms. Reports published by many persons responsible for national programs demonstrate with little question that significant differences are found among institutions, in some cases the top students in some groups are no better than the bottom students in others. In the face of this kind of evidence, schools still tend to make use of the national norms reported in the large scale programs and they rarely develop their local norms.

Along with the possibility of discouraging the development of local tests and norms, these large scale testing programs also may discourage locally sponsored and conducted research on the problems toward which these testing programs are directed. The recently published research of Ralph and Taylor in medicine and of Weiss and of Layton in dentistry demonstrate the need for such research.

The nature of the sponsorship of these programs also may have the effect of encouraging uncritical acceptance on the part of the local schools. A dean in a school of dentistry, knowing very little about psychological testing, is apt to accept without question any battery of tests bearing the stamp of approval of the American Dental Association, quite apart from the merits of the tests themselves. Similarly, high school teachers may feel that the name of a university attached to its testing program may serve to guarantee the quality of the product without seriously considering the nature of the program as it relates to the needs of the school.

In general, I suppose we can say that one of the chief disadvantages of these large scale national and regional testing programs is that they may tend to discourage a diversity of approaches in meeting educational problems where tests have a role to play.

Now obviously these advantages and disadvantages are not inevitably attributes of large scale testing programs. Many factors determine whether these tests are used effectively or not. We can identify some of the conditions conducive to the effective or ineffective use of tests.

First, let us look at the preparation of the persons using the tests. In only a minority of schools are there persons

technically trained to use psychological or educational tests and only our larger school systems have a person who may have the equivalent of a Master's degree in psychometrics. In most schools, tests are used primarily by teachers and, in general, teachers are not particularly well prepared to use tests.

For instance, the United States Office of Education "Manual on Certification Requirements for School Personnel in the United States," reports that only twenty-seven states even clearly mention tests and measurements in relation to teacher certification requirements. For elementary teachers, in only two states are courses in tests and measurements required and in six states such courses are optional. For secondary teachers, in three states such courses are required and in eight states they are optional. Woellner and Wood, in their "Requirements for Certification of Teachers, Counselors, Librarians, Administrators," report that in only six states are such courses required for administrators and in seven states such courses are optional. For counselors, only eight states specify a testing course as a requirement. These figures cannot give an accurate picture of the situation, but certainly they do suggest that training in testing is not perceived as a major requirement for teachers in most states. We must recognize that in our schools are relatively few persons well trained in the use of tests.

Another condition conducive to the ineffective use of tests is the limitations of the tests used in programs. Sometimes these tests have poorly demonstrated validity, if any at all. Occasionally, inadequate norms are used and more often, although the norms may be adequate, inadequate descriptions of the norms are provided. Sometimes the tests selected are not appropriate for the purpose for which they are presented.

Another condition related to the ineffective use of tests involves the limitations of the interpretative materials and aids available to those who use the tests. Sometimes the descriptive materials and test manuals are not legible or provide ambiguous presentations. Frequently the material presented is not complete, occasionally such important information as that concerning validity is omitted. Frequently illustrative material is not provided to help teachers and counselors make better use of test scores.

Encouraging and Improving Local Use of Tests

Undoubtedly, while trying to increase our awareness of some of the dangers accompanying the continuing development of large scale testing programs, I have presented a somewhat pessimistic picture. I think, however, that the varied development of these programs has demonstrated adequately that in the long run they will benefit all of us, but both we who are responsible for the administration of some of these programs and the persons who are directly using the obtained results can do much to increase the efficient use of test scores.

First, those directly responsible for such programs can continually arouse and sustain interest in the programs and convince persons that if they acquire the necessary skills, the programs provide data that can be of unlimited value. We have the responsibility for repeatedly clarifying the purposes of these programs. Recognizing the great turn-over in educational personnel, to state our purposes only once is not enough. Then we must recognize that many, if not all, of these regional and national programs can serve a variety of purposes, not only a single purpose. Although we may be concerned with one purpose, we must be aware that test users will be making additional

uses of these tests and we have a responsibility for helping them. This calls for effective and continuing communication between those who are administering the programs and the test users. Cooperative planning and review among these persons is essential and only in this way can the necessary interest be maintained.

Not only must interest be aroused and sustained, but relevant information must be provided to those who are using the tests. The persons responsible for the administration of the programs must make available adequate manuals and bulletins describing the purposes of the programs, the tests included, possible uses of the test, and illustrative materials to serve as training aids. A constant flow of research publications should be maintained between those administering the program and the test users, research done by the central organizations and by other persons. Periodically, reviews of the current status of the program should appear. Summaries of the research completed, summaries of clinical experiences, and carefully considered judgments of informed authorities should be communicated to test users.

Next, the periodic assembling of test users provides an opportunity for reviewing problems in the application of tests and allows for those administering the programs to maintain awareness of the revealed needs. Similar purposes frequently are accomplished by having consultants visit test users, perhaps actually to observe the operations during which the tests are used. In the same way, the effective use of correspondence between program administrators and test users accomplishes these purposes.

Relevant information frequently can be provided to *prospective* test users before they begin the job, particularly in programs involving elementary and secondary schools. Test users at one time or another are students in teachers' colleges or colleges of education and in these colleges courses in testing usually are taught. These courses should include information about testing programs.

Program administrators frequently can encourage more effective use of tests by assisting users to collect data, both for large scale research and for local research purposes. Cooperative research frequently can be encouraged among the various schools making use of tests provided in a program. Comparative validities can be determined, new tests and techniques experimented with, and interpretive aids tried out. Local research projects can be encouraged by test administrators; program administrators might even encourage research in local settings by providing personnel and financial assistance and by facilitating the distribution of the results of such local research.

Thus, many things can be done to increase the effectiveness of the use of these national and regional testing programs. Most of my attention here has been given to things that the program administrator can do. Naturally, he does not bear all of the responsibility—the test user himself, must share in this responsibility, but the program administrator usually is the person who must assume the initiative.

In conclusion, I would like to draw these generalizations concerning national and regional testing programs as they relate to local users. First, these programs currently provide a tremendous amount of useful, objective data to local schools. More data is supplied at present than the test users are capable of using. Next, the presentation of test scores can be improved by program administrators, first by gaining more personal involvement on the part of the test users, and secondly, by providing them with more training. Next, more research is necessary if test users

and program administrators are to have the information needed to make adequate use of these data. Next, as with most problems related to this, more imaginative development of techniques and methods are needed, particularly on the part of the program administrators. Finally, we must recognize that these testing programs do not exist in a vacuum and that ultimately the effective use of data obtained through these programs depends upon the improvement of both counseling and instructional programs in our schools.

Making Test Results Meaningful

MAX D. ENGELHART

MAKING TESTING MEANINGFUL TO TEACHERS THROUGH LOCAL TEST CONSTRUCTION AND ANALYSIS OF TEST DATA

PARTICIPATION IN state, regional, or national testing programs or the use of standardized tests in local testing programs have certain definite values in measurement and evaluation. It seems evident to the speaker, however, that local test construction is also an important means of making testing and test results meaningful to teachers. Through participation in state, regional, or national programs and through the use of standardized tests, data may be collected which make possible comparisons of pupil or student aptitudes and attainments in the local situation with the aptitudes and attainments of pupils or students in other school systems, or higher institutions. Local test construction can, however, be a means of providing evaluation instruments more closely related to local instructional objectives. It can be a means of stimulating greater interest in testing and in use of test results by teachers. Teachers experienced in test construction are usually better able to select standardized tests for local use and to understand data collected in state, regional, or national programs. Just as an amateur artist or musician may have a better understanding and appreciation of the works of professional artists or musicians than the person who has never tried to paint or play, the teacher experienced in test construction may have a greater understanding and appreciation of professionally prepared tests.

The test expert working with teachers in local test construction encounters a variety of problems. Teachers differ widely in their knowledge of testing and their attitudes toward it. Where a group of teachers are asked to cooperate in the production of an examination to be given at the end of the course which all of the teachers are teaching, the members of the group can be expected to react in a variety of ways. Some members of the group will deplore the idea of preparing objective exercises. Other members, if asked for contributions of exercises, will prepare series of true-false or multiple-answer exercises relevant to facts of importance only to themselves. Often such exercises are contributed by the teachers who would have preferred essay questions, or who contend that attainment of worthwhile objectives can only be measured by essay questions. Frequently, however, the test expert is pleasantly surprised by finding one or more teachers in the group who have a considerable understanding of what should be done and ability in preparing effective exercises.

The logical first step in constructing an achievement test by a group of teachers is to have the teachers define their instructional objectives. It has been the experience of the speaker,

however, that to begin by asking teachers to list their instructional objectives seldom results in a useful compilation of objectives. Adequate definition of objectives develops slowly and through continued cooperative efforts in test construction and in analysis of test data. In planning the construction of a test with a group of teachers initiating test construction, discussion of objectives and of exercises useful in evaluating attainment of the objectives needs to be quite general and elementary. It is helpful to provide the teachers with examples of various types of objective exercises and with notes explaining how to write such exercises. These notes may call attention to such things as having the problem of a multiple-answer exercise set in the stem of the exercise, having the answers of parallel construction, having all of the distractors plausible, and avoiding the making of the correct answer the longest and most involved one. It is desirable to attempt some definition of objectives. It is important on the level of knowledge and understandings to identify the common content of instruction otherwise many exercises will be contributed by individual teachers which the group as a whole will later reject. In initiating cooperative test construction one cannot obtain the emphasis on production of exercises designed to measure intellectual skills that can be reached after several semesters of experience. As a matter of fact, emphasis on such exercises in the earliest tests prepared may be quite unjustified since such skills are not likely to be among the real objectives of instruction. One of the major advantages of local test construction is that it can contribute to widening the scope of instructional objectives to include such skills. It seems worthwhile, however, even when initiating local test construction, to seek some departure from evaluation of facts or information alone. Progress has been made when teachers accept the idea that an objective exercise will measure more than factual knowledge when the exercise represents to some extent a problem situation and requires application of knowledge in context other than that in which the knowledge was taught.

After a group of teachers has met for the purpose of planning an examination and various members of the group have accepted responsibilities for the writing of exercises, it is desirable for the test expert to seek consultations with the individual teachers. Such conferences may be an effective device in stimulating the teachers to begin work prior to the deadline set. Where a teacher has written some, or even all of his exercises, the conference can be a means of giving the teacher helpful advice with respect to exercise writing techniques. Such conferences are also an opportunity for the test expert to become much more familiar with the real objectives of instruction.

After the exercises have been prepared it is desirable to have the contributions of the various members of the group evaluate each others work. Where the group of teachers of a course is so large that exercise writing may be done by a committee, it is desirable to have all teachers participate in the evaluation of the exercises. The test expert may synthesize these evaluations or it may be done in a later meeting of the entire group of teachers or of the committee. It would be ideal, of course, if the exercises finally selected are all approved by all of the teachers. Unfortunately, this ideal can seldom be realized. It will usually happen, however, that many of the exercises rejected are judged faulty by most of the teachers. This is especially true of exercises which are factually incorrect or which pertain to content uniquely taught by one of the teachers. The opportunity to participate in the evaluation of exercises is appreciated by

teachers, and given this opportunity, use of exercises not approved by one or more of the teachers, but approved by the group as a whole, are much less likely to provoke resentment when the test is given.

The planning of an examination, the writing of exercises and their critical evaluation all take time. No single teacher should be asked to contribute an inordinate number of exercises. There should be plenty of time for evaluation, editing, and duplication of the finished test. Hence, it is desirable to initiate the work early in a semester and to keep the work moving during the semester.

It was said above that adequate definition of objectives requires time and that this is particularly true of objectives pertaining to intellectual skills. The following list of such objectives was written for use in the production of social science exercises pertaining to quoted material and as a guide for instruction in the social science general course of the Chicago City Junior College. The list was prepared by a social science instructor* long interested in evaluation with the help of other instructors most of whom have had some years of experience in local test construction.

The kinds of intellectual abilities we should attempt to measure include:

1. The ability to identify the central issue or problem.
2. The ability to recognize and understand underlying assumptions.
3. The ability to identify the hypothesis, or hypotheses, tested by the data presented in the quoted material.
4. The ability to analyze an argument with respect to bias, emotional factors, and propaganda devices.
5. The ability to distinguish between facts, opinions, points of view, and value judgments.
6. The ability to analyze the logic of an argument.
7. The ability to evaluate data or evidence in terms of relevance to the problem and with respect to their adequacy to prove or disprove conclusions or generalizations based upon the data.
8. The ability to identify conclusions or inferences definitely supported by the data, probably supported by the data, or conversely, definitely disproved, or probably disproved by the data.
9. The ability to recognize and understand relationships and aspects of similarity and difference.
10. The ability to identify legitimate predictions of effects or of social consequences of given courses of action.

The above list was incorporated in a statement entitled "Suggestions for Writing Exercises for Social Science Examinations" which was mimeographed for distribution to all of the teachers contributing to the social science examinations. In order to increase the effectiveness of the list for exercise writing, a briefer statement of each objective was followed by two or three exercises from past examinations judged to be useful in evaluating the skill defined. Time will permit the quoting of only a very few examples. All of these exercises are from series of exercises pertaining to selections presented in the examinations, or distributed to the students for study prior to the giving of the examinations. It is to be hoped that something of the nature of the quoted material can be inferred from the exercises.

Identification of central issue or problem.

The paragraphs quoted above discuss a condition which is most relevant

* Hymen Chausow of the Wright Branch of the Chicago City Junior College.

to which of the following basic problems of government?

A. Which is better, a parliamentary or presidential form of government?
B. Which is better, a democratic or a totalitarian government?
C. Which is better, government by experts, or government reflecting the will of the masses?
D. Which is better, a government based on direct or indirect democracy?

Recognition and understanding of assumptions.

Regarding the role of government in society, the author apparently assumes that

A. "that Government is best which governs least."
B. "there should be more business in government and less government in business."
C. government should direct economic activities for the good of the people.
D. government should do for the people what they, as individuals, cannot do for themselves.

Analysis of an argument with respect to its logic.

One can make a very good case to show that Mr. Brown is inconsistent when he

A. is against expanding governmental functions but is for a St. Lawrence Waterway.
B. is speaking as a private citizen while at the same time speaking as a representative of the John Jones Company.
C. tries to show that the aims of the Anti-Tax Group are compatible with the aims of his firm.
D. advocates the elimination of farm subsidies, but also wants to reduce government expenditures.

Identification of conclusions or inferences supported or not supported by the data.

On the basis of evidence presented by the author one can logically conclude that

A. employers are usually right in the disputes which arise between management and labor.
B. the laboring class can not be expected to know what is best for society.
C. the laboring class uually demands that which is detrimental to society.
D. the methods used by the laboring class to secure their demands are sometimes detrimental to society.

Prediction of consequences or effects.

If the sales tax advocated by the author of the above quotation were to be adopted, which of the following would be likely to happen?

A. Less money would be collected by mean of the income tax.
B. People with large incomes would suffer more than people with small incomes.
C. People with small incomes would suffer more than people with large incomes.
D. The burden of the tax would be proportionate to income.

The exercises just quoted are all of the multiple-choice type. Other types frequently used include items to be classified according to such categories as "A. if the first speaker would agree, B. if the second speaker would agree, C. if both speakers would agree, and D. if neither speaker would agree" or "A. definitely true, B. probably true, C. insufficient evidence, D. probably false, and E. definitely false." In any given examination series of exercises of various types are also included which do not refer to selections of quoted material.

Many instructors exhibit increasing enthusiasm for exercises relevant to quoted material feeling that such exercises minimize recall of information and do evaluate important intellectual skills.

Other teachers, however, deprecate such exercises and claim that students can answer them successfully without serious study of the course. Apart from hoping that such instructors will soon retire, the test expert working with teachers can make an effort to change such attitudes. It is helpful to call attention to the fact that such exercises are becoming increasingly used in the newer and better standardized tests and in the tests of important regional and national testing programs. It is sometimes convincing to point out that such exercises have functions similar to the essay questions the teacher has claimed measure the worthwhile objectives. The argument that no study is necessary has less force if the selections on which exercises are based are chosen wisely. They should be of sufficient difficulty that experiences gained during instruction contribute to understanding of the selections. Certain of the exercises in the series which follows a selection may measure background knowledge of terminology, facts, principles, or conditions relevant to the quoted material, but not defined or explained therein. For example, suppose a paragraph on some labor problem includes such terms as "closed shop" and "union shop" and that those terms are *not* defined in the paragraph. An exercise may concern the selection of the best statement of the difference between the closed shop and the union shop, or two exercises may be written involving separate definitions of these terms. Suppose that a paragraph represents a proposal for the adoption of some form of sales tax. While other forms of taxation are not discussed, it will be legitimate to have some of the exercises deal with the relative merits of different kinds of taxation. Such efforts to increase the relevance of the exercises to the subject matter taught definitely make the exercises more valid with respect to instruction and do tend to promote acceptance of exercises in the same series of exercises more concerned with evaluation of intellectual skills.

It is the belief of the speaker that while teachers are engaged in the activity of exercise writing they are devoting thought to instructional objectives. The teacher asks himself whether the knowledge required in solving an exercise is within the scope of the knowledge taught. If the exercise is one which requires application of intellectual skills, the teacher may ask himself whether or not most of the students can make such application of intellectual skills. In other words, the teacher attempts to forecast the student behaviors and to imagine what will go on in the mind of the student confronted with the exercise. Even in writing distractors the teacher may use experience gained in instruction to phrase distractors which owe their effectiveness to the fact that they are the kind of answers he can expect from less able students. Similar thinking occurs in the mind of the teacher carefully evaluating exercises prepared by his colleagues. Finally, while participating in the selection of exercises for the test as a whole, the teacher must think about whether or not the examination will sample the content of the course in a representative manner and whether there is an appropriate balance with respect to the evaluation of various instructional objectives. The teacher who has participated in the whole process of test construction and evaluation comes to accept testing as a major and essential function of instruction.

The preceding discussion has been largely concerned with some of the problems of local test construction and I hope has indicated, at least indirectly, how participation in test construction can make testing more meaningful to teachers.

After the tests have been given and scored a carefully written report can be

a means of making the test results meaningful to the teachers. Most instructors understand tables presenting frequency distributions of scores and such averages as means and medians. Standard deviations, standard scores, and percentile ranks need repeated explanation both in reports and in discussion with teachers. The same is true of coefficients of correlation. Understanding of these things comes slowly, but as semesters pass it is gratifying to find more and more teachers interested in such data and increasingly familiar with their meaning. Participation in the construction of the tests is surely a factor.

In addition to reporting summaries of test scores, testing and test results become more meaningful to teachers if the teachers are given the item analysis data relevant to the tests to which they have contributed. We enter, adjacent to each exercise in several copies of each test, the item difficulty and the item test correlation. The item difficulties, percents of correct response, make it possible for each teacher to evaluate achievement in terms of instructional objectives. Both the item difficulties and the item test correlations are of great help to teachers when revising a series of exercises for future use. The instructors of one of our departments are not satisfied with analysis pertaining to only the correct answers. They have insisted that revision of exercises requires knowledge of the proportions of students choosing each distractor. It is evident from subsequent item analysis that revisions based on such data have materially improved the exercises.

Participation in cooperative efforts in test construction is a factor in the improvement of instruction and in the improvement of evaluation. The effects are not restricted, however, to courses and examinations in which the cooperative effort occurs. The informal testing done by the teachers in their other courses is affected. The teachers cooperating in the construction of tests acquire more interest in the testing done for guidance and placement purposes. They become interested in measurement of objectives other than knowledge and understandings and of intellectual skills and want to know how to construct or select instruments for the measurement of attitudes or other traits. They become enthusiastic about participation in evaluation studies. The examiner's problem finally becomes one of keeping up with the teachers with which he has to work.

Making Test Results Meaningful

ROGER T. LENNON

THE TEST MANUAL AS A MEDIUM OF COMMUNICATION

THE FACT that this discussion of the test manual as a medium of communication is part of a program devoted to the general topic "Making Test Results Meaningful" dictates that we concern ourselves with problems of transmittal, via test manuals, of those knowledges and skills that will enable test users to comprehend, evaluate, and make more effective use of test results. The inclusion of the topic "Making Test Results Meaningful" in this conference betrays our awareness that test results in many instances are less meaningful to users than they might be, and our hope that at least some increase in understanding can be effected through improved test manuals. Few will question the need for better understanding of test results by the user. Durost, in summarizing trends in testing a year ago, noted that "Technical developments in test construction have outrun practice and there is a widening gulf between the test maker and the test user."[1] Traxler, reporting on a recent survey of testing practices in large cities, states that the major problem in testing is now that of communication, including the communication of the test maker's knowledge about the uses and misuses of appraisal instruments to the actual users of tests.[2] How soundly based is the hope that we shall be able to improve this situation appreciably by improving test manuals is another matter, and the one with which this paper will be largely concerned.

In the remarks that follow, I have in mind particularly manuals for achievement, intelligence, and aptitude tests; and when I speak of the "test user," I am thinking generally of the classroom teacher, the guidance counselor, and the supervisor, rather than of the psychologist, the director of research or the test specialist.

The problem of the test manual as a medium of communication has two aspects: what is to be communicated, and how it may best be communicated. Let us consider each of these in turn.

Content of the manual. What *must* a test manual include? What information is necessary in order that test results be meaningful? There is substantial agreement at least as to the topics that ought to be covered, if not as to the detail with which each should be treated. This agreement has been fairly well summarized in the APA's set of technical recommendations for test manuals.[3] As a minimum, a manual should include specific directions for giving and scor-

[1] Walter N. Durost. "Modern Trends in Testing and Guidance" in *Modern Educational Problems*, edited by Arthur E. Traxler. American Council on Education, 1953.

[2] Arthur E. Traxler. "The Status of Measurement and Appraisal Programs in Large City School Systems" in Educational Records Bulletin No. 61. Educational Records Bureau, 1953.

[3] APA Committee on Test Standards. "Technical Recommendations for Psychological Tests and Diagnostic Techniques: Preliminary Proposal. *The American Psychologist*, Vol. 7, No. 8, August, 1952.

ing a test, and information on how to interpret results, validity and reliability, and on the purpose and construction of the test.

Few will deny that all these kinds of information have a contribution to make toward a better understanding of test results and toward enhanced meaningfulness of the results. We may well wonder, however, whether the information usually included under these headings is what the ordinary teacher or counselor needs to make the test results meaningful for him. For test results to be meaningful to a test user, it is necessary that he be able to see connections between the test results and problems which are real and urgent for him—to perceive how the test results are related to the goals for which he is striving. In my judgment, information about item validities, errors of measurement, selection of norm groups, and the like, does not help the user to see these connections—in fact, may set up a barrier to such understanding by focusing undue attention on the technicalities.

When we ask what information is necessary in order that test results be meaningful, we must also ask "Meaningful for whom?" For the teacher? for the director of research? for the counselor? for the administrator? Each is seeking something different in the results, each has his own purposes in mind. Information that will add to the meaningfulness of the results for one may not for another. The effort to meet the needs of these various audiences, differing as they do in training, understanding, and test sophistication, greatly complicates the task of preparing a satisfactory manual, both from the standpoint of providing sufficiently for all their needs, and with respect to the level of presentation.

If it be true, as some feel, that manuals do not provide as much technical information as the specialist needs for adequate evaluation of a test, it is far more certain that they do not provide enough assistance to the teacher or counselor in the proper use of the tests. How much help can a teacher derive from the average achievement test manual with respect to using results in such matters as marking and grading, improvement of instruction, motivation of pupils, diagnosis and remediation? The teacher needs more specific recommendations, more concrete illustrative material than is ordinarily provided on these applications of test data.

Lest you misinterpret what I am saying as a recommendation to reduce the amount of technical information in favor of the how-to-do-it type of content, let me hasten to say that the solution certainly does not lie in the direction of providing less technical information, even though we may suspect that such information goes unread or uncomprehended by the majority of test users, but rather in the provision of more of the kind of information that relates the test results to the user's own needs and problems.

The attempt to provide all the kinds of assistance that all types of users need has one inevitable consequence—the unending expansion of test manuals. This poses a most difficult problem, the possible solutions to which we cannot even begin to consider here. Whatever the proposed solution—whether it be reorganization of material within the manual, relegating technical data to an appendix, publication of extensive manuals not included in test packages, or publication of several items to take the place of the traditional manual, all of which have been resorted to—it leaves something to be desired for adequacy of communication. Until it is safe to assume considerably more training in measurement than the typical teacher or counselor possesses, at the present time, the need for more extensive accessory material for tests than can conveniently be provided will persist.

Let us now turn from consid[illegible]ther than a reflection of the literary of *what* the manual should c[illegible]ent of the author. Technical concate, to the *how* of its commu[illegible]pts, usually of a quasi-mathematical Can the ordinary user read, [illegible]aracter, and specialized vocabulary, hend and apply information as [illegible]not and should not be avoided enset forth in a test manual? [illegible]y in any adequate treatment of the manuals be improved in this r[illegible]lopment of a test and interpretaof results. We must be willing to presuppose enough training in urement to equip the user with nate background for understanding this technical information, or to acknowledge that this information in manuals will not contribute greatly to the meaningfulness of results.

Readability of test manuals. We h[illegible] explored the matter of readability o[illegible] manuals by computing Dale-Chall readability indices for the manuals for a group of achievement, intelligence, and aptitude tests. The tests involved—all widely used—are those of five different publishers, and I may say that there was remarkably little variation among them in the matter of readability. I believe these manuals to be reasonably representative. The average manual was found to be, on the whole, of about 11*th* or 12*th* grade reading difficulty, which would hardly appear to be beyond the level of the ordinary teacher. Within any manual, readability of content varies appreciably from one part to another. Those sections that consist of the specific directions for administering and scoring tend to be the easiest; those sections that have to do with technical aspects of a test—standardization data, reliability, item and test validity, etc.—are, as we would expect, more difficult. In this sample of manuals the more difficult sections were of college graduate level of reading difficulty, according to the Dale-Chall values. We may safely assume that material of this level of difficulty will be skipped or inadequately comprehended by substantial proportions of typical test users.

Can test manuals be written more simply, so that they will communicate more effectively to users of relatively limited measurement background? I have no doubt that they can—within limits. The fact that the various manuals which we studied were so similar in readability causes me to suppose that the level of readability is pretty much inherent in the nature of the material,

Comprehensibility. The fact that material is readable, of course, does not guarantee that it will in fact be read and comprehended. In an effort to obtain some insight into how well typical users actually understand the content of a test manual, we have recently conducted a study of the extent to which teachers could read, comprehend and apply information presented in the manual for the new edition of *Stanford Achievement Test.* We prepared a 68-item test covering information specifically set forth in this manual and administered it to two groups of teachers taking summer-session courses in tests and measurements. Prior to taking the test the examinees had been directed to read the manual carefully, as if they were going to give the tests, and were further told that they would be tested on their knowledge of the content of the manual. In reading the manual they would, therefore, probably have been at least as highly motivated as the ordinary teacher preparing to give and score the test. They took the test on the manual first as a closed-book examination without access to the manual and later with the benefit of the manual at hand.

The results under either condition evinced a disappointing level of mastery of the content of the manual. The average per cent correct on the items of the

test under the open-book condition was a little over 60. Only one item was answered correctly by all examinees, despite the fact that the answer to actually every item was readily obtainable in the manual.

We have studied the results of this test item by item in order to get some clues as to the sources of difficulty and nature of misunderstandings. I must confess that I have been unable to arrive at any generalizations in which I would repose much confidence—in fact, some of the results I find inexplicable. But I venture at least these few observations:

a. Items requiring some operation—*e.g.*, conversion of scores to grade equivalents or percentile ranks, calculation of months above or below norm — were harder than items requiring merely location of information. This finding we interpret as evidence of the limitations of the printed word as a medium for developing skills. Verbal directions for carrying out operations that are essentially quite simple frequently may create an impression that these operations are complicated and, by an odd paradox, the more complete and careful the directions, the greater the likelihood that they will seem hard. An operation which can be demonstrated very easily may in verbal presentation appear disturbingly complex. This fact highlights the importance of workshops, teachers' meetings, and in-service training programs, in which the operations of administering, scoring, and interpreting test results can be demonstrated, and in which teachers who are to be using tests may have an opportunity to perform the necessary operations under supervision. It further suggests the desirability of additional visualization in manuals themselves.

b. Items requiring the reading of tables proved difficult. For example, an item reading "The most reliable of the nine subtests in the Intermediate Battery is *P*," to be answered by reference to a simple table of reliability coefficients, was answered correctly by fewer than half the examinees; and other table-reading items were similarly hard. This indicates the necessity of textual decription of tabular data, and of provision of examples of how tables are to be used. Again, *show-how* and supervised practice seem called for.

c. Some items answered almost verbatim in the text of the manual were missed by surprisingly large percentages of the subjects. To illustrate, consider the item:

> "An examiner should never cut short the time specified for a test even though all pupils have finished." T or F

which was answered correctly by only 60 per cent of the examinees. The pertinent statement in the manual reads, "If all pupils, or all but two or three in a class finish before the stipulated time has elapsed, time may be called."

It is impossible to suppose that the problem here is one of readability or comprehensibility of the material as presented. One must seek the explanation in the attitude, or set, or motivation, which the user, or in this case the examinee, brings to the task of reading the manual. It is hard to believe that the average teacher, if sufficiently interested, could not locate in a test manual information of the kind called for in the question just cited. Perhaps our real problem is that of discovering why they are not interested and how

this interest can be stimulated—which brings us back to the principle earlier stated, that testing and test results take on meaning when the user sees them in relation to his own needs and problems.

Information of the kind revealed in this study is disconcerting, if not entirely unexpected; but it is necessary for any realistic appraisal of what test manuals can reasonably be expected to do, and as a basis for their improvement. I am glad to note, in this connection, that Roger Allison of Educational Testing Service has been making a somewhat similar study of the usefulness of the manual for the ACE Psychological Examination. He has prepared two forms of a test intended to reveal the extent to which teachers of counselors can make effective use of ACE results on the basis of information presented in its manual. He has, moreover, solicited their opinions as to the adequacy of various sections of the manual, further information they would like to have included, etc. His findings, as far as I know, have not yet been made public.

To summarize: we have considered the test manual as a means of transmitting to the user the information and skills he needs if the results are to be meaningful and useful to him. The following points have been advanced.

1. There is unquestionably an urgent need for better communication from test maker to test user, if test results are to be made more meaningful to the average user.
2. The test manual can be made a more effective communication medium, and can contribute more to better understanding of test results.
3. The meaningfulness of test results to the average user will be enhanced if the manual concerns itself to a greater extent with relating test results to the user's problems and needs, and demonstrates specifically and concretely the kinds of actions indicated by the results.
4. The readability of test manuals, and presumably their comprehensibility, can be improved; but it is neither possible nor desirable to avoid technical material.
5. Dependence on the test manual alone for proper understanding and use of test results is inadequate. Not only must more formal training in measurement be encouraged as vital for proper use of test results, but such formal work should be supplemented by workshops and other in-service training, if teachers, counselors, and supervisors are to derive maximum benefit from tests.

The Teaching of Educational Measurement

VICTOR H. NOLL

THE INTRODUCTORY COURSE IN EDUCATIONAL MEASUREMENT

A RECENT SURVEY[1] of psychologists employed in schools or departments of education in 339 institutions provides among other things, a list of the courses which they teach together with the frequency of mention. Among these are many courses whose title includes the words measurement or testing. In fact, such courses constitute one of the first ten in frequency of mention in a list of 43. The most common titles of the measurement courses are Tests and Measurements, Educational Tests and Measurements, Mental Testing. In addition there is a substantial number of measurement courses whose titles are not readily classified under these three.

Since these courses, however named, are all offered in schools or departments of education it may be taken for granted that most of them, if not all, are designed and offered for prospective teachers, counselors and school psychologists or those already in such positions. It seems likely also, that if schools or departments of education offer any course in testing or only one such course it probably would be an introductory course in educational measurement. In order to check this line of reasoning the catalogs of four types of institutions were examined. These four were (1) large publicly supported colleges and universities, (2) large privately supported institutions, (3) state teachers colleges, and (4) better known liberal arts colleges. The number of each type of institution was approximately the same. All told, the catalogs of 68 institutions were studied.

Although the analysis is incomplete certain interesting facts are fairly clear. First, it appears that an introductory course in educational measurement is offered in nearly every institution of the first three groups, but of the liberal arts colleges only about one half list it among their offerings. The course is usually either undergraduate, or open to both undergraduates and graduates; in four institutions the course is offered in the department of psychology; in all others it is offered in the department or school of education. The usual number of credits is 2 or 3 semester hours. It is generally an elective course though 16 institutions require it of undergraduates in teacher education and about half that many (7) say it is required of graduate students in education. Data on these matters as well as on prerequisites, purposes of the course, requirement for a teacher's license and other considerations are being collected.

It seems evident that the introductory course in educational measurement is a matter of importance in programs

[1] Symonds, Percival M., Samuel Z. Klausner, John E. Horrocks and Victor H. Noll. Psychologists in Teacher Training Institutions. The American Psychologist. 7: 24-30. January, 1952.

for the education of teachers and to those concerned with measurement courses as a field of specialization.

Purposes of the Course

What should be the major purposes of this first course in educational measurement? Naturally, these will vary somewhat from situation to situation, but the following seem important and practical.

(a) *Orientation*—The student has undoubtedly heard of various types of tests such as achievement, intelligence, personality, etc. He has taken some constructed by his teachers and perhaps some standardized tests. He probably has heard terms like validity, norms, standardized and performance but this is his introduction to a systematic presentation and definition of such concepts. In one sense, the function of orientation may be the most important for this course. Certainly it is contributed to by practically everything that we do in it since so much of it is new to the student. However, orientation is a very broad term with many possible facets. It is necessary to choose the lines along which it can most profitably be directed since we should not attempt to do too much in this course, which for many is their only one in measurement. Nevertheless, it seems that there should be orientation with respect to the following, at least:

Understanding of basic concepts of measurement such as, validity, reliability and other criteria of good measuring instruments; different kinds of tests and other evaluative methods and devices; statistical concepts and some practice in simple statistical techniques; some acquaintance with typical standardized tests of the principal kinds; acquaintance with several standardized tests in the students' own major field; and knowledge of sources of standardized tests and information about them.

(b) *Improvement of teacher-made tests*—Every teacher must test and evaluate, if nothing more than at least the status and progress of the learner in subject-matter. It would appear, therefore that an important function of the introductory course in educational measurement should be the improvement practices of teachers, that is, to help them do better what is an integral part of the work of a teacher.

(c) *Introduction to the use of standardized tests in the school*—This includes knowledge of sources of tests and critical information about them; the ability to read test manuals and make judgments between standardized tests in the light of the conditions existing in the school where they are to be given and of objective data about them; how to administer, score and analyze results of standardized tests; and how to use this knowledge in planning and carrying out a practical testing program which contributes to the solution of educational problems and furthers the purposes of the school.

To recapitulate, the major purposes of the first course in educational measurement are conceived to be (a) *orientation* to appropriate elementary phases of the subject, (b) *improvement of teacher- or locally-made tests* (c) *introduction to the use of standardized tests in the school.*

Procedures

Whether one accepts the purposes as stated or holds to a different set of goals, the logical question in any case becomes that of how to achieve them. The remainder of this paper is devoted to discussion of methods for doing this. No claim is made that the ideas are unique, or outstandingly successful. They represent merely the results of a number of years of experience in teaching such a course and an interest in experimenting with various methods in the hope of improving it.

(a) Orientation

During the orientation phase the main activities are lecture, class discussion and problems. The lectures and class discussions center around basic concepts, different kinds of standardized tests, criteria of good measuring instruments, and statistics. Specimen sets of standardized tests are brought to class, studied and discussed. Problems are assigned in making a frequency table, calculating means and medians, semi-interquartile range and standard deviation, rank difference correlation, and setting up an ogive curve. Percentiles, standard scores and quotients are also introduced and discussed at this time. An attempt is made to bring out meanings and uses of statistical measures rather than to emphasize calculation, but each member of the class works out the assigned problems which are graded and returned. There is no illusion that these ideas and techniques are mastered by all students or even the majority. However, an attempt is made to give them enough understanding so that they can at least read reports and articles based on results of testing with fair comprehension and to develop some understanding and appreciation of statistical methods as an essential tool for deriving meaning from raw scores or other test data.

(b) Improvement of Teacher- or Locally-Made Tests

As was said earlier, every teacher must test and evaluate. It is felt that improvement in this phase of the work can best be brought about by actual practice in setting up objectives, making test items, and building a test. Immediately following the mid-term examinations mimeographed sheets are handed out specifying in detail the nature of the assignment. Briefly stated, it includes the construction of a 100-item objective test, together with instructions for giving and scoring it, and a scoring key. A statement of the objectives of the course in which the test is designed to be used and an indication of which objectives the test has been designed or is thought to measure are also required.

The full time of the class is devoted to this project for approximately three weeks. One of the first activities is examination of standardized tests in fields requested by the students, usually their major fields or subjects, and of tests made by students in previous classes in meeting this same assignment. The class is divided into groups according to fields of specialization such as reading, science, homemaking, etc. A large room, preferably with tables and chairs, is best for this purpose. The instructor brings to class specimen sets of different selected standardized tests in each of the desired fields and subjects. The various subject matter or interest groups then work over them. Next, student-made tests are brought in and used in the same way. Three or four class meetings are usually enough for this part of the project. Students who want more time for it may use the departmental test files during office hours. In general, the students are not encouraged to take these tests out of the classroom or office but they may be permitted to do so by individual request.

While this is going on, students are given some help in thinking about objectives. Suggestions are made on different ways of stating them and the desirability of stating them in terms of pupil behavior rather than subject matter is stressed. They also examine the manuals of both standardized and student-made tests for such statements. Visits to the so-called Juvenile Collection of elementary and secondary school textbooks in the library for ideas as to objectives and content are encouraged. Students are strongly urged to plan and

construct a test at the level at which they expect to teach. This is considered desirable because the students' subject-matter competence is usually much better suited to this than to a college or university level test. Equally important is the consideration that if he makes a test for elementary or high school level he has something he may find quite useful in the not-too-distant future.

If students wish to attempt construction of something else such as a rating scale or readiness test they are encouraged to do so provided they and the instructor are satisfied that they have the necessary knowledge and experience to make the effort worthwhile.

Following these activities, consideration is given to each of the commonly used types of objective test items. Each type is illustrated, and its advantages, disadvantages and particular usefulnesses are briefly discussed. Students are invited to try making sample items which they may hand in for criticism and suggestions from the instructor.

Elementary principles as to format of a test, grouping and arrangement of items, setting up of scoring key or stencils, directions for taking the test, for administering and scoring it, and use of answer sheets are also taken up.

After these activities in class are concluded students generally have about a week or ten days longer to complete the project and hand it in, usually a week before the end of the term. They are urged to type the entire paper and make a carbon copy. There are two reasons for this. First, if the test is typed it is much simpler to control spacing and alignment. In addition, the student-author usually has more pride and satisfaction in a neatly typed paper than one written in longhand. The second reason was alluded to earlier, in that the carbon copies are filed and used for instructional purposes with succeeding classes. This has proved to be much appreciated by the students.

Both the original and the carbon copy are turned in to the instructor. The original is carefully criticized, evaluated, and returned with a mark to the student-author. The carbon, unmarked, is added to the collection to be examined for ideas and for criticism by classes in succeeding terms. When the papers are returned the hour for that day is generally used to discuss common or outstanding strengths and shortcomings. Students are also invited to ask for individual conferences with the graduate assistant or the instructor on the details of their project and there are always some who do so.

Over the years in which this assignment has been used we have come to feel that it is one of the most rewarding activities of the course. Students occasionally complain that it takes too much time and effort but the much more common reaction is one of—"It took a lot of work but we learned a lot and it was worth the trouble." It often opens up an entirely new concept of measurement to the student. It gives him some know-how in formulating objectives of instruction and trying to devise test items to measure these objectives. It gives him an appreciation of the care and effort required to construct a good, objective test. Finally, it gives him a sense of pride in authorship and in having produced something tangible and substantial which he takes with him, and possibly finds use for, long after most of what we try to teach in the course may be forgotten.

Although it is believed that the benefits to the student of this exercise far exceed its disadvantages we recognize that it has some of the latter. For one thing it is time-consuming. A good share of the time of the last month of the course, both in class and out of it, is given over to the project. Some may question whether the practice in test planning and construction merits that

much time out of a term. We definitely feel that it does for reasons already stated.

Another possible criticism is that the student may, in his inexperience and lack of interest, be impelled to lift items or at least, ideas for items, from sources available to him and not do much original work or thinking at all. Every effort is made to prevent this. Every opportunity is taken to emphasize the practical value of the exercise and to point out that students can obtain the maximum benefit from it only by going through the entire experience for themselves. Also, they are told that it is not difficult for the instructor to distinguish between items taken from existing tests and those original with the student. There is, of course, an element of bluffing in this, but it has been possible in some cases, from time to time, to detect such cheating and have the student confronted with the evidence, own up to it. In the great majority of cases, our impression is that they do not knowingly borrow from other sources for test items.

Another disadvantage is the great amount of time required to read the papers. It takes on the average, one half hour per paper to read, criticize and evaluate. The procedure followed is to have a graduate assistant sit in the course, especially during the second half, so that he knows the assignment and emphases thoroughly and also gains some familiarity with the areas in which students are working. By the time the papers come in he has been briefed on procedures in evaluating them, what to look for, etc. He reads each one carefully, makes notes, and gives a tentative evaluation. Following this the instructor looks them over, adds or modifies comments and, if he disagrees with the assistant's evaluation the paper is discussed with him and a decision is reached as to the final mark. All this takes time and energy but it is felt that something into which students have put so much time and effort deserves a careful, conscientious reading and the benefit of any suggestions we can offer.

One of the interesting sidelights on this project is the variety of tests that have accumulated in the instructor's files over the years. Besides all the commonly taught branches in which standardized tests are available the list includes tests in every major sport; the various phases of industrial arts and agriculture; speech, radio, and journalism; commercial subjects; and tests on two Bibles—the King James version and the Freshman Handbook.

(c) Introduction to the Use of Standardized Tests in the School

In planning a testing program for a school in such a course as this it is not uncommon to set up a hypothetical situation as a basis for the discussion. The instructor supplies such data as size of enrollment, number of grades and pupils per grade, data available from tests previously given, if any, amount of money to be spent, and similar pertinent matters. Beginning with these data a testing program for the first year and for succeeding years may be planned. Practical considerations such as selecting tests in the light of the purposes to be served, ordering the tests, securing cooperation of the staff, administering and scoring the tests, and analyzing and using the results are among the points usually considered. While this exercise is not like the actual experience it may provide some insight into the requirements and the problems of an effective use of tests in the schools. Nevertheless, it cannot take the place of actual participation in planning and carrying out such activities.

It is quite often possible to give students actual experience and participation in a testing project or program. Requests are often received from schools for advice and assistance in se-

lecting tests, planning for their administration, administering and scoring them, analyzing the results, and studying them for ways in which they can be used. If it can be arranged without too much inconvenience or loss of time we try to give students taking measurement courses an opportunity to participate in such activities in a way that will be beneficial to them and to the school seeking help. Students in the introductory course in educational measurement as well as those in later courses in individual examining and group testing have frequently taken part in various phases of testing programs with which we have been involved. A more detailed report on such participation than time permits today has been given elsewhere.[1]

Perhaps a school has decided to give a readiness test to first-graders; or intelligence tests in all elementary grades; or reading tests in secondary grades. Whatever the nature of the tests to be given, the purpose for using them usually arises out of a problem or need somewhere in the school situation which is described and explained to the class in measurement.

Generally, the test or tests to be given have been selected by the principal of the school or by a committee of teachers. We ask for enough copies of the tests so that every member of the class may have one. We also collect all the manuals and other accessories that can be spared. Then the class is organized into teams of 2 to 5 members each, depending on the size of the class on the one hand and the number and size of pupil groups to be tested on the other, the tests to be administered, and similar considerations. In the grades we generally work with the classroom as a unit, while in secondary schools a whole class may be tested depending on its size and the facilities available. Occasionally, if the measurement class is a large one the pupils are divided into small groups so as to give more students experience. A team is assigned to a group and is responsible for testing that group. The team elects one of its number as the examiner and the rest serve as assistants or proctors. Copies of the test and manual are studied and duties of each team member are determined. The examiner practices his part such as reading directions, timing, etc. and proctors familiarize themselves with the procedures and with their responsibilities. Much of this goes on outside the classroom but at least one class session is usually devoted to answering questions, and a general briefing by the instructor.

During this time of preparation the detailed arrangements have been made for the administration of the tests. The date and time are set, permission is obtained for participating students to make up work of classes that will be missed and, if necessary, transportation by college bus or other means is arranged. It has been found preferable, if the class is large, to furnish transportation to insure that all participants will arrive at the appointed time and place.

It is always suggested to the principal, as diplomatically as possible, that the regular teachers be given a free period while the testing is going on. It has been found best, generally speaking, not to have the teachers in the room during the testing. The children are usually thrilled at being visited by college students, some of whom may be well-known for athletic prowess or other achievements, and there is seldom a problem of lack of cooperation. The instructor has assigned teams in advance, sees that they get to their stations, and stays on the job until the testing is completed.

After the tests have been given they

[1] Noll, Victor H. and Marvin D. Glock. Functional Courses in Measurement and Evaluation. School and Society. 70: 339-340. November 26, 1949.

are taken by the teams and scored. Papers are exchanged between teams and the scoring is checked. The results are analyzed by use of simple statistical procedures. Early in the course some problems have been worked out by the members of the class in which such procedures were practiced. Students who have collected data by administering tests to real children and scoring these tests, are much more strongly motivated toward such work than they were previously. With the help of the instructor the class studies the results, attempts to interpret them, and perhaps formulates some tentative conclusions and recommendations. If it can be arranged, the principal and perhaps a teacher or two meet with the class to discuss the results. This is usually a very interesting session. Students gain insights into the situation which they would otherwise not get and representatives of the school receive assistance in interpreting the results of the tests. In addition, of course, they have had the substantial help of having the tests administered and scored for them and of having a preliminary statistical analysis completed.

Students always seem to enjoy this project and feel that it is practical and valuable. Most students like to give tests and they consider this an opportunity to gain some actual experience. Scoring presents no problems when the work is divided among thirty or more students. Sometimes they have developed a personal interest in some of the children tested and are most anxious to see how certain ones came out on the test. Likewise, analysis and study of the results to see what use can be made of them by the school presents a challenge to which they react quite favorably. One of the objectives which students in measurement courses consider most important is that of learning to interpret and to apply results of testing.

From the school's standpoint the assistance received is always favorably regarded. The average teacher in service often knows little about the use of standardized tests. When the project is carried out cooperatively as here described teachers learn along with students. They also may change from a negative or skeptical attitude toward tests to a more realistic one which recognizes the benefits that testing may provide and not just its limitations.

Because of its value to the greatest number, the test construction project previously described is always a part of the first course in educational measurement. The class participation in testing in the schools is included whenever it is feasible to do so. The major considerations are time available, whether it can be fitted into the schedule, size of the class, and above all, whether the experience is such that it will be educationally valuable. We do not enter into such arrangements merely or primarily for the assistance and benefit of the school requesting it. Sometimes the tests to be given or the grade level at which they are to be given do not appear to be such that the experience seems very appropriate or valuable for a particular class. Again, the class may be too large to handle easily, although as many as 40 have been used. Also, the maturity of the students and their readiness to undertake the project may be a determining factor. And, of course, if it appears that there is not enough time left in the term to do a good job, the class is not brought into it.

It may be appropriate to speak briefly concerning the general requirements of the course lest some listeners gain the impression from what has been said here today that it consists solely of the projects or activities which have been described. This would be a rather incomplete picture. During the orientation phase, which occupies the first

month or so of the term, major emphasis is placed on assigned readings in the textbook covering basic theory and statistical methods. We have not generally required much outside reading in undergraduate sections because we feel that reasonable attainment through the study of one basic textbook is perhaps all that can be expected of beginners. Students who evince interest in particular phases of the work are given a supplementary reading list and assisted in finding additional material.

To measure achievement we use three objective examinations in addition to the problems and papers described earlier. The first of these deals entirely with elementary statistical concepts; the second is the mid-term examination on basic theory and statistical methods; the final examination covers the reading and lectures of the entire course.

In closing, I should like to make explicit two ideas that have been implied in much that has been said here today. The first is that a great deal of what we try to teach in introductory courses in measurement probably has little significance to the student unless he has opportunity to make use of the ideas we present. Measurement theory, statistics, use and interpretation of standardized tests and test construction can mean little to the person who has not actually applied them or worked at them. This is not to say that verbalization of terms and concepts, and classroom discussions and practice in educational measurement have no value; however, it seems reasonable to assume that application of these ideas and procedures in a more realistic situation would increase both motivation and understanding. The value of our courses in educational measurement probably depends for most students on the extent to which what is taught carries over into practice. If this is so, every effort should be made by instructors in such courses to facilitate and increase this transfer.

The second point, which is closely related to the first, is that it is possible to carry out activities, even in beginning courses in educational measurement, that will provide practical opportunities for students to apply what is presented in the course. The projects here described provide experience and practice in the application of measurement theory and techniques to situations which are like those many prospective teachers will face after graduation. Through activities of the types discussed, we aim to give students something useful and basically sound in the field of educational measurement. No claim is made that what we do is very original or unique. What has been described here today is offered merely as an account of some procedures that have been used successfully, with the hope that it may be interesting and possibly helpful to others responsible for the same type of course.

The Teaching of Educational Measurement

W. C. KVARACEUS

IN-SERVICE TRAINING IN MEASUREMENT BY MEANS OF UNIVERSITY EXTENSION COURSES

ANNUALLY THE Harvard-Boston University Extension Division conducts fifty to sixty courses for teachers throughout the New England area. During the 1952-1953 school year, sixty-one different courses were offered in forty-seven communities. Over the thirty-one year period during which the Extension Division has been functioning, a total of 794 courses has been given in 159 communities enrolling 21,998 teachers.[1] If these adaptive training programs are effectively organized, they should result in improved school practices on a wide scale.

Several courses have been called for with high frequency in the area of educational measurement and evaluation. During the past nine years, the writer has personally conducted fourteen measurement workshops in the following New England communities: Portland, Maine; Barre, Vermont; Manchester, Peterboro and Hampton in New Hampshire; Framingham, Milford, Malden, Norwood, Stoughton, Rockland, Fall River, New Bedford, and Quincy in Massachusetts.

All the courses are set up upon request of the local school community. The courses generally carry two points of credit, although a third point can be earned by special arrangement with the instructor. Most courses carry graduate credit and run for fifteen weekly meetings of two hours each. The instructor commutes to the community, which is generally within two or three hours' riding distance from the Boston University campus.

Other than for the obvious saving and convenience in transportation for the class members, where do the advantages lie in offering an off-campus course? In what way does the off-campus measurement workshop course differ from the course with the same title offered within the University? The major strategy lies in the following: (1) the course is offered only on community invitation or demand, thus indicating that teachers and school officials are bothered by measurement problems and are, perhaps, in a state of readiness for learning; (2) the teachers are in close contact with real measurement problems around which learning experiences may be planned; and (3) course evaluations are now possible in terms of actual changes and improvements in the classroom behavior of the teacher. In view of these circumstances, the off-campus course in measurement and evaluation can be expected more readily to affect actual school practices.

[1] W. Linwood Chase, "Field Service," *Boston University Graduate Journal*, Vol. I, No. 10 (March, 1953), pp. 151-152.
Donald D. Durrell, "They Are Popular With Teachers," *The Journal of the National Education Association*, Vol. 35 (December, 1946), pp. 572-573.

Disadvantages in off-campus courses center around the lack of adequate library resources. However, except for bound periodical references, this has been largely overcome through the establishment of a separate extension course library.

Based on the experiences of the past nine years during which the writer has offered fourteen workshop courses in as many New England communities, the following generalizations have been drawn concerning methods and procedures which condition the effectiveness of an off-campus course in improving measurement practices in the local schools and communities.

While all measurement courses have been offered only upon invitation of the local community, much depends upon who initiates the request or who sponsors the course. Courses conducted because a supervisor or superintendent feels that his teachers are in dire need of this type of in-service training may result in an "increment-happy" captive audience lured on by monetary stimulus rather than by any desire for self-improvement. Only as the teachers have been consulted and have themselves elected to sponsor such an extension course will there develop a wholesome learning situation. Frequently the local teachers' association or a professional improvement or in-service training committee initiates the invitation and sponsors the course, thus representing a true consumer demand. Under such auspices there is more promise for ultimate improvement of school practices.

The Harvard-Boston University Extension Division has lent considerable assistance to a number of communities in the sequential planning of courses requested over the years. For example, the measurement course has frequently served as a starting base from which have stemmed courses in guidance, curriculum planning, etc. If an off-campus course does not fit within the mosaic of the over-all in-service training program in the community, it may represent a sporadic and spontaneous flight in adaptive training. Such isolated course offerings eventually prove of limited value.

Prior to the first meeting with the local workshop group, the writer has followed the practice of spending at least a day in the schools of the community in order to get the "feel" of the present status of testing, with particular reference to the existence of cumulative records, testing programs already in progress on an all-community basis, local leadership responsibilities for testing, curriculum revision programs underway, nature of the present forms for reporting pupil progress, and attitudes of local principals and supervisors toward current measurement and evaluation practices. For example, prior to offering a course currently being conducted in Quincy, Massachusetts, an afternoon was spent in the schools of the community. A conference with the local superintendent, assistant superintendent, guidance director and elementary school consultant was held. This conference brought out the following types of information: much thought was being given to problems of report cards and other forms for reporting pupil progress; the Stanford Achievement Test, Partial Battery, Form J, was being administered to all children in grades four through nine; and that tests of ability were being administered in grades one and five. This type of information enabled more effective planning for meaningful learning experiences related to the testing program already in progress.

While most of the workshops have been conducted on a one-term basis and run for approximately a fifteen-week period, a promising adaptation has been made in several communities by extending the course throughout the

year. In this way, the course meets every three or four weeks for the same number of fifteen meetings, thus enabling projects to be planned and carried out. A measurement course confined to a fast sequence of fifteen meetings seldom enables the teachers to do more than plan for the construction and use of tests. A course spread out over the year insures the opportunity to administer, analyze, and use the various instruments and techniques discussed in the course.

In one community, Framingham, Massachusetts, such a workshop was planned on a full-year basis and included all elementary teachers. This course met one afternoon a month on school time. Classes were dismissed for this afternoon, enabling a three-hour work session. General meetings were held the first half of the period, followed by small group meetings with committees working on different problems in the second half of the session. Meeting on school time gave added prestige to the importance of educational tasks undertaken in this in-service training program. Special arrangements were made by the community to bear a share of the expenses of the course, since many teachers were enrolled on a non-credit basis.

In the first years of the writer's experience in conducting these courses in the field, a comprehensive inventory test covering the measurement and evaluation area had been prepared and was used at the first session of every course. These inventory tests revealed that teachers had little knowledge of the field of measurement and that they labored under a great deal of misinformation. One other effect was notable: use of the inventory instrument always tended to cut down on the enrollment in the class. Those people with particularly low scores dropped out of the course in greatest numbers. Teachers most in need found the initial testing a traumatic experience and decided not to take the course. Since repeated use of this instrument always seemed to result in these two phenomena, the writer now assumes that the teachers by and large enter the course as beginning learners. At the same time, it is acknowledged that the size of the enrollments are larger. However, there is considerable merit in discovering those persons who have had recent courses in measurement and who do have more knowledge and skill than others. Attempts are now made to uncover these people through non-test techniques in order to capitalize on the proficiencies which they bring to the class.

It is vital to discover and work with the individual who has been assigned responsibility for overseeing the testing activities in the local school system. In one community this may be the superintendent; in another, it may be the supervisor; and in still another, the principal of the school may make decisions related to test construction, purchase and use. If the measurement workshop is to be effective, this individual should be a participating member of the group. It is easy to see how a course conducted by an outsider may run at cross-purposes with a testing program planned by another individual working within the school system.

It is significant to note that many communities had not assigned this responsibility to any one person and that, by and large, no qualified school person was available to assume such over-all school responsibilities. During these workshop courses, this problem has always been faced by attempting to locate those teachers or administrators who had had some previous training and experience in the use of tests, and through the establishment of a Local Evaluation Committee consisting of representative members of the school staff. One outcome of the course observed in a number of communities, as

in Malden, Massachusetts[2] was the establishment of a position of Director of Measurement and Research which was eventually filled by a person trained for the task.

Whenever possible, the attempt was made to limit the course to the teachers in one town or city in an effort to plan the workshop around the unique problems of a particular school system. This has not always been possible, since some courses have been set up initially on a regional basis and have drawn from many small towns that by themselves did not have sufficient enrollment to support a course. Currently, such a course is being offered in Hampton, New Hampshire, composed of twenty-three teachers representing a half-dozen different nearby school systems.

In each community the course content tended to center around the same core found in any beginning measurement course. At the same time, each community's requirements have called for some local patterning and for varied emphases on particular topics. However, in every community, the learning experiences that were set up for the group varied considerably. For example, one community may place emphasis on local test construction, going through the process of item analysis and local norm-building; another community may center its efforts on the administration, analysis and classroom use of standardized tests; still another may spend a large share of time in revising its cumulative records and pupil report forms; and still another community may take off the failing pupils in their class lists and do a thorough individual testing job with each youngster. No school system showed any shortage in testing tasks to provide a rich source of learning experiences. The particular advantage of an off-campus workshop comes in this rich opportunity for providing real-life classroom situations involving knowledge of, and skill in, testing.

In a number[4] of communities, the superintendent of schools has made available a testing budget which gave the workshop members an opportunity to plan and carry out a school testing program. Such administrative cooperation tended to insure a desirable realism in the course, since the group did more than analyze tests that might sometimes be used in a theoretical school system.

In evaluating learning as a product of the course, use has always been made of a carefully constructed, item-analyzed achievement test prepared by the instructor, together with an appraisal of the individual and group activities carried on by class members. However, this final examination has always been drawn from a pool of available items in the instructor's files, which are selected to cover the areas of emphasis made in the particular course. By evaluating the actual use made of the tests constructed or purchased and which have been used in the classroom, and through school and classroom visitations, some subjective impressions of the successful implementation of testing practices have been obtained in all fourteen communities. However, the problem of evaluation remains a serious one. The writer looks to the day when courses such as these will be given on a non-credit basis which will, in a sense, evade the necessity of marking. This assumes in advance a professional maturity on the part of the in-service teacher and adheres to the concept that the activities engaged in are oriented in the direction of need and are basically desirable school activities.

Generally, a person working in the field discovers that testing is somewhat

[2] W. C. Kvaraceus, "Adapting a University Extension Course to the Needs of the Local Community," *School and Society*, Volume 68, No. 1754 (August 7, 1948), pp. 81-84.

in bad odor. The teacher is convinced that someone wants to test her—not the pupil, since this has frequently been her experience in the past. At the same time, most teachers show a severe allergy to matters statistical. Furthermore, local leadership in the area of measurement on an administrative and supervisory level is either ineffective or lacking. Through the off-campus workshops, the measurement man has a rich opportunity to help local administration face the problems of adaptive training in measurement and evaluation. It is in this area that the critical requirements in measurement and appraisal are such as to make the difference between what is an outstandingly effective or a definitely unsatisfactory teaching performance.

The Teaching of Educational Measurement

HAROLD GULLIKSEN

TRAINING FOR RESEARCH IN PSYCHOLOGICAL MEASUREMENT

BEFORE OUTLINING the type of *training* which I would consider appropriate for research in psychological measurement, let us consider briefly the nature of the problems of psychological measurement, in order to indicate the type of goal toward which this training is directed.

If one surveys the different problems which have engaged the attention of psychologists, one finds that there is a certain class of problems which are met with repeatedly and are characteristically psychological. The major features of these problems are that the psychological objects being studied are basically qualitative in nature. However, it seems reasonable and valuable to describe them quantitatively.

For example, intelligence is an important psychological trait of individuals. It is basically a qualitative trait, yet it seems reasonable both to laymen and to psychologists to speak of one person as being *more* or *less* intelligent than another, implying quantitative differences in intelligence. The measurement of intelligence and other mental traits, and the determination of their interrelationships, is a large field of psychological investigation. It is important both for psychological theory, and for practical applications of psychology.

Correspondingly, *sensations* are basically qualitative. The differences between different odors; between different tones, or between different colors are clearly qualitative differences, yet we have a two dimensional representation of tones, a three-dimensional color pyramid, and attempts to represent odors in three-dimensional space.

Attitudes of persons toward various social problems, or individual judgments of value, are additional illustrations of psychological objects which are basically qualitative. Yet the study of problems of attitude and value is aided tremendously, when attitudes are quantified and measured by various attitude scales. It is necessary to know how to measure values before laws regarding value can be experimentally investigated.

There are a great many more illustrations of the importance of quantification in the approach to numerous psychological problems that initially appear to involve a field that is essentially qualitative. These problems are the basic problems of psychometrics. One method for recognizing such problems in common-sense conversation is to note that a person discussing the problem is likely to say, "It is all a matter of opinion and you can't even get agreement among the opinions of different persons." This statement could readily be made about attitudes, values, sensory qualities, etc. Many of the methods

of Psychological Measurement were developed precisely to deal with this situation, where the scientist must deal with opinions or judgments and these opinions disagree.

A slightly different way of viewing these problems is illustrated in Table 1. We may say that many if not most psychological problems are concerned with some aspects of the responses of per-

TABLE 1.

		STIMULI			
		S_1	S_2	... S_g ...	S_K
PERSONS	P_1	R_{11}	R_{12}	... R_{1g} ...	R_{1K}
	P_2	R_{21}	R_{22}	... R_{2g} ...	R_{2K}
	.	.	.	. .	.
	.	.	.		.
	.	.	.	.	.
	P_i	R_{i1}	R_{i2}	... R_{ig} ...	R_{iK}
	.	.	.	.	.
	.	.	.	.	.
	.	.	.	.	.
	P_N	R_{N1}	R_{N2}	... R_{Ng} ...	R_{NK}

sons to stimuli. Each person may be designated P_i (where $i = 1 \ldots N$). Each stimulus may be designated S_g (where $g = 1 \ldots K$). Such a set of psychological data for the responses of N persons to K stimuli can in all, or practically all cases be represented as a matrix, with (say) a column for each stimulus, and a row for each person. The response of the i-th person to the g-th stimulus would then be designated R_{ig}. Table 1 illustrates such a matrix, which can represent a large number of different sets of psychological data. For example, in a psychophysical experiment, liminal stimuli may be used, and the R_{ig} is a 1 or a 0 depending on whether the person responded, or did not respond to the stimulus. Instead of being a simple yes, or no response, R_{ig} in the matrix representing another experiment might be the reaction time in thousandths of a second of each person to each stimulus. Mental test data also fit the same general format, since the S_g may represent either an item, or a test. The R_{ig} then represent the person's correct or incorrect answer to each item, or represents his score on each of the tests. The fact that the basic data of test theory, and of psychophysics can be represented as matrices, indicates that thinking precisely about these problems would probably be facilitated by matrix algebra. Many of the developments in both psychophysics and test theory over the past ten years or so demonstrate the usefulness of matrix algebra in developing the solution to various problems.

If one regards the stimuli, the tests or the items for example, as standard and attempts to measure inter-individual differences we have the problems of mental tests. If one assumes that the individuals constitute a somewhat standard group and is interested in evaluating the *stimuli*, i.e. the test items, we have the problems of item analysis. If the S_g represent tests, then the R_{ig} are test scores, and we might wish to determine the number of intellective factors represented in the set of tests. In this case we have the problems of factor analysis. As is well known, matrix theory has been useful here in determining the minimum number of factors needed to account for a given set of tests.

If the stimuli are tones, or colors, or lights, the persons are regarded as constituting a representative normal group, and the scientist is interested in scaling the *stimuli*, we have the problems of psychophysics. In the method of Paired Comparisons, the S's would be "pairs of stimuli," the persons would judge "which is the larger of the two." In this case a matrix of 1's and 0's is the basic matrix of experimental data for the method of paired comparisons, and the law of comparative judgment. In studying attitude statements, or sensation for

example, if one uses the method of triads for determining interpoint distances, then another type of matrix representation as used in multidimensional psychophysics gives the solution for the dimensionality of the set of objects used in the experiment.

In summary of the foregoing, we may say there are a large number of important psychological problems which involve the quantification of qualitative material. These problems in which "it is all a matter of judgment and judgments disagree" are the problems which have been classified as psychometrics or psychological measurement. Having indicated the nature of this field let us consider the training needed for such work.

Adequate training for research in psychological measurement can be grouped under five major headings.

A. Basic training in Psychology.

This training should include work in experimental psychology, social, theoretical, comparative and other areas of psychology. The purpose of this training is to give the person a good knowledge both of the experimental procedures used in psychology, and of the present status of various psychological problems.

Depending upon the division of labor among departments in a given university, and the interests of the student, such training might be obtained in a Department of Psychology, a Department of Education, a Department of Social Relations, or in some other department.

B. Basic training in mathematics.

The problem of precisely what mathematics is most valuable cannot be answered with finality. Probably different people should study different areas of mathematics in order to see which becomes more useful as various psychological problems are explored by psychometric methods. It is also possible that a re-arrangement of the contents of some mathematics courses would make it possible to obtain the necessary training in a shorter time than the current arrangement does. Prior to such a reorganization of the mathematics curriculum, which may take several years (or even decades), it would seem that the student of psychological measurement should have the following courses now in the college mathematics program.

First year—college algebra, analytical geometry, trigonometry
Second year—differential calculus, integral calculus
Third year (one term)—differential equations.
Fourth year (one term)—matrix theory

Probably a person with a foundation in mathematics similar to the one indicated above would be adequately prepared to explore various other specialized mathematical topics which seemed to be of special interest. It might be noted that such a set of courses would probably not ordinarily be regarded as a major in mathematics. It would perhaps constitute a strong undergraduate minor in mathematics.

C. Basic training in modern statistics.

It should be recognized that the contributions of statistics to psychometric problems are essentially different from those of mathematics. One needs the mathematics in order to *formulate* the scientific laws. Statistics is concerned with determining if the data and the equations are in agreement or in disagreement. Probably the student should have two years in statistics, including such topics as correlation theory, multiple and partial correlation, analysis of variance and co-variance, various significance tests, confidence limits and estimation procedures, probability theory including stochastic processes, and certain topics in multivariate statistics such as discriminant analysis and canonical correlation theory. These

techniques should be known at the operational level so that the person is competent to apply them, and also is qualified to select the technique appropriate to a given problem. Insofar as possible the theory on which the statistical procedures are based, and by means of which they are developed would also be known since part of the psychometricians problem would be to develop, or at least to initiate the development of new statistical techniques appropriate to current problems.

D. A set of courses in procedures of psychological measurement should be a central part of the curriculum fo psychometric training. There is also the further problem of differentiating the artifacts of the measuring techniques from the scientific findings. This problem arises in many other areas, but perhaps not in the same degree as in psychological measurement.

For example, work on the physiology of the cortex has involved the use of operative, fixing, sectioning, and staining techniques. These constitute an area of study in themselves. Those not thoroughly versed in such techniques will frequently confuse artifacts of the technique with scientific findings.

Correspondingly, studies of cortical potentials, or brain waves, involve complex electronic equipment. The problems of properly designing and handling amplifiers, oscilloscopes and other equipment frequently seem to eclipse the functioning of the cortex which the instruments are supposed to record. However, it is essential to know when one is recording cortical functioning, and when one is recording various artifacts produced by the equipment.

A similar situation obtains when the research involves the techniques of psychological measurement. Almost anyone feels that he can frame a poll question, administer it and draw conclusions; can construct a test of a few items, or devise an attitude scale and draw valid conclusions without exercising other than a few common-sense precautions. There is perhaps a great deal of misleading work in these areas due to the fact that the properties of the measuring instrument are misinterpreted as properties of the phenomena supposedly being measured.

During the past fifty years an elaborate body of theory has been developed to deal adequately with psychological measuring instruments. Probably a minimum of three one-year graduate level courses would be necessary to cover such material.

1. A course in the theory of tests, including work on reliability, validity, error of measurement, adjustment of time limits, problems of group heterogeneity, of scoring, scaling, and weighting methods, procedures for improving tests by item analysis and item selection, the best methods of combining tests into a battery, etc.
2. A second course would deal with methods of factor analysis. Matrix theory would be a prerequisite for such a course, or included in the early part of the course. The course would deal with the theory of factor analysis, methods for manipulating large matrices of test scores, and also with the various uses of factor analysis in fields other than tests.—Fields such as abnormal psychology, sensory psychology, Perception, Learning, etc. Probably it would also include some introduction to modern high speed electronic computing procedures.
3. A third course would deal with the psychological scaling procedures, including the logical foundations of such measurement, the experimental and logical distinctions between ordinal, extensive and other scales, the distinction

between fundamental, derived and defined magnitures, as well as various modern developments in psychological scaling. These latter would include, for example, the method of paired comparisons and law of comparative judgment, the method of successive intervals and the law of categorical judgment, the method of similar reactions, scale analysis, latent structure analysis, and the multidimensional procedures in psychophysics.

Such a set of three one-year courses, if compactly and carefully organized, and if the students enter them with a good mathematical preparation could cover the basic material now available in psychological measurement and could give the students a foundation so that by means of independent work he could move ahead, and keep up with new developments in the field, as well as make contributions of his own to the area.

E. The Development and the Function of Mathematical Models.

Since in the present state of psychology the development of mathematical models for various areas is a somewhat new and undeveloped idea, it is desirable at present to have a course dealing explicitly with this area. Eventually a mathematical treatment will probably be a routine part of most psychology courses, however for the next five to 20 years, such material will probably be covered in a separate course, if covered at all.

This course on mathematical models should include a survey of models already developed in various psychological areas, and practice in the development of new models. It should begin with some consideration of well-established mathematical models in other areas in order that the student could become acquainted in detail with the nature of such models and the functions which they serve in science. For example, the derivation of the orbit of a planet, or the trajectory of a projectile from Newton's laws of motion might constitute appropriate introductory material for such a course. One might also utilize some illustrative derivations and theorems from field theories in electricity and magnetism, and possibly some mathematical models in the field of biology from mathematical biophysics. Such an introduction from mathematical models in more established fields would serve to illustrate various important general points, such as that the basic postulates may be unverifiable directly, sometimes are merely definitions of terms, and may occasionally be contradicted by ordinary experience. It might also be desirable to include some attempts to verify certain theorems experimentally in the laboratory, in order to demonstrate how easy it is to set up a poorly designed, or a poorly controlled experiment and thus "disprove" some of the basic theorems of physics. Doubtless many elementary physics students have "demonstrated" that falling bodies do not obey the law of gravitation, for example.

The fact that experimental conditions must agree approximately with the assumptions of the mathematical model is another important consideration frequently overlooked in criticisms of present attempts at mathematical models in psychology and in social sciences.

The introductory material might also include a brief review of differentiation and integration, non-linear curve fitting, Pearson's method of False Position, and graphic methods such as rectification and translation of a master curve.

The body of the course, however, would deal with various mathematical models within the field of psychology. These might include illustrations from the psychology of learning, as de-

veloped by Thurstone, Hull, Bush, Mosteller, Estes and Burke. Others would be from the field of social behavior as illustrated by some of the work of Nicholas Rashevsky, J. Q. Stewart, and G. K. Zipf. A treatment of epidemic theory as developed by Lowell Reed and others might also be interesting.

The minimum necessary work in this area could probably be covered by a one-year course provided there were two prerequisites for this course,

1. Differential equations
2. Some knowledge of experimental psychology, such as a course in the psychology of learning.

During the latter half or third of this course, students should be given an opportunity to develop new models which might be tested experimentally.

F. Miscellaneous Special Courses.

It might be that one would want to consider different special courses which would be particularly useful for students with special types of interest in Psychological Measurement. For example, time series, auto-correlation, or spectral analysis might be topics which one would wish to have available and to encourage some students to take, but which would not be particularly suitable to require from all students in psychological measurement.

With such training in:

A. Basic psychology,
B. Mathematics
C. Statistics,
D. Psychological Measurement Theory,
E. The Development of Mathematical Models, and
F. Other Special Courses,

it seems to me one would have students who would understand current Psychological problems, be equipped to utilize measurement techniques in various areas where such methods were relevant.

All of this material could not be covered as a three-year program of graduate training. However, if the students began such work as an undergraduate, this program allows a great deal of freedom for other electives and even for another major interest. In terms of one-year courses at either graduate or undergraduate level the program may be summarized as follows:

	one-year course or equivalent
Mathematics	3
Statistics	2
Psychology (say)	6
Psychological Measurement	3
Mathematical Models	1
Special	1
Total in Program	16

Assuming four years of undergraduate and three years of graduate work, (the last of which is spent on a theses) gives six years of study. Assuming a load of four courses which is usually not considered heavy gives a total of twenty-four one-year courses in the six years. This program specifies two-thirds, leaving one-third of the students time free for other sorts of work. There would thus be considerable time for study in related fields of special interest to each student.

Another way of looking at the total load is to say that a student taking a Ph.D. in Psychology at Princeton and specializing in Psychometrics (provided he had a strong undergraduate *minor* in mathematics and the equivalent of an elective course or two in Psychology and Statistics) would have essentially the program outlined here.

In order to obtain well-trained persons in Psychological Measurement, it is necessary that superior high-school students and college freshmen be informed regarding the nature of the field and the opportunities it offers so that

they may direct their undergraduate work in this direction if the field is of possible interest to them.

It is important to emphasize that such undergraduate training in mathematics and statistics does not constitute *specialization*, but instead constitutes a *good foundation* training for entrance into a number of different fields, including both physical sciences and social sciences. It is highly desirable that this fact be brought to the attention of superior high-school students and college freshmen.

The Teaching of Educational Measurement

JOHN T. COWLES

SUMMARY OF DISCUSSION

Dr. Cureton, Chairman of the discussion, opened with two comments; one concerned the apparent limited population mobility of people trained in educational measurement—people trained in the midwest, for example, apparently do not migrate to New England. The other comment concerned a rather significant occurrence at his institution: A college freshman who was doing poorly in two of his courses was asked by the counselor about his difficulty. It appeared that he was bothered only by those courses in which essay exams were given. He had never had *that* kind of exam before reaching college!

Dr. Davison emphasized the value of first analyzing the methods of teaching a subject before trying to test it; this is a fruitful approach to the problem of relating test content to future objectives. Dr. Noll agreed that the gap between the formulation of objectives and the test item itself is a major problem.

Major Carlson, commenting on Dr. Kvaraceus' presentation, noted that people conducting courses are often averse to measurement, and suggested that future instructors should be taught what they can get from measurement. He explained a technique used at the Air University: after students have taken a test, they are sometimes placed in the role of instructor and asked how they could improve the test, or what use, as an instructor, they could now make of it. Dr. Kvaraceus pointed out that a formal course in measurement might not answer the need; he would prefer a "functional analysis of failures" and continuous working out of appropriate evaluation devices with teachers, rather than relying on short-term consultants on evaluation methods after a curriculum is designed. In response to the question whether it is desirable to have students construct their own examinations, Dr. Noll felt that in general students would not make up good enough tests for evaluative instruments but that the experience of making up tests could serve as an instructional aid.

Dr. Schweiker commented on the problem of conflicting attitudes and experiences among faculty on testing, which seriously hamper efforts to introduce better evaluation. He cited the principle stated by a humanities faculty group that "anything that can be measured is not worth measuring." Dr. Schweiker suggested that an analysis of the examinations of such a group would be most enlightening.

The Interview as an Evaluation Technique

E. LOWELL KELLY

AN EVALUATION OF THE INTERVIEW AS A SELECTIVE TECHNIQUE

I. Introduction

In the broadest sense, an interview is nothing more than a conversation between two individuals, directed by one of them, toward a specific end or purpose. The particular purpose served by this directed conversation is a function of the situation in which it is being used and the interviewer's conception of the possibilities of the interview as an appropriate procedure in the situation. For example, an interview may be directed primarily at eliciting information from the interviewee. The information may be desired as a basis for making a decision regarding the interviewee or it may be collected and pooled with parallel information elicited from other interviewees to provide tables of normative data for a large group. In other interviews, the purpose may be primarily therapeutic or educational, i.e., carried out for the purpose of making changes in the interviewee.

Regardless of the purpose of the interview, it may also vary on another important dimension, the degree to which it is structured. At the one extreme, it may be almost completely structured, in which case the interviewer's task is essentially that of orally administering a questionnaire and recording the responses given. At the other end of the continuum, the interview may be almost without structure—the interviewer's task is simply that of encouraging the interviewee to talk.

There is no doubt but that the interview may be used effectively to elicit certain kinds of information from human subjects. There is also considerable evidence to suggest that the interview may be used as a basis for reasonably valid appraisals of certain personality variables. As a psychotherapeutic technique, the interview has no rival, although parenthetically it should be noted that evidence for its validity in this domain is extremely scant.

II. The Selection Interview

I have been asked to discuss the interview as a selective technique. Even with this limitation, however, the topic is a broad one and also complex.

Without doubt, the interview is the oldest and the most widely used of all selection techniques. Crissy in a recent paper (2) notes that it continues to be the most popular personnel selection method in private industry. Swenson and Lindgren (11), in a survey of Minnesota industries, found it to rank first among personnel selection procedures. And as a reminder that the use of the selection interview is not limited to industrial settings, Stalnaker and Eindhoven (10), in a survey of medical school requirements, report that applicants are required to report for a selection interview in 53 of the 80 medical schools. The 27 remaining ones only urge the applicant to appear for a selection interview!

In these widely varied settings, the persons responsible for the selection interviews vary widely in ability, in personality and in type and amounts of training for and experience in interviewing. They also vary greatly with respect to the particular type of interview techniques employed. Wherein may we seek to find the communality among various users and proponents of the selective interview? I suggest two areas of communality: (a) a common function or definition of the task and (b) a shared confidence or belief in the validity of the technique as used in the local situation.

Common Function: In describing what seems to be the common function of the selection interview we shall also define it for the purpose of this discussion: Regardless of the setting, the characteristics of the interview or of the techniques used, the interview as a selection technique involves:

1. A situation in which a limited number of persons from a larger number of candidates are to be selected for available appointments (job, position, scholarship, or group membership).
2. An assumption that individual differences among the candidates are correlated with successful performance in the role to which the candidates aspire.
3. A conversation between an interviewer and candidates for the appointment.
4. An assessment of each candidate by an interviewer leading to a prediction (explicit or implicit) regarding the probable relative success of the candidate in the performance situation. This prediction may be implicit and expressed in categorical form, e.g., "accept" or "reject" or it may involve a more explicit rating of probable success or failure either in adjectival terms or on a rating scale. It is not essential here to distinguish between those instances in which the prediction (implicit or explicit) is made by the same person who does the interviewing or by a second person who bases his prediction on information or a personality appraisal growing out of the interview. In either instance, the use of the interview as a selection technique involves the prediction of future behavior of the candidates with respect to a selected criterion.

Confidence in the Validity of the Selection Interview:

The most eloquent evidence of the widely shared belief in the validity of the interview as a selection technique is its continued widespread use; in most situations, the confidence in the technique is so high that neither the interviewing staff nor administrators of the organization even consider the desirability of determining the actual validity of the technique in the local situation. Furthermore lack of evidence to support such confidence in the validity of the selection interview does not keep intelligent people from believing in or even testifying to such beliefs. Thus we find the authors of *Assessment of Men* (7) making the bold statement that "the interview is probably the best and only indispensable method of assessment." In a similar vein, Alec Rodger (8) in a recent article entitled "The Worthwhileness of the Interview" states "the interview is the standard means whereby people are judged for many purposes, and is likely to remain so, maybe till the end of time."

Further evidence of the high esteem in which the selection interview is held by psychologically trained persons derives from the Michigan Assessment Project (3) on the selection of clinical psychologists. In this project, we used

two interviews as parts of a week-long series of assessment procedures which included also an extensive battery of objective tests, a battery of projective tests, several situation tests, etc. Near the end of our major assessment program, each staff member was asked to rank order the procedures used in terms of their usefulness to him in arriving at judgments concerning the future performance of the candidates. Practically all of the staff of 25 persons ranked the interview either in first or second place.

III. Evidence Regarding the Validity of the Selection Interview

The selection interview may be a valid technique of personnel selection in some situations. However, as I think most members of this audience know, the accumulated evidence is such as to throw the burden of proof on the proponents of the method. Nearly 35 years ago Scott (9) and Hollingworth (3) in independent pioneer investigations reported surprisingly low interjudge reliabilities and validities of interview judgments regarding sales ability of prospective salesmen. Since that time, many comparable studies have been conducted and most of the research findings point to similarly low reliability and validity of interview judgments. Occasionally, if one searches long and hard, he can find a study with somewhat more promising results. One such is Vernon and Parry's (13) report of a war-time research on the selection of trainees for commissioned rank in the Royal Navy. Predictions were based first on a series of paper and pencil cognitive tests; second, on an interview carried out by a conventional officer selection board which had access to the test results and also to reports by commanding officers on the candidates; and third, an interview by one of three psychologists who also had access to the test results but not to the reports. In this study it was found that the cognitive tests alone were better predictors of the criterion than the judgments of the officer selection board, but the psychologists did somewhat better than either. Similarly Bobbitt and Newmann (1) report promising validities by interviewers in the prediction of success of candidates in the officer training program of the U. S. Coast Guard Academy. In this study the combined judgment of two interviewers correlated with the pass/fail criterion .49. This is most encouraging until one notes that the test scores alone, which were known to the interviewers, correlated .47 with the same criterion. In this particular study a statistical combination of interview judgments and test scores yielded a validity of .56, but whether these findings could be replicated with other interviewers is not known. Likewise, Hunt, Wittson and Hunt (5) report low useful validities for even a brief psychiatric screening interview when judged against criteria of later adjustment in Navy life.

Such results, however, are exceptional. Even such an ardent proponent of the selection interview as Rodger (8) admits the "general tendency of experimental findings has been to show that even where the interviewing has been done by psychologists, interview judgments can be very variable and very wide of the mark." Furthermore, individual differences among interviewers with similar training appear to be sufficiently great to suggest the need for some procedure of selecting interviewers to conduct selection interviews!

Let us look briefly at some of the more recent findings concerning the validity of the interview as a selection technique. In our own work on the selection of clinical psychologists (6), we used two interviews. The first was one hour long, conducted by a staff member who had previously made

judgments of the candidate on the basis of his credential file only. The second interview was two hours long conducted by a different staff member and carried out only after the interviewer had previously made an intensive study of the candidate's credential file, his scores on an extensive battery of objective and projective tests, a biographical information inventory and a long autobiography. These interviews were carried out by trained professional persons who were permitted to structure the interviews in the manner which they believed most useful for the task at hand. The validity of judgments made before and after each type of interview was estimated against a dozen different criteria obtained four years later. The results were such as to force us to conclude that neither of these interviews made an essential contribution to our assessment program. Actually, the median validity of the judgments made after each type of interview was only .01 greater than the median validity of judgments made before the interview.

At least the validity of these interview judgments was not negative, as was true for those reported in a recent study by Thayer (12). Thayer, in a doctoral dissertation at the University of Pittsburgh, attempted to predict the subsequent field success of missionaries to whom a battery of psychological tests had been administered some 20 years earlier. In this study, an admittedly fallible criterion was predictable with a correlation of .53 with a battery of three psychological tests but the ratings made by the Secretary of the Missionary Selection Board, presumably based on interviews, references and other papers, showed a negative correlation with the criterion measure.

Perhaps the most cogent evidence for doubting the validity of the conventional selection interview appears in a recent article by Holt and Luborsky (4), which reports on the Menninger Foundation research project on the selection of psychiatrists. In contrast to the Michigan assessment program, the Menninger project relied most heavily on judgments based on selection interviews and a battery of individually administered psychological tests. In this project, each applicant was independently interviewed by three psychiatrists, each of whom made a global prediction regarding the candidate's probable success in psychiatric training at the Menninger School of Psychiatry. In many ways, this study would seem to have been almost ideally designed to yield a maximum estimate of the validity of selection interviews. First, it should be noted that the interviews were conducted by staff psychiatrists, presumably expert in the art of interviewing; secondly, the predictions were made with respect to criterion performance in the local situation about which the interviewers were maximally informed; finally, although Holt and Luborsky do not provide evidence on this point, many of the interviewers, in their role as staff members of the Menninger School of Psychiatry, must have contributed to the later criterion evaluations of the candidates. In spite of these seemingly optimal conditions, the validities of these interviewer judgments were shockingly low: the median being .06 for the 14 interviewers. Furthermore, only one of these 14 presumably expert interviewers had a validity large enough to achieve statistical significance.

Such evidence does not prove that the interview is valueless as a technique of personnel selection. There may be some situations in which some interviewers are able to use the technique and arrive at judgments with high predictive validity. If so, it is most unfortunate that they have not been reported in the literature. On the contrary, all evidence available suggests

that the technique is apt to have sufficiently low validity even under optimal conditions to make doubtful its general utility as a selection device.

IV. Paradox

We are thus forced to conclude that the most widely and confidently used technique of personnel selection is one for which there is surprisingly little evidence of predictive validity. This curious situation appears to have its parallels in the clinical field in the current popularity of projective techniques largely unvalidated for predictive purposes and in the field of education with the continued widespread use of essay examinations. In all three instances, the choice of the technique is obviously based on factors other than evidence of predictive validity. I do not pretend to know what all of these other factors are, but I have a hunch as to what is going on. Note that in each of the three cases, the technique is chosen and used by professional persons confronted with the necessity of making decisions about people—decisions which are significant to the organization of which the professional person is a part and/or to the persons about whom the decisions are made. One cannot take lightly such responsibilities as deciding whether it is A or B that is hired for a particular job, C or D is sent to a mental hospital, or E or F that gets into medical school. Ideally, such decisions should be made on the basis of tested techniques with high predictive validity. Unfortunately, as we all know, such techniques simply do not exist. The best of our tools lead to but fallible predictions of later criterion behavior; they enable us to guess right much more often than not but there are still many errors of prediction. Furthermore, in the domain of psychometric tests, we have reasonably precise estimates of the accuracy of our prediction and of the magnitude of our errors. And in the impersonal situation of a selective program based on psychometric procedures, we can tolerate the truth of our fallibility—knowing that our errors have at least been reduced.

Now many persons do not appear to be able to accept the inevitability of errors of decision and prediction inherent in even the best of our tools for evaluating and predicting human behavior. Such persons idealistically search for an instrument with more sensitivity to the subtle nuances of human behavior and personality, and noting evidences of such sensitivity in the writings of poets, dramatists, and philosophers, conclude that the best instrument for the task is another human being, perhaps themselves! They next proceed to try out this newly discovered instrument by interviewing a few people, by interpreting their handwriting or reading their essays. Lo and behold, the instrument appears to work and work well. Since the instrument, i.e., the human being and the technique works so well, it should obviously be used for some practical purpose. Let's try it out in some practical selection situation. And so the interviewer, the projectivist or the proponent of the essay test goes to work. In each case, the technique is employed by a person who is already pretty well convinced of the validity of his tool and thus reasonably confident of the correctness of his decision regarding individual cases. Under the circumstances, it is not surprising that the user of the technique rarely finds occasion to submit himself and the technique to a true validity check. Instead, as the result of each decision made (a decision which just dares not be wrong!) he becomes even more convinced of the validity of his technique and himself. And if someone else insists on investigating the validity of the technique, he finds many good reasons why the results of the study are not to be taken seriously; for example, the criteria used were not appropriate

or the users of the technique were not really competent!

One having committed oneself to the position that the human being is the most essential part of the assessment process, it follows naturally that the choice of specific techniques to be used will be one which enhances the role of the human being who uses it; it should be one which provides for maximal flexibility, one which requires the extensive use of good judgment. Furthermore, it should be one which provides a maximum of information to be integrated by the human mind since from a common sense point of view it would seem that the best decisions are those made on the basis of the most information. The interview, especially the unstructured interview rates high in each of these respects and hence, I suggest, serves admirably to reduce the threat of anxiety which would otherwise be present in persons who accept the responsibility for making important decisions regarding the lives of others. If it serves this function as well as I have suggested, it is hardly surprising that the interview is so confidently believed in by its proponents—evidence or no evidence.

V. Another Look at the Problem

The more I think about the selection interview, the surer I become that it is a task which no human being can be expected to carry out. Let us look at the task or rather the series of tasks involved in the making of valid predictions of criterion performance. What would the interviewer need to know and to do?

1. He would need a thorough knowledge of the performances demanded in the criterion situation. If a previous job analysis does not provide such information, he should really make one. And let us not forget that, although the criterion may be an overall or global one, a more careful analysis is likely to show it to be a weighted combination of uncorrelated dimensions.
2. He would need to know the relevant variables, i.e., what abilities, characteristics and traits of people are related to performance on the criterion. This would include a knowledge of which variables have no correlation with the criterion.
3. He would need to know how to select and to elicit behavior in the interview which are valid indicators of the relevant predictor variables.
4. He would need to know how to evaluate or weight each of the behaviors elicited in the interview in order to arrive at the most valid score or rating or the relevant variables.
5. After having arrived at this point, the interviewer would next have to weight or combine each of the variables evaluated during the course of the interview or in order to arrive at an overall predictive "score" for the candidate. Incidentally, it should be noted that the preparation of this implicit regression equation would necessitate a knowledge of the intercorrelation of all the predictor variables evaluated as well as the correlation of each with the criterion.

Looked at in this way, it will be seen that the selection interviewer is expected to combine in one person the role of the job analyst, the test constructor, the test administrator, the test scorer, and the statistician. Furthermore, successful performance would demand that he be at least minimally competent in playing each of these several roles. The end product, his judgment regarding the probable performance of a candidate in the criterion

situation, would be vitiated as the result of poor performance in any of the several roles. For instance, even though a person should be highly competent at appraising personality from the interview, he might fail miserably in an effort to predict success in the criterion situation by virtue of a lack of knowledge concerning the relevance of each of the personality variables to the criterion performance.

It seems unlikely that selection interviewers ever think of their task in this way. In fact, were they to do so, like the thousand-legged worm, they would probably be so traumatized by the complexity of their task as to be unable to function at all. Furthermore, it must be remembered that most selection interviewing is carried out by people relatively unsophisticated with respect to the psychometric steps enumerated above. Even those psychologists most given to using the selection interview are likely to be those less sophisticated in psychometric procedures. And if confronted with the true complexity of the task, such persons are likely to take refuge in the belief that the job is really not as we have analyzed it above but rather one requiring 'the artistic or intuitive appraisal of the total individual by processes not completely communicable and certainly not statistically manipulable.' Furthermore, such persons are likely to argue that the resulting global judgment has more validity than can possibly be arrived at by the use of objective psychometric devices and statistical equations. Certainly this is a tenable hypothesis and a testable one. Unfortunately, as of the present date, I know of no evidence to support it.

Since the task of the selection interviewer, when conceptualized in psychometric form, is clearly an impossible one, it is only natural for the interviewer to proceed on what seems to him a most common sense basis, namely secure as much information concerning the candidate as possible, and then weight these items of information intuitively on the basis of whatever theory or biases he has picked up regarding the functioning of human beings in the situation for which he is making the selection. The absence of tested knowledge concerning the actual relevance of bits of information or specific variables to performance in the criterion situation does not bother him too much. He merely substitutes his subjective impressions of the appropriate weights. Furthermore, this intuitive weighting of evidence goes on without reference to the interrelationships among predictor variables. Each item of information is weighted as it is elicited with the result that even relevant variables may be overemphasized. In general, I fear that the whole situation is one which encourages the interviewer to use each bit of additional information in a manner which introduces as much error variance as true variance in his judgments.

VI. The Future of the Selection Interview

In closing, I am going to "stick my neck out" and make a number of predictions:

First, I predict, with a very high level of confidence, that the selection interview will continue to be a widely used and highly respected technique. No amount of negative evidence regarding its validity seems likely to change the situation. Second, I predict that its popularity will decrease only when and to the degree that more valid techniques and devices are developed to do the practical jobs of selection in our complex society. These must be done by somebody and in some way.

Thirdly, I predict that improvement in the selection interview itself will come about only by fractionating the total task, i.e., dividing it into a se-

quence of tasks, each of which can be reasonably carried out by an intelligent and trained human being. Thus, one person might be assigned the task of eliciting information, another that of assessing key personality variables, another that of weighting the assessed variables in an empirically determined manner to arrive at the prediction of a criterion. In all honesty, I doubt that any amount of training of the kind currently described in texts on personnel interviewing will enable a single individual to do the whole series of things now expected of the selection interviewer. Since we give him an impossible task, perhaps we can forgive him for developing unjustified confidence in the work which he does. Furthermore, since if our analysis is right, he must develop the confidence in order to reduce his anxiety, we can hardly expect him to do the research needed to improve the technique. Such research must be initiated and supported by persons who believe the problem sufficiently important to justify considerable expenditure of research effort in order to change a currently doubtful practice. The research itself must be done by persons trained in the skills demanded by the complexity of the task but who are also capable of operating in a practical world in which selective techniques are now evaluated largely in terms of their face and faith validity.

REFERENCES

1. Bobbitt, J. M., and Newman, S. H., Psychological activities at the U. S. Coast Guard Academy. *Psych. Bull.*, 1944, *41*, 568-79.
2. Crissy, W. J. E., The employment interview, research areas, methods and results. *Personnel Psychol.*, 1952, *5*, 73-86.
3. Hollingworth, H. L., *Vocational Psychology and Character Analysis.* New York: Appleton, 1929.
4. Holt, R. R. and Luborsky, L., Research in the selection of psychiatrists: a second interim report. *Bull. of the Menninger Clinic*, 1952, *16*, 125-35.
5. Hunt, W. A., Wittson, C. L. and Hunt, Edna B. The relationship between definiteness of psychiatric diagnosis and severity of disability. *J. Clin. Psychol.*, 1952, *8*, 314-15.
6. Kelly, E. L., and Fiske, D. W., *The Prediction of Performance in Clinical Psychology.* Ann Arbor, Mich.: Univer. of Michigan Press, 311 pp. 1951.
7. O. S. S. Staff, *Assessment of Men*, New York: Rinehart and Co., 541 pp. 1948.
8. Rodger, A. The worthwhileness of the interview. *Occupational Psychol.*, 1952, *26*, 101-08.
9. Scott, W. D. Selection of employees by means of quantitative determinations. *Annals of Amer. Academy of Political and Social Science*, *65*, 1916.
10. Stalnaker, J. N., and Eindhoven, J., in *Admission Requirements of American Medical Colleges.* Chicago: Association of American Medical Colleges, 106 pp., 1953.
11. Swenson, W. M., and Lindgren, E. The use of psychological tests in industry. *Personnel Psychol.*, 1952, *5*, 19-24.
12. Thayer, C. R., The relationship of certain psychological test scores to subsequent ratings of missionary field success. *Univer. of Pittsburgh Bull.*, 1952, *48*.
13. Vernon, P. E., and Parry, J. B., *Personnel Selection in the British Forces.* London: Univer. of London Press, Ltd., 324 pp., 1949.

The Interview as an Evaluation Technique

W. J. E. CRISSY

INTER-PERSONAL ASPECTS OF THE INTERVIEW PROCEDURAL TECHNIQUES AND RESEARCH PRACTICES

SOMEONE DEFINED an interview as a conversation with a purpose. The primary purpose of the interview used as an evaluative technique is to enable one or more interviewers to evaluate the interviewee—e.g. in industry, with respect to his fitness for a given vacancy. Now, in order to bound the scope of the assessment, one or both of two frames of reference are usually furnished the interviewers.

1. A description of the tasks expected to be performed by the interviewee.
2. A description of the qualities or traits presumed to characterize an incumbent.

Interviewing as a personnel assessment technique varies not only from interviewer to interviewer, but within the same interviewer over a period of time. If it is to be used, then several problems of both inter-personal and intra-personal consistency must be dealt with. Most of these, fortunately, admit of identification through research and of at least partial solution through training.

In any but the smallest organizations, several persons share the responsibility for selecting new members, thus raising the problem of inter-personal consistency. I should say, parenthetically, this is as it should be. On many counts, I think the selection of a single applicant or candidate should be based, as a matter of policy, upon consensus judgment rather than upon the decision of one person in the organization. This, however, raises questions: Are the several interviewers using a common frame of reference? To what extent do they agree among themselves? How much inter-interviewer agreement is really desired? It is my purpose now to discuss some of the ramifications of these problems and to outline briefly some ways of focussing interviewer training to promote improvements in this aspect of interviewing.

We would, I think, hypothesize that the more structured the interview, the more inter-interviewer agreement. This is in part due to the closer prescription of what may be asked—in the extreme case, the actual wording of questions and sequence of questions are set. In part it is attributable to the detailed nature of the rating structure within which the interviewer is asked to quantify his judgments. If inter-interviewer consistency were our only concern the way to accomplish it would seem to be, in the main, by structuring the interview as much as possible. But here's the dilemma, if our interviewers' quantifications of judgments correlate highly it can mean: (1) that we can accomplish about the same results using fewer interviewers (if multiple interviewing is being practiced—either board or consecutive type): (2) that individual interviewers are not making unique contributions of judgment. If

our interviewers' quantifications of judgments correlate low it can mean: (1) that they share no common frame of reference for their judgments; (2) that their individual judgments as well as their concensus judgments are unreliable; (3) that each is making an individual contribution—verifiable by validity procedures noted below. If the interviewers are asked to cite evidence substantiating their judgment quantifications a content analysis of the evidence, interviewer by interviewer, will shed light on the tenability of the conclusion of no commonframe of reference for their judgments. If the content analysis is pursued over a period of time, we can determine inferentially how reliable the judgments are. If little overlap of evidence content is found from interviewer to interviewer, training should include a discussion of what the prescribed traits mean and what comprises varying amounts of each of them.

Turning back for a moment to the case of high inter-interviewer correlation I suspect, on the basis of scant research data described by Sternberg, that individual differences do exist among any group of interviewers on sharpness of discriminability of judgment, trait by trait, as well as over-all. It seems to me we lose the opportunity to make potentially valid use of this if we structure the interview so much that these individual contributions of judgment cannot be made manifest.

As a pragmatic solution to this dilemma, I have espoused in my industrial work semi-structured interviewing. By this I mean arming each interviewer with as precise a job description as can be obtained, furnishing him a carefully prepared set of manpower specifications couched in behavioral terms; requiring a quantification of judgment on prescribed traits, these, too, behaviorally described, as well as a citation of evidence to substantiate the judgments made. However, the interviewer asks what questions he deems desirable in order to arrive at his judgments. Continuing research, along the lines to which I have made reference, I view as essential.

It is outside the scope of my assigned topic to treat intra-interviewer consistency. Obviously, in the over-all design of research on reliability, provision would be made for exploring this aspect of the problem. A facet of intra-interviewer consistency comprises sporadic and systematic errors of perceptual process, attitude, and judgment. Two former graduate students of mine, Regan (7), and McCandlish (5), have concerned themselves with research in this area.

Now then, if the interview is to be used as an evaluative technique, what of its validity?

It is not my purpose to get into a general discussion of validity. I should rather like to point up some ideas which may properly be subsumed here and which have a direct bearing on my assigned topic of inter-personal aspects of the interview. I should state at the outset that I believe it is more appropriate to speak of the validity of the individual interviewer's judgments than it is to talk about the validity of the interview. As a case in point my attention was called to a paper at the recent Paris meetings. Husen (4), the investigator, presented data not only concerning the validity of individual interviewers' judgments but the effect of age, sex, training, etc. on such judgments.

In passing, we should remind ourselves that the unique nature of the interview among screening and selection techniques raises an immediate issue of whether it should be conducted independently of any knowledge of data from tests, references, application, etc., or whether it should be an inte-

grative step. This is a moot point in our present state of knowledge. However, when it is used integratively many obvious complications arise in the determination of validity. Many investigators, including Husen, have shown that relatively little is added to the prediction when interviewer judgments are made integratively. It is not within the scope of this paper to discuss this issue but it should be borne in mind in what follows.

As a first note under validity, it would appear to be good practice to apply traditional correlational analysis to the individual interviewer's judgment quantifications, predictor variables, and such interviewee performance data as may be subsequently made available, criterion variables. The difficulty of course is that a particular interviewer may not do enough interviewing to yield a sufficient number of cases for this approach. Certainly it would be a rare situation where sufficient data of this kind were available for each of the several interviewers. A poor substitute for this kind of validity check is to use a "position offered—position not offered" dichotomy as a quasi-criterion and to analyze the judgment quantifications of the individual interviewers against this measure.

As a second point under validity I should like to mention the place of factor analysis in research on the interview. Sternberg (8), Barnes (1), and more recently, Chillian (3), graduate students of mine, used direct factor analysis on interview data. They factor analyzed matrices obtained by intercorrelating traits on which interviewers had made quantifications of judgments. Studies of this type help in shedding light upon the concepts underlying the superficial judgments made by the interviewers. It would be interesting to do this kind of study on individual interviewer's data. For example, would the data from a "good picker" yield the same factor construct as those from a "poor picker"?

Cash (2), another former student, factor analyzed a matrix obtained by intercorrelating applicants for the same vacancy. For factoring such a matrix of inter-personal correlations, the most parsimonious hypothesis, he reasoned, should be the existence of a single man-factor-type. He found the hypothesis untenable. Further, he found three man-factor-types which he then attempted to identify by means of subsequently collected on-the-job performance data. This type of study has potential value in defining the molar man-power requirements for specific openings. Also, a similar design might be used with on-the-job performance data to shed light on "types' needed to fill job needs.

Where multiple interviews are accorded applicants and where many interviewers are used, it may be enlightening to do an inter-personal type factor analysis of interviewers. It might be found that several interviewer "types" existed. If such were the case, it might provide a basis for setting up combinations of interviewers to maximize discriminability of judgments trait by trait.

For my third point under validity, I must refer back to the use of content analysis mentioned earlier. It seems to me that both reliability and validity are likely to be increased when interviewers have clearer concepts of the traits or qualities which they are trying to judge. Content analysis seems to be a technique which is useful in clarifying these concepts. What shall we content analyze then?

1. The evidence cited by interviewers to support their judgments, as previously mentioned.
2. Definitions of the traits formulated by interviewers.
3. Sound-scripts of interviews.
4. Qualitative statements made about

candidates in memos, etc. Incidentally, as a by-product of such analyses inferences can be made concerning the dependence of the interviewer's judgment upon other sources of selection data, e.g. test results, college record. For instance, the writer encountered an interviewer who cited little else except rank of candidates in their graduating class to substantiate his judgments of their Oral Expression.

As a fourth aspect of validity, I should like to mention something kindred to content analysis, namely, an analysis of evidence check-off lists if such exist. A graduate student of mine, Walrad (9), is presently making this kind of an analysis of the same cases as were included in Chillian's study mentioned earlier. It seems to me there are two useful ways of doing this.

1. By analyzing the evidence items as you would test items against internal and, if available, external criteria.
2. By seeing if "good pickers" show different patterns of checked evidence than do "poor pickers."

Unfortunately Walrad's data are limited to (1) above and are at this stage incomplete.

Now in order to have reliable and valid interviewing done, what should the interviewer know? Along what lines should he be trained? He certainly should have an intimate knowledge of the job to be done and the man-power specifications as mentioned earlier. In addition, he should be informed on the organizational structure, management policies, company practices, and the like. He should also have some knowledge of the persons with whom the interviewee is going to have to adjust if he is selected. He should also be acquainted with other phases of the evaluative procedure, especially if it is his responsibility to synthesize the total picture of the candidate. He should have some knowledge of the nature and extent of individual differences. He should be trained in interviewing techniques and specifically in those used by the company. This should, of course, include acquainting the interviewer with forms used and the procedural manual if such exists.

What are some feasible training methods? Books, pamphlets, and reprints on individual differences, screening and selection tools and procedures, and on interviewing techniques should be available. These are more likely to be used if a reading reference guide is prepared pointing up these materials. Lecture-discussions by specialists as well as experienced operating personnel comprise a useful source of information. Role-playing sessions may be conducted during which each participant takes both the role of the interviewer and that of interviewee. "Do's" and "don't's" are demonstrated dramatically by this means. Actual cases from the personnel files can be used retrospectively to high-light areas of success or failure in properly assessing candidates. Also, in this connection I have found play-backs of soundscripts particularly helpful. As mentioned before, oral or written "feed-back" of research findings augmented by conference method discussion is another training aid that has value. Finally, the more the interviewers themselves participate in modifying procedures, changing forms, revising manuals, the more they are likely to learn and the more seriously they may handle their responsibility for doing the best job of interviewing they know how. Incidentally, Husen's paper, previously cited, contains data to substantiate the contention that improvement in judgment can be accomplished through training.

Time limitations prevent a discussion of these additional interpersonal aspects of the interview. All of them are worthy

of research investigation and "feedback" in the course of continuing training of interviewers.

1. The degree of projection of personal qualities of interviewers into their judgments of interviewees.
2. The degree of influence of interviewee's personal qualities on the judgments of interviewers.
3. Board versus consecutive interviews—the multitudinous reliability and validity ramifications. (Oldfield and other British investigators have published in this area.)
4. Time allocation and expenditure as it effects performance of individual interviewers. (My staff and myself did some work along this line.)

I should like to conclude my paper by quoting from that gem of a book, Oldfield's *Psychology of the Interview* (6, pp. 138, 139):

"If interviewing is an art, the lines along which improvement may be sought can best be imagined by considering those by which other arts have been governed. In music, painting and the drama we find in varying degrees the gradual growth of a body of agreed, communicable technique and method. Additions and improvements occur, in part by the accumulation of experience, in part by the absorption of knowledge gained in other fields, and sometimes by deliberate effort. There are times when preoccupation with technique threatens true achievement. There are others when the stream of traditional method runs thin and weak. The occasion may even arise when, for certain purposes, the practice of an art is abandoned and its functions are taken over by mechanical devices—when the gramophone is substituted for musical execution, pictorial representation is given over to the camera, and the film deputises for the drama. These are all possibilities inherent in the development of the art of interviewing. But for the present, as a special branch of the general art of conducting human relations, it lacks all but the beginnings of an agreed, explicit technique. As this begins to emerge, the results may be surprising. And in its formation a scientific psychology will have a large part to play."

BIBLIOGRAPHY

1. Barnes, Paul J. A factor analysis of interviewer judgments of executive trainee applicants. Unpublished M.A. dissertation, Fordham University, 1950.
2. Casil, Harold C. An experimental investigation of "executive types" by the method of inverse factor analysis. Unpublished M.A. Thesis, New York University, 1948.
3. Chillian, Ralph F. Isolation and evaluation of critical traits in the screening and assessment of enlisted candidates for the submarine service. Dissertation in progress, Fordham University, 1953.
4. Husen, Torsten. "The validity of interviews as related to the sex and training of interviewers." (presented orally at Paris meetings, 1953)
5. McCandlish, L. Alexander. An investigation of interviewing ratings to check the feasability of measuring the extent of halo and central tendency errors. Unpublished dissertation, Fordham University, 1951.
6. Oldfield, R. C. The psychology of the interview. Published by Methuen and Co., Ltd., London. 4th edition, 1951.
7. Regan, James J. An investigation of evidence cited to substantiate quantifications of judgments on the part of hiring interviewers. Unpublished M.A. dissertation, Fordham University, 1951.
8. Sternberg, Jack J. An analytical study of a selection interview procedure. Unpublished M.A. thesis, Syracuse University, 1950.
9. Walrad, Leo. An investigation of evidence cited by medical interviewers to substantiate their qualifications of judgment of candidates for the submarine service. Dissertation in progress, Fordham University, 1953.

The Interview as an Evaluation Technique

NEVITT SANFORD

THE INTERVIEW IN PERSONALITY APPRAISAL

EVEN WITHIN the relatively limited area of personality appraisal the conduct of the interview and the handling of the data it yields will depend upon the aims pursued and conditions under which the work is undertaken. The present paper is concerned with the use of the interview in a fairly comprehensive, longitudinal study of personality development in normal late adolescents or young adults, the interviewers being trained at a fairly high level in clinical psychology, and the subjects being volunteers. My intention is to describe a procedure that promises to be useful despite the numerous problems and difficulties to be anticipated. From time to time the contrast with what might better be done under different conditions will be pointed up.

Purposes of the Interview

Fifteen or twenty years ago the inclusion of the interview among the techniques to be used in such a study would have been taken for granted. Today one must seriously question whether it is wise to use the interview at all. There seems to be general awareness of the filing cabinets about the country, bulging with unanalyzed interview material. Knowledge of research methodology has become so widespread that the interview is often perceived as a vast tenderloin of sin and error. The young investigator, trained to conceive of Heaven in quantitative terms, shrewedly induces other unsuspecting souls—teachers, psychiatrists, employers, supervisors and the like—to provide the unreliable criterion measures while he with unassailable innocence provides the predictor measures. Actually, this approach has been so successful, the objective standardized tests, empirically validated, have so far outstripped the projective techniques and other more global clinical devices as predictors of external criteria, as to give rise to the hope that a new day of righteousness might be at hand. Still, there are sinners among us, those who believe that something has been hidden from our view and who exhibit an insatiable curiosity about the inner workings of things. They are apt to regard the interview as an effective probe. But, today they must face the question: are there any scientific purposes to be achieved by the interview than cannot more efficiently be achieved by other methods? (Discussion is limited to the type of research undertaking outlined above. It seems obvious that where certain immediately practical aims are involved the interview is indispensable. If one has to decide whether to undertake extended psychotherapy with a given individual, or which of a number of applicants to accept as members of a small research team or to take along on an extended cruise, he will do well to interview them no matter what the battery of tests they have taken.)

It seems safe to say, in the first place, that in the field of personality there is

still place for the *exploratory* interview. In many areas the need is still to *find* the important variables, to give them preliminary definitions and to gain some notion of the signs by which they are to be known.

The exploratory interview also offers a means for seeing relations among variables, and thus for setting up hypotheses for future testing. One of the great difficulties in personality research springs from the fact that in order to understand a given phenomenon it is necessary to take simultaneous account of numerous factors. It is because the interview, or the series of interviews, makes this to some extent possible that it has been the major source of fruitful hypotheses in the field of personality.

In the second place, it seems at the present time that the interview may provide estimates of variables for which no objectve tests are as yet available. Although efficient objective instruments are produced at an increasingly rapid rate, it cannot be claimed that they have so far embraced more than a small section of the total sphere of personality variables, or that they even keep pace with the finding and definition of new variables. There is a fairly new scale for measuring rigidity; but we now know that there are several different kinds of rigidity and of nonrigidity for which objective measures are needed. A study of personality with any pretentions of comprehensiveness cannot wait for the development of these tests.

Again, if one wishes to perform case studies, to exemplify in detail the common patterns found in his data, to add to our understanding of the organization of personality, or to seek for new relationships of the kind which appear only when numerous variables are considered together, he is almost bound to make use of interview material. Even after the most comprehensive testing program, with the fullest use of projective techniques and situational tests, it will be found that in order to make the case "come alive" or "hang together," interview material will be needed. This is not merely because the use of the case study involves some commitment to an ideographic approach, that is, to the assumption that in each individual the organization of variables is in some sense unique; the number of common traits that will enter the picture will undoubtedly exceed the number of available objective tests, and the use of available ones will involve more testing than is ordinarily feasible. More than this, the study of an individual personality, properly carried out, will require attention to sequences and continuities in time; the interview is well designed for getting the life story, and in lieu of the most comprehensive longitudinal study, it will have to be used.

Finally, there is an area that may be approached either by means of the interview or by means of the questionnaire, a decision in favor of the one or the other depending on circumstances and a variety of rather subtle factors. The life-history, matters of fact about childhood, education, work and the like has long been the special province of the interview. But the questionnaire has made enormous inroads in this area; and this, very probably, is why Professor Kinsey must be at such pains to show that the questionnaire would not have been suitable for his purposes. One argument for the interview is that certain significant facts of the life-history are, for most people, sufficiently embarrassing or otherwise painful so that we should not expect an accurate account except under very favorable conditions. The anonymous questionnaire can go quite a long way toward establishing these conditions; it is by no means to be scorned by those who desire such facts as are contained in the Kinsey Report. If, however, one is un-

dertaking a longitudinal study, or any study of personality that requires repeated contacts with the same subject, that subject is bound to become known to the investigators, and account must be taken of his sensibilities. In these circumstances it would be foolish to go on asking, by means of the questionnaire, highly personal questions about past events and current practices without obtaining information about how these questions were understood and reacted to. With respect to certain significant matters we should expect a certain amount of conscious withholding of information and a great deal of unconscious resistence or distortion. There is no guarantee that even the most careful interview will overcome these obstacles, but it would appear to be at the present time the best instrument we have for the purpose. It must be emphasized that the concern here is with matters of fact, with what actually happened or what is going on now; where one is interested only in the subject's thoughts and attitudes, his conception of his childhood, his imagery of his parents and the like there is no question but that questionnaire and projective techniques may even now go a long way, particularly when there is knowledgeable use of indirection.

The Role of Theory

Probably the main reason why so much interview material is collected without its ever being analyzed is because the interviewing was undertaken without thought being given to what was significant, and in what way. We have commonly proceeded on the notion that if one asks about education, jobs, hobbies, relations of parents, expectations for the future and the like he will get material of which something can be made later on. We shall probably have to admit that what can be made of it in the end is directly proportional to what was put in at the beginning, in the way of theory and hypotheses. The chances are that no one is going to theorize about one's interview material except oneself; I should argue that this theorizing were better done before the actual work of interviewing rather than after. Every question, or at least every line of questioning, should have its theoretical rationale; one should even have thought about the kinds of answers likely to be elicited and about how these were going to be categorized in the analysis. It might be objected that this procedure would interfere with the exploratory function of the interview. The answer is that, at this stage of our knowledge, more will be discovered by conceiving hypotheses and following them through in the interview or by the study of material systematically gathered, than by floating about in a sea of undifferentiated verbal material.

It might also be objected that the direction of interviewing and of interview analysis by theory, by conceptual schemes and hypotheses, opens wide the door through which the investigator's biases will intrude. The answer would be that, while bias has always to be contended with in interviewing as elsewhere in personality research, it is no more of a problem in theory-directed interviewing than in any other kind, and that in any case bias is not to be controlled by the avoidance of hypotheses but rather by meeting certain methodological standards in the conduct of the interview, and in the analysis of the material it elicits.

The Conditions of the Interview

The fact that an interview is undertaken with normal volunteers, the subjects being sought by the interviewer rather than vice versa, has immediate implications for the conduct of the interview. It here becomes necessary to act in such a way as to maintain the subject's motivation on as high a level

as possible, to encourage positive attitudes toward the research both before and after the interview. It has, I believe, been sufficiently demonstrated that a wish to take part in an investigation that promises to make a contribution to science is sufficient motivation for most subjects, provided the demands upon their time are kept within reasonable limits. (Allowances may have to be made for "volunteer error," however; and non-volunteers will have to be offered additional inducements.) This being so, it is ordinarily a mistake to promise that the subjects "will get something out of it too," meaning counselling of one sort or another. (But this does not mean that counselling may not in fact have to be given, a matter to be gone into later.)

The necessity for maintaining positive attitudes toward the research calls into serious question the use of non-directive interviewing techniques. Since the subject has not sought the interview, it is of course out of the question for the interviewer to signify that he is ready to listen while the subject tells "what's the trouble" or "what's on his mind." And even though the subject may be brought to accept the task of telling about himself in his own way, we may expect difficulty and resistence to mount as time goes on. The situation is well-calculated to arouse anxiety and hostility even in the healthiest individual. Intelligent, educated subjects will, to be sure, go to work on some such question as "What influences have had most to do with your becoming the kind of person you are?", but an interviewer should have very good reasons for imposing so difficult a task—and, of course, some extraordinarily efficacious way of exploiting the material he collects. There is no denying that much can be learned by permitting a subject to bring up whatever topics he likes, in whatever order he likes, but it is difficult to justify the investment of time, either that of the subject or that of the researcher. And this is assuming that some way has been found to analyze the material.

Non-directive interviewing, or that kind of directed interviewing that begins with the most innocuous topic and proceeds warily in the direction of the touchy ones, usually assumes that the subject will be seen a number of times, perhaps over an extended period. But in a longitudinal study the need is for an appraisal of the personality *now*—this month or this fall—at the time that other tests or diagnostic procedures are being administered. Although several interviews may be considered to belong to the "now," one must be careful not to mistake "on better acquaintance" for an actual change in the personality. The need is for reliable estimates of the present state of affairs, something with which later states of affairs may be compared. Let us accept this requirement then, and ask how we might proceed when the task is a fairly comprehensive appraisal of the personality *now*, when not more than two or three interviews can be given and these within the space of a few days, and where there is the necessity of maintaining sufficiently good relations with the subject so that he will come again six months, or a year, hence and after that. Interviewing under these conditions requires a schedule based upon theory, it seems better carried forward according to some procedures rather than others, and it must yield material that can be handled in an objective or quantitative fashion.

The Technique of the Interview

Let it be assumed, then, that the interviewer begins his conversations with his subject with a well-conceived interview schedule on his desk. And let it also be assumed that he intends to cover the ground within the time allotted. Covering the ground here

means getting enough material bearing on a given topic so that judges, independently examining that material later on, can make an appraisal of the variables which were supposedly brought into focus by the discussion of that topic. The interview schedule is not, however, a list of the questions which will scrupulously be put to every subject. If this were the case, one might as well have used a questionnaire in the first place. No, the schedule will be made up of the questions which the interviewer is asking himself; ideally, the area in question has been sufficiently differentiated in advance, well enough mapped out conceptually, so that he knows just what he wants to find out. The matter of how to draw the subject out on the points in question will have to be left to the interviewers judgment. It is just here, of course, that art or skill or experience will enter the picture. The task is to get the subject to go far enough into the significant matter at hand, without arousing too much anxiety or hostility or causing him to regret it later. This calls for some ability to be aware of the subject's feelings, considerable knowledge of how anxiety and hostility are expressed and, particularly perhaps, the ability to judge what the traffic will bear. If one knows about these matters he can persist in his purpose despite some discomforture on the subject's part, because he knows that this discomfiture is within reasonable limits and that he can dispel it later on. By and large we should expect from the less skillful or less experienced interviewer errors of two kinds: that through over-cautiousness he has failed to bring out enough significant material or that he has elicited the material at the cost of our having disgruntled subjects on our hands.

The experienced interviewer may demand a completely free hand in the matter of how he is going to draw out the subject and manage the emotion that is generated, but he will rarely refuse to discuss with his colleagues the way he proposes to proceed in approaching the particular topics of the interview schedule. A result of such discussion might very well be that a number of suggestions about questions to put to the subject will be included in the interview schedule.

It should be emphasized that having accepted the general procedure being described here the interviewer, whatever his gifts, must accept the discipline of the interview schedule. Although he need not maintain any particular order in taking up the topics, he must hold himself to the task of covering them and not be led off in directions which might seem to him more important in some particular case. The point is emphasized because in other types of situation the interviewer is more than willing to be led in directions that seem significant at the moment. If one has to make a practical decision based on a thorough-going understanding of a subject, for example, when the question concerns psychotherapy — when and what kind, or when one is responsible for a tentative formulation of the central dynamic structure of a case, as in certain types of assessment work, the best procedure is quite different from that indicated above. In these latter instances, the interviewer must formulate hypotheses as he goes along, find the means for rejecting quickly the unpromising ones, spend the bulk of his time following up leads that point to the centrally important. Of all types of interviewing this is the kind that calls for the greatest skill and knowledge. There will probably always be a place for this kind of interviewing, but the research setting with which we are concerned is not one of them.

Contrary to what appears to be a common opinion, the interview schedule is, in the research interview, an

actual aid to the management of anxiety and the maintenance of morale. It helps to lend an air of impersonality to the proceedings, and it may serve as a convenient scape-goat. It is not that the interviewer wants to pry into personal affairs; the schedule is the task-master; the interviewer and subject have no choice but to do its bidding. With an interview schedule clearly in the picture the interviewer may put the touchy questions in the same even tone, in the same routine manner, that he employs with other questions; and he may, if he notices signs of embarrassment, relieve the situation by passing on to the next question. (In the questionnaire or inventory that touches on affect-laden topics this control is lacking. An item may start a chain of free associations which leads the subject, amid mounting anxiety, deep into fantasy. These are the subjects who seek out the investigator to protest the whole proceeding.) In the systematic interview the subject may be spared this kind of painful rumination. Thus, there is something to be said for the famous "rapid-fire" questioning of the Kinsey interview. However much the psychologist might object to the neglect of the psychology of sex in the Kinsey researches, it probably must be admitted that the data on tumescence-detumescence were fairly accurate and collected without undue hardship on the subjects.

There is, of course, a question about the after-effects of this kind of interviewing. I have conducted, in different settings, quite a large number of rather probing "one-shot" interviews with volunteers and with paid subjects without becoming aware of any undue repercussions, and one may well believe that this was true of the Kinsey investigations. The trouble is, one may never know. It is my impression that we are safe in conducting research interviews of this kind provided certain precautions are taken. Obviously we must, in the beginning, establish confidence in the whole investigation and particularly in those who do the interviewing, and obviously the conduct of the interview must be such as to keep any painful emotions within bounds. Beyond this, there are steps that may be taken to leave "good feeling all 'round" at the end. For example, a lapse into a non-directive approach, or an indication that the formal part of the interview was over and that what was now said was off the record, might enable the subject to bring out his doubts or misgivings or unresolved tensions so that they might be dealt with constructively. Finally, the door should be left open for another conference at some future time. Although, as stated above, it is a poor policy to "sell" a research interview by suggesting that the subject will get something out of it, it is wise to provide the possibility of counselling, or of telling something of the results, for those subjects who make this request. This is not only wise, it is a minimum requirement of a human approach. It is not our wish to regard the subject's problems, and private thoughts and deeply felt sentiments as nothing more than grist for our statistical mill. He has the right to expect human reactions to these human expressions. Experience shows, however, that few volunteers avail themselves of opportunities for counselling, and that those who do are not very demanding.

In a longitudinal study, the matter of counselling the research subject assumes very considerable importance. When a subject is seen from time to time over long periods some kind of relationship between him and the investigator, or investigators, develops inevitably. Thus the subject of study changes as a result of being studied. (This consideration makes it all the more essential, in a longitudinal study, that all possible effort be directed to achieving a formulation of the case at

the time the study begins.) The investigator has no choice but to assume responsibility for these changes, and to act in a way that is in the best interests of his subject. A developing relationship between subject and investigator has serious implications for experimental design. If for example, one is studying psychological changes accompanying physical growth, or changes resulting from education, the effects of repeated contacts with the investigator may be very considerable. Such effects are probably minimal when techniques are limited to objective tests, even though these be administered individually, but the repeated use of the interview would seem to be quite another matter. We cannot solve the problem by avoiding all counselling or by striving to keep everything on an impersonal basis. The wisest course, it seems to me, is to make a virtue of a necessity, to hypothesize that subjects interviewed over a period of months or years will as a result of that experience be somewhat different in the end from similar subjects who did not have the experience, to test the hypothesis through the use of the control group, and to keep close watch upon the whole interaction process, trying to observe which aspects of this interpersonal process have which effects with which subjects.

A device that would seem to offer several advantages for research is that of having the same subject interviewed, within a relatively narrow space of time, by two or more investigators. It is interesting to note here that in the Tavistock Institute of Human Relations, which carries on in the spirit of the War Office Selection Board (the British "Assessment Center"), assessment in selection work has been boiled down to just two procedures: the group discussion (in which the performances of 10 or 12 candidates are evaluated by the staff) and the interview (in which a given candidate is interviewed by four people in the course of an afternoon). Thus a high premium is placed upon judgment, that of the assessment experts and that of the executive with whom the successful candidate will be working. It would seem that if the ability to judge people counts for anything in personality research—and I believe that it does—the interview offers the best means for enabling us to take advantage of it. Why not avail ourselves of the judgments of the most experienced people we have, and let them serve as checks upon one another? Let us divide the task of covering the interview schedule among several interviewers, and let them, after covering the topics assigned, offer their judgments concerning some of those pervasive aspects of personality that would be subtly manifested as well in a discussion of future plans as in a discussion of childhood.

The Exploitation of Interview Material

The whole conception of the research interview, as set forth here, has been worked out with an eye firmly fixed on the problem of how to handle the interview material. The device just mentioned, that of having several interviews for each subject, is an obvious attempt to avoid throwing out the expert altogether and yet to meet the minimum requirements of sound methodology. A check list of several hundred adjectives, to be filled out for each subject by each of, say, four interviewers should provide ample opportunity for setting forth the uniqueness of each subject even while it afforded data for quantitative treatment. Ratings on some of the so-called generalized traits of personality can still be very useful; as a matter of fact, for many common traits of personality they are still the best measures we have. Moreover, if we are going to have experts, and if they are going to undertake to penetrate the outer layers of the personality, then we should have the

benefits of their judgments concerning the more central, or the more subtle, aspects of the personality. These judgments, too, may be rendered by means of a rating scale, the interviewers having agreed in advance upon the form and content of this device.

But the general procedure of interviewing that has been described here has been designed with particular attention to the objective, quantitative analysis of the subjects own reports. It is a part of the interviewer's task to record, under appropriate headings, what the subject has to say about the topic in question. During the interview and during the recording, the interviewer bears in mind that judges other than himself are going to be required to classify the material bearing on a given topic, and that they are going to do this without reference to other parts of the record. The proposal here is that the interview protocol be divided, cut-up, into a number of sections, corresponding to the topics of the schedule; that the productions of all the subjects, on a given topic, be then assembled and handed, without names or identifying data being attached, to two or more judges; and that the judges, working on one topic at a time, categorize the material or make estimates of variables in accordance with the conceptual scheme upon which the interview was based. This procedure should eliminate halo effect and insure that judgments on particular issues are not contaminated by any other kinds of information about the subject; it should yield data that are ready for quantitative treatment. Measures obtained in this way, like the measures based on ratings by the interviewers themselves, may be used either as predictors of external criteria, or as measures against which tests can be validated.

The Interview as an Evaluation Technique

HENRY RICCIUTI

SUMMARY OF DISCUSSION

Comments on Dr. Kelly's paper

Dr. MacNamara asked whether it would be practical in terms of time and money, to carry out Dr. Kelly's suggestion that the complex task of the single interviewer be fractionated into smaller units, each to be assigned to a different specialist. Dr. Kelly expressed the opinion that this was an empirical question, in which the relative economy of various approaches would need to be determined and evaluated. However, if the only way to do the job properly was to use teams of experts, then the job would have to be done that way. Dr. Kelly mentioned that it was interesting to note that he and Dr. Sanford were not too far apart in regard to this matter of fractionating the total interview task.

Comments on Dr. Crissy's paper

Dr. Speer asked Dr. Crissy to comment further on the matter of experimental designs for studying interviewer consistency. Dr. Crissy pointed out that the ideal design for studying *intra*-rater consistency would involve keeping both interviewer and interviewee free from any carryover effects from one interview to the next. For studying *inter*-rater consistency, ideally one would like to keep the interviewee uncontaminated by successive interviewer effects.

One illustration of the type of experimental design which has been used in practice is seen in some recent work carried out by Dr. Crissy and his students. These investigators recorded interviews by each of several individuals, and subsequently played back these recordings to a small group of interviewers, including the original interviewer in each case, who didn't recognize his own voice. When this group was asked to evaluate the interview material, promising results were obtained in regard to both inter-judge agreement and consistency of the judgment of the original interviewer.

Another possibility would involve setting up machinery for routinizing interview research in a particular situation over a period of years. Some of these ideas have been outlined by Dr. Crissy in a recent paper in Personnel Psychology. (1952, 5, 73–85)

Dr. Peterson wondered why so little attention has been given to the possible use of classical psychophysical methods in dealing with interviewer judgments. Dr. Crissy indicated that there has been at least one rather encouraging study in which "paired comparison" judgments were made by supervisors in evaluating sales trainees, each trainee being compared with every other trainee. Such methods might be used profitably in situations where a constant flow of interviewees is available.

Comments on Dr. Sanford's paper

Dr. Symonds wondered if there might be some value in having two interviews by the same interviewer, so as to reduce the subject's resistance, along the lines sometimes followed in repeated projective testing. Dr. Sanford indicated that in his current work, it was hoped that resistance was resolved in the first interview, and if not, that it would be dissipated by the time the subject reached the second or third interview. However, it might be desirable if the same interviewer and the subject got together several times in the course of a few days. One should always be aware of the possibility of resistance, and one should try to deal with it right away. When resistance is observed, it would be very valuable for interviewers to compare notes on it and to attempt to identify its causes.

At the request of Dr. Symonds, Dr. Sanford made a few more comments regarding the interviewer's analysis of the subject's responses. In the approach described by Dr. Sanford, the interviewer induces the subject to talk about various topics, e.g. his parents, etc., according to the plan of the study. He takes notes on the material gathered in each area, dictating a more complete summary later. (This approach is preferred to that of recording the entire interview.) Then, independent judges check one another's interpretation of the interview material in each of the areas covered.

General Comments

Dr. Zubin made the observation that a most unusual situation seems to exist with regard to the interview—a situation warranting some close examination. Very strong criticism has been directed at the interviewing technique, yet most of us continue to use it, to varying extents. He feels that one of the main factors involved in this situation is that we tend to view the interview without proper historical perspective. Actually, the interview was among the earliest methods used in the physical sciences as well as in the psychological sciences. In first studying physical phenomena like heat, for example, people were asked to report whether something was "hot" or "cold." Before the work of Binet, a person's intelligence was often estimated by an "interviewer" who probably judged the person "intelligent" if he liked him, and "unintelligent" if he disliked him. Some of the earliest studies evaluating the interview as a technique, yielded negative results because they were concerned with the relation of interview material to entirely inappropriate variables.

Dr. Zubin feels that for studies of attitude and motivation, the interview is our best technique. However, we need to objectify the interview, identify the factors underlying it, and then systematically vary these factors and study their consequent effects. With intensive research of this sort, the interview will eventually be developed into a basically sound technique.

Impact of Machines and Devices on Developments in Testing and Related Fields

ARTHUR E. TRAXLER

THE IBM TEST SCORING MACHINE: AN EVALUATION

Introduction

IT MAY SAFELY be assumed that this audience needs no explanation of the nature of the test scoring machine, nor of the answer sheets on which it bites rhythmically when duly stimulated, nor of the basic principle on which it operates. I trust that it will meet with your approval if I omit all that from my remarks.

Perhaps, however, I may be permitted one small, faintly nostalgic historical note before I get down to work. At the outset, perhaps we may remind ourselves that just about twenty years ago the proponents of objective testing were fighting a somewhat uphill battle to obtain general acceptance of newer procedures of measurement in schools and colleges. As for the industrial field, these procedures had as yet hardly created a ripple in that area. Except for a few novel departures, the almost universal way of recording responses was in test booklets. Many so-called objective tests were of the semi-objective kind which required a certain amount of judgment on the part of the scorer. Admittedly, these tests could be scored more rapidly and with a higher degree of agreement among graders than could essay examinations, but to many persons the difference was not impressive.

At that time, in the early 1930's, the International Business Machines Corporation was convinced that mechanical scoring of tests was possible and was actively interested in producing such a machine. Various experiments were being carried on in different places, and fearful and wonderful devices were appearing here and there throughout the country. It was at this propitious time that Reynold B. Johnson, now head of the West Coast Developmental Laboratory of IBM, devised the first crude model of what was to become the IBM Testing Scoring Machine. This was based not only upon the fact that a lead pencil mark would conduct an electric current; it included the much more ingenious idea that by inserting high resistors of equal value into the circuits closed by the lead pencil marks the amount of current allowed to flow through each of these tiny circuits could definitely be controlled and need not be affected by the length or thickness of the lead pencil mark, except within very minute limits.

In recent months, there have been persistent rumors throughout the testing fraternity concerning astonishing—even imagination-defying—new devices that are just over the horizon, and no doubt we will hear about some of these this afternoon. Regardless of what the future may bring, however, I feel sure that everyone here, in fact every measurement specialist in the country, acknowledges that we are greatly indebted to IBM for pioneering machine

scoring, and for persistent interest which will no doubt lead to important improvements in the present machine.

I insert this brief historical note into my introduction because there does not seem to be much, if anything, in print about the genesis of the scoring machine, notwithstanding the importance of this event to the field of measurement. I had this somewhat forcefully brought to my attention recently in conversation with a young man whose first and somewhat casual contact with the scoring machine was during World War II. The young man seemed slightly saddened and disillusioned to learn that this novel gadget did not spring full blown from the inner mysteries of the Pentagon Building!

Some General Effects of Machine Scoring

During the seventeen years of its existence, the scoring machine has, of course, contributed enormously to large-scale testing. It has done this not only through its own potential for rapid, inexpensive scoring, but even more important through its influence upon the use of separate answer sheets. Since separate answer sheet use was a requirement, if tests were to be scored by machine, users rapidly accepted the separate answer sheet as a supplement to the test booklet. This favorable climate, sychologically speaking, without doubt made for more ready acceptance of large-scale hand scoring programs in which separate answer sheets were used.

I like to think that the test scoring machine has been to the testing business what the Model-T Ford was to the automobile industry. If the Model T put America on wheels, the test scoring machine has put the youth of America on objective-test answer sheets.

Some cynically-minded individuals have regarded each of these two phenomena as a not unmixed blessing. In all fairness, it should probably be conceded that the influence of the test scoring machine upon the kinds of examination situations set for young people and the kinds of responses required may not have been entirely fortunate. The use of the kind of answer sheet required by the fixed response-position and the fixed fields of the scoring machine has tended to force objective testing into a kind of strait jacket—in truth, a somewhat loose fitting and benigh strait jacket—but a strait jacket nonetheless. The four- or five-choice, discrete test item has become virtually standard so that, except for differences in content, the parts of many of our standard tests are almost as interchangeable as the housing units in a Levittown. The test scoring machine is not, of course, wholly responsible for this development, but I think it has accelerated a trend that might have been present regardless of mechanical means of scoring.

From the standpoint of measurement, this close similarity in kinds of test items usually is not a serious limitation, although it does occasionally foist upon test content an unnatural test situation. For example (if I may take a friendly, roundhouse swing at a test of which I am really very fond), in one section of the Cooperative English Test A: Mechanics of Expression, an exercise in punctuation is forced into the multiple-choice IBM answer sheet form in such a way that a considerable per cent of the junior high school pupils simply do not understand what is to be done. There is introduced what seems to be a kind of closure factor, which, in all probability, influences the results of this English test considerably.

From a broader educational standpoint, there may be more reason to deplore the forcing of objective test items into a comparatively small number of kinds in order to satisfy the require-

ments of the test scoring machine. This procedure has given rise during the last fifteen years to an expression which is sometimes used by the unregenerate die-hards among the critics of objective measurement as a kind of epithet —"the multiple-choice mind." Moreover, there is reason to believe that there are many teachers who have accepted objective measurement, perhaps with quiet resignation, who are uncomfortable over the thought that regardless of the resourcefulness, skill, imagination, and stimulation to thinking that they bring to their teaching, in the end the achievement of their students, and indirectly their own success, is going to be judged largely by how well their students respond to a single type of test item.

Fear that the almost exclusive use of the multiple-choice kind of test item will have a very deleterious effect upon learning may have little foundation in fact; yet it does seem probable that more flexibility in the kinds of questions, even to the point of introducing an occasional open-ended question into the measurement situation, might have a salutary influence upon instruction and upon the psychology of student preparation, and perhaps upon measurement as well.

So far as measurement is concerned, it seems to me that it has been a choice of values. In adopting machine scoring and in accepting the limitations to test construction imposed by the machine scorable answer sheet, we have perhaps tacitly chosen to bring to the many, tests not quite as good as could be prepared under optimum conditions for the few to whom measurement might be restricted in many places if separate answer sheet procedures were not available.

There are also the question of the extent to which the necessity of adjusting the answer sheet to the test booklet introduces extraneous perceptual and spatial factors into the measurement situation, the question of the lowest age level at which separate answer sheet tests may successfully be used, and the question of the influence of limitations of desk space on results of tests administered with separate answer sheets. There remains a need for more research on all these questions.

But let us turn from questions of these kinds to take a more down-to-earth, practical look at the values and limitations of the IBM Test Scoring Machine as it is used in a test service organization.

Values and Problems in Machine Scoring

At the Educational Records Bureau we have three test scoring machines, one of which is equipped with an item counter. I'll comment first on the use of the machine in test scoring and then very briefly on its use in item counting.

The Bureau's experience in machine scoring a few million tests over a period of some sixteen years indicates that, when the machine is in good working order and properly set and when the answer sheets are marked with heavy, black, glossy marks with the recommended pencils and are free from stray dots and marks, machine scoring is highly efficient and extremely accurate. A first-rate machine operator can score tests yielding a single score on one side of an answer sheet at the rate of 500 or more an hour and can maintain such a speed for a period of at least two hours. When several scores are obtained through one insertion of each answer sheet, the number of tests scored per unit of time is, of course, lower but the number of scores obtained is much higher.

At the same time, it should be noted that a test service organization such as the Educational Records Bureau seldom finds it practicable to offer IBM ma-

chine scoring services on certain tests yielding scores on numerous scales, particularly the Strong Vocational Interest Blanks, because of the extremely large number of insertions of the answer sheets that are required in order to obtain scores on all scales.

From the beginning, the *accuracy* of the scoring machine has been deliberately played down in order to avoid having users expect too much. Accuracy within one point is all that has ever been claimed for the machine. As a matter of fact, in the vast majority of the cases, the scoring is wholly accurate. When the scoring formula is $S = R$, and the conditions of machine and answer sheet optimum, and the operator reasonably alert, there should practically never be a scoring error. When $S = R - \frac{W}{N-1}$, there will occasionally be a one-point error because no one has ever been able to teach the machine to round fractions to the nearest whole number; but experienced operators become very adept at making fine distinctions so that there should not be an error more often than once in a hundred papers. I repeat that when all conditions are right a very satisfactory job is done even with the present machine, which is virtually unchanged from the first working model brought out by IBM.

Frequently, however, not all conditions are right. Occasionally, a machine will get out of order unbeknown to the operator and so systematic errors will appear in the scores of a group of papers. But this is a fairly minor difficulty, since such errors will, of course, be caught in checking either with another machine or by hand.

Another comparatively minor drawback is that on a warm, humid day the answer sheets tend to collect enough moisture to affect the reading slightly and thus to slow down the scoring process. Through the use of a "heater" in the machine, the answer sheets may be dried out so they will yield correct scores, although this procedure likewise tends to slow down the scoring process to some degree.

The main difficulty is in the marking of the answer sheets. It seems almost impossible to persuade groups of individuals to mark their answer sheets well enough so that all of them will yield correct scores when the scoring is done by machine.

Departures from what is desirable in answer sheet marking may take several forms—use of wrong pencils, light marks, many stray marks, and the use of different kinds of marks such as the drawing of a single slanting line across the response position. Several years ago, Doris Bretnall of our staff and I made a little study in which the answer sheets of a group of pupils on one of the Cooperative tests were recopied to form five identical sets except that a particular kind of undesirable marking, such as very light marks, was simulated throughout each set. Each set was scored by machine, and the scores thus obtained were correlated with the correct scores. Means and standard deviations were also computed. The paper was reported at the IBM Research Forum in Endicott in 1948. It created no great stir among the assembled experts, but I think it did bring out very clearly the fact that while most kinds of departure from desired marking are not of much importance, light, gray looking marks lead to numerous and rather large inaccuracies in machine scoring.

There are several ways in which a machine scoring unit can undertake to meet this problem of poorly marked answer sheets. In the first place, there is a nice question whether anything at all should be done about this other than to score and return the answer sheets. It may be argued that, since it takes

longer for a subject to mark an answer sheet carefully with heavy black marks than it does to mark lightly and carelessly, the fixing up of poorly marked answer sheets before scoring them constitutes indirectly a penalty upon the careful, conscientious individual. Since the purpose of testing, however, is measurement, not character training, and since the best approximations of relative achievement probably are to be obtained if all answer sheets yield correct scores, a procedure for obtaining correct scores from poorly marked answer sheets seems a requirement.

Various procedures are possible. The packs of tests may be inspected, and those that contain many poorly marked answer sheets may be scored and rescored by hand, or portions of them may be handled in this way. Or the scoring unit may score all answer sheets by hand and use the machine for rescoring, or the answer sheets may be scored in one machine and rescored in another and the ones having scoring discrepancies then hand scored. If the machines differ in sensitivity, as in one experiment at the Educational Testing Service, this procedure no doubt has considerable merit. Scoring machine operators, however, tend to take a dim view of this procedure, for when poorly marked answer sheets, particularly those with many stray marks, are run through the machine, the pointer has a tendency to creep. A few hours of experience in trying to read a creeping pointer is hardly conducive to emotional stability.

Everything considered, the least unsatisfactory procedure seems to be to scan and re-mark carefully all sets of answer sheets before putting them into the machine. This is by far the most time-consuming and costly aspect of machine scoring. I believe our machine scoring department spends at least twice as much time in scanning and remarking as it does in the actual operation of the machine. This work ordinarily takes a large share of the time of the very persons who are trained to operate the machines, for we have never succeeded in hiring a crew of answer sheet re-markers and getting them to stay with their job. Perhaps for this particular job specification, remarker of poorly marked answer sheets, which for some reason has never found its way into the Dictionary of Occupational Titles—perhaps what we need to do is to raid a mental institution and hire some employees with IQ's of about 35 to 50!

Notwithstanding the great advantage in speed of machine scoring over hand scoring when no scanning is required, I doubt that, in a situation calling for much scanning and re-marking, the rate of the total machine scoring process will equal the best hand scoring speed, such as Lindquist's workers have achieved at the University of Iowa.

The real cure for the major difficulty of scoring with the present machine is to strike at its source, the marking of the answer sheets by the persons taking the test. I think that the main reasons why many answer sheets are so poorly marked are (1) that many test makers and test publishers have either failed to stress the necessity for well marked papers or they have buried these instructions in the fine print; (2) that thousands of persons administering tests in our schools know nothing about how machine scoring is done and care less; and (3) that great numbers of school pupils seem impervious to specific directions unless the directions are reiterated several times and made graphic by illustration.

For at least fifteen years, the Educational Records Bureau has tried repeatedly to persuade schools to send in well marked answer sheets. With a few noteworthy exceptions, these attempts have met with little success. Year after year, we have sent special instructions

accompanied by a sample well marked answer sheet to all examiners contemplating machine scoring, but none of this has done much good. Not until the current fall program. This fall we introduced two small innovations into our instructional materials. Now the tests are coming back to us for scoring, and for the first time in all these years we seem to be getting somewhere. At any rate, the answer sheets received from a number of the schools are amazingly improved.

Now the changes in special instructions were very simple. The specimen answer sheet, which heretofore had served to illustrate good, black marks only, was divided into two parts. One-half was well marked and the other half poorly marked. Apparently this was the first time that some of our examiners found out what an answer sheet too poorly marked to yield correct scores looked like. Such an illustrative sheet, incidentally, must actually be marked by hand to get the idea across to the examiner.

Copies reproduced by the duplicating process simply will not do the job.

The second change was an urgent request that each school plan its testing schedule to allow an extra five minutes in which the response positions on the answer sheets would be darkened by the pupils themselves after the test booklets were collected. One of my more realistic colleagues pointed out that, if the test were a timed one, an occasional pupil might welcome the extra period as an opportunity to mark responses to a few additional items more or less at random. Naturally, I was quite shocked at the thought that any of our young people might engage in such shenanigans, but we did think it advisable to include a suggestion that this operation be well supervised by alert proctors!

I am aware that this kind of device is frowned upon by some authorities on the administration of tests, and they may be right in their disapproval. It is what is sometimes referred to as a calculated risk. I have a feeling that the test wiseness of pupils in our schools and their knowledge of the existence of correction formulas will cause them to be fairly ambivalent about wild guessing.

The Graphic Item Counter

There is time for only a brief comment on the item counter. This attachment to the scoring machine makes possible a graphic count of responses to the various items as the answer sheets are run through the machine. On a single graphic item count record, it is possible to count the correct responses to ninety items, or to count all responses to eighteen five-choice items, and so forth. The responses on 115 answer sheets may be counted on any one graphic record. The graph is made with carbon paper, and thus it is not always as legible as one would desire.

The use of the item counter unit is much slower than scoring, but still it does count responses to items many times as fast as they can be counted by hand. We have found that the item counter saves an enormous amount of time in making item analyses of new tests or of tests on which further item difficulty and validity data are needed.

An aggregate weighting unit is also available and is useful in research, as well as in different kinds of scoring, such as the scoring of rating scales where the various characteristics to be rated have different values. However, this unit is not often used, and since we have had virtually no experience with it at the Educational Records Bureau, I will omit comments on that aspect of the scoring machine.

Conclusion

In summary, some seventeen years of experience with the IBM Test Scor-

ing Machine in the Educational Records Bureau program indicates that this machine is, in itself, rather highly satisfactory, even without further refinement. Our difficulties arise mainly from failure to educate test publishers and test users to recognize the crucial importance of *well marked* answer sheets. If this one problem can be fully solved, the efficiency of the use of the scoring machine will be increased at least threefold.

My remarks have been confined to an evaluation of the machine in its existing form. Although, as I have indicated, the present model is giving good service, so far as the Educational Records Bureau is concerned, I should like, in conclusion, to suggest very tentatively four or five kinds of improvement that might be made in mechanical devices for the scoring of tests. Without doubt, various ones of these items, as well as others, will be covered more authoritatively in Mr. Bradley's talk.

First, the scoring of poorly recorded responses. A machine that would accurately score answer sheets in their natural, uninhibited, unrepaired, raw state as they are received from the average student in almost any school or college would contribute enormously to efficient scoring.

Second, the speed of operation of the machine. The actual process of scoring with the IBM machine is amazingly fast in comparison with hand scoring. It does, however, depend upon several comparatively slow mechanical movements within the machine, and, of course, it involves a number of movements on the part of the operator. Much scientific progress has been made even in the few years since the machine was developed. It may be assumed, even by a non-mechanically-minded observer, that it may be possible to introduce fundamental changes into machine scoring that will put these procedures thoroughly into step with a streamlined, supersonic, electronic, atomic age. For example, automatic feeding of the answer sheets into the machine and automatic recording of the scores would noticeably increase speed of operation and, at the same time, reduce the possibility of error.

Third, the scoring of multi-scale tests. Any change in the basic structure of the machine that would make possible the rapid scoring of multi-scale tests, such as the Strong blank, after the manner of Hankes, would enable units equipped with the IBM machine to render somewhat more inclusive service.

Fourth, provision for greater flexibility in test construction. At present, test authors who desire to adapt their tests to the IBM Test Scoring Machine must observe certain requirements, such as the arrangement of the number of items in a given part in multiples of fifteen or else the omission of several numbered items, and so forth. If it were possible to eliminate the fixed fields in the scoring machine and also to make refinements so that a single answer sheet would accommodate a larger number of responses, these innovations woud assist both test authors and test users.

Fifth, the elimination of various minor annoyances in the use of the scoring machine. These include the influence of humidity on the scores, pins that sometimes stick, occasional batches of answer sheets printed so they do not register perfectly, and the sticking of scored papers in the machine. Each one of these is of no great importance, but they add up to lost time and frayed nerves. Their elimination would make for faster machine scoring services with a smile and would help appreciably to safeguard the accuracy of the scores obtained.

Finally, the expense. Thus far, use of the scoring machine has been limited largely to service centers, large school

systems, large industrial organizations, and military establishments. The annual charge for the rental and service of the machine, while moderate, has been beyond the budgets of most small, and many medium sized schools. Either an inexpensive "desk model," or an electronic, centrally located model designed to render nationwide service, would contribute greatly to increased use of standardized tests, to the scoring of teacher-made tests in numerous schools, and to research. (Incidentally, either of these developments might put the ERB and kindred organizations out of business, but happily this dreary topic may be omitted because my time has expired!)

Impact of Machines and Devices on Developments in Testing and Related Fields

JOHN E. ALMAN

THE UNIVERSITY SERVICE BUREAU

THE UNIVERSITY SERVICE BUREAU may be defined in general as an agency within a university operating punched card equipment and making its services available to any department within the university. Though such agencies have existed in universities for twenty-five years or so, the number of such prior to World War II was quite small, and the growth and spread of this type of service during the past seven years has been remarkable. There are now about 200 colleges and universities in the country which meet this general definition; in addition there are another 150 or so schools utilizing punched card machines that do not in that the availability of the machines for other than their basic mission is non-existent or very sharply limited. Such installations are mostly in the University business offices, with some in the Registrars' offices.

Among the institutions which meet the above general definition there are those in which the machine installation is administered by the particular department that provides the basic workload for the machines, e.g., the Office of the Registrar. In these cases the service function is more or less secondary to the primary mission; we might refer to such administrative entities as quasi service bureaus. On the other hand, in about 30 of the larger universities machine services are centralized in an administrative unit whose primary mission is the service function. The number of such institutions has increased four-fold in the past seven years, and the trend is still upwards. Since such service bureaus are expensive and require sizeable workloads to justify the administrative organization, it is likely that such will be restricted generally to the larger institutions. But there is no doubt that the concept of an autonomous bureau that provides machine services for the entire university is here to stay.

What does the centralized machine service agency have to offer to testing programs and research? Those in the testing field have long been eager customers of punched card equipment. Testing characteristically requires large samples with a relatively small amount of information per individual in the sample, a situation tailored to the punched card. Testing makes extensive use of frequency distributions and order statistics, of bivariate distributions from coded or grouped data, of second moments and product moments. All of these can be handled on the basic accounting installation—the sorter, the reproducer, the collator, and the accounting machine. These machines, together with key punches, verifiers, and the interpreter, for many years constituted the facilities of the university service bureau. In very recent years many have acquired punched card com-

puters, IBM's slow speed 602A, the medium speed 604, or Card Programmed Calculator. The availability of these machines enormously increases the potential of the service bureau in test processing and research. They allow the direct computation of first and second moments and product-moments in large numbers as well as many other manipulations of the data that involve multiplication and division. The CPC, particularly, permits for the first time relatively rapid computation of the inverse of a correlation matrix, an operation fundamental to many areas of multivariate analysis. Without these computers many areas of research are closed to the researcher because of the gargantuan computational task. To complete the machine potential of the present-day university service bureau there is IBM's 101 Electronic Statistical Machine which permits the high-speed preparation of contingency tables from categorical or coded data, item analyses, editing and sequence checking, selective sorting, and many other tasks. It is a highly versatile machine that has been available for too short a time to come into general use in test processing.

No mention of services of a university installation would be complete without pointing out that the machine scoring of tests is properly a function of such a bureau. Test scoring is a subject of itself, and is treated by other speakers; however, it is pertinent here to mention that scoring is so often followed directly by card punching that it is advantageous to the entire process to keep all operations on the data within a single organization.

The complete university service bureau incorporates all of the above equipment, and can provide all manner of statistical and processing services within a single organization—test scoring and item analysis, the preparation of listings and frequency distributions, order statistics, the computations of measure statistics, and finally the involved matrix computations of multivariate analysis. Desirable as this picture is, it must be recognized that it all costs money and requires a large workload. Furthermore, a sizeable machine installation requires a type of ongoing workload that is repetitive in nature, one that provides the financial support to absorb most of the fixed costs of operating the installation. With this as a base, it is possible to schedule on top of the regular load the many "one-shot" research jobs that arise in a university. Since the support for a university service bureau must primarily be derived from income for services rendered, the users must share the costs on the basis of machine and personnel time required. This requires cost accounting, scheduling, and careful planning of research jobs to use available set-ups and procedures. For these reasons I feel that the service bureaus will less and less be in a position to allow the consumer to handle his own data processing; that is, the trend is toward providing *machine services* for the researcher, not just the machines.

The complete service bureau can easily require an annual budget of $100,000 or more; it is obvious that many institutions simply cannot find the workload or financial support for such an undertaking. So the kind I have described above will necessarily be restricted to the larger institutions, and many will have to be content with a basic installation of accounting machines. Much can be done with such an installation, of course, and the way is opened for greater use of available machine potential by the larger service bureaus opening their services to the smaller institutions, to supplement the limited capacity of a small installation when a sizeable job arises, and to provide the more expensive computing services that the small institution cannot afford. I feel that the large institu-

tion with the complete facilities has an obligation to provide this service on a regional basis, and the trend toward this kind of sharing of machine services is definitely indicated.

While it is true that large scale studies are most natural for punched card machines, the university service bureau must handle many small jobs. Since oftentimes planning and set-up time may far exceed running time, it is important to have basic statistical procedures and machine set-ups permanently planned and boards wired. Both cost and time can be substantially reduced by careful planning for the *small* job. For example, the cumulations necessary to compute the correlations among three variables on a hundred or so cases can be done on a desk calculator, and one is tempted to say that such a job is too small to be handled efficiently on expensive machines. However, if *all* the set-ups required—from key punching to the final accumulations—are immediately available, the job can be done in an hour or so, and the cost is kept low. At Boston University we maintain such a set-up for the 602A that will allow from one to four variables per card for computing and accumulating squares and products. We maintain a permanent tabulator set-up for producing frequency distributions and corresponding percentile equivalents. Permanent set-ups for both the tab and the 101 Statistical Machine allow the preparation of a two-way frequency table with a single column spread horizontally. Permanent boards for matrix computation on the 602A are maintained, though this machine is too slow and too limited in storage for matrix computations; we are looking forward to transferring these to the CPC when it is delivered. With the permanent set-ups mentioned above we can handle a rather wide variety of statistical computations on small jobs without requiring personnel time for extensive planning and board wiring. Indeed we frequently find that it is advantageous to force the desired computations into one of these pre-set patterns rather than make a special machine set-up even though running time might be saved thereby. During the past year we handled about 100 research jobs of various kinds, only about 10 of which could be considered "large jobs." For the many small jobs the permanent set-ups paid off handsomely in lowered costs and personnel time saved.

The machines themselves, of course, do not make a service bureau. More important are the people who plan the jobs and run the machines. An adequate staff of machine operators is essential; however, the key figure is the supervisor who handles the communication between the consumer and the machines. The need for persons with competency in testing procedures and analysis, in statistical methodology, as well as in machine methods, is a high-order need, and one that is rarely met except in those institutions that can afford a centralized machine service agency. There is great need for the availability of such persons in the planning stages of testing projects, with particular emphasis on those aspects of the project that have direct impact on punching the cards and processing the data. These considerations are bound to involve experimental design and statistical methods as well as the mechanics of collecting and recording the data; they involve the design of answer sheets, and the statistical methodology of item and test reliability and validity, as well as the mechanics of test scoring.

It is my feeling that a true university service bureau must be able to provide personnel who can carry out the role of statistical and research consultants, and that the effectiveness of a service bureau depends to a considerable extent upon the integration of its personnel into research planning. Many are

the cases where unnecessary data transcription, recoding, and machine processing could have been avoided had machine specialists been consulted in the planning stages of a project. For the person with the small job—often a graduate student—the service bureau must provide personnel who can take the time to consult with and advise him, to steer him clear of the errors of omission and commission that make eventual machine analysis awkward or difficult, to assist him in devising data record forms suitable for the use of key punch operators, to make sure that codes are in line with machine requirements and limitations, and to avoid designs in which the necessary statistical methods are not amenable to ordinary machine procedure. For the graduate student the service bureau personnel should supplement—but not supplant—his adviser.

Machine specialists with a good background in mathematics, statistics, testing, and research methods, are not readily available, nor will they become so unless training programs produce them. Hence one of the functions of the larger service bureaus is to provide a program of instruction in machine methods to tie in with backgrounds in other fields. Courses in basic punched card methods, in numerical analysis, in computer programming, and graduate assistantships for the service bureau are ways in which this training program can be organized. Courses and programs are now in existence, but the large need for such training—greatly increased, it should be pointed out, by the availability of electronic computers—has become apparent just in the past few years. The next decade will see enlargement of the general role of the service bureau such as to actively integrate it into the academic program of the university.

To summarize, I have described the University Service Bureau as an autonomous agency within a university whose primary mission is to provide machine services to all departments. The distinction between this and other agencies that provide machine services in a university lies in the administrative organization and the primary mission for which that administration is responsible. Machine services today include not only the routine statistical operations that can be carried out on punched card accounting machines, but the more complex operations that can be carried out by electronic punched card computers. Machine services include a consulting function provided by personnel trained in statistical methods and research procedures as well as in machine methods; and it includes the function of training persons to operate at this consulting level. The cost of staffing and equipping such an agency is high, not all institutions can afford it. From this stems the notion that the service function must ultimately extend outside the University to other institutions. Finally it is my personal conviction that the concept of a machine services agency is firmly implanted and that it is now a most necessary and vital part of the total research facilities of a University.

Impact of Machines and Devices on Developments in Testing and Related Fields

LEDYARD R TUCKER

USE OF ELECTRONIC COMPUTING MACHINES FOR TESTING PROBLEMS

IN PREPARING this paper concerning possible impacts of electronic computers on testing problems and practice I had some difficulty in framing the introductory section. I wanted to insert a disclaimer about being expert on these machines without implying that I might not have anything more to say in the allotted ten minutes. Progress in the development of these electronic devices is the result of masterly work of an expanding group of mathematicians, physicists and engineers. Some of this progress has been spectacularly reported in the press as development of "giant" brains." Having several problems in psychology and testing for which I was unsatisfied with our present methods and answers, I have approached the field of these "giant brains" to try to elicit some help. I have succeeded in being introduced to a few of these electronic wizards and have the beginnings of a speaking acquaintance with a couple. In talking today I wish to make no pretense at being an expert on electronic computers, but rather to approach the topic of possible impact from the view point of a psychologist interested in quantitative methodology.

I am going to limit my remarks to general purpose machines as contrasted to such special purpose devices as differential analysers or scoring machines.

An initial point I wish to make about the electronic computers is one with which many computer experts will agree. If these devices are "brains," they are simple minded in the present state of development. They perform such automatic logic as the computational processes or addition, subtraction, multiplication, and division and make simple decisions based on inequalities in intermediate results. Consider a single clerk who could perform the computational processes and could follow explicit, coded, step by step directions including conditional directions based on inequalities of numerical results. This clerk is provided with a slate on which to write. Consider also that a file of three inch by five inch cards is available to this clerk. These cards are to be numbered so that each one may be referred to by its code number. At any patricular time, one, and only one, ten digit number may be recorded on each card. Whenever our clerk records a result on one of the cards, he erases any previous number on the card.

There will be two types of numbers, direction codes and problem numbers. Directions to our clerk will be coded in such a fashion that each of the several numerical and decision operations has a unique code number occupying

the first two digits of the directions code numbers. Code numbers for the cards having the problem numbers to be used are recorded in the later section of the direction code numbers. In the case of decision steps, the card code number part of the directions code will be the code for a card containing directions for a conditional step. Normal progress of the clerk in directions executed will be consecutively coded cards.

Our clerk is now set up to perform the functions performed by an electronic computer. Before our clerk can go to work, however, a program of directions has to be devised. This is the step where brains are required. I am tempted to say plain, ordinary human brains, but this would not be quite true because special talents are involved. Extensive, detailed, and painstaking analysis of the general problem to be solved is required in the preparation of a program of directions. This program has to be self sufficient so that once the machine is started it can continue through the successive steps without human intervention. Much time is taken checking each program, trying it out on sample problems, and checking results so that program errors may be detected and corrected.

Once a program has been devised, the electronic machine has a terrific speed advantage over our clerk and file of cards. It is also less likely to make random errors. These advantages along with reduction of tedious labor performed by human beings provide the major assets of these machines. Probably the most valuable, as well as spectacular, consequence of these assets is the utilization of electronic computers for solution of problems which were otherwise practically insoluble because of the extensive computations required. With the use of electronic machines, scientists may extend the horizon of problems attempted in the direction of problems requiring extended computations as well as certain other automatic logical operations.

In my analogy of a clerk and file of cards I omitted mention of two important matters: the original recording of the program and data on the cards and the preparation of a report of results. Input and output aspects of present machines are serious problems. Insofar as psychological and educational testing involve much data, simple computations, and considerable results to be reported, the input—output problems are most aggravated. This raises the problem of proper balance between various machine facilities so as to provide the most usable services for testing problems. Is a large, fast, and expensive computing unit justified when coupled with relatively slow input—output units? Would several smaller and slower computing machines, each coupled to the present type of input—output units, be more economical? I have been pleased by information concerning development of such machines under a title of data processing machines. An alternative solution is the connection of parallel input—output units to one large machine. This is also done.

Those of you who are acquainted with the electronic machines will note that I have not emphasized the size of "memory" problem. This is the size of the card file and relates to the number of numbers that can be maintained in the machine. This is also critical for many testing problems involving much data. Machine developments, however, seem to be solving this problem.

Let us try to evaluate the impact of these machines on testing. Some work has already been done, especially in factor analysis. Charles Wrigley and Jack Neuhaus at the University of Illinois have reported determination of principal components. Frederic Lord of Educational Testing Service is performing a maximum likelihood method fac-

tor analysis on the computer at Massachusetts Institute of Technology. Extensive operations similar to those encountered in testing have been conducted on an electronic machine by the United States Bureau of the Census. Some consideration has been given to possible use of these machines for operations now being performed in testing studies. It seems possible that large tables of intercorrelations can be computed more economically by large electronic machines than by present methods. Whether other analysis operations can be performed more economically must be determined by further study.

It is my opinion that electronic computers will have their main impact on testing in terms of new procedures. Some of these procedures are now automatically discarded, others are not even conceived. We may now extend the horizon of our conceptions. New, and more distant, limits on feasible analysis now exists. We may now use theoretically preferred methods which have been discarded as too complex and attempt solution to problems now left either only partially solved or unsolved entirely. For example, selection of item pairs, triplets, and larger sets which will yield a maximum test correlation with some given criterion involves such extensive computations that approximate methods are now used. John Keats and David Saunders of the Educational Testing Service are investigating the feasibility of explicit solutions on an electronic computer.

As a general summary, it is my opinion that the impact of electronic computers on testing is a matter for the future. It will depend on the fertility of our imaginations.

Impact of Machines and Devices on Developments in Testing and Related Fields

HARRY H. HARMAN* AND
BERTHA P. HARPER

AGO MACHINES FOR TEST ANALYSIS

DR. TUCKER has spoken to you about electronic high speed computers and suggested the extent to which they might be adapted to the special needs of testing in the field of psychology. Of the machines in The Adjutant General's Office, Department of the Army, that I will discuss this afternoon, one is in the family of the electronic computers. A second group of machines, while unique, involves more conventional punched card methods. Although my topic will cover heterogeneous types of machines designed for quite distinct problems, nevertheless, they comprise a single entity with respect to problems of test analysis.

For the sake of our present discussion, these problems might be put in two classes, namely (1) study of batteries of tests for the determination of basic psychological factors, and (2) study of individual tests via item studies for the development and improvement of such tests. The former type of problem often leads to extensive analysis employing methods of factor analysis. While we do not have especially powerful facilities for the reduction of a matrix of correlations among many tests to an adequate factor matrix, we do have special facilities for the formulation of psychological hypotheses resulting from factor analysis.

Our special machine, known as the Factor Matrix Rotator, is designed to facilitate the work of going from an initial factor solution to a desirable final factor solution (frequently described as "simple structure"). It may be of interest to some of you that the machines Dr. Tucker discussed are of the digital variety the AGO machine is of the analog type, that is, the readings are in the nature of displacements along a scale so that the figures have to be estimated from calibrations on the scale rather than read as precise digit values.

The initial factor weights are set into the machine by means of a series of dials; then the positions of the points representing the tests are viewed as points of light on a scope equivalent to a 17" television screen. At the beginning of the work, the dials representing the transformation matrix are set for the identity transformation so that the researcher can first view the plot of the points as they appear in the initial (mathematical) factor solution. The axes are rotated by a simple manipulation of a dial; when desirable positions of the axes are located, the researcher can cause the appropriate elements in the transformation matrix to be reset to take the new positions into account.

*Since October, 1953 associated with Rand Corporation, Santa Monica, California.

Of course, the plots of points viewed on the scope are in two-dimensional space, that is, plots are exhibited for a pair of axes at a time. An immediate advantage over hand methods arises from the fact that while a decision is being made regarding the rotation of a particular pair of axes, it is possible to view each of the remaining factor axes, in turn, with each of the two under consideration. The usual procedure employed by researchers using this machine is to view the plots in relation to all possible pairs of axes and on a schematic chart to note those planes in which rotation is most desirable and those that might be considered in order to keep in number of rotations at a minimum.

When final decisions have been made about the location of factor axes to exhibit simple structure, the elements of both the transformation matrix and the final factor matrix can be read out as fast as the experimenter can turn the dial and read a scale, since the computations are done electronically.

The Factor Matrix Rotator is designed primarily to handle problems of up to 50 tests and 12 factors and involving orthogonal rotations. The machine has already been applied to problems involving up to 130 tests and 24 factors. Further, it has been employed in connection with oblique rotations. However, applications beyond the basic capacity of the Rotator necessarily involve extensive additional computations off the machine.

The primary advantage of the machine is that of speed, sometimes as much as 50 to 1 in comparison with hand and desk calculator methods. As a result of this speed, other advantages accrue, such as the ability to achieve a more satisfactory final product because rotations and examinations of them can be carried out with such ease.

As indicated earlier, this paper is concerned with two types of problems. The second, covering several machines, involves item analysis of tests.

The source machine for our item analysis work is the Document-to-Card Punch. This device consists of several components—first, a test scoring machine chassis containing a sensing unit and plug board. The IBM answer sheet is fed into the hopper of this machine. Then a tabulating card, with punched holes corresponding to the item responses, is prepared by the second component, which resembles a modified IBM reproducer. Unique, identifying information for each answer sheet is transcribed into the punched card concurrently by means of the third component, a manually operated keyboard.

The Document-to-Card Punch is designed to accept item information in 60 columns, allowing the 20 remaining columns of the punched card for identification, background, and criterion information. Of course, many more than 60 items can be put in the 60 columns. For a typical Army test involving four alternative responses per item, a total of 180 items can be placed in the card at a single pass of the papers through the machine, by means of triple punching each column. Within this standard of four alternative positions, special checks have been built into the machine for the detection of double responses and omits. For tests involving five-choice items, true-false items, or any type other than four-choice, the Document-to-Card Punch may still be used but not all the checking features can be applied.

Considering speed of operation, it should be recognized that the machine is controlled by a human operator. Optimum exploitation of the machine arises from the reading of a maximum of item information in one pass of the answer sheets. Thus, recording the responses to a test of 150 four-choice items by means of this machine, we estimate that the task can be done in

1/7th the total time it would take a crew of key punch operators to do it manually.

While item information would not ordinarily be multiple punched in cards when using conventional key punches, having such information prepared on the Document-to-Card Punch does not at all distract from the usefulness of the tabulating cards. At the Personnel Research Branch, the multiple punched cards are used for the study of item difficulty and discrimination on such conventional IBM machines as the Type 101 Electronic Statistical Machine, Type 602A Calculator, and Type 407 Tabulator.

The Document-to-Card Punch has been found to be a very useful machine for problems of item analysis in the AGO over the past five years, first in a rather crude experimental form, and during the past two years in an improved model. During the course of our experience with this machine, we found a great need for another piece of equipment—something that would provide an efficient means of getting a score into a punched card which already contains item responses. A machine was designed by us and built by IBM which does precisely that. A stack of cards, punched with item responses, is placed in the hopper; the key for the test is punched in a master card; then, completely automatically, the individual cards are fed, matched against the key card, and punched with the number of instances of agreement with the key.

This special machine is known as the Card Scoring Punch. It is designed primarily to supplement the Document-to-Card Punch, but, of course, can also be used to score punched cards that are prepared by other means. We do not propose this machine as a substitute for a test scoring machine. Its efficiency in scoring arises when item responses of a test are already in punched card form. Then, of course, the automatic scoring is very fast. It operates at a speed of about 400 cards per hour, with no more attention on the part of an operator than that of placing item cards in the hopper and removing them after scoring. The score obtained for each card is punched automatically in any one of six fields selected in advance by merely setting a switch. By replacing the key card it is possible to rescore the same cards, thus obtaining several different scores and recording these in the same cards. "Rights" and "wrongs" can be obtained separately; then any scoring formula can be applied by means of a pass of the cards through the 602A Calculator. The ease with which various scores can be obtained from a given set of item responses opens the way for new avenues of research in test development in fields such as personality measurement. When trial scoring can be accomplished at such low cost, the researcher can easily afford to experiment with various methods of keying non-cognitive data.

In summary, the Personnel Research Branch has a broad mission, that of developing techniques to aid the Army in its problems of selection, classification, and evaluation of men. To accomplish the statistical analysis inherent in this research most expeditiously, we have developed and put to use certain unique machines. Among such are those discussed here: the Document-to-Card Punch and Card Scoring Punch, which are useful in large volume item-analysis, and the Factor Matrix Rotator, a special electronic device to expedite an important phase of factor analysis. All of these have already proven extremely useful to The Adjutant General's Office in test analysis.

Impact of Machines and Devices on Developments in Testing and Related Fields

ELMER J. HANKES

NEW DEVELOPMENTS IN TEST SCORING MACHINES

I WISH TO THANK the sponsors of this meeting for the opportunity to tell you about Testscor's scoring machines. We have several kinds and I would like to tell you what they do and something of how they do it. Please bear in mind that these are all existing machines.

First, let's consider the machine for scoring the Strong Vocational Interest Inventory. This is a completely automatic device—the operator inserts an answer sheet which is scanned by a resistive type pick up. This pick up transfers the pattern of responses to a battery of 1,200 double throw switches which are in turn connected to a grid of resistances. This grid is a fixed memory containing all the scoring weights for the Strong Inventory.

The 1,200 switches indicate "selected" or "rejected" for each of the inventory questions and put voltage on the resistance grid. Current proportional to the student's score is fed from the grid to a group of 50 Electronic Algebraic Analysers. The Analysers then operate a printing mechanism which plots the profile of the student's interests.

The time required to score a test is thirty seconds. It has been computed that this machine is equivalent to over 70 of conventional type scoring machines for doing its particular job and is an illustration of the value of the special purpose machine where conditions make such a device economically feasible.

This machine is approximately 4 feet wide, 8 feet long, and 4 feet high. It has 300 vacuum tubes, over 100,000 soldered connections, and requires 3 kilowatts of electricity. It also has built in checking systems but frequent accuracy checks are made. Being an analogue device, accuracy is limited but the error is usually only one to two standard points.

I mentioned that a resistance type pick up was used—we have found that the dependence upon the student for making a usable conductive pencil line is extremely unsatisfactory. Therefore, we have solved this problem by remarking all Strong tests with a conductive ink made of a colloidial suspension of graphite in water and marketed by Acheson Colloids Company under the trade name of Aquadag.

Our experience with the above and knowledge of the problems inherent in the resistive pick up lead to the development of a very simple and reliable optical pick up which was incorporated into our next major machine. We call this one TUSAC for Testscor Universal Scorer and Computer. TUSAC is really designed to score test batteries, rather than single tests. It is too big and complex to do simple tasks. Its economic worth lies in its ability to take an answer sheet, optically scan *both* sides

simultaneously, gather the scores, counting both rights and wrongs, convert these scores, weight and combine the conversions and print out the resultant index scores. Single or multiple answer sheets can be used—all the clerical work is done in the machine. Fatigue and resultant errors are eliminated. TUSAC is a digital counter and is completely accurate.

A study made at a large military installation showed that a single TUSAC machine running ten hours a day with two operators could replace a staff of 100 men actually required for an emergency situation. Under normal operations, in addition to the man power savings, the savings in paper alone was estimated at $14,000 a year with considerable savings in space needed for storage of records, supplies, etc.

The TUSAC answer sheet is modelled after the IBM. The vertical spacing of the response positions has been retained, but the horizontal spacing has been doubled from 8 per inch to 6 per inch. Marking has been simplified, as any kind of soft pencil or ink can be used by the student. Because of the optical pick up there are none of the blank spaces required by machine structural consideration and the entire surface of the answer sheet is usable. This gives over 2,100 available response positions per side. Therefore, several tests can easily be placed on one answer sheet.

At this time I would like to suggest the adoption of a standard format for answer sheets so that the various machines now building or under consideration will be universally useful. TUSAC is designed for flexibility in this regard and scores both Testscor and IBM answer sheets.

TUSAC is essentially a simple machine. It counts marks on a sheet of paper. It counts them one at a time, just as you would when hand scoring. As it scans each mark it determines, by comparison with previously aquired instructions, whether the response is correct or not. If it is right, it adds one to the accumulating total—if wrong, the count will be affected as per previous instructions. A fourth or a third of a point may be deducted from the score —Simple—You've done it yourself. The big difference is accuracy and speed. TUSAC is not affected by humidity or cheated by stray marks or smudges. It doesn't tire and make clerical errors. Its scores are converted and printed automatically. The output of TUSAC can be readily coupled to punch card equipment if desired.

The scoring rate of TUSAC is 100 complete test batteries per hour at present and we expect to double this rate as we gain experience with the machine.

Dimensionally, TUSAC is not large. The size is only 3 feet by 6 feet by 7 feet high. Power consumption is 4 kilowatts and 800 vacuum tubes are used.

TUSAC is a special machine and here again we have the case of a machine being economically advisable only when there is a large volume of work suitable to its abilities. Through cooperation with various test publishers, we hope to make TUSAC a boon to you through reduction of test scoring costs.

In my analysis of the test scoring field I reached the conclusion that the primary consideration for scoring educational tests was *accuracy—Dependable accuracy equivalent to that obtained by repetitive hand scoring.* Next, was simplicity and reliability of the scoring equipment. Repairs, when necessary, should be effected by a typewriter or adding machine service man. Third, scores should be printed. Too often a tired operator transposes the score, and reading a meter is not an accurate procedure.

Answer sheet and student marking requirements should be simplified to permit schools to make answer sheets

for tests of their own and to eliminate the need for special pencils.

The cost of the device should be under $500 to permit even smaller schools to take advantage of the time savings of test administration and scoring and remove some of the teachers' burden.

Portability would be an advantage, as it could then be used in any room or passed from school to school in a district.

Automatic conversions to standard or percentile scores would be desirable as would be accumulation of number of tests scored and accumulation of totals for a group of tests of both converted and raw scores. Could this be done too at a reasonable cost?

The answer is DUS—which stands for Digital Universal Scorer. Like the soap of a similar name—DUS does everything. Yes, DUS does do all of the things outlined. Lets review them:

Accuracy—DUS is a digital device with optical pick up.

Service—DUS is made up of simple rugged components, most of which have stood the test of time. Any typewriter shop can service it. Scores are printed—Raw or Converted. Uses any kind of answer sheet and they may be mimeographed from special stencils. No special pencils are required. It's portable—weight is under 50 pounds cased. Separate visual registers accumulate the number of tests scored and the total of scores.

DUS will be available for distribution in September. The price will be $385 to $650 depending upon the accessory items desired.

DUS does not do item counting. We are making a separate device for this work which will be digital in accuracy and have a capacity of 30, 60, or 90 items per pass through. The basic price of this unit will be around $2,000 and it will be ready in February or March.

That covers our work and I hope will cover your requirements in the test scoring field. I want to thank you all for your kind attention and invite all of you to visit Testscor when you are in our area. Our staff will be delighted to see you. Thank you.

Impact of Machines and Devices on Developments in Testing and Related Fields

E. F. LINDQUIST

THE IOWA ELECTRONIC TEST PROCESSING EQUIPMENT

THERE IS NOW being installed at the State University of Iowa a new "electronic brain" for the processing of objective tests and test data. This equipment is designed to perform at one time practically *all* of the clerical and statistical operations—scoring, tabulating, computing, and reporting—involved in wide-scale testing programs using relatively long multiple-test batteries, but it is capable also of performing many other clerical and statistical tasks involved in educational and psychological measurement and research in general. So far as I have been able to discover, the performance specifications for this new equipment are far in advance of those for any other test scoring machine or test processing equipment now in existence or now being developed elsewhere. Indeed, the specifications are such that I believe that equipment of this type will exercise a very important influence on the direction of future developments in the entire objective testing movement. My purposes in this paper are: (1) to tell you what this equipment will do, (2) to discuss the possible significance of this type of equipment for the testing movement in general, and (3) to explain the basis on which the services of this equipment will be made generally available to any testing organization or agency that may wish to make use of them.

Perhaps I should say, before going further, that the installation of the basic equipment in Iowa City will not be completed until almost a year from this date, and that considerably more time will be required to install all of the presently planned extensions to the basic equipment. In certain respects, therefore, I may be "jumping the gun" a bit in making this announcement here today. Had Walter Durost not selected the general topic that he did for this afternoon's program, and had he not asked me to appear on the program—to discuss quite a different subject—I would surely have waited at least several months yet before making news of this project public. However, I felt that if I were to appear on this particular program at all, I was professionally obliged to make the most interesting and significant contribution to it that I could—and that meant that I was obliged to tell you about the Iowa electronic test processing equipment.

The Iowa electronic test processing equipment consists essentially of a high-speed automatic test scoring machine, coupled with a special-purpose digital electronic computer and a fast output printer. A highly flexible and compact answer sheet design permits the answers to as many as 960 multiple-choice items—organized into as many as 14 different tests of any relative lengths—to be recorded on two sides of a single 8½" x 11" sheet. The exam-

inee marks the answer sheet by simply making heavy black dots in boxes corresponding to the answers he thinks are right. Any ordinary soft lead pencil may be used, and no more than ordinary care in erasing is required. The machine will automatically detect and disallow double or multiple marks, so that no preliminary scanning of the answer sheets will be needed.

The feeding of the answer sheets through the machine will be completely automatic. A stack of as many as 5000 sheets may be placed in the machine at once, and the operator will only have to stand by until the sheets have been processed. As the sheets pass through the machine, the marks will be sensed photoelectrically from both sides of the sheet simultaneously, the answers will be compared with the "right answers" stored in a magnetic memory, and the scores will be recorded in digital electronic counters. The scoring will thus be serial and digital in character, rather than of the analogue type, and will be extremely accurate.

If desired, the machine will count the number of "rights" and "omits" separately for each test, and will compute a weighted composite of "rights" and "omits" according to standard "correction-for-guessing" formulas. If desired, also, for a single test of not more than 120 items, the machine will provide a score in which the individual items are given different weights—as many as 14 different weights, with integral values of from 1 to 50, being permissible. Again, if desired, the machine will obtain scores, either corrected or uncorrected, for odd and even numbered items separately for each of as many as seven tests, so that reliability coefficients may be readily computed.

As the scoring proceeds, the machine will convert the raw score on each of as many as 14 tests on each answer sheet into a derived or scaled score, according to a different conversion table for each test. Thus, the machine will convert raw scores to normalized T-scores, to percentile ranks, to age- or grade-equivalents, or to scaled scores of practically any type now in use. Furthermore, the *same* raw scores may be converted to two or more different types of scaled scores simultaneously, so that, for example, the machine may report both grade-equivalent scores and within-grade percentile-ranks on the same tests at the same time. The machine will also secure weighted totals and subtotals of the test scores for each examinee, and will convert these totals and subtotals in turn into derived scores comparable to those for the individual tests.

An especially valuable feature of the machine is that it will "read" the examinees name and other information, either alphabetic or numeric, from the answer sheet. It will then produce printed reports of the names of the examinees and all their scores. The output printer will print a line of 93 characters, any 18 of which may be either alphabetic or numeric, for each answer sheet, or, if necessary, will print two lines for each answer sheet. The reports may be either in list form or in the form of individual report cards or profile forms for each examinee separately. Unless very many scores are to be reported for each answer sheet, the machine will print both list and individual reports simultaneously.

All necessary headings in the list reports may be printed automatically from blank answer sheets on which the desired headings have been entered in place of the examinees' names, these heading sheets having previously been placed in the stack of answer sheets at the appropriate points. Thus the machine works exactly like an IBM alphabetic printing tabulator in this respect, the answer sheets taking the place of the punched cards.

As the answer sheets are being

scored, the machine will *cumulate* each of the scores for groups of successive answer sheets, and will compute and print the *means* of these scores at the foot of the list report for each group. If desired, the machine will also cumulate and compute the means of the *squared* scores and of all possible *cross-products* of the scores for a limited number of tests, thus providing the basic terms needed for variances and for inter-correlation and reliability coefficients. Furthermore, the machine will tabulate graphic frequency distributions of the scores on each test for groups of successive answer sheets, so that percentile ranks may be readily computed as soon as the scoring is completed. Finally, the machine will prepare a graphic record of the number of times each response to each item has been marked on a number of successive answer sheets. Not only that, but these graphic item-response counts may be separately obtained for examinees making high and low scores on the test being scored, so that item-test correlations may be readily computed. All of these operations—scoring, transforming, cross-footing, cumulating and computing of means of scores, squared scores and cross products, printing of complete reports, tabulating of frequency distributions and item counting—all will be performed together during a single original run of the answer sheet through the machine.

I shall tell you in a minute at what rate these operations will be performed, but first let me say a few words about the scoring and converting process. Preliminary to scoring a group of answer sheets, a "key sheet" and a "conversion sheet" must be prepared. The key sheet consists essentially of a standard answer sheet on which the right answers have been marked in pencil in the ordinary fashion, and on which certain control information is entered in coded form by marking boxes. The conversion sheet looks just like the key sheet, but on it the conversion tables are entered in coded form, again by marking boxes. The key and conversion sheets are placed on top of the stack of answer sheets to be scored by them, and these are placed in the machine in the same pile with similar stacks for other test batteries. The machine "memorizes" each key and conversion sheet as it comes along, and scores the succeeding answer sheets accordingly, each key and set of conversion tables being automatically erased from the memory as new key and conversion sheets come along. Thus, practically no machine time is required for make-ready or for change-over from one job to another, and the machine is capable of practically continuous operation for long periods, even though many different jobs are to be handled during any period.

The basic rate of the equipment is 6000 answer sheets per hour. However, the number of tests scored or scores reported is greatly in excess of this number. In the case of the Iowa Tests of Basic Skills, for example, which constitute a 14-test battery with three subtotals, the machine will obtain, convert and print 102,000 separate three digit scores per hour. Even this, however, grossly understates the capacity of the equipment, since so much more than merely scoring is performed. Perhaps I can make the true capacity of the equipment more meaningful to you in terms of a specific comparison.

In the Fall Testing Program for Iowa High Schools we administer in September of each year a 56 per battery of nine tests to approximately pupils in about 400 high schools. All marked answer sheets are sent to Iowa City, where we do the scoring and statistical work and prepare three types of reports for each school. One is a typewritten report of the names of the pupils and their scaled scores, another is a

set of four profile cards for each pupil, with his name and scores typed across the top of each card, and the third is a school summary report of the mean of the scores on each test for each grade in the school. For these purposes we require a large temporary staff, provided with all of the presently available specialized equipment-such as electric typewriters, comptometers, automatic computing machines, electric accounting machines, and punched card tabulating equipment—that it is possible for them to use efficiently. I cannot take time now to describe our procedures, but I do not hesitate to claim that we do this work as accurately, quickly, and economically as the same kind of work is being done anywhere in the country, regardless of the type of equipment used. The staff consists of about 40 unskilled workers who are trained on the job, and about 20 skilled and experienced workers, such as typists, computing and accounting machine operators and punched card tabulating equipment operators. It takes this staff of 60 about five weeks to handle our program. The new equipment will do their work in three days, actually in 12 hours of continuous machine operation. I have estimated that to do the same work in three days by present methods would require at least 400 workers, provided, of course, with a correspondingly large amount of the kind of specialized equipment we are now using. Even this example, however, seriously understates the capacity of the new equipment, since in this example the equipment does not perform many of the operations of which it is capable, such as tabulating frequency distributions or cumulating variance and correlation data.

Perhaps this is a good time to remind you that what I have been telling you is what the equipment is *designed* to do, and not what it has yet done. Each of the component operations has been performed independently, either by our experimental "breadboard" units, or by units performing nearly identical functions in other electronic equipment now in operation elsewhere, so we have little doubt of the outcome. The construction of the finished units is well along, but their installation at Iowa City has just begun, and many problems of integration remain to be solved. Our engineers are extremely confident that the basic equipment will be ready for practical use a year from this fall—but they do admit that unanticipated "bugs" may cause delays. The basic equipment, incidentally, does not include the "omits" counter, nor the sums of squares and sums of products cumulators, nor the graphic item counters and frequency distributor. These will be added as soon as possible after the basic equipment is in operation.

You can see then why I said I may be "jumping the gun" a bit in making this announcement now. However, there are other reasons for doing so than the one I have just given. As I have already indicated, we are planning to make the services of this equipment available at a low cost to testing and research agencies generally. We hope to work out many of the details of this plan during this coming year, so that the plan will be ready when the machine is ready. This means that during the year we will wish to consult many of you who are here today, so as to give adequate consideration to all interests and points of view, and we want to give you a chance to think about the implications of this equipment before we consult you. Another reason for an early announcement is that it will give any of you who are interested more time to get ready for an early utilization of the equipment. I should guess that most test publishers or program directors who may wish to adapt their answer sheets and related materials to this equipment will require

from 6 to 18 months for the purpose, not including the time needed for careful consideration before taking such a move at all, so that for this purpose this advance announcement is none too early.

I shall tell you more later about the plans we have thus far worked out for making this equipment generally available. First, however, I'd like to provide you with a background for my later remarks by discussing briefly the probable significance of electronic test processing for educational and psychological measurement and research in general. These remarks will be concerned not just with this particular equipment, but with any equipment of this general type that exploits present possibilities in electronics. What I shall next have to say, then, will be said on the assumption that the services of equipment of this type can somehow be made available to anyone who needs them.

I don't think you will require much convincing that there is a real and urgent need for the kinds of services that electronic equipment can render. There are now in operation in this country a large number of wide-scale testing programs, in each of which relatively long multiple-test batteries are administered to large numbers of examinees on the same day or days or within a relatively short period of time, and in which it is highly desirable to report the results to the participating examinees and institutions in the shortest possible time. These include not only state, regional and national testing programs, but also programs conducted within individual institutions or individual city school systems, such as freshmen placement or entrance examination programs in large universities and colleges, or achievement testing programs in large city school systems. Practically all of these programs involve a very heavy peak load of clerical work which can usually only be handled by a temporary staff—a staff hired, trained, and used for only a short time and then dismissed, upon each recurrence of the program. The recruiting, training and supervising of these temporary staffs usually constitutes the most difficult problem in administering a large-scale testing program, and in many instances very definite limits have had to be set on the program on this account. In Iowa, for example, for our Basic Skills Testing Program for elementary schools, which is of twice the scope of the high school program I have described, it is utterly impossible for us to recruit in Iowa City as large a temporary staff as we need to process the test results centrally. We have, therefore, been forced to require the schools to do their own scoring and preparing of list reports—thereby placing a heavy and unwelcome burden on the teachers, who are required to do this work at no extra compensation. Similar situations exist, I am sure, in many other states. With electronic equipment the services provided to the schools in this and other programs could be greatly extended, and the cost and administrative inconveniences to the schools could be greatly reduced.

Not only would electronic equipment make possible great improvements in existing programs, but it should result in the inauguration of many new programs as well. The advantages to be gained through cooperative organization in educational testing are extremely important, and should be made much more widely available to the schools of the country than they are today. Through the cooperative administration of uniform batteries of tests at the same time and under the same carefully controlled conditions to large numbers of pupils and schools, it is possible for the schools to secure much more meaningful, more specialized and more up-to-date norms, to insure greater comparability in the test results from school to

school and from pupil to pupil, and to profit from the many economies made possible through large-scale operation, not only in test processing but in the construction, production and distribution of test materials as well. In the Iowa Fall Testing Program, for example, we are able to give the schools, at a cost of only 35¢ per pupil, services that they would be unable to secure at several times the cost if each school were to plan and conduct its own local program independently. I am sure there are many state agencies and educational institutions that would now like to make the advantages of cooperative testing available to the schools in their own states, but that have been prevented from doing so primarily by the lack of the personnel and equipment required. Electronic test processing, if generally available, should provide a new and powerful stimulus to the launching of many new testing programs of a wide variety of types, and should thereby greatly increase the scope of testing throughout the country.

Electronic test processing could also have a profound influence upon the methods of marketing standardized test batteries that are primarily intended for independent use by individual school systems — particularly relatively long batteries such as the Metropolitan and Stanford Achievement Tests, the Differential Aptitude Tests, and the Primary Mental Abilities Tests. Electronic processing may make possible the marketing of such tests in much the same fashion as Eastman Kodak now markets its Kodachrome film. That is, the tests could be sold with machine-scorable answer sheets on which the processing charges are paid by the consumer when he buys the tests. After administering the tests, the schools would simply send the answer sheets to a central service agency and quickly receive back again individual and summary reports of the results at no extra trouble or cost. Copies of these reports, or of the frequency distributions of scores, could be sent to the test publisher, who would thereby be enabled to establish and maintain very stable and up-to-date norms on the basis of the results from *all* current users of his tests, and at practically no extra cost to him. At the low processing costs which electronic equipment and this method of marketing should make possible, it seems likely that most schools would jump at the chance to avoid the annoying clerical burden which is otherwise placed on their teachers and administrative staff, and with attendant delays in reporting. Both in cooperative and in independent testing, then, electronic test processing could greatly increase the scope and popularity of objective testing in general.

Electronic test processing will perhaps be of greater interest to many of you here today because of the ways in which it will make possible new *types* of tests and testing services, or types that have heretofore been regarded as impracticable. Tests that involve very complicated scoring procedures, for example, such as weighted-item or weighted-response scoring, or tests in which a number of different scores are obtained from the *same* items or from overlapping groups of items by the use of different scoring keys and different sets of weights, have heretofore been generally impracticable for wide scale use, both because of high scoring costs and high costs of test development. Electronic processing will make these and even more complicated scoring procedures and developmental procedures as practicable as any others. Consider, for example, the possibility of providing tests, inventory blanks, attitude survey blanks, etc., on which the report to the consumer consists primarily of counts or average scores on individual items, rather than of over-all scores. What might be the market, for instance, for

an employee inventory blank for large scale industrial employers, or for a highly diagnostic curriculum survey instrument on which percentages of correct responses to individual items are reported for all individual classes and schools in a large school system? Such instruments and services cannot now be provided at a sufficiently low cost to make them saleable, but could easily become practicable with the aid of electronic processing equipment.

I have saved for the last what I suspect is the most intriguing subject of all for most of you here today, but which because of lack of time I shall only be able to mention. That is the possible influence of electronic equipment upon educational and psychological research in general. What will it mean to you, for instance, to be able to secure item analysis data, frequency distributions, group variances, reliability coefficients, and inter-test and inter-item correlations on experimental tests as the tests are being scored, at only a small fraction of the cost and, what is more important, in a very much smaller fraction of the time, than would now be required? What will be the result of similar savings in time and cost in the analysis of widely administered questionnaires, personality inventories, check lists, opinion survey field reports, etc.?

In making these comments, I have been assuming that the advantages of electronic processing can somehow be made available to everyone that can profit from them. How this may be done I can now only guess—but I do have some informed guesses to offer. I do *not* believe it will come about through any widespread duplication at other centers of the kind of equipment we are now installing at Iowa. There are two good reasons for this. One is that this type of equipment is too costly, as any of you will know who have given any attention to the cost figures for existing electronic computer installations. The other is that no *need* for more than a very f[illegible] stallations of this type.

It is very difficult to estimate what the final cost of the Iowa equipment will be. We have been very fortunate in having had available the resources and facilities of a large university, including the use of the machine shops of our colleges of engineering and the technical and consultation services of members of their staff, at practically no cost for the project, and the basic logical design of the equipment has cost us nothing in cash. I think I can safely say, however, that had we turned over our present performance specifications to any large electronic engineering firm and asked it to start from scratch to meet these specifications, we would have had to spend in the neighborhood of half a million dollars.

Quite obviously, therefore, it is not to be expected that this kind of equipment will be duplicated in very many other places in the country. It happens that the Iowa programs are large enough to justify the heavy expenditure required on their own account, without counting on outside business, but, as I have already indicated, there are very few other test agencies that can justify a similar expenditure on the basis of work already in hand. If the Iowa equipment is made available to them, even these agencies may find it difficult to justify similar installations for themselves. Accordingly, until electronic engineers find very much less costly ways of doing these things, if this kind of processing service is to be provided to the country as a whole, it will have to be on a highly centralized basis, from a very small number of installations.

Actually, in consideration of its amazing capacity, a single installation like that now being made at Iowa can; for a considerable time, take care of practically all of the wide-scale testing

programs in this country that are now able to take advantage of it. I have already indicated that this equipment can handle the Fall Testing Program for Iowa High Schools in about three days time, but it should be noted again that this is one of the largest of all testing programs now in regular operation. Most prevailing programs can be handled by this equipment each in a few hours time. For example, it should be possible to do all the scoring and reporting for the Medical Aptitude Testing Program of ETS in a single day, and yet have time for one or two smaller programs as well. From the point of view of the whole testing movement, therefore, it will for some time be difficult to justify even one additional installation of this kind, unless it be on a purely competitive basis.

I think I have said enough now to make it clear to you why we have been willing at the State University of Iowa to gamble so heavily on the development of this equipment, and why we have been so concerned about the problem of how to utilize this equipment to the best interests of educational and psychological measurement and research in general, rather than in our own interests alone. With these general interests in mind, we have established, for the management of this equipment, an independent non-profit corporation to be known as Measurement Research Center, Inc. The original assets of this corporation are to be the "free time" of the basic equipment which I have described in this paper. This basic equipment is being paid for and will remain the property of the State University of Iowa. The University and the Iowa Testing Programs will have prior claim on the use of the equipment, but will almost certainly not require more than ten per cent of its time. The remaining 90% or more is the "free time" which belongs to the Measurement Research Center.

As is stated in its charter, the purposes of the Measurement Research Center are exclusively beneficial, scientific and educational, and no part of the net earnings of the corporation may inure to the benefit of any private shareholder or individual. Incidentally, this is definitely not the kind of "non-profit" corporation in which large salaries are paid to its officers in lieu of profits. The principal officers of this corporation are all full time employees of the State University of Iowa, and there is now no intention that they will draw any salaries from the corporation.

The more specific immediate purposes of the corporation are research into, development of, and the beneficial utilization of labor and time-saving devices in the field of educational and psychological measurement. The first claim upon the net earnings of the corporation will be for research and development leading to the improvement and extension of the original equipment, or to its duplication elsewhere if other service centers seem desirable. We hope in this way to build up a general purpose computer laboratory facilitating many types of educational and psychological research and development.

What is left of the net earnings of the corporation after these immediate purposes are served will be devoted exclusively to educational research. However, it is by no means our intention to maximize the net earnings of the corporation for these research purposes. On the contrary, our main ultimate objective is to encourage the growth and improvement of measurement in general by *reducing its cost to the consumer* as much as possible. We hope from the beginning to offer our services at no more than half the cost at which the same services can be obtained by any other present means, and we expect to reduce the costs still further as the scope of our operations increases.

You may, perhaps, be interested in a few facts about the development and construction of the equipment. The electrical and electronic components are being constructed in Cambridge, Massachusetts, under a contract with the Educational Research Corporation, and under the direct general supervision of Professor Philip J. Rulon of Harvard University. Fortunately for us, ERC is small enough to be able to bypass the usually expensive and time-consuming red-tape of larger organizations, and yet is able to devote to the project the very best of engineering talent, of whom I should mention particularly Mr. Earl Krohnand, Mr. George Hite, the senior engineers of ERC, Mr. Andrew Veranais, and Mr. Robert Edberg, the project engineer for the State University of Iowa. The mechanical components have been constructed or adapted by the Mast Development Company of Davenport, Iowa, and in the shops of the College of Engineering of the State University of Iowa. We have been very fortunate also in having secured the sympathetic cooperation of Remington-Rand, Inc. Remington-Rand has made available to us an output printer almost ideally suited to our purposes, and has helped us in the procurement of other vital equipment also. The basic logical design of the equipment is my own, but I am very deeply indebted to Professor Rulon for having induced me to look for a thoroughgoing electronic solution to the problem, and for having suggested certain general design features. Major credit must go, of course, to our engineers who have succeeded so well in translating this logical design into workable electronic terms.

I should like to say, in closing, that I have felt justified in utilizing my time on this program as I have, because, to the best of our intentions at least, the Measurement Research Center is to be regarded as the common property and in the common interests of practically all of you here today. Next to hoping that the equipment will work at all, I hope that it will work to the lasting and significant benefit of educational and psychological measurement in general.

Impact of Machines and Devices on Developments in Testing and Related Fields

PHILIP H. BRADLEY

SPEAKING FOR INTERNATIONAL BUSINESS MACHINES

I APPRECIATE very much the opportunity of attending this conference and particularly of appearing on this panel to tell you of IBM's interest and activity in the test scoring field.

We have heard many outstanding discussions and papers today on test scoring machines, testing procedures, new approaches to the test scoring problem, and high speed electronic computing machines. Following so many fine presentations, my remarks will be limited to a few statements of fact regarding our position in the test scoring field.

First of all let me restate a basic factor which establishes IBM in the testing field. We in IBM are not testing experts, guidance counsellors, or psychologists. We are a manufacturer of technical precision machines designed to relieve the individual of many manual repetitive processes involved in computation, accounting, and record keeping. We are no more trying to set a pattern for testing procedures and scoring than we are attempting to limit the applied scientist in computation methods. All of you are aware, I am sure, of the many developments in this latter field which we have made in the last few years. To name some of them — the 604 Electronic Calculating Punch, the Card Programmed Calculator, the 701 Electronic Data Processing Machine, and the 650 Magnetic Drum Calculator.

Many of these machines can be and are being used in processing certain operations in the test scoring field. The capacities of these machines are tremendous, their speeds extremely high, and their practical applications almost without limit. But we fully realize that these machines, in their present form, are not the answer to the majority of your problems. Most significant of the restrictions which make them impractical for your needs are, first—the tremendous expense of the machines, and second—the time involved in setting up and programming work for them.

Many of you are using some of our standard punched card units in your scoring, item analysis, and evaluation operations. But we still know that these machines, at least at present, are not the ideal solution to the problems.

We have studied the needs of the large testing agencies and the smaller ones as well. We think we know what it is that you want. At least, we know what you have asked for.

Before going any further with some problems and ramifications involved in designing a new scoring machine, I would like to make an announcement. You may have heard rumors—I have talked with many of you over the last

two years. We have encountered some delays due to defense engineering projects. But—in our laboratories now, our engineers, are working on a vastly improved test scoring machine. The specifications for this machine have been drawn up after studying the problems with you. Those of you with whom we have discussed these have given us your approval of them.

We sincerely believe that this machine will give you the needed results. You will find that it will incorporate those features which are presently lacking in our existing machine with some additional ones to expedite the scoring operation. In principle, it is similar to our present machine. But the improvements which will be made will give you a far superior product and most assuredly you will find that its results will give you the accuracy you need.

The first question you will probably ask is why have we decided to improve the present type machine rather than come up with an entirely new operating principle. Several influencing factors have decided this. First—you need and want an improved machine within the foreseeable future. Although we are working on new sensing and recording principles, they are not yet at a point where they can be made available. Second — improved circuitry design makes possible now highly reliable and accurate analogue computing mechanisms. This is evidenced by the fact that some of the latest high speed and large capacity computing machines in use today are analogue machines. Third —it is essential that we produce a machine which will be economically and financially acceptable to you. I need not review with you the fact that since we first introduced a test scoring machine in 1937 engineering and production costs have increased to a point where comparisons are staggering. Most of us would like to be the proud owners of a Cadillac but I dare say that probably most of us are using an automobile which costs half as much to buy and operate. It serves our needs quite well even though the horsepower rating is about half of the other and in general dimensions it is considerably smaller.

IBM is perfectly capable of building a machine which will meet practically any requirements specified. However, there are definitely some factors which must be taken into consideration. Large capacity machines are very costly not only to build but also to operate. Similar to large production equipment, these so-called "giant brains" must be able to support themselves. Can the testing field as a group support the development of a machine which affects primarily only the larger organizations? What about the smaller organization currently using one or two of our machines, or perhaps sharing one machine with several others? It is interesting to note that these smaller users of test scoring machines constitute the greater majority of the total number of machines in use today.

Even though the argument might be presented that a few large scale machines, strategically located, would be able to process all machine scoring throughout the country to better advantage, we can point out many impracticalities in this idea which more than offset the advantages. We and other organizations have established large computing and accounting centers throughout the country with the most powerful machines known to handle every type of problem. However, we have had no indication that these machine centers are replacing any numbers of smaller installations. So it is with our test scoring machine. Even though many of you are offering test scoring services on a wide basis, our sales indicate that the single machine users list continues to increase. We are enthusiastic with you about the

future growth of the testing field. We, therefore, believe that the need for an improved low cost machine that can serve the large user and the small is the first important step in the development program. I would like to reemphasize also that this machine is what *you* have asked us to make.

May I emphasize also that this machine does not represent our final interest in this project. If it were I would not have recommended that we even consider building it. IBM will build almost any machine asked for on a special contract basis where funds for such a project are available. We may have discouraged some of you on such projects at times when engineering and production facilities were not available. However, we are interested in such projects for your field as well as others.

But our own research and development projects continue in our laboratories, and we know that some of them will definitely affect your interests. We are working on new devices and principles which will be standard equipment in the future. A machine capable of functioning as a scoring machine alone and also which could be tied in with other standard data processing units is certainly the ideal. We believe that future scoring machines will be forthcoming in our normal document scanning and sensing, computing, and data processing machine development program. These will be the answers to your many needs, both practically and economically.

Now, what are some of the improvements of our proposed new machine?

First and most needed — answer sheet mark discrimination. The ability of the machine to determine those answer sheets which are not machine scoreable along with other sheets in a group.

Second—an improved scoring key.

Third—improved calibration and stability of circuits to eliminate drift and continuous checking.

Fourth—increased meter capacity and scoring combinations.

Fifth—improved feeding and faster operation.

Sixth—simplified operation, reading and marking of scores.

Seventh—overall improvement in reliability, both as to operation of the machine and service.

There are others which have to do with the engineering technicalities, but all of these will, I assure you, lead to a tremendously improved machine.

We have considered other features such as automatic feeding of answer sheets and printing of scores, as well as automatic programming for formula scoring. We will be happy to consider these further with any of you who feel you would like to have them. But for the present they will not be added as standard units. This is in the interests of cost to you and because we are optimistic about how some of our future developments will be able to serve you. We are also desirous of getting the new machine to you as soon as possible. When laboratory models are completed they will be placed on test in the field and at that time we will ask your assistance in evaluating the machine.

We are glad to see some other new test scoring machines being introduced to the field. It would be most presumptuous of us to think that we were the only ones capable of building a marketable scoring machine. With these new machines more people will be working on and suggesting machine improvements and design changes. More creative ideas will be offered to produce scoring machines which will better serve you. And naturally, the more different operating principles employed the more complete will be the coverage of the machine scoring field.

Finally, we in IBM want you to

know that we will continue to be interested in assisting you. We hold our association with the testing field in high regard. We appreciate the help and assistance you have given us and we look forward to continuing our fine relations with you. Your suggestions will always be welcome and we are confident that with the close cooperation which exists between us we will be able to produce for you the machines which will answer your future needs.

It has been a great pleasure for me to be with you today and although my acquaintance with members of this group has been growing rapidly for the last two years, this meeting has been of tremendous assistance to me. I am looking forward to my visits in your offices and I hope you will continue to call on us for assistance when problems arise and when you feel that we may be able to improve our machines and service to you.

Appendix

Participants—1953 Invitational Conference on Testing Problems

ABBOTT, Frank C., American Council on Education
ADAMS, Joe Kennedy, Bryn Mawr College
ADKINS, Dorothy C., University of North Carolina
AFFLERBACH, Janet, Professional Exam. Service
AHMANN, J. Stanley, Cornell University
ALLEN, Margaret E., Portland Public Schools
ALLISON, Roger B. Jr., Educational Testing Service
ALMAN, John E., Boston University
ANASTASI, Anne., Fordham University
ANDERHALTER, O. F., St. Louis University
ANDERSON, Paul R., Pennsylvania College for Women
ANDERSON, Rose G., The Psychological Corporation
ANDERSON, Roy N., North Carolina State College
ANDREW, T. G., University of Maryland
ANGELL, George W., Educational Testing Service
ANGOFF, William H., Educational Testing Service
ANTHONY, C. William, Maryland State Department of Education
APPEL, Valentine, Richardson, Bellows, Henry, & Co., Inc.
AQUINO, José G., University of Puerto Rico
ARMSTRONG, Charles M., New York State Education Department
ARMSTRONG, Fred, Lehigh University
ARNOLD, Samuel T., Brown University
ARSENIAN, Seth, Springfield College
AVAKIAN, Rose, Columbia University
BAKER, Paul C., Purdue University
BALTHAZAR, Earl E., Science Research Associates
BANNON, Charles J., Crosby High School (Waterbury, Conn.)
BARDOCK, Herbert D., New York State Dept. of Civil Service
BARNES, Paul J., World Book Co.
BARTNIK, R. V., Educational Testing Service
BAYROFF, A. G., Personnel Research Br. AGO
BEARDSLEY, Katherine, USDA Graduate School
BEARDSLEY, Seymour W., John Martin Associates
BEELER, Nelson F., State University Teachers College (N. Y.)
BEMENT, Dorothy M., Northampton School for Girls
BENEDICT, G. G., Phillips Academy
BENNETT, George K., The Psychological Corporation
BENSON, Arthur L., Educational Testing Service
BERDIE, Ralph F., University of Minnesota
BERGER, Bernard, Municipal Civil Service Commission (N. Y.)
BERGESEN, B. E., Personnel Press, Inc.
BISHOP, Ruth, National League for Nursing
BLAESSER, Willard W., University of Utah
BLANCHARD, Carroll M., The Signal School, Fort Monmouth
BLAUL, R. Elizabeth, Highland Park High School (Illinois)
BLOOM, B. S., University of Chicago
BOBBITT, Joseph M., National Institute of Mental Health
BOLLENBACHER, Joan, Cincinnati Public Schools
BONNER, Hubert, Columbia University
BRACA, Susan E., Archdiocesan Vocational Service
BRADLEY, Philip H., International Business Machines Corporation
BRANDT, Hyman, American Occupational Therapy Association
BRANSFORD, Thomas L., New York Civil Service
BRASTED, F. Kenneth, National Association of Manufacturers
BRAY, Douglas W., Columbia University
BRIDGES, Claude F., World Book Co.
BRISTOW, William H., Bureau of Curriculum Research
BROENING, Angela M., Department of Education (Maryland)
BROLYER, Cecil R., N. Y. State Civil Service
BROOKS, Richard B., College of William & Mary
BROWN, Frederick S., Great Neck Public Schools (N. Y.)
BRYAN, Miriam M., Silver Burdett Company

BRYAN, Ned., Rutgers University
BUCKTON, LeVerne, Brooklyn College
BUNKER, Harris F., University of Puerto Rico
BURDICK, Eugene I., Carnegie Corporation
BURKE, James M., Darien Public Schools (Connecticut)
BURKE, Paul J., Bell Telephone Laboratories (N. Y.)
BUROS, Oscar K., Rutgers University
BYRNE, Richard Hill, University of Maryland
CAPPS, Mrs. John W., South Carolina State A & M College
CAREY, Robert E., Board of Education (Yonkers, N. Y.)
CARLSON, C. Raymond, Air Command and Staff School, Alabama
CARLSON, Harold S., Upsala College
CARROLL, John B., Harvard University
CASE, Ethel E., Polytechnic Institute of Brooklyn
CAYNE, Bernard S., Boston University
CAYNE, Helen M., Massachusetts Institute of Technology
CHASIN, Joseph B., New York State Department of Civil Service
CHAUNCEY, Henry, Educational Testing Service
CHURCHILL, Ruth, Antioch College
CLARK, Willis W., California Test Bureau
CLIFF, Norman, Educational Testing Service
COBB, William E., Pennsylvania State College
COCKLIN, John H., Temple University
COFFMAN, William E., Educational Testing Service
COHEN, Philip S., Montclair State Teachers College
COLEMAN, William, University of Tennessee
CONGER, Louis, N. Y. State Education Department
COPELAND, Herman A., Atlantic Refining Company
COWLES, John T., Educational Testing Service
CRAWFORD, Barbara, Educational Testing Service
CRISSY, William J. E., Queens College
CRONBACH, Lee J., University of Illinois
CROOK, Frances E., World Book Company
CROWLEY, Harry L., State Teachers College (Massachusetts)
CUMMINGS, Allana, The Psychological Corporation
CURETON, Edward E., University of Tennessee
CURETON, Louise W., Knoxville, Tennessee
CYNAMON, Manuel, Brooklyn College
DAILEY, John T., Bureau of Naval Personnel
DALY, Alice T., N. Y. State Education Department
DAMRIN, Dora E., Educational Testing Service
DANIELS, Harry Waller, Richardson, Bellows, Henry, & Co., Inc.
DAVIDOFF, Melvin D., U. S. Civil Service Commission
DAVIDSON, Helen H., The City College of N. Y.
DAVIS, Fred B., Hunter College
DAVIS, J. Sanford, Research Branch, ONR
DAVIS, Hugh M., Pennsylvania State College
DETCHEN, Lily, Pennsylvania State College for Women
DIAMOND, Lorraine Kruglov, Teachers College, Columbia University
DIEDERICH, Paul B., Educational Testing Service
DIERS, Helen A., Vocational Advisory Service
DION, Robert, California Test Bureau
DOBBIN, John E., Educational Testing Service
DOPPELT, Jerome E., The Psychological Corporation
DRAGOSITZ, Anna, Educational Testing Service
DRAKE L. E., University of Wisconsin
DRESSEL, Paul L., Michigan State College
DRY, Raymond J., Life Insurance Agency Management Association
DUBOIS, Philip H., Washington University
DUNN, Frances E., Brown University
DUROST, Walter N., Test Service and Advisement Center
DYER, Henry S., College Entrance Examination Board
EBEL, Robert L., State University of Iowa
ENGELHART, Max D., Chicago Public Schools
ENRICK, Ralph, Michigan State College
EPSTEIN, Bertram, The City College of New York
FAN, Chung-Teh, Educational Testing Service
FAY, Paul J., New York State Department of Civil Service
FELDT, Leonard S., State University of Iowa
FELS, William C., College Entrance Examination Board
FERGUSON, William C., World Book Company
FINDLEY, Warren G., Educational Testing Service
FINKLE, Robert B., Metropolitan Life Insurance Company

FLANAGAN, John C., American Institute for Research
FLEMING, Mae, Educational Testing Service
FLEMMING, Edwin G., Burton Bigelow Organization
FLETCHER, Frank M., Jr., The Ohio State University
FOLEY, John P., Jr., The Psychological Corporation
FOURATT, Jean F., Educational Testing Service
FOX, William H., Indiana University
FRANKFELDT, Eli, Personnel Research Branch, TAGO
FREDERIKSEN, Norman, Educational Testing Service
FREEMAN, Paul M., Educational Testing Service
FRENCH, Benjamin J., New York State Civil Service
FRENCH, John W., Educational Testing Service
FRENCH, Sidney J., Colgate University
FRIEDMAN, Sam D., New York State Department of Civil Service
FRIEDMAN, Sidney, Bureau of Naval Personnel
FURST, Edward J., University of Michigan
GALLAGHER, Henrietta L., Educational Testing Service
GARDNER, Eric F., Syracuse University
GARRISON, Harry W., Educational Testing Service
GAWKOSKI, Roman, World Book Company
GEKOSKI, Norman, Temple University
GELINK, Marjorie, The Psychological Corporation
GERBERICH, J. Raymond, University of Connecticut
GLASS, Albert A., Fort Monmouth, N. J.
GREENE, Edward B., Chrysler Corporation
GREENE, Harry A., State University of Iowa
GREER, George D., Human Resources Research Office
GRIMM, Elaine R., Professional Examination Service
GROVES, Kenneth J., Air University
GULLIKSEN, Harold, Educational Testing Service
GUSTAD, John W., University of Maryland
HAGEN, Elizabeth, Teachers College, Columbia University
HAGGERTY, Helen, Personnel Research Section, TAGO
HAGMAN, Ernest R., Board of Education (Greenwich, Conn.)
HALPERIN, Joseph, New York State Department of Civil Service
HANKES, Leslie J., Testscor
HARBOUR, Paul B., Milton, Massachusetts
HARPER, A. E., Educational Testing Service
HARPER, Bertha P., Personnel Research Branch, TAGO
HART, May E., Babcock and Wilcox Company
HARVEY, Philip R., University of Connecticut
HASTINGS, J. Thomas, University of Illinois
HEALY, [illegible] Center for Psychological Service
HEATON, Kenneth [illegible] Richardson, Bellows, Henry and Co
HEINEMANN, Rich [illegible] D., Richardson, Bellows, Henry [illegible] mpany
HELMICK, John S. [illegible] cational Testing Service
HIERONYMUS, A. N., S[illegible] University of Iowa
HITTINGER, William F., Penn[illegible] State College
HOBERMAN, Solomon, Civil Se[illegible] mission (New York)
HORNBOSTEL, Victor O., Nati[illegible] tion Association
[illegible]BOWITZ, Milton W., Queen
[illegible]ON, Clark W., Dartmouth
[illegible]ESTON, Edith, Educati[illegible] ice
[illegible] Thelma; George Washington Uni[illegible]
[illegible] Elwood C., Tulane University
[illegible], Genevieve P., Archdiocesan Vo[illegible] al Service
[illegible], Robert W. B., Ontario College of [illegible]
JACOBS, Robert, Educational Records Bureau
JASPEN, Nathan, National League for Nursing
JEFFREY, W. E., Barnard College
JENNINGS, Helen H., Brooklyn College
JOHNSON, A. P., Educational Testing Service
JOHNSON, Lewis W., Personnel Department, City of Philadelphia
JONES, Harris, U. S. Military Academy
KABACK, Goldie Ruth, The City College of New York
KALIN, Robert, Educational Testing Service
KARON, Bert, Educational Testing Service
KEETS, J. A., Educational Testing Service
KELLER, Robert J., University of Minnesota
KELLEY, DeCourcy, Educational Testing Service
KELLEY, H. Paul, Educational Testing Service
KELLEY, Truman L., Professor Emeritus, Harvard University
KELLY, E. Lowell, University of Michigan

KELLY, Margaret, University of the State of New York
KERNAN, John P., White Plains, New York
KERR, Colin H., Boston University Junior College
KIDD, John W., Michigan State College
KING, Richard G., Harvard College
KIPNIS, David, New York University
KLEINMAN, Ruben, Brooklyn College
KLINE, William E., The Choate School
KOGAN, Leonard S., Institute of Welfare Research
KOLKEBECK, R. F., Educational Testing Service
KOSTICK, Max, State Teachers College (Boston, Mass.)
KUDER, G. Frederic, Duke University
KUSHNER, Rose E., City College of New York
KUTCHER, Charlotte, American Public Health Association
KVARACEUS, W. C., Boston University School of Education
LAMKE, T. A., Iowa State Teachers College
LANGMUIR, C. R., Syracuse University
LANNHOLM, G. V., Educational Testing Service
LAYTON, Wilbur L., University of Minnesota
LAZO, Elizabeth, American Public Health Association
LEACH, Kent W., University of Michigan
LEIMAN, John M., Lackland Air Force Base
LENNON, Roger T., World Book Company
LEV, Joseph, New York State Department of Civil Service
LEVERETT, Hollis M., American Optical Company
LEVINE, Richard, Educational Testing Service
LEVY, Charlotte, National League for Nursing
LINDQUIST, E. F., State University of Iowa
LITTERICK, William S., The Fund for the Advancement of Education
LIUTKUS, Stanley, Temple University
LOHMAN, Maurice A., New York State Education Department
LONG, Lillian D., Professional Examination Service
LONG, Louis, City College of New York
LORD, Frederic, Educational Testing Service
LORD, Shirley, Educational Testing Service
LORGE, Irving, Teachers College, Columbia University
LUBIN, Ardie, Army Medical Service Graduate School
LUCAS, Charles M., Cedar Crest College
LUSK, Louis, Norwalk, Connecticut
MCARTHUR, Charles C., Harvard University
MCCALL, W. C., University of South Carolina
MCCAMBRIDGE, Barbara, Educational Testing Service
MCCANN, Forbes E., Personnel Department, City of Philadelphia
MCCOLLUM, Joyce E., New York State Department of Civil Service
MCCORD, Richard, Personnel Department, City of Philadelphia
MCGILLICUDDY, Marjorie, New York State Department of Civil Service
MCGUIRE, Anne, New York State Department of Civil Service
MCGUIRE, John P., New York State Education Department
MCKINNEY, Lidie, Educational Testing Service
MCLAUGHLIN, Kenneth F., Florida State University
MCNAMARA, Walter J., International Business Machines Corporation
MCQUITTY, John V., University of Florida
MACPHAIL, A. H., Brown University
MAGNUSON, Henry W., California Department of Education
MALTBY, Jane, Public School System (Hamden, Connecticut)
MANUEL, H. T., University of Texas
MARQUIS, L. K., Arthur C. Croft Publications
MARRIOTT, John C., Boston University
MARSH, Mary, Educational Testing Service
MARSTON, H. M., Educational Testing Service
MARTIN, Priscilla Clark, New York City
MATHEWSON, Robert H., Board of Higher Education (New York City)
MAXWELL, Eleanor, Hempstead, Long Island
MAYHEW, Lewis B., Cooperative Study of Evaluation in General Education
MEDLEY, S. Donald, Educational Testing Service
MENDELSON, Martin A., Test Development Division, Mitchel Air Force Base, N. Y.
MERRY, Robert W., Harvard University
MESSICK, S. J., Educational Testing Service
METZ, Elliott, Queens College
MICHAEL, William Burton, University of Southern California
MIKLE, Stephen R., Educational Testing Service
MILLETT, Esther, Westover School
MITCHELL, Blythe C., World Book Company
MITZEL, Harold E., City College of New York
MOLLENKOPF, W. G., Educational Testing Service

MORRISON, Thomas F., Milton Academy
MOSELY, Russell, Wisconsin State Department of Public Instruction
MUIRHEAD, Peter P., New York State Department of Education
MUNGER, A. M., Standard Oil Company of New Jersey
MURRAY, John E., Special Devices Center, ONR
MYERS, Charles T., Educational Testing Service
MYERS, Thomas I., Human Resources Research Office, Fort Ord, California
NILL, Kathryn Fisher, Rinehart and Company
NOBLE, Lawrence M., Groton School
NOLAN, Edward G., Educational Testing Service
NOLL, Victor H., Michigan State College
NONKIN, Abraham, New York State Department of Civil Service
NORTH, Robert D., University of Kentucky
NOSOW, Sigmund, Michigan State College
O'CONNOR, Virgil J., Headquarters, United States Air Force
O'KANE, Marianne, Educational Testing Service
OLSEN, Marjorie, Educational Testing Service
ORLEANS, Beatrice S., Bureau of Ships, Navy Department
ORLEANS, Joseph B., George Washington High School (New York)
ORR, David, World Book Company
ORSHANSKY, Bernice, Mitchel Air Force Base
OSTREICHER, Leonard, The College of the City of New York
PACE, C. Robert, Syracuse University
PAGE, Maureen, The Psychological Corporation
PALMER, Orville, Educational Testing Service
PASHALIAN, Siroon, Queens College
PITON, James B., Virginia State Department of Education
PEARSON, Richard, Educational Testing Service
PECKHAM, Dorothy T., The Bancroft School
PERLMAN, Mildred, New York City Civil Service Commission
PETERSON, Donald A., Life Insurance Agency Management Association
PETERSON, Shailer, American Dental Association
PHILLIPS, Laura M., Silver Burdett Company
PHILP, Hugh, Harvard University
PIERSON, George A., Queens College
PINZKA, C. F., Educational Testing Service
PLUMLEE, Lynnette B., Educational Testing Service
POLIN, A. Terrence, Mitchel Air Force Base
POLLACK, Norman C., New York State Department of Civil Service
POTTS, Edith Margaret, The Psychological Corporation
PRICE, Hilda, The Psychological Corporation
QUICK, Robert, American Council on Education
RABINOWITZ, William, Bank Street College of Education
RAPPARLIE, John H., Owens-Illinois Glass Company
RAYMOND, Thomas J., Harvard Business School
REGAN, James S., Special Devices Center, ONR
REPPERT, Harold C., Temple University
RICCIUTI, Henry, Educational Testing Service
RICHARDSON, Ruth P., Richardson, Bellows, Henry, and Company, Inc.
RICKS, J. H. Jr., The Psychological Corporation
RIESSMAN, Frank, The City College of New York
RIMALOVER, Jack K., Educational Testing Service
RIMOLDI, H. J. A., Educational Testing Service
ROBB, George, Educational Testing Service
ROBBINS, Irving, Queens College
ROCA, Pablo, Department of Education, San Juan, Puerto Rico
ROESSLE, Robert L. B., The Standard Oil Company of New Jersey
ROGERS, Miles S., Harvard University
ROSENBLATT, Frank, Cornell University
ROY, Howard L., Personnel Research Branch
RUBIN, Clara, Personnel Department, City of Philadelphia
RULON, P. J., Harvard University
RUNDQUIST, E. A., Personnel Research Section, AGO
SADACCA, Robert, Educational Testing Service
SAIT, Edward, Rensselaer Polytechnic Institute
SANFORD, Nevitt, Vassar College
SAUNDERS, David R., Educational Testing Service
SCATES, Alice Yeomans, American Council on Education
SCATES, Douglas E., University of Florida
SCHEIDER, R. M., Educational Testing Service
SCHRADER, William B., Educational Testing Service

SCHULTZ, Douglas, The Pennsylvania State College
SCHWEIKER, Robert, Educational Research Corporation
SEASHORE, Harold, The Psychological Corporation
SEIBEL, Dean, Harvard University
SFORZA, Richard, New York State Department of Civil Service
SHARP, Catherine G., Educational Testing Service
SHAYCOFT, Marion F., American Institute for Research
SHERMAN, A. W. Jr., Mitchel Air Force Base
SHIENBLOOM, Charles, Philadelphia Board of Public Education
SHIMBERG, Benjamin, Educational Testing Service
SHREWSBURY, Roy R., Pingry School
SMITH, Alexander F., University of New Hampshire
SMITH, Allan B., University of Connecticut
SMITH, Densel D., Office of Naval Research
SMITH, Muriel, Educational Testing Service
SOLOMON, Herbert, Teachers College, Columbia University
SOLOMON, Robert, Educational Testing Service
SOUTHER, Mary Tayloe, Tower Hill School
SPANEY, Emma, Queens College
SPAULDING, Geraldine, Educational Records Bureau
SPEAR, Arthur, World Book Company
SPEER, George S., Illinois Institute of Technology
SPENCE, Ralph B., Teachers College, Columbia University
SPENCER, Douglas, U. S. Military Academy
SPRAGUE, Arthur, Hunter College
SPRAGUE, Marjorie R., Columbia University
STALNAKER, John M., Association of American Medical Colleges
STATE, Mary J., South Portland High School (Maine)
STERNBERG, Carl, Queens College
STEWART, Mary, Institute of Physical Medicine and Rehabilitation
STEWART, Naomi, Educational Testing Service
STOCKHAMER, Nathan N., Cornell Medical College
STODDARD, George D., Princeton, New Jersey
STONE, Paul T., Huntingdon College
STOUGHTON, Robert W., Connecticut State Department of Education
SULLIVAN, Richard H., Educational Testing Service
SUPER, Donald E., Teachers College, Columbia University
SWANSON, Margaret E., American Dental Hygienists Association
SWIFT, Everett L., The Peddie School
SWINEFORD, Frances, Educational Testing Service
SYMONDS, Percival M., Teachers College, Columbia University
TABB, Charles A., American Cyanamid Company
TAYLOR, Calvin W., National Research Council
TAYLOR, Justine N., Educational Testing Service
TCHORNI, Bernard L., Educational Testing Service
TERRAL, J. E., Educational Testing Service
THIBAULT, Paula, Educational Testing Service
THOMPSON, Albert S., Teachers College, Columbia University
THORNDIKE, Robert L., Teachers College, Columbia University
THURSTONE, L. L., University of North Carolina
TIEDEMAN, David V., Harvard University
TRAXLER, Arthur E., Educational Records Bureau
TRENT, Richard D., The City College of New York
TRIGGS, Frances, Commission on Diagnostic Reading Tests
TUCKER, Anthony C., Walter Reed Army Medical Center
TUCKER, Ledyard R, Educational Testing Service
TURNBULL, William W., Educational Testing Service
TWYFORD, Loran C., Special Devices Center, ONR
TYLER, Matilda, Yale University
UPSHALL, Charles C., Eastman Kodak Company
VIOLA, F., New York Civil Service Commission
VITELES, Morris S., University of Pennsylvania
WADELL, Blandena C., World Book Company
WAGNER, E. Paul, Teachers College (Bloomsburg, Pennsylvania)
WALLACE, Wimburn L., The Psychological Corporation
WALSH, John J., Boston College
WALSH, Thomas, Personnel Department, City of Philadelphia
WALTON, Wesley W., Educational Testing Service
WAND, Barbara, Educational Testing Service
WATT, Edwin, College of the City of New York

WANTMAN, Morey J., University of Rochester
WATKINS, Richard W., The Pennsylvania State College
WATSON, Walter S., The Cooper Union
WEINBERG, Sol, Freeport, New York
WEITZ, Henry, Duke University
WELLCE, A. A., University of New Mexico
WELTZ, Paula, The Psychological Corporation
WENZEL, Bernice, Barnard College
WESMAN, Alexander G., The Psychological Corporation
WEST, Elmer D., American Institute of Research
WEYBREW, Benjamin B., Medical Research Laboratory, New London, Connecticut
WHITLA, Dean K., Harvard University
WHITNEY, Alfred G., Life Insurance Agency Management Association
WILKE, Marguerite M., Greenwich Public Schools (Connecticut)
WILKE, Walter H., New York University
WILKS, S. S., Princeton University
WILLIAMS, Robert J., Columbia University
WILSON, Finis, Human Resources Research Office, Fort Ord, California
WILSON, Phyllis C., Queens College
WINANS, S. David, New Jersey State Department of Education
WINTERBOTTOM, J. A., Educational Testing Service
WOLMAN, Benjamin, City College of New York
WOLZ, Charles G., New York State Department of Civil Service
WOOD, Ray G., Ohio State Department of Education
WOODBURY, Max A., University of Pennsylvania
WRIGHT, Wilbur H., Geneseo State Teachers College
WRIGHTSTONE, J. Wayne, Board of Education, New York City
YOGMAN, Harvey J., University of Denver
ZUBIN, Joseph, Columbia University

One can proceed in this manner, which is very closely analogous to Hotelling's procedure.

We have found that the convergence in several problems is fairly rapid. If the convergence should turn out to be slow here there is no sense in carrying out fifteen or twenty or more iterations. If the next two dimensions are nearly circular, so that the next two latent roots are nearly the same, then the convergence is known to be slow. When that happens, we should be practical enough to realize that it really does not make any difference where you put the next axis. You can stop after about the third iteration. If you are talking to a mathematician, he might not accept it, but for practical work, it is useful.

DR. TUKEY: I should like to raise several specific questions. One: if we are going to do this on the least square basis, what is the philosophy that guides us to use unweighted squares of deviation? If we were looking solely at sampling variations, which we are not, then we would want to weight this quite a lot differently for large correlations than for small. For the purpose of factor analysis as a whole, I am not clear how we should weight it. I am just asking for guidance.

DR. THURSTONE: You are not referring to the A's as weights.

DR. TUKEY: No, I am referring to the possible weights outside the parenthesis.

The second question I should like to raise is this: if we are going to give up the communalities and go ahead in this direction, putting aside the difficulties of mathematics in computation which might be really serious, if we decide to take out two factors, why should we take out for the first one, the one that comes first and singly by this method? Would we not be better off, in principle, maybe not in practice, to accept this formula and minimize with respect to the A's and the B's at the same time? If one could do this, in this example you cited of Rank R, you would get everything if you could do it for all R's together.

DR. THURSTONE: Yes, I would rather write the equation with R plus one or two or three terms. You could then discard some if you want to. However, suppose you gamble that there are, say, ten factors. You write ten terms. It gets a little unwieldy to handle. That is why we do not do it.

DR. TUCKER: I have one question about the examples you have tried. Did some examples have somewhat equal communalities while other examples had quite variant communalities? I suspect the range of communalities makes quite a bit of difference in the closeness of approximation by your method. If the communalities were to come out the same, your method would yield an exact solution.

DR. THURSTONE: That is a good point to make. As a matter of fact,

we set up several empirical cases. For small variation in communalities, the convergence was much more rapid. In a case with wide difference of communalities the convergence was slower. I think you are quite right in that respect.

DR. LORD: You pointed out the procedure was equivalent to putting zeroes in the diagonals. What is the advantage in having zeroes in the diagonals rather than ones? Is it an approximation to something else?

DR. THURSTONE: That is not the same problem. You are raising the question of why we should use communality instead of unity in the diagonals.

DR. LORD: No. If you are not going to use the communalities, are you better off using zeroes rather than ones?

DR. THURSTONE: My objective is to ignore the diagonals, and they do not participate in the equation. That is the principle of the method. If the diagonals are unknown, let us leave them blank. We do not have to write zeroes—just leave them blank, they do not participate. However the computations would be the same as if the diagonal correlations were zero.

SECTION 2

Evaluating Group Interaction

A New Technique for Measuring Individual Differences in Conformity to Group Judgment

RICHARD S. CRUTCHFIELD

Central to research both in personality and in group dynamics are methods for the measurement of the individual's behavior in situations of group-interaction, for example, conformity behavior to group pressure.

Such measurement is made difficult by certain demanding requirements. (1) Ideally, the behavior should be measured *directly* in actual group situations, rather than indirectly by questionnaires about group situations. (2) The group situations should be psychologically *relevant* for the individual. (3) There should be *standardization* of the group situation so that measurements of different individuals may properly be compared. And (4) there should be adequate *economy* of the test method, so that substantial numbers of persons may be tested without unreasonable cost in time and money.

The standardization problem is the most acute. In genuine groups, involving face-to-face interaction of several subjects, the stimulus situation confronting each person is *unique,* being dependent in part upon what the others in the group do. This leads to an undesirable confounding of personal and situational factors in the measurement of the individual's group behavior, and there is no simple way to disentangle the two.

One fruitful attack on this problem is what in previous research applications I have called the "quasi group-interaction method." Its essence is simple. Subjects to be measured are placed together in a group situation which, though perceived as genuine by them, is actually so contrived by the experimenter that he wholly controls and manipulates the conditions of group-interaction. This serves to standardize the situation for each subject in an identical fashion, so that observed individual differences in behavior may properly be ascribed to differences among the persons rather than to situational differences.

The quasi group methodology is by no means new. Earlier variants are found throughout the field of experimental social psychology.

Little, however, has been made of it in personality measurement, and in general the method has not been widely exploited.

In recent work I have applied it to what gives promise of being a powerful new technique for measuring individual differences in conformity to group judgment. Five persons are tested simultaneously. They are seated in a row in front of an electrical switchboard apparatus, consisting of five identical panels. Each panel is shielded from the other four. It contains eleven numbered switches by which the person may signal his response. It also contains five rows of signal lights which can display the responses being made by the other four group members. In short, there is a simple electrical communication system among the five persons. No *direct* communication is permitted.

The task for each person is to make judgments pertaining to a large number of slides which are projected on a wall facing the group. Each slide offers a set of numbered alternative answers among which he is to choose. He records his choice by closing the appropriately numbered switch and this causes his response to be displayed on the panels of the other four members. He is also instructed to respond in a specified serial *order* within the group, person A going first, then B, and so on, E being last. The designation of his letter position—A, B, C, D or E—is indicated to him by the experimenter. Such letter designations are rotated from time to time, permitting each person to respond in each of the five serial positions. The experimenter offers no further explanation of the purpose of this procedure.

The slides offer a mixture of materials—simple *perceptual* comparisons, such as of relative length of lines; *logical* problems, such as the completion of a number series of the kind found in standard mental tests; expression of one's own *opinions* and *attitudes* on various issues, etc.

On the first set of slides the person finds the judgments fairly easy to make, and he observes that there is a sensible agreement between his judgments and those of the other four members. But when later in the series he is for the first time designated as E, so that he must respond in last position, he experiences something new and disturbing. On this slide, calling for a simple judgment of relative length of lines, he sees the other four members unanimously agree on a choice which clearly contradicts his own. This throws him into a severe conflict between the clear evidence of his own senses and the unanimous contradictory consensus of the rest of the group. How he chooses to resolve this conflict, either by yielding to the group pressure and conforming to its judgment or by remaining independent of it, is the basic measure of conformity behavior in our procedure.

There is, of course, more than one such critical slide. While responding in position E, he is presented with more than twenty other critical slides—pertaining variously to matters of perception, of logic, of opinion, and of attitude—on each of which he is confronted with a serious disagreement between his own judgment and that of the rest of the group.

As you will doubtless have surmised, the situation is not really what the persons have been led to understand. They have been grossly deceived. The five panels are not connected to one another at all, but to a control board operated by the experimenter. It is he who signals the responses which allegedly come from the other group members. The wiring is in parallel so that the pattern of signals he chooses to simulate appears simultaneously and identically on all five panels. Moreover, the designations of serial order of responding—A through E—are likewise identical for all five persons at every moment.

By this deception, therefore, a quasi group-interaction situation has been contrived, permitting each person to be exposed to a standardized set of simulated group judgments and to be confronted at predetermined points with identical external pressures toward conformity.

Since the Spring of 1953 when this technique was first developed, three studies have been made using different populations of subjects. The first study was of 50 men, averaging 34 years of age, all members of a profession in which leadership is one of the most important qualifications. They were tested at the Institute of Personality Assessment and Research as part of a larger assessment program. The second study was of 59 college students, mostly sophomores. The third was of 50 women, all college alumnae in their early forties being studied under the auspices of the Mary Conover Mellon Foundation.

The results of these studies, which attest to the technical success of the method and throw light on the nature and determinants of conformity behavior, are summarized with great brevity in the following eight points.

1. *The general amount of conformity behavior exhibited is large.* A single fairly representative item will serve to illustrate. A circle and a star are exposed side by side, the circle being about one-third larger in area than the star. The false group consensus is on *star* as the larger, and as a consequence 46% of the men give this same false judgment.

2. *The degree of conformity shown depends in part on the kind of material being judged.* Although some amount of conformity can be elicited on every one of the critical items, the range in effectiveness among them is extremely wide. At the lower end, a simple expression of *personal preference* for one of two line-drawings is very little suscepti-

ble to a contradictory group consensus, the degree of conformity varying from zero to ten percent in the several studies. At the upper end, the most effective item is one which has been deliberately maximized for ambiguity. The subjects are asked to complete a number series, for which there is actually no logical solution. When the simulated group judgment agrees unanimously on an obviously illogical completion, 79% of the men express agreement with this answer.

3. *Substantial conformity is elicited on socially important judgments as well as on more abstract judgments.* A critical methodological issue is whether the conformity effects so far mentioned merely represent rather superficial readinesses of the person to agree with the group on matters of little real importance to him, or whether instead they do reflect more basic conformity tendencies in the person. Support for the latter interpretation is found by the introduction of new critical items in the studies of college students and mature women. These items called for expression of the person's attitudes on matters of civil liberties, subversion, ethics, crime and punishment, etc. Pronounced conformity effects are found on these socially and psychologically relevant items. Take a single example from the study of college students. The question was asked, "Which one of the following do you feel is the most important problem facing our country today?", and these five alternatives were offered: economic recession, educational facilities, subversive activities, mental health, crime and corruption. Among control subjects tested alone, only 12% chose "subversive activities" as the most important. But when exposed to a spurious group consensus unanimously making this choice, 48% of the subjects gave this answer.

4. *There are pronounced individual differences in amount of conformity shown.* A total conformity score for each person may be readily computed by summing the number of the critical items on which he exhibits conformity to the false group consensus. Virtually the entire possible range of such scores is found in all three studies. Among the men, for instance, at the lower end several subjects showed practically no conformity, being influenced on one or two items at the most. At the upper end one man was influenced on 17 of 21 critical items. The scores are well distributed between these extremes, with a mean score of about 8 items and a tendency for greater concentration of scores toward the lower conformity end. As estimated from the correlation of sub-scores on two matched halves of the critical items, the reliability of the total score for the sample of men is .90.

5. *There are both generality and specificity in the conformity tendencies among individuals.* Although there is a generally positive matrix of intercorrelations of degree of conformity effect for the various items,

there are also some useful differentiations to be made among the items. A cluster analysis yields one principal set pertaining to highly structured items, i.e., those involving clear judgments and unambiguous stimuli. A second main cluster consists of poorly structured items, i.e., those involving uncertain judgments and ambiguous stimuli. The individuals can be scored separately on these clusters. There is evidence that these cluster scores as well as individual performance on single items must be taken into account in the analysis of conformity behavior.

6. *The degree of conformity behavior relates significantly to relevant dimensions of personality.* Validity of the conformity measure is attested to by its substantial relationships with numerous ratings, objective test scores, and other personality determinations in assessment of the sample of men. Those low in conformity, that is, those who successfully resist the group pressure, can be clearly characterized as having intellectual effectiveness, ego-strength, self-acceptance, leadership ability and maturity of social relations. The high conformers, on the contrary, reveal inferiority feelings, rigid and excessive self-control, intolerance of own impulses and lack of self-insight, authoritarian outlook, emphasis on external and socially approved values, and disturbed attitudes toward other people. This general picture coincides well with prior theoretical and empirical studies of the personal traits associated with conformity, but, of course, it represents a gross oversimplification of the complex relationships yet to be explored.

7. *There are significant differences in average amount of conformity exhibited by the several populations of subjects.* The college student sample was made up of both males and females. The male students exhibited just about the same average level of conformity as previously found in the adult, professional men. But the female students exhibited significantly *more* conformity than these male groups. On the other hand, the sample of mature women—college alumnae in their early forties—showed significantly *less* conformity on the average than that found in all the other groups.

8. *Experimental variations in the testing situation can change the amount of conformity shown.* We have seen that there are differences in degree of conformity relating to personality determinants, to the nature of the populations of subjects, and to the kind of items being judged. We can also show differences resulting from certain experimental changes in the situation itself. As one example, when a group of student subjects were exposed to additional pressure, namely, that given by having the experimenter later confirm the "correctness" of the false group consensus, conformity effects were markedly increased.

To revert now to our initial statement of requirements for proper measures of individual behavior in situations of group-interaction, namely, directness, relevance, standardization and economy, it would appear that the present method fulfills all these requirements. The very essence of the method is such as to guarantee standardization of conditions. It is clearly economical, providing for the testing of five persons at once in a period of approximately one hour. It is direct, in that the persons perform in a situation which is for them one of real group-interaction. On this point it should be emphasized that the genuineness of the situation is virtually never challenged by the subjects. Of the total of 159 persons already tested in the procedure, only a small handful when questioned immediately afterwards expressed doubts of its genuineness; of these only two or three seemed to have felt such doubts while actually in the situation.

Finally, as to psychological relevance, we may judge from both the manifest behavior and the retrospective reports by the subjects that the situation was deeply involving and that the group pressure created anxiety, often acute. A substantial number of persons freely admit on later questioning that they had violated their own inner convictions in order to give responses in accord with those of the group.

The findings attest to a rich potential use of this measurement technique when applied to a variety of problems—personality assessment, group dynamics research, sociological and cross-cultural comparisons of groups, experimental and theoretical study of the psychological conditions of conformity. A number of such studies are contemplated and several are now in progress.

The Russell Sage Social Relations Test: A Measure of Group Problem-Solving Skills in Elementary School Children

DORA E. DAMRIN

The test I am going to describe to you this morning represents an attempt that has been made by myself and certain of my colleagues in ETS to develop an instrument of measurement that will get at certain of the intangibles in education. As all of you know, over the past several years there has been increasing emphasis, particularly in the field of elementary education, given to the idea that children should be taught to work together cooperatively, to share things with each other, to participate in the planning and carrying out of certain classroom activities, and to conduct themselves in ways compatible with a free democratic society.

The test I am going to show you today is designed to measure the extent to which some of these things have been accomplished. Since the test is a little unusual, I thought that it would be best if I would pretend that you are a group of students and give it to you in the same way that we give it to elementary school classes. The score sheets you have will be explained after this demonstration.

The examiner enters the classroom and sets up a card table in the front of the room. He then says, "Today, boys and girls, I am going to give you a test. Now, you know usually when you take a test, each one of you sits in his own seat, does his own work, and you are not allowed to help each other. Isn't that correct?"

The students usually nod rather solemnly in the affirmative at this point. The examiner goes on to say, "But this test I am going to give you today is exactly backwards. In this test you are *supposed* to help each other, you are *supposed* to work together, and you are *supposed* to talk to each other while you are taking it." Here the children usually evidence some amazement and seem to wonder what in the world is coming next.

The examiner then says, "Let me show you the first problem on the test so that you will understand what it is about." The examiner opens the testing case and holds up this rather attractive form, built of different colored plastic blocks, and says, "Now, what does this look like?"

The children respond with, "That's a house."

The examiner continues, "Yes, that's right, this is a house, and I am going to put it here on the card table for you to look at all during the time you are taking the test."

"Now here in this box, boys and girls, I have all of the blocks that you will need to build this house, and what I am now going to do is pass around the room and give each of you one or two of these blocks. Then I want all of you to get together and see how quickly and how well you can build this house."

The examiner goes around the room, giving each child one or two blocks, depending on the number in the class, and at that instant we start scoring the test, continuing on until the children have completed the problem.

After the blocks are passed out, the children are told, "Now before you start you can take all of the time you need to plan how you are going to get together and do this, but after you tell me that you are ready, then I am going to begin timing you, and you will be given fifteen minutes in which to build the house. If you complete your house *before* that time, your score on the test will be much higher."

The examiner then works with the children while they build a plan, and all during this time their comments and suggestions are scored. When the children indicate that they are ready to begin, the examiner scores them on how well they carry out their plan. At the end of the first test problem the examiner presents them with the second, and finally with the third problem.

The second and third problems are progressively more difficult than this one. If you would be interested in examining them, I will be glad to show them to you afterward, since our time is somewhat limited.

Unfortunately, in my discussion of the scoring of the test, I cannot present data on the validity or the reliability of the instrument. It is of too recent development to do that. We are, however, presently conducting a major validity study in cooperation with Dr. Orleans and his research staff in New York City.

On the basis of the data which will be collected within the next six months or so, I hope that I will be able to support some of the things that I am going to hypothesize about this morning.

The test has two separate and relatively independent parts. The first is called the *Planning Stage* and the second, the *Operations Stage*. The Planning Stage begins at the time the examiner starts passing out blocks to the children, and ends when the children inform the examiner that they are ready to begin constructing the model. The Operations

Stage begins when the examiner tells the children "Begin" and starts his stop watch, and ends when the 15 minute time period has elapsed or when the children finish building the model. A set of six scores is obtained for the Planning Stage and a set of three scores for the Operations Stage.

The first score given in the Planning Stage is termed *Group Participation*. It is based upon the four specific behaviors listed on your Observation Sheet at the top of the page. The Observation Sheet is checked as follows:

1. Individual: A check is entered each time an individual class member makes a problem-relevant comment.
2. Buzz Groups: A check is entered each time one-half or more of the class members spontaneously turn to each other in small groups and begin discussing various elements and aspects of the problem. Many children speak at one time, comments are not individually distinguishable, but it is clear that the children are engaging in problem-relevant remarks.
3. Chorus: A check is entered each time one-half or more of the class members spontaneously chorus their opinion about a previously made individual comment. This chorus may indicate either agreement or disagreement with the idea.
4. Noise: A check is entered each time one-half or more of the class members become noisy and undisciplined. Many children speak at one time, the examiner is unable to hear individual comments, but it is clear that the children are engaging in many problem-*irrelevant* remarks.

The pattern of checks made by a group of children over these four categories of behavior is translated into a numerical score which reflects the amount and quality of the children's total participation. A high score on this variable indicates that all members of the group participated actively and freely in the planning stage, whereas a low score indicates a minimum of free participation. It is our hypothesis that the Group Participation Score is closely related to the amount of freedom which exists in the classroom of a particular elementary school teacher.

The second score is *Communication Pattern*. It is concerned solely with the intent of the verbal comments made by the individual children who contribute to the planning discussion and who are checked in the Group Participation Score. All comments made by individual children are classified in one of the four categories listed second on your Observation Sheet. These categories are defined as follows:

1. Suggests discrete idea: A check is made if the idea is one which has not previously been stated and which does not incorporate any part of or build on a previously mentioned idea.
2. Repeats idea: A check is made if the comment is either a rephrasing or a restatement of a previously mentioned idea.
3. Evaluates idea: A check is made if the comment takes a previously mentioned idea and points out why it is wrong, why it won't work, or why it is a good suggestion and should be adopted.
4. Improves idea: A check is made if the comment takes a previously mentioned idea and builds on to it so that the end product is more comprehensive, more precise, and more detailed than the initial idea.

The pattern of checks made by a group over these four kinds of statements is translated into a numerical score which reflects the way in which children converse with each other in building a mutually agreeable plan of action. A high score is obtained by the group whose members improve and build upon each other's ideas and progressively work toward a final plan that is more precise, more detailed, and more comprehensive than the initial idea for a plan with which the group started. A low score is obtained by the group whose members' comments are largely discrete and repetitive. In this score a series of discrete and repetitive checks indicates that children are competing with each other to give "different" ideas rather than cooperating with each other to build a constructive solution. Thus the significance of the Communication Pattern Score resides in the fact that it provides information about whether the members of a particular elementary school classroom function as a collection of single individuals or as a cooperative group.

The third score is termed *Organizational Efficiency* and is based upon the content of the ideas which children propose during the planning period. Fortunately the major ideas which children can have about ways of solving the test problems are finite, and all possible suggestions can be classified into one of the ten categories listed in the third section of your Observation Sheet. These categories have the following definitions:

1. Construction Authority: Having one or a few children perform all of the actual building
2. Supervisory Authority: Having one or a few children supervise and help others who are building the puzzle
3. Simultaneous Groups: Organizing the class into groups on some basis and having all groups working simultaneously

4. Sequential Groups: Organizing the class into groups on some basis and specifying the order in which the groups are to take turns in building the puzzle
5. Problem Groups: Organizing the class into groups on the basis of some structural element of the puzzle, e.g.:
   a. By colors: red, white, blue
   b. By parts: neck, body, legs, tail
   c. By shape of block: square, triangle
   d. Any combination of the above
6. Non-problem Groups: Organizing the class into groups on the basis of something *external to* the puzzle, e.g.:
   a. By sex: boys, girls
   b. By seating arrangement: rows, tables
   c. By numerical groups: 1-at-a-time; 2-at-a-time; 3-at-a-time
7. All at once: Having everyone in the room go up to the problem table at one time, each person putting in his own piece
8. Same as last time: Having the group follow the plan which was followed in the preceding problem
9. Group discipline: Making suggestions about how the group ought to behave while building the puzzle, e.g.:
   a. Don't push
   b. Don't crowd around
   c. Don't jerk pieces away from other people
10. Change personnel: Suggesting new builders, new leaders, or suggesting a different order of sequential groups

The Organizational Efficiency Score provides information about how much knowledge a group has regarding ways and means of organizing itself to attack a particular problem. It will be observed that the items on the list of Organizational Techniques are, on the whole, arranged in descending order of efficiency. The first two items concern the idea of depositing leadership authority in certain group members; the next two deal with the idea of how the class should proceed if and when it is organized; the next two deal with the bases which the class can use to organize itself into groups; and the idea "all at once" represents complete lack of organization. The last three items on the list of Organizational Techniques apply only in the second and third problems. These ideas, if checked, must be evaluated in terms of the plan used in the preceding problem.

The Organizational Efficiency Score, if high, indicates that a class recognizes the need for leadership and for some means of arranging its members into a structure that will function efficiently in the present problem situation. A low score indicates that children lack the neces-

sary knowledge of organizational procedures and can think of nothing better than to have each child put in his own piece himself.

The fourth score is termed *Task-Centeredness* and is based upon the same list of ideas as is the Organizational Efficiency Score. However, whereas Organizational Efficiency concerns *all* of the suggestions children have about ways of solving the problem, the Task-Centeredness Score is based upon only those ideas which are selected and incorporated into the final plan. The logic of this score resides in the notion that a group skilled in techniques of cooperative group action will select, from among all the ideas presented, those few ideas which are "best" in terms of the demands of the particular problem at hand. A group which is not skilled in these techniques will tend to disregard "good" ideas in favor of poorer ones.

A high Task-Centeredness Score indicates that of all the ideas a particular group of children had about ways of solving the problem the "best" ideas (e.g., those highest on the list of organizational techniques) were selected and incorporated into the final plan. A low score indicates that although a group had several good ideas they rejected them in favor of ideas which capitalized upon individual desires and minimized the total group good.

The fifth score is termed *Independence* and concerns the group's ability to plan by themselves without external assistance from or interference by the examiner. The elements of this score are based upon the kinds of actions the examiner is forced to take during the planning period and these are listed in the fourth section of your Observation Sheet.

The first action, "Prods Group," means that the individuals in a class respond with stony silence to the examiner's initial request: "Now what ideas do you have about ways the class might go about solving this problem?" When this occurs, the examiner repeats certain of the rules and the demands of the test and again asks, "Now, how are you going to get together and build this model? Does anyone have an idea of how the class might go about it?" "Prods Group" is also checked whenever the examiner has to urge the group to give *more* ideas after the first one or two.

The second action, "Suggests Vote," means that the class, after presenting a few ideas, comes to an impasse and does not know how to proceed. That is, children keep repeating ideas that have already been presented, or they again fall into a complete silence and look with bewilderment upon the examiner when he asks, "Which of these ideas are you going to select?" In either case the examiner suggests that they vote on one of the ideas which they have presented.

The third action, "Reminds Group," means that the examiner is forced to recall to the class the rules of "good group behavior." This is necessary when the children become noisy and undisciplined, when they all talk at once or shout out their ideas to the examiner. In such a case the examiner is forced to say, "Now, boys and girls, let's speak one at a time for I'm afraid I can't hear your ideas and I'm sure you can't hear each other's."

The fourth action, "Summarizes Ideas," is taken by the examiner whenever the class reaches an impasse because of the complexity of their ideas. That is, the discussion eventually comes to revolve around three or four different plans, each of which is so involved and detailed that the class cannot pull together and hold in mind all of the different elements and cannot come to any kind of decision. In such an instance the examiner names the plans in a one-two-three fashion, briefly summarizing the details of each.

The fifth action, "Conducts vote or election," means that the class has suggested that they vote on a plan or that they elect a leader. The examiner then proceeds to perform the mechanical task of conducting the vote or election.

A high Independence Score indicates that a group has the ability to discipline themselves and to carry on successfully a discussion from the beginning to a point at which decisive action is taken. Conversely, a low score indicates that a group lacks the several skills necessary for independent planning and must rely heavily on outside help to reach a point where decisive action can be taken.

The sixth and final score which is given during the planning stage is termed *Decision Method* and refers to the way in which the group decides upon a particular plan of action. These four methods are shown in the fifth section of your Observation Sheet.

The method of active consensus is that in which all members of the group come to a more or less spontaneous and mutual agreement about a plan of action. The method of active vote indicates that the class is split upon two or more suggested plans and must use the method of voting to achieve the kind of consensus necessary to group action. The method of passive vote indicates that there is no *real* concern on the part of group members about which of several plans of action is adopted. Voting is artificial and meaningless and tends to be engaged in as "something we are supposed to do." The method of passive consensus indicates that the class as a whole is uninterested in both the problem at hand and the plan which ought to be selected in resolving it.

Our hypothesis is that use of the method of active consensus indicates a warm, cooperative, cohesive and skilled group, whereas use of

the method of active vote indicates a skilled but less cohesive one. Use of either of the passive methods seems to be associated with very low degrees of skill in group planning and group problem-solving.

The scores obtained during the Operations Stage have not yet been as precisely defined as those given during the Planning Stage. The reason is that the behavior which occurs during this second stage is extremely more complex and far less "controlled" than the behavior in the planning period.

The first score in the Operations Stage is termed *Execution Pattern* and indicates whether and how well the group puts its adopted plan of action into effect, once the examiner has said, "Begin." A high score on this variable indicates that the children move easily and efficiently into the mechanics of their plan and proceed without hesitation to carry it out in detail. A low score indicates that the plan breaks down immediately and that children rush up to the problem table in a helter-skelter fashion with absolutely no regard for the rules of group behavior they decided upon during their planning session.

Our hypothesis is that children who are used to formulating the rules which are to govern their behavior and, more important, who are accustomed to *acting* in accordance with these rules will score high on this variable. Conversely, children accustomed to engaging in group planning as a relatively meaningless mental exercise will tend to exhibit some skill in the planning period but will fail completely in their ability to actually govern their behavior in accord with the rules they themselves have outlined.

The second score on the Operations Stage is as yet a highly subjective, and accordingly, a highly unreliable one. It is a measure of the overall psychological tone-quality of the group engaged in working on the puzzle. The descriptive terms listed on your Observation Sheets are checked on the average of once a minute by the observer on the basis of his judgment about what the group is like. It is our hypothesis that the "psychological tone" of the group is to some extent related to and hence indicative of the psychological quality of the classroom atmosphere which is created and maintained by the particular teacher involved. Our findings here are suggestive, but as yet so tentative that it is not possible to discuss them in detail at this time.

The final score obtained during the Operations Stage is termed *Group Interest* and provides an indication of the concern for the problem which is exhibited by those children who are not actively engaged in the construction task. We are not completely sure what is responsible for observed differences in interest in this test between various classroom groups of children, but we think that the group that main-

tains high interest in the task is more cohesive, more skilled in group work and more ego-involved in the test because of its being a "group project" than is the group whose interest is solely dependent upon immediate personal participation in the task.

Observer_____

RSSR OBSERVATION SHEET Date_____

Teacher_____ School_____ Grade___ No. in group___ Problem___

PLANNING STAGE

1. Group Participation

Individual																	
Buzz Groups																	
Chorus																	
Noise																	

2. Communication Pattern

Suggests Discrete Idea																	
Repeats Idea																	
Evaluates Idea																	
Improves Idea																	

3. Organizational Techniques

Construction Authority																	
Supervisory Authority																	
Simultaneous Groups																	
Sequential Groups																	
Problem Groups																	
Non-problem Groups																	
All at once																	
Same as last time																	
Group Discipline																	
Change Personnel																	

4. Independence

Prods group																	
Suggests vote																	
Reminds group																	
Summarizes ideas																	
Conducts vote or election																	

5. Decision Methods

Active Concensus																	
Active Vote																	
Passive Vote																	
Passive Concensus																	

6. Description of plan: Time Planning_____

RSSR OPERATIONS STAGE

Approach

1. Plan put into effect smoothly and without hesitation _____
2. Plan put into effect immediately in a wild excited rush _____
3. Plan put into effect after some hesitation _____
4. Plan breaks down immediately _____

The Plan

In Operation																
Changed—Inefficient																
Breaks rules																

Number at Table

Plan																
Total																
1/2 — 3/4																
1/3																
Small Group																

General Atmosphere

Quiet and Intent																
Warm and Cooperative																
Tense and Excited																
Bickering and Squabbling																
Whining and Complaining																
Angry and Hostile																
Instructor																

Non-Builder Group Interest

Good																
Fair																
Poor																

Remarks:

Time _____
No. of Errors _____

Description of Group Characteristics

JOHN K. HEMPHILL

The study of social groups and the behavior of group members is rapidly becoming an active field of endeavor for psychologists. Both empirical research and theory building are receiving attention. However, if the scientific study of social groups is to continue to be vigorous and effective, one area of problems that has been largely neglected must be given increased attention. It is imperative that a taxonomy of group characteristics be developed in a form that permits accurate measurement.

Research work on problems of group taxonomy may be less exciting than investigating the intriguing processes of group functioning, but if the many discrete research findings that are accumulating in the literature are to be brought into some order or relationship, we must use common concepts for designating major differences among groups. If we could specify the characteristics of different samples of groups by using more comparable terms than are now available, much confusion that exists at present in group research would disappear. Questions about seemingly contradictory research findings and about the generalizability of the results of laboratory research would be more susceptible to answer.

This paper reports the development and use of a tentative set of dimensions devised for use in describing group characteristics.

Development of the Group Dimension Description Questionnaire

The following four criteria served as guides in the selection of characteristics to compose the set of group dimensions: (a) each characteristic should be meaningful within a sociological or psychological framework; (b) each characteristic should be conceived as a continuum varying from the lowest to the highest degree; (c) each characteristic should refer to a relatively molar rather than a molecular property of a group; (d) each characteristic should be relatively orthogonal or independent of all other characteristics in the descriptive system.

As a first step, and after a thorough review of relevant literature in social psychology and sociology, fourteen group dimensions were selected and defined as follows:[1]

[1] A more complete definition of each dimension and a detailed account of the development of this instrument is given by Hemphill and Westie. (3)

1. *Autonomy* is the degree to which a group functions independently of other groups and occupies an independent position in society.
2. *Control* is the degree to which a group regulates the behavior of individuals while they are functioning as group members.
3. *Flexibility* is the degree to which a group's activities are marked by informal procedures rather than by adherence to established procedures.
4. *Hedonic Tone* is the degree to which group membership is accompanied by a general feeling of pleasantness or agreeableness.
5. *Homogeneity* is the degree to which members of a group are similar with respect to socially relevant characteristics.
6. *Intimacy* is the degree to which members of a group are mutually acquainted with one another and are familiar with the most personal details of one another's lives.
7. *Participation* is the degree to which members of a group apply time and effort to group activities.
8. *Permeability* is the degree to which a group permits ready access to membership.
9. *Polarization* is the degree to which a group is oriented and works toward a single goal which is clear and specific to all members.
10. *Potency* is the degree to which a group has primary significance for its members.
11. *Size* is the number of members regarded as being in the group.
12. *Stability* is the degree to which a group persists over a period of time with essentially the same characteristics.
13. *Stratification* is the degree to which a group orders its members into status hierarchies.
14. *Viscidity* is the degree to which members of a group function as a unit.

As the second step in the development of the instrument for measuring group dimensions, over 1000 items, each referring to some specific characteristic of relationships among group members or to some manner of group functioning were assembled. These items were secured (a) by administration of open-ended questions about the nature of their groups to members of over 500 groups (2) and (b) by selecting excerpts from sociological, psychological, and literary writings in which groups were described.

Data for an index of "homogeneity of placement" were secured for each item as the third step.

This index was adopted to give a single numerical evaluation of each item with respect to its homogeneity, not in the usual sense of inter-item variance, but in relation to the other thirteen dimensions

composing the descriptive system. We sought to find items that were clearly relevant to one of the set of dimensions and also clearly irrelevant to the remaining thirteen. Each of five judges sorted all of the items fourteen times. After carefully reading the definition of one of the fourteen group characteristics the judges placed each item into one of three categories. One category consisted of those items considered to apply directly to the characteristic. A second category included those items about which the judges were undecided. The third category was for items considered irrelevant or inapplicable to the characteristic. Upon the completion of one sort, the items were shuffled and the same procedure was repeated on the next dimension. Agreement among judges that an item applied to a dimension and agreement that it did not apply to other dimensions in the system are given approximately equal weight in the value of the index.[2]

The fourth step in the development of measures of the group characteristics consisted of a conventional internal consistency item analysis. For this purpose a preliminary questionnaire composed of the 350 items that had highest index values was administered to 200 individuals who were members of 35 different groups. These respondents indicated how well each item described their group by selecting one of five choices: Definitely True; Mostly True; Undecided; Mostly False; or Definitely False.

As a further refinement, a smaller sample of the respondents who completed the 350 item questionnaire were interviewed and questioned about why they had selected specific choices to specific items. Ambiguities in reasons for selecting responses were noted and then each item was edited to eliminate sources of ambiguity, or removed from further consideration.

[2] The index of "homogeneity of placement" was computed using the following formula:

$$I_i = \frac{N \sum_{j=1}^{n} X_{ij} - \sum_{i=1}^{N} \sum_{j=1}^{n} X_{ij}}{2\,[\,2n\,(N-1)\,] + \sum_{i=1}^{N} \sum_{j=1}^{n} X_{ij} - N \sum_{j=1}^{n} X_{ij}}$$

where: j = any judge
i = any dimension in the system
n = number of judges
N = number of dimensions
X = score assigned to item placement as follows:
Definitely applies = + 1
Undecided = 0
Definitely does not apply = − 1

The final instrument entitled "Group Dimensions Descriptions" is composed of 150 statements about groups each assigned to one of thirteen dimensions.[3] A respondent's description of his group is summarized by thirteen group dimension scores. Each score is the sum of the weights given choices 'Definitely True' through 'Definitely False,' for the items assigned to each particular dimension.

Experience with the Questionnaire

The final form of the questionnaire has been used in several empirical studies since its development in 1950. Group descriptions have been secured for office and shop groups in industry (1), departments of instruction in universities and colleges (5), high school and elementary teaching staffs of 26 Ohio public school systems (8), various military units (7), religious organizations (6), sports teams, informal social groups (4), etc. We have had opportunity to examine some of the properties of the dimensions scores that are available from these studies.

Perhaps the most important question that can be asked about the group dimensions scores is whether or not different members of the same group agree about the characteristics of their group. Unless some agreement can be shown there would be cause to question seriously the existence of such an entity as a group. We have examined three sets of data where it was possible to compare the variance of the group dimensions scores for the descriptions supplied by members of the same groups with estimates of the variance of these scores based on mean scores for different groups. Significant F ratios have been found for each of the thirteen dimensions but not for all dimensions in all three studies. In a study of eight groups that were described by a total of 65 of their members, the extent of agreement among descriptions supplied by the members of the same groups, expressed as unbiased correlation ratios, ranged from .47 for the dimension Polarization to .74 for the dimension Permeability. In a second study involving 134 faculty members and their respective 20 departments, significant between-group variance was found for the dimensions Flexibility, Hedonic Tone, Intimacy, Participation, Polarization, Stability, Stratification, and Viscidity. In a third study the staffs of 26 elementary public schools were shown to differ significantly on Autonomy, Flexibility, Homogeneity, Intimacy, Stability, and Viscidity. It appears evident that the descriptions supplied by members of their groups with the use of the Group

[3] Size was dropped from the final form of the questionnaire, not because it was considered an irrelevant dimension but because there appeared to be more direct methods of obtaining this information.

Dimensions Description Questionnaire provide differentiations among groups.

Some evidence of the validity of such differentiations has also been observed. For example, descriptions of a military training department (ROTC) were contrasted with descriptions of a committee from a college of education and the expected large differences along the dimensions of Control, Participation, and Stratification clearly emerged.

Not all of the variance among respondents who describe the same group is to be attributed to errors of measurement. The dimension scores have been shown to be related to characteristics of the respondents' status in the group. For example, in the study of the departments of a liberal arts college it was discovered that those members of the faculty who possessed greater status, i.e. held the rank of Assistant Professor or better, described their departments with lower scores on the dimension Control than did other members of their department, who were largely Instructors. In the same study it was also found that those individual members of the faculty who expressed above average satisfaction with their jobs tended to view their departments as higher on the dimensions Hedonic Tone, Participation, Polarization, Potency, and Viscidity, but as lower on the dimension Stratification. In a study involving aircrews, which used a modified form of this questionnaire, Rush (7) found that Aircraft Commanders differ from other crew members in the way they describe their crews. No study has been made relating measures of the personality variables of respondents to their descriptions of their group, but such relations might be expected. Thus we see that the group dimensions scores are sensitive both to commonly seen group characteristics and to the respondent's particular relationship to his group.

Considerable effort in the development of the Group Dimensions Descriptions questionnaire was expended in attempting to secure independent dimension scores. In order to examine the extent of overlap among the dimension scores empirically we have computed intercorrelations for three sets of data; a sample of 100 descriptions of miscellaneous groups, a sample of 136 descriptions of 22 departments of a liberal arts college, and a sample of 315 descriptions of the staffs of 26 Ohio public school systems. In general the correlations among the dimensions are quite small—approximately three quarters of them below .3 in magnitude. Certain dimensions are frequently found to be correlated more highly with each other. For example, the relationship between Hedonic Tone and Viscidity was found to be .33; .64 and .57 for the three samples respectively; between Participation and Polarization, .09; .34 and .47; and between Stratification and Autonomy, −.54;

— .31 and — .24. The size of the correlations between specific dimensions appears to be sensitive to the institutional setting from which the groups are selected. Experience accumulated to date has provided no clear-cut evidence for the abandonment of any one of the dimensions if they are to be considered applicable to all varieties of groups. However, it does appear that if the investigator is interested only in differences among a relatively homogeneous sample of groups, certain of the dimensions might not be pertinent. For example, in adapting the instrument for use with aircrews, we eliminated the dimensions of Size, Autonomy, and Permeability, because the assumption of variance on these dimensions appeared far-fetched.

Finally, the problem of establishing normative data for these group dimensions has been given attention. I frankly do not know how to establish defensible norms for group characteristics. The problem appears to be one of lifting yourself by your bootstraps. Certainly, securing representative samples of individuals would not insure a representative sampling of their groups, since individuals have numerous and complicated multiple group membership.

Conclusions

The development of a tentative set of dimensions of group characteristics has been described. Some of the experience that we have had with the use of the dimensions has been cited briefly. Much remains to be done. We need defensible norms. We need to reconsider the selection of group characteristics. Perhaps the list of dimensions should be enlarged, or perhaps it can be reduced. But unless we solve some of these elementary problems of describing the basic entity of research on groups, confusion will reign.

References

1. Gekoski, N. The relationship of group characteristics to productivity. Unpublished doctor's dissertation, Ohio State Univer., 1951.
2. Hemphill, J. K. Situational factors in leadership. Bureau of Educ. Res. Monog. No. 31, Columbus: Ohio State Univer., 1949.
3. Hemphill, J. K. & Westie, C. M. The measurement of group dimensions. *J. Psychol.*, 1950, 29, 325-342.
4. Hemphill, J. K., Seigel, Ann, & Westie, C. M. An exploratory study of relations between perceptions of leader behavior, group characteristics, and the expectations concerning the behavior of ideal leaders. Unpublished manuscript.
5. Hemphill, J. K. & Bentz, J. Leadership behavior associated with the administrative reputations of the departments of a college. Unpublished manuscript.
6. Knight, R. A study of thirteen group characteristics of selected religious organizations at Ohio State University. Unpublished master's thesis, Ohio State Univer., 1950.
7. Rush, C. Group dimensions of aircrews. Unpublished doctor's dissertation, Ohio State Univer., 1953.
8. Seeman, M. A sociological approach to leadership: The case of the school executive. Unpublished monograph.

DISCUSSION

PARTICIPANTS

WILLIAM E. COFFMAN, RICHARD S. CRUTCHFIELD, DORA E. DAMRIN-ERICH M. DAVISON, LINDSEY R. HARMON, JOHN K. HEMPHILL, IRVING L. ELLIOTT M. MCGINNIES, EMMA SPANEY, S. DAVID WINANS

DR. DAVISON: I have a question for Dr. Hemphill. Say you have a group, and it has dimensions on your test over here to the right on a distribution. You have another one over here to the left on a distribution. Now, how do I know whether or not, if I traded people around, they would not coincide? Is it possible that we are all making their own frame of behavior and therefore, might it be impossible to really compare groups?

DR. HEMPHILL: You are asking a question which I do not know that I can answer. I would suggest one thing, however, that might approach such a problem by using the questionnaire that I have described. Off hand, I would not attempt to answer the question of whether we change the group when we change some or all of the persons in it. I would have to know much more about the group before I could even make a guess. I think you need a tool or instrument like the Group Discussion Questionnaire.

DR. LORGE: I was hoping that the other question would be asked; if the individual could identify the groups to which he belonged, could you then specify the groupness of the group of which he is a member, and if there be homogeneity in his group memberships?

DR. HEMPHILL: I think I would have to give the same answer to your question, Professor Lorge. We should study the problem. I do not know what the answer might turn out to be. I should like to re-emphasize, however, that we are quite convinced that we are studying something involving groupness when we are able to show that the members of a group give descriptions of their group which differ from descriptions of other groups by their members.

DR. WINANS: We have two rather simple questions. First of all, I would like Dr. Damrin to show us the other problems she mentioned; and second, I would appreciate it if Professor Hemphill would comment on the difference between his items *permeability* and *stability*.

DR. DAMRIN: As I told you, the problems are designed so that they increase logically in order of difficulty. In the first problem, the house, you will notice that the colors are all together in straight lines and that the square and triangular blocks are not inter-mixed.

In the second problem, the bridge, the colors are still separated, but the triangles and squares are mixed, and are thus more difficult to put together.

In the third problem, the dog, both the colors and shapes of the blocks are inter-mixed, thus making it the most difficult of all. On this problem, by the way, the children are asked to beat their time on the bridge problem. That is, they do not get fifteen minutes to do it, they get only the number of minutes it took them to build the bridge. This forces them to evaluate and improve their procedure.

DR. HEMPHILL: I presented very short definitions of these dimensions. I think a little elaboration would make clear the difference between stability and permeability. Permeability refers to a characteristic of the group which is often reflected in membership requirements, or the difficulty of getting into the group—what you have to do or be in order to qualify for membership. Stability refers to whether the group remains the same with respect to all significant characteristics over a period of time—in other words, a stable group in 1950 might be exactly like it was in 1899.

DR. SPANEY: I should like to ask Dr. Damrin whether she plans to extend her experiment to include verbal, or does she plan to stay with what she has. I am thinking particularly of adult problems.

DR. DAMRIN: At the present time, we do not plan to extend the construction type of task and develop a set of tests that are more verbal. The chief reason is that verbal tests depend rather heavily on intelligence, and we are trying to keep intelligence out of the picture as much as possible. These problems are sufficiently simple that even classes of very dull children can do them. In fact, we have successfully administered them to classes with an average I.Q. as low as 90.

In the adult area we have found that the test as it now exists is quite appropriate, although the present scoring system is not completely so. I have given the test to my research board, all of whom are Ph.D.'s, and in general the obtained results were quite satisfactory in terms of our present scoring system. It is our feeling, therefore, that to go into the area of verbal problem solving would serve primarily to add unnecessary complexity, and the test, as it now stands, already has so much complexity that it is extremely difficult to score.

DR. HARMON: I should like to ask Dr. Crutchfield if he has any information on the validity of these conformity scores against criteria

of importance to the individual's adjustment on the job, or anything of that kind, other than ratings or other test scores.

DR. CRUTCHFIELD: We do have some external criteria of performance in the study of professional men. Here, the zero order correlations with conformity scores do not seem very promising. There are, however, questions of whether the criteria are adequate and how they should be predicted to relate to conformity. All that I can really say is that we have some work in progress on this point, but no clear results as yet.

DR. COFFMAN: I should like to ask Dr. Damrin a question. She has made the statement in the general presentation that she has no data on validity or on reliability and did not add the statement, "of the traditional sort." There are three questions involved here: one is, does the behavior of the classes on this test relate to any other evidence of what the group problem-solving ability of the class is? Second, does the test differentiate among classes? And third, can scorers see the same thing? Do you have any evidence on those questions?

DR. DAMRIN: My statement that I have no data of a formal type referred to the fact that I have no tables of correlation coefficients to show you.

In answer to the first question, the pretrial study on this test was done in Trenton, New Jersey. The elementary school supervisor there knew her teachers very well and selected for us, but unknown to us, pairs of teachers. That is, from any particular school she would pick one teacher who she thought was trying to get across in her classroom the things which the Russell Sage test is concerned with, and for the other she would select a more formal type teacher. In every instance but three, the tests discriminated these two types of teachers. In other words, in the more formal class there would tend to be very low group participation; in the other, greater participation. In one class, student leaders would tend to emerge; in the more formal class this would not occur, and so on. Therefore, we have those kinds of evidences about the validity of the tests.

To the question of does the test discriminate between classes, the answer again is yes. Preliminary scaling of the scores that I have described to you reveals that classes differ very markedly from one to another, and yet the same class is relatively consistent in its behavior throughout the three problems. This gives us one type of evidence about test reliability.

To the third question, do scorers see the same thing when observing children on the test, I can only answer that I hope they do. I have just finished training two teams of examiners in New York. For the

planning stage we seem to be getting reliability that I would roughly estimate is about .9. This is on the basis of one week's training. On the operations stage I am afraid the reliability is about .5. This is because factors such as the "psychological tone quality" of a group of children require much more actual experience to evaluate accurately than do factors such as "amount of group participation."

My present feeling is that the data for the operations stage are not yet usable because of this fact.

DR. MCGINNIES: Dr. Crutchfield, I was interested in your description of some of the parameters that seemed to influence conformity behavior, particularly those measures of personality characteristics, and it sounded as though the nonconformists had all the desirable personalities while the others had the undesirability traits. If this is true, I wonder whether you could say a little bit more about some of the instruments that you used to arrive at these conclusions.

DR. CRUTCHFIELD: I should emphasize that the particular way the studies have been conducted has not been such as to measure extremes of nonconformity, that is, negative suggestibility effects, so to speak. This could easily be done. There is some incidental evidence in the results which would lead one to suppose that overinsistence on independence may at times be associated with undesirable personality features. But this has not yet been fully investigated.

Coming to the main point of your question, the instruments used in the personality assessment are very varied. The assessments of the professional men and of the mature women lasted three full days, and involved such procedures as standard tests, specially designed personality inventories, group interaction procedures of various kinds, interviews, projective techniques, physiological and medical measures, and so on; in short, the usual range of things found in large-scale assessment. The correlations I was reporting are drawn from all of these.

One representative finding is from the Terman Concept Mastery Test, a newly devised instrument measuring superior intellectual performance, which correlates approximately .50 with independence.

Some of the scales of the MMPI and similar personality inventories were also used. Another very important source of information came from Q-sort descriptions of the assesses made by the assessment staff in complete ignorance of their performance in the conformity procedure. Here there were found quite striking relationships along the lines I outlined.

Finally, since these days the famous or infamous F-scale studies of personality are much in the literature, I should report an appropriate finding there, that correlation with conformity is of the order of .40.

DR. LORGE: I take the Chairman's prerogative in summary. The three papers, by design, make a beautiful unity in this important area. They give a sagacious attempt to get at a concept of the nature of the group, with the possibility that it will enable workers to interpret the degree to which published evidence about groups really represents integrated and functioning groups. Perhaps I am too disturbed by *ad hoc* groups created in classes by instructors saying, "The next five individuals will constitute group 1 and the next five group 2, etc." for purposes of measuring group performance or group productivity. They show that in some process observations it is quite clear that group behavior is related to the fact that some members may not work at all.

And the papers show interesting dual relationships in attempts to evaluate group process as well as the group product. The complexity of relationships exhibited are: (1) between kinds of groups, (2) between kinds of members, (3) the nature of the processes related to groups and members, and (4) the quality of the product, whether it be a solution, an action program, etc.

A large number of theoretical issues come to the fore, because as far as I can tell, the evidence about the relationship between group process and group product is just emerging. More observations and data are needed, not only in the area of action, but also in the area of the quality of the group solutions as opposed to individual solutions. The distinction betwen the productivity of individuals and of groups must be made.

ADDRESS

. . . . And Have Not Wisdom

DANIEL STARCH

When your chairman invited me to address you here today, he suggested that I need not necessarily talk about technical problems but that I might discuss anything I thought relevant to the purposes of your conference. I shall, therefore, deal with some problems which have been forced upon us by the events of the last fifty years. They are especially vital to all of us engaged in the study of human behavior. The solution of these problems will most assuredly affect our well-being. Failure to solve these problems aright is certain to affect volcanically the culture and welfare not only of ourselves but of many generations to come.

The problems to which I refer have become particularly acute in our time because they arise out of the wide, yawning gap that exists today between the tremendous advance in science, particularly applied science, and the appalling backwardness still manifest in human relations. Our advance in scientific knowledge has been so rapid that the dreams of Jules Verne have seemingly come true, but our backwardness in humane behavior, in dealings between individuals, between groups, between peoples and nations, has been so retarded that it is open to question whether there has been any progress at all in the general spiritual level of human relations since the Greeks took Troy with their wooden horse.

First then, let us look for a moment at the growth of science. Interest in scientific method began to germinate, to be sure, as far back as Aristotle, but more especially three centuries ago when Francis Bacon wrote his *Advancement of Learning* (1605). But it is only during the past fifty years that the great burst in the growth of science has occurred. I think it is fair to say that more advance in scientific knowledge has occurred during the last fifty years than in all previous centuries combined. We are all aware of this in a general way, particu-

larly those of us whose personal recollections go back to the turn of the century. We know from our own observation the fantastic changes in the number of telephones, automobiles, machines, and gadgets of all kinds. But let me deal briefly, though in reasonably specific detail, with a few fields that may serve to dramatize this unprecedented growth.

Take, for example, travel and transportation. When George Washington was President of the United States, it took longer to travel from his home at Mount Vernon on the Potomac to New York than it takes today to fly from New York to Bombay. If we visualize the earth in 1800 as a globe eight feet in diameter, this globe in terms of travel time had shrunk by 1900 to the size of a basketball and by 1954 to the size of a golf ball.

This phenomenal advance in scientific know-how is fundamentally associated with the development and use of energy. Man has been able to increase his productivity, not by increasing the strength of his arm and back, but by the use and control of energy in the form of tools and machines. Around 1800 steam engines were just coming into use. By 1850 the United States had two million horse power, or one-tenth of one horse power per capita. By 1900, it had risen to one-half a horse power per capita and as of today the total energy in all prime mover engines in the United States is over six billion horse power, or about 39 horse power per capita. (Estimated by John A. Waring, Jr., *Steelways Magazine,* August 1954.) This is eighty times as much mechanical power as was available only fifty years ago. Productivity per man hour in manufacturing industries has increased fivefold since 1900.

This enormous advance in technology has been stimulated by, and made possible by, the manifold increase in technical education. Around 1900 there were 1,150 electrical engineers. Today there are 46,000, a forty fold increase. In all fields of engineering combined, there were at the turn of the century about 13,000 members in the various professional engineering societies. Today there are 170,000, or thirteen times as many.

But this is not all of the broad picture of rapid change. Take the study of man himself, his characteristics, behavior, and motivations. Here also in the last fifty years, a tremendous acceleration in scientific research and knowledge concerning human behavior and human relations has taken place. To cite merely one example, around 1900 there were only about 125 members in the American Psychological Association. Today there are more than 12,000, a growth in technically trained personnel exceeding even the relative gain in members of professional engineering societies. Add to that similar growth in number of trained

economists, sociologists, political scientists, and historians and you have a truly impressive army of professional talent devoted to human activities and relations, whereas fifty years ago the picture was largely one of solitary pioneers.

Take another aspect, channels of communication, messages and materials for influencing people's behavior have increased at an unprecedented rate both in numbers and in speed of transmission. Today there are in the United States one and one-fourth copies of newspapers per day per family, two and one-half copies of magazines per week per family, two radios per family, and one television receiver per two families.

Mass communication has so speeded up that it is almost instantaneous. An important event occurring in Washington at breakfast time will be known in all important capitals of the world before your toast has cooled off. By contrast, when Lewis and Clark arrived at St. Louis in December 1803 to start their westward exploration, the Spanish commandant stationed at St. Louis refused to let them cross the Mississippi because he had not yet received official word that the territory west of the Mississippi had been ceded to France and sold to the United States months before.

Look at still another area—education. School enrollment has taken an immense upward surge. The number of pupils and the number of years in school have increased by leaps and bounds in the last fifty years. In 1900, the total enrollment in our high schools was 700,000. Today it is 7,000,000. Around 1900 the number enrolled in our colleges and universities was 115,000. Today it is 2,500,000. In other words, high school enrollment has increased tenfold and college enrollment more than twentyfold. Put still another way, 80% of our young people of high school age are enrolled in our high schools and 30% of college age are enrolled in our colleges. Hence, as of now, of all adults 25 years or older, 60% have attended high school of whom one-fourth have also attended college.

Now let us take a searching look at what has happened in the management of human affairs. Has there been any improvement at all in the spiritual level of society in the fifty years during which these marvelous advances in the physical and cultural standards have occurred, or even during the three centuries since the days of Francis Bacon, or if you will, since the days of Aristotle, Moses and Confucius? Here we are in a field that is without a generally accepted measure of value. But speaking in all humility, I for one doubt that we have made any real progress. I doubt whether we have learned to live more wisely. We certainly have not learned to live in peace. On the contrary, wars have not

diminished in frequency and have become more brutal and destructive than ever. The first World War cost the United States alone 30 billion dollars and the second World War 380 billion dollars. But far worse, of course, is the destruction of human life. During the Thirty Years War three centuries ago, one-third of the population of Europe died. During the first and second World Wars lasting nine years, twenty-five million people were killed. It may well be that more people have perished through war in the last fifty years than in all previous known history, and these are the fifty years during which has occurred our boasted rise in technology and know-how, both in the physical sciences as well as in the human behavior sciences—in schooling and in the means of mass communication for influencing men's minds. Can it be that we have not realized that with increase in technical knowledge there must be a corresponding increase in responsible humane behavior if wholesale brutalization, nay even universal destruction, is to be avoided?

Turn to the behavior of persons as individuals in their family and community relations. According to United States government statistics, there was, in 1900, one divorce to thirteen marriages. Today there is one divorce to four marriages, a ratio three times as high today as fifty years ago. According to F. B. I. records, offenses including murder, manslaughter, rape, robbery, assault, burglary, larceny, and auto theft increased 37% in the thirteen years from 1939 to 1952.

Why is there so wide a gap between what we know and what we do with what we know, between knowledge and the wisdom of how to use that knowledge? There are at least three reasons. The first is that many people do not recognize that knowledge and use are two different things. Knowledge as such does not guarantee right use, in fact it does not necessarily assure use of any kind. Only purpose can determine use. The invention of a method of making fire was a landmark in the march of man, but knowing how to make fire did not determine the purpose for which the fire was to be used. It could be used to warm the hut and cook the food, or it could be used to burn down the neighbor's hut and destroy his crops. Whether nuclear energy will be used for man's benefit or for his destruction will depend on purpose. Science provides technology but not teleology.

The second reason is a by-product of the first. Technological advance makes comfortable, luxurious living an easy choice. This tends to push aside long range consideration of wise living. We feel so comfortable, so luxurious, so smart that we falsely feel self-sufficient, superior and trouble free. As necessity is the mother of invention, so trouble is sometimes the mother of humility.

The third reason is that knowledge generates power, and power over the forces of nature, and particularly over the behavior of man, whets the appetite for more power, for going rough shod over the sensibilities and rights of others. Unless power is accompanied by an increase of sensitivity to the common weal, it ends in abuse, oppression and brutality. As Lord Acton put it, "Power corrupts, absolute power corrupts absolutely."

The net effects of our fabulous advance in science are three things: we work less, about half as many hours as formerly; we live longer, about 70 years against an average of 25 years; and we live more comfortably. *Quaere,* however, whether we live more wisely and more happily. We worship the god of science in the laboratory but unfortunately neglect the goddess of wisdom in the temple. As an interesting side-light, consider the case of the five employees, all foremen, sent in 1953 by the N. V. Nederlandsche Linoleumfabriek of Amsterdam to work at regular American wages three months in the Armstrong Cork Company plants in Pennsylvania. They came not to learn new skills or to obtain technical knowledge of linoleum production but solely to get the "feel" of America. These men were greatly impressed by the facilities, the ownership of automobiles, and particularly the amount that they were able to earn and save during their short working period. One man commented that he saved enough to buy a good used automobile and would have bought one if he could have driven it all the way back to Amsterdam. He questioned, however, whether an American is any happier with his automobile than a Dutchman with his bicycle.

What remedy is there? How can advances in science be made to produce advances in wisdom? The remedy is not, I am sure, as someone suggested, to halt research in order to give ourselves a chance to catch up with our knowledge. That would merely halt the progress of science without any gain whatever in wiser living. H. A. Overstreet has remarked, "Better a dish of illusion and a hearty appetite for life than a feast of reality and indigestion therewith." But this alternative so vigorously expressed is by no means a necessary alternative. By all means let us continue research and more of it, but let us also search to find out what needs to be done to live more wisely. Therefore, I wish to suggest for your consideration a three point plan of action to close the gap between the advance of science and our discouraging delinquency in human relations.

Having recognized that knowledge of itself does not automatically lead to wise use, let us then first of all *make it a specific goal to seek wise use of scientific knowledge.* Psychological research has shown that training in precise, logical thinking with mathematical concepts

does not automatically lead to more accurate, logical thinking in other areas and with other data and concepts.

It is essential, therefore, that all people of goodwill keep their goal of constructive use of science bright and shining both for themselves and for others. Those with evil intent keep their minds firmly fixed on their goals. But people of goodwill are more likely to go along passively without much thought about their direction. The goal-conscious persons, whether their goal be good or evil, have this dynamic advantage: They know where they want to go and they keep going.

What *is* wise use of knowledge and how can we determine whether a given use is wise? The decent, self-respecting life, whatever else it may require, has this one absolutely unfailing requirement: *Freedom of choice,* freedom of choice in thought, attitude, speech, action and occupation, and with it freedom to seek satisfying reward according to effort, ability and achievement. In order to give proper regard to the rights and opportunities of others, in order that everyone may have maximum freedom of choice in our complex world, let us call it the *optimum freedom of choice.*

The second point of action that I submit for your consideration is that *we must learn and teach how to make sound value judgments.* Discoveries of science and uses of knowledge can and must be evaluated by their contribution to the optimum life, toward life at its best with its maximum freedom of choice. Specifically, how can we do this? Let me suggest three criteria.

1. The *test of cause and effect.* Human behavior events are linked to each other by cause and effect. They do not just happen from nowhere. They arise out of previous events and actions combined with human motivation. The first test, then, that I wish to propose for sound evaluation of ideas and actions is to ask this question, What will be the effect of the idea and of the proposed course of action based on it—not merely the immediate effect but especially the long range effect? One of the distinguishing characteristics of man is to be able to think ahead, to project himself into the future. But so much of our thinking is entirely in terms of immediate effects when in fact the important dynamic effects are the delayed, long range consequences. The immediate effects may be "sweet as honey in thy mouth" but the delayed effects may be "bitter in thy belly."

Two things are essential in visualizing the chain of cause and effect: full information and objective interpretation. These are particularly urgent when dealing with current problems and events as they are reported through the many channels with all the screening and slanting along the way. As an editor said, he did not care whether people read his editorials on the masthead page, he put his editorials into the

headlines on the front page. To test the extent to which news is slanted and screened, compare the same events as they are reported in different newspapers, magazines and over the air. Note especially the differences in the headlines and lead paragraphs. You will see what is omitted, what is emphasized, and what is slanted. In a free society we cannot and must not restrict news. Read and listen to all sides and sources and then ask, What seems nearest the truth and the right?

2. The second criterion for making sound value judgments is the *test of spiritual survival value*. Will the proposed course of action in the long run help or hinder, build or destroy, the optimum life? That cannot be done perfectly since it means looking into the future, but certainly it can be done far better if we set our minds to it.

3. The third and crucial criterion is the *test of personal substitution*. Are you willing to put yourself in the place of anyone else affected by the proposed course of action? Theorists and leaders may very well think they know what is good for everybody else so long as it is the other fellow who is affected, but no course of action is good unless it is good for both leader and led. In a dictator state, would the driver be willing to take the place of the driven? When the court in a recent trial gave the persons convicted of communism the choice between going to jail or going to Russia, they chose going to jail. This third test is obviously the golden rule, with the phrases reversed. Is the leader willing to have done to him what he proposes to do to the led. Human experience has found no better test.

I make no apology for stressing these points. The events of our times have forced them upon us. It is the especial province of psychologists, of all students and researchers in the human behavior sciences, economists, sociologists, political scientists, historians, journalists, but especially psychologists, to assume responsibility for closing the gap between right and wrong use of scientific knowledge. In his recent annual report, Dr. James R. Killian, president of Massachusetts Institute of Technology, urges a course half engineering and half humanities and social sciences and stresses the clear need for a larger place in the curriculum for psychology. But promoting research and providing instruction in these fields alone will not achieve the result. The key lies in the spiritual survival values, in understanding, appreciating and cultivating them.

And this brings me to the third point in my program of action: *Train young people in making value judgments as part of our educational system*—not necessarily as separate courses, although one course might well be established, but as parts and aspects in all courses dealing with human problems—and what courses do not? It may well be not only a part of courses in the human behavior areas—psychology, economics,

literature, sociology, history, and mass communication, but also in the physical, chemical and biological sciences. Training in making value judgments may follow procedures such as I have just outlined. Right value judgments and attitudes will not be developed unless conscious effort is made, and made with the materials in all areas of study and research. As psychologists, you know that we learn little from passive exposure.

Our day-to-day behavior is a mixture of (A) voluntary, purposely directed thoughts, attitudes, and actions, (B) involuntary, externally, and forcibly directed thoughts, attitudes, and actions, and (C) conditioned, habituated thoughts, attitudes, and actions. When "A" and "B" are repeated often enough, they drop down to the "C" level and automatically condition our lives. "A" is the avenue used in a free society. "B" is the whip used in a dictatorial society. It is at the "A" level that we as free individuals have maximum control of our own lives.

If the present rate of school enrollment continues for another generation, eight out of ten adults will have gone to high school and three out of ten will have gone to college. With this upsurge in school attendance, it is all the more urgent that education in long range spiritual survival values become a part of the cultural climate of our time. It is shocking to discover the utter illiteracy and immaturity in the thinking concerning the durable human values on the part of so-called educated, intelligent people.

Here is one example. Why should anyone get more than $7500 income a year? Analyze this through. Apply the criteria I listed a moment ago.

1. What would be the effect on society as a whole? On individual effort and incentive?

2. Would these effects help or hinder life at its best with optimum freedom of choice for everybody, not merely for the individuals concerned?

3. Would the enforcer of the idea be willing to exchange places with anyone else affected by the plan?

To think it through thoroughly and objectively would be good training in value judgments not only for teenagers but for many adults.

Take another idea—not prevalent today but held stubbornly a hundred or more years ago. Owners and managers of textile mills made it so difficult for adults and children to hold their jobs in the mills that it was next to impossible to get time off for children to go to school.

Again apply the same three tests. The stubborn obstinacy on the part of the mill owners led to the formation of labor unions—and that was a good thing. It was possible thereby to throw off this oppressive yoke. In turn, however, power of the union leaders led to extremes on

the other side, as always happens. Excessive power at one extreme generates counteracting forces which in turn lead to excessive power and abuse on the other side.

As I look back upon my [illegible] and teaching days in several large and highly respected [illegible] recall how professors would shy away from value judgments, moral and spiritual considerations as something on a level with [illegible]es.

Science has in its hands [illegible] of life and death. This is literally true not only in the physical [illegible] the psychological area as well. Look at the subtle mind conditioning and the brutal brainwashing and the frustration of the free spirit in [illegible] that we have seen going on in the last 25 years. Wars today are not won on the battlefield. They are won in men's minds before the fighting in the field even starts. I propose, therefore, to you psychologists especially interested in educational testing problems, two things:

1. That you urge the need for training in value judgments as part of our educational processes, and

2. That you devise tests for measuring value judgments and attitudes and changes produced by such training.

The far-reaching effects of many ill-considered, not to say irresponsible, decisions of men who govern so greatly outweigh the passive goodwill of the millions who suffer therefrom that the only cure can be in the goodwill of the millions to become directly active in understanding and cultivating value judgments and attitudes. A dictator can do with the stroke of a pen what millions know to be wrong but are helpless to do anything about. As Dr. Albert Schweitzer recently said, "Scientists must speak up."

And now to summarize briefly:

1. Recognize that knowledge of itself does not automatically lead to wise constructive use and that therefore we must set about specifically to seek and promote wise use.

2. Determine what is wise use by evaluating proposed ideas and courses of action against three criteria such as I have outlined:

   a. The test of cause and effect.

   b. The test of spiritual survival value.

   c. The test of personal substitution between leader and led.

3. Make the training of young people in value judgments a consciously recognized part of our school and college teaching.

It is evident what will decide our destiny, our ways of living, our happiness, our well-being or our misery, will not be the miracles of science by themselves but the stamina of our wisdom in using this knowledge. "Wisdom and knowledge shall be the stability of thy times, and strength of salvation."

GENERAL MEETING

New Developments in the Education of Abler Students

Acceleration: Basic Principles and Recent Research

SIDNEY L. PRESSEY

Sometimes a topic becomes so involved with situations emotionally charged that dispassionate consideration of it seems almost impossible. Occasionally an unfortunate label leads thinking astray. Educational acceleration is a subject which has suffered greatly from these handicaps—and more. It was closely associated with the war and two clumsy expedients then: college entrance without completion of secondary school, which aroused bitter antagonism in public school people; and lengthened college year, which burdened and antagonized college faculties. "Acceleration" implies hurry with probable superficiality. And every experienced educator has known bright youngsters double-promoted into an older group who there felt miserably out of place—perhaps he had been such a case.

In contrast, this paper will beg your open-minded consideration of the possibility that for abler students to progress through school at faster-than-average pace is normal *for them*, not hurrying; that there are ways of facilitating their progress which help rather than hinder good social adjustment; and that such steps can lessen the load and facilitate the work of our overcrowded schools. An additional gain should be very timely. Russian universities and technical schools appear now to be graduating about three times as many engineers and twice as many scientists a year as American institutions.[1] Multiple evidence indicates that facilitating the progress of able students leads more of them to complete collegiate and professional training; also probably the occasional notable genius (the Edisons or Einsteins) will thus be more likely to reach full fruition.

Two Major Neglected Facts of Human Development

Of basic importance are two conclusions from recent developmental studies regarding the gifted—in childhood and youth, *and* in adult career. First, gifted children tend to develop more rapidly than the average youngster, not only intellectually and educationally but also

[1] Meyerhoff, H. A., United States Shortage: Scientists, *U. S. News and World Report*, January 15, 1954, 46–9.

physically and in personality. The bright six year old is likely to be not only in intelligence, but also in reading and physique and social assurance, more like a second than a first grade child. The bright sixteen year old is probably not only in ability and in general knowledge up with the general run of eighteen-year-old high school seniors or even nineteen-year-old college freshmen; he probably reached puberty earlier than average and is in physique, interests, and social adjustment more mature than the average for his age. To have him in the eleventh or twelfth rather than the tenth grade, or to start the bright six year old in the second grade, is not to hurry him but rather to have him progress according to his real growth rate. Terman remarks that, "The exceptionally bright student who is kept with his age group finds little to challenge his intelligence and too often develops habits of laziness that later wreck his college career. I could give you some choice examples of this in my gifted group."[2] There is also evidence that holding a bright youngster back with his age group is less favorable to good social adjustment than carefully advancing him into a group more like him in ability and maturity of personality.

A second major question of educational policy regarding the able student is usually not faced squarely: if he is to be most productive in career and most adequate as a citizen, at about what age should he be through with full-time school and really begin his adult life? Vital statistics show the late teens and early twenties to be the healthiest years; physical tests and athletic records show them to be physically the most vigorous. And Lehman's very extensive findings regarding ages of most brilliant scientific discoveries, most important inventions, best writing, most remarkable paintings, all indicate that the most outstanding creative work is done early in the adult years—often in the twenties, sometimes even in the teens.[3] The total of such evidence is impressive. Nevertheless, American educational programs for our most able young people are being more and more lengthened; they must have not only college but also graduate or professional training, perhaps an "intern" experience—possibly even some post-doctoral work. Before the second world war, the median age of receiving the Ph.D. degree in this country was 30; now, it tends to be a little later. By the age of 25, Edison and Einstein were doing important creative work. If they were of that age in this country now, they would instead probably be worrying about their language requirements! However,

[2] Terman, L. M. The discovery and encouragement of exceptional talent. *Amer. Psychologist*, 1954, 9, 221–30; also *The gifted child grows up*. Stanford Univer. Press, 1947.

[3] Lehman, H. C. *Age and achievement*. Princeton Univer. Press, 1953.

Edison might have been spared such difficulties; he would probably not have been admitted to an American university, not having even been to high school!

The argument thus is that *able students should progress more rapidly than the lockstep rate through school and college because they develop more rapidly than the average young person, and should get into their productive careers earlier than occurs with the lockstep.* But what about possible social maladjustment, gaps in training, or damage to health?

Means of "Acceleration"

By a historical perversity, the worst means for rapid progress—grade-skipping in school and the lengthened school year in college—have been so much more used than better methods that, to many people, these worst ways are synonymous with "acceleration." The boy who is skipped from fourth grade to sixth (obviously a half-grade skip is less risky) may suffer at least briefly from ignorance of some arithmetic process taken up in the fifth grade. If he is a conspicuous only one thus to be advanced into a close-knit little sixth grade social unit, he may initially meet hostility—and parents may talk. But even this clumsy method — grade-skipping — need not cause much trouble. Bright children are usually ahead of their age in their reading and other subjects; a little help from a teacher or parent usually takes care of any omissions. If grade-skipping of bright pupils is made fairly common and if (as is common nowadays) changes in the membership of a grade group are frequent for other reasons, social difficulties are usually minimal. Trouble is likely only if the "accelerated" child is not really of superior ability but is pushed ahead because of parental pressure, or school work neurotically superior in compensation for a social maladjustment already existing. Grade-skipping is the easiest method of advancing a pupil, especially in a small school; and if "skippers" are selected carefully on the basis of adequate measurement (and a little help is given in adjusting for the skipped work and to the new group), outcomes are usually good—as will be seen shortly.

However, there need be no skipping over possibly needed school content, or moving an occasional child into a strange group. Children may be admitted to the first grade (or transferred from kindergarten) on the basis of mental age and reading readiness rather than chronological age; and some half-dozen investigations have all shown bright five year olds, so admitted, thereafter excellent in school work and in relations with other children. Or bright six year olds with some initial skill in reading may be started in the second grade. Or, a "primary

pool", may throw together all children usually in the first three grades, and move each on into the fourth grade when he is ready. Large junior high schools may have rapid progress sections, made up of bright youngsters with excellent school records to date and good social adjustment, which do the usual three years' work in two. And several careful investigations have shown these rapid-progress pupils doing as well in senior high school, academically *and* in relations with other students, as pupils of the same general ability and total record at the beginning of junior high but who spent the usual three years there. Similar rapid progress sections (three years in two) in senior high have shown similar success in college.

A few college students may be really fatigued by a lengthened school year, though a total of six weeks or more of vacation usually still remains; both fatigue and protest seem greater in the faculty. A few students who need the earnings or the experience of summer employment may miss these benefits. But the questionableness of a double assumption in the method must be stressed: that bright students can learn only in college courses, and that they must spend as much time in these courses as average students. A substantial number of studies have shown that students who obtained credit for a course by passing a comprehensive examination in it instead of going to class (preparation having been obtained from superior or extra work in secondary school, from independent study, travel, or otherwise), do excellently in later courses in that subject and also in total college record. Streamlined sections of college courses, with methods adapted to the superior student and with reduced hours in class to facilitate acceleration, have been reported as successful, also honors programs combining or replacing several courses, and guided independent study. *In short, the progress of superior youngsters may be facilitated in a variety of ways; most of them are better than grade-skipping or lengthened school year.* It need hardly be mentioned that wise use of each requires initial testing to assure superiority, and continuing measurement thereafter to guide progress and determine outcome.

Whatever the methods for rapid progress, modifications should be attempted of the present powerful social pressures for the lockstep. Thus, from the day of entrance, the college student is designated a member of the class which is to graduate *four* years later; the entrant in '54 is a member of the class of '58 (English universities are wiser—there, an entrant in '54 is a member of the class of '54). Each of the four years is given a distinctive name and status; special opportunities (as for presidencies of student organizations) are open only to students in their fourth year. A student who finishes college in three years soon

becomes a social anomaly who belongs nowhere. Instead the feeling should be fostered of belonging to the school not the class, associations and friendships should be freely formed within the total student body without class distinction. And status should be determined by academic progress and other accomplishments, not time served. That would be the democratic way; and it would benefit not only the gifted but also those who for some reason take longer than the usual time to finish an academic program.

Outcomes of "Acceleration"—and "the Cold War"

As already indicated in passing, the evidence is that most able students do not suffer from acceleration. But much evidence goes further. *They actually seem to gain!* When Terman compared those in his gifted group who graduated from high school young (under 15½) with those graduating near the usual age (over 16½) he found that 16% more of the younger group graduated from college, and 19% more took one or more years of graduate work. The young group married over a year younger, and had fewer divorces. And twice as many of the younger group were highly successful vocationally. Results to date of the Fund for the Advancement of Education, and of the Ohio State University investigations, are similar; these last studies show both those entering young and those taking a four year program in three years doing better in college, *and* having more successful careers after, than matched cases of the same ability but entering at the usual time or taking the usual four years for a degree.[4] And in most of the investigations to date the accelerates were not carefully selected as capable of rapid progress, or given guidance in acceleration: also, they moved forward mostly by these least satisfactory methods—grade-skipping or lengthened school year. There is every reason to suppose that better selection, guidance, and methods would bring yet better outcomes. For instance, 25% more of a group of bright freshmen selected as capable of acceleration and given guidance in so doing, finally obtained a degree than a controlled group not accelerating. A more effective education in *less* time—how can this be explained? Presumably because more rapid progress is normal for the gifted, and the beginning of adult career (and marriage and responsible citizenship) without long delays is biologically sound. There are of course certain obvious advantages, as that saved time and funds leave more of both for advanced training. There are also motivational gains of

[4] The Fund for the Advancement of Education. Bridging the gap between school and college, 1953. Pressey, S. L. Educational acceleration: Appraisals and basic problems. Ohio State Univer., 1949.

possible pervasive importance. It is in the American tradition, and stimulating to the bright and ambitious youngster, that there be opportunity to get ahead. The lockstep negates all that. You may have heard the jibe that a certain state penitentiary had a major advantage over the neighboring university: you could get out of the pen sooner if you did well. *Wise means for "acceleration" are legitimate opportunities for the able to forge ahead, which encourage a general climate of enterprise and lively effort.*

In Résumé and Final Application

Numerous studies of human development thus show that able youngsters should progress in school faster than the lockstep rate of a grade a year, not only because they develop faster but to prevent too long drawnout education from delaying productive careers. Wise methods of acceleration expedite progress without hampering social adjustment. With acceleration, more able students complete advanced training; they get into productive careers earlier. And these careers tend to be more fruitful. In addition, congested schools are relieved to some degree. It seems a not unreasonable estimate that each year there remain in the secondary schools perhaps 300,000 bright youngsters who ought well have been graduated. *More trained men, sooner, at less cost,* to counter the Russian technological threat! Clearly such a program calls for frequent appraisals from the primary years to graduate school, continuing guidance, and informed educational statesmanship. Surely, then, the topic is most appropriate to this conference.

College Admission with Advanced Standing

WILLIAM H. CORNOG

The School and College Study of Admission with Advanced Standing originated in discussions of the faculty of Kenyon College regarding the possibility of revising some of the rules governing requirements for the bachelor's degree, in order to enable very able students to save time and yet not lose the essential values of a liberal arts education. President Chalmers of Kenyon described the plan to friends and associates in schools and colleges and in 1951 a group of twelve institutions formed a Committee on Admission with Advanced Standing. The committee consisted of the administrative heads and representatives of the following institutions: Bowdoin, Brown, Carleton, Haverford, Kenyon, Massachusetts Institute of Technology, Middlebury, Oberlin, Swarthmore, Wabash, Wesleyan, and Williams. At a meeting in the spring of 1952, the college presidents and deans agreed to invite into the Central Committee twelve headmasters, principals, and superintendents and in May, 1952, the full committee of twenty-four met to organize what was thenceforth known as the School and College Study of Admission with Advanced Standing.

The Central Committee agreed that it would limit the study to eleven subject fields on the college freshman level, and in the summer of 1952, we proceeded to organize eleven sub-committees, of four colleges and three secondary school teachers each, in the following subjects: English composition, literature, Latin, French, German, Spanish, history, mathematics, biology, chemistry, and physics. The subject matter committees were given the task of defining in their respective subjects the standard of achievement of intensive courses in secondary schools which could be offered to the ablest high school students and for which the twelve colleges could give partial or full first-year credit toward their bachelor's degrees. In addition the Central Committee established from its own membership a Committee on Individual Development to examine and define the qualities of the able student himself and to advise the other sub-committees concerning elements of development not commonly included in the mere acquisition of knowledge.

To insure close contact with the college faculties at every point in the study, the college representatives on the Central Committee named correspondents in their institutions in each subject field. The sub-committees continually sought the advice of the correspondents and kept them informed of the progress of committee work. Committee assignments alone involved the participation of more than a hundred and thirty school and college teachers and administrators.

As is the case with all experiments, the School and College Study began with a series of hypotheses. We made some conjectures about American education and constructed an apparatus to test them and to search for solutions to the problems derived from our postulates. We assumed 1) the continuity of education in school and college, 2) the virtues of the liberal arts continuum, 3) the mutuality of interest and understanding of school and college teachers of the same subject matter, 4) the immediacy of the need to revise the timetable and better to utilize the time of the ablest students, 5) the possibility of describing desirable revisions and means of better utilization in terms of specific subject matter definitions in eleven college freshmen course areas.

Some interesting and challenging propositions may be derived from these basic assumptions. We have been committed to the testing of such concepts as the following:

1. That able students can and should be given more intensive instruction in secondary schools and have the opportunity to qualify for admission to college at a level higher than freshman entrance in specific subjects in which evidence can be shown of strong preparation, the equivalent of first-year college work.

2. That committees of school and college teachers can define requirement for admission with advanced standing acceptable to our twelve participating institutions and broadly applicable in institutions of similar standards and aims throughout the country.

3. That acceleration of able students out of high school after two years or three is less desirable than enrichment of the high school curriculum and admission to college with advanced standing at the normal college entering age of seventeen or eighteen after high school graduation.

4. That the advancement of American education demands the strengthening of our secondary schools, particularly in those divisions in which the ablest students are enrolled, and that the colleges have an obligation to encourage the secondary schools who strive to establish and maintain high standards of academic achievement.

5. That sound learning is respectable and that academic subjects, the content of a liberal arts education, constitute worthy intellectual

and spiritual nourishment for young minds, if these disciplines are liberally and wisely taught.

6. That, finally, this study reaches beyond parochial considerations of departmental regulations and self-interest and even of college degree requirements that it may offer a challenge to American education truly commensurate with the dynamism of our culture, the wealth of our resources, and the still unawakened powers of our highly endowed youth.

Our committees recognized from the start the many difficulties and pitfalls involved in drawing blueprints and specifications by which our principles and propositions could be put to the test of action. They produced definitions of standards which present a broad range of topics, in order to give the secondary school teachers freedom for imaginative teaching and adventurous planning. We have avoided standardizing subjects within narrow limits and have encouraged variety in secondary programs for the gifted. We have tried to plan courses which are liberal also in terms of the academic achievement expected. The cumulative extracting of the best practices and highest standards of our twelve colleges could easily result in an unrealistic, "premium" type of advanced credit. Perhaps this has happened in some of our subjects, perhaps we have priced them out of the market, despite the rich endowments of our gifted students. Nevertheless, we shall not allow definitions and examinations to be the sole criteria for the granting of college credit for school courses. Much weight will also be given to other evidence of unusual ability and valued personal qualities, such evidence as will be found in the recommendations of principals, guidance officers, and teachers. It is, finally, important above all, in our opinion, to encourage and challenge the secondary schools to exercise freedom and a bold use of imagination in their teaching and planning, and to avoid as the plague "cram" courses for the bright student.

In June, 1953, the subject matter committees and the Committee on Individual Development submitted their final reports, which were published and circulated to the twelve college faculties for their vote. In the autumn of 1953, the faculties held discussions of the recommendations of the study and, with no college dissenting, voted approval of the experimental plan to consider for admission with advanced credit able students who had received instruction at the level defined by the committees and had met the standard set by the advanced credit examinations to be administered in the spring of 1954.

In the fall of 1953 also, schools associated with the study indicated their readiness and intention to present candidates for examination and advanced credit in accordance with the conditions of the experiment

outlined in the Central Committee's published *Reports of Committees, 1952-1953*. In 1953-1954 about 550 students in eighteen schools were enrolled in courses preparatory to application for advanced credit. In May, 1954, these students were examined in one, two, and in a few instances three and even four subjects. The total number of examination papers was about 1000. In the first year of the experiment, the number of examinations which we have been able to administer has necessarily been limited. This year we have expanded the number of experimental schools to forty-two.

No one can say at this point how many gifted students in how many secondary schools over the country will eventually be offered the opportunity and challenge of advanced courses. We know only that there are probably thousands of very able students who should be given more intensive instruction in our schools. We know that many secondary school teachers and principals are eager to do more than they are doing for their gifted students. We also know that education of the gifted as they should be educated costs money. We shall need, here and there, some allotments of extra teacher time in those high schools which undertake intensive preparation. Many of us already receive extra teacher time for remedial work. Certainly education of the handicapped, of the retarded and the slow learner, costs money and money which we are all glad to spend. We of this study believe in this type of expenditure; we further believe that the provision of such special services for the handicapped is not to be regarded as merely the discharge of humanitarian obligations, nor even to be justified to taxpayers by the arguments of sentiment. The handicapped are members of this society of free men and should receive these services as a birthright, and for those services this society may not take undue credit, as for some singular or added grace. By the same token, we hold that it is no less democratic to provide special educational services for the gifted. This provision is also *their* birthright, for democracy has the responsibility to afford opportunity for full personal development to all its citizens, and to each of them in ways and degrees commensurate with the person's endowment and his aspiration. If this is truly the extent of our society's commitment in education, we face the task of discovering to boards of trustees and boards of education how far short we are falling in meeting that commitment, and what necessary and moderately expensive steps we must take to give our gifted students as full a measure of their educational birthright as we give to their less endowed, and no more than equally deserving fellows.

The advanced examinations which our study administered last May were given not only to the advanced credit candidates in our eighteen

secondary schools but also to college control groups who had had comparable instruction in our selected subjects in the twelve colleges. These examinations have been read and scored, and while correlations of secondary school and college performance must be based on somewhat meager numbers it is a fair general conclusion that the secondary school candidates made very respectable showings. This is the more remarkable since it was necessary in many of our schools, in order to get the experiment under way, to make rather hurried revisions and to reach some compromise solutions. The schools are confident that with a few more years' experience a much more satisfactory job can be done. No information is yet available regarding the numbers of our school candidates who have received advanced credit in colleges this fall. It is perhaps significant that many of our schools report that their candidates will probably not seek advanced credit, but are content with the satisfactions of their enriched secondary school experience. This fact recalls the statement made in the report of our Committee on Individual Development:

"The Committee wholeheartedly supports the values of a liberal education and would caution school advisers not to let these opportunities for advanced work become vocational in primary intent. The gifted student should focus on the academic challenge and the possibilities for personal enrichment, and not be urged to think and aim for possible future professional fields too early in his development. We recognize the fact that gifted students are apt to feel a sense of urgency to get on with professional study, and may therefore look upon a plan for advanced credit as merely a time-saving, corner-cutting device toward that end. We have been aware of a tendency to regard both college preparatory courses in school and liberal arts courses in college as mere stepping stones of credit to the happy shore of a professional course of study and a career. The saving of time in school and college, though often important from the point of view of family finances and also by reason of the extraordinary pace of development of an individual student, is not the essence of this program for advanced credit. The essence of what we seek may rather be expressed in terms of compressing the school and college experience of the very able in such a way as to make its impact and meaning continuously more challenging and rewarding, and to awaken and constantly to stimulate the reason, imagination, ethical sensibilities, and capacities for moral commitment and action in our highly endowed youth.

"We deplore but accept the necessity of measuring scholastic progress by credit counts, but we do not accept the mere acquisition of credit as the aim of our plan. We do, however, hold that credit demon-

strably earned should be awarded and not be withheld because of scruples of academic bookkeeping. Our chief concern is that the college credit earned in a secondary school course represents an experience in personal growth as well as a mature grasp of the subject itself."

It should, finally, be emphasized that our examinations are but part of the attempt to define courses and standards. The central purpose of this study is to promote and develop courses and teaching in the schools. To this end, we have extended the responsibilities of our examining committees to visiting schools and to the holding of evaluating sessions of school and college teachers. One of the major affirmations of this study has been that almost incalculable good has accrued from the prolonged and frequent association of our school and college teachers. In late June of this year we held three conferences of our teachers, one in chemistry at Kenyon College, one in biology at Wabash, and one in history at Williams. The enthusiasm expressed by the teachers who participated lends support to the final statement of our committee on Individual Development:

"We are confident that school and college teachers who undertake to share the responsibility and delight of educating our gifted students as they should be educated will not only contribute to the founding of strong and mutually beneficial school and college relations, but will help to establish a standard of education in which the nation can take pride."

In summary, The School and College Study plan will, we hope, afford one means of achieving a balance of opportunity in American education by making a more nearly equitable provision for a comparatively neglected minority in our schools, our ablest students. It will also, we hope, help to place a new emphasis on certain neglected or unfashionable necessities in education. Among these are: the necessity of patient effort in learning, the necessity of mastery of one's subject in teaching, the necessity of addressing ourselves to the study of worthy things, the necessity of producing thoughtful men. It may be that we can produce thoughtful men by a curriculum that does not require our students to think, or that conditions them merely to agile adjustments to the trivialities of daily living. The burden of proof of this possibility lies upon those who maintain that the proper study of mankind is how to live by bread alone, in personal security and unassailable eupepsia. I am unpersuaded that the majority of our students are fit only to be taught to earn a living, salute the flag, make orderly exits from burning buildings, and come in out of the rain. To base education in any large scope on the assumption that common men cannot be touched by greatness or uplifted by genius is to conspire

with degradation and embrace despair. What is required to bring men into touch with greatness is, of course, superb teaching. For men must be worthy of the things they teach as well as of the things they learn. The School and College Study would, finally, express the hope that education in America can re-dedicate itself to the making of master teachers who can make us thoughtful men.

In March, 1884, in his twenty-first year, a young man who aspired to be a teacher wrote in his diary, "Out of the shabby, squalid, starving life if I come with scholarship about me and in me and on me, these are the things I should long to do." He meant, to lecture on Wordsworth, Emerson, Shakespeare, Dante, Goethe. The young man joined the faculty of Central High of Philadelphia in September, 1884, and for twenty-three years was a master teacher. He taught Shakespeare, and wrote the definitive ten-volume life of Benjamin Franklin. He delivered the oration at the dedication of Franklin's statue in Paris. When news of his untimely death reached Stratford-on-Avon, where he had spent many of his holidays, the local schools closed for the day. His influence upon a generation of men has endured with them, and the mark he made upon the school is ineradicable. His name was Albert Henry Smyth, and he left a legend and a legacy. I suppose the moral of the story is this: He was a graduate only of the Central High School; he never went to college. He was the son of an unsuccessful carpet salesman, but much was expected of him. At the High School, he had to study, was given copious opportunity to study, high and serious things, under teachers who were worthy of the noblest calling. We still have such teachers today. May their tribe flourish and increase!

Special Treatment for Abler Students and Its Relation to National Manpower

DAEL WOLFLE

In the year 1820 the professorship of astronomy at Trinity College, which carried with it the title of Royal Astronomer of Ireland, fell vacant. After due deliberation, the electors chose William Rowan Hamilton to this distinguished post, and to the research opportunities it provided. The appointment was a notably successful one for Hamilton served with distinction until his death and came generally to be acclaimed as one of the two or three greatest mathematicians of the English speaking countries.

The astonishing feature of Hamilton's appointment was his age and status: he was 21 years old and still an undergraduate student at Trinity College. The great boldness of suddenly transforming an undergraduate into a professor marks this as one of the outstanding examples of the early recognition and reward of high ability. But there have been many other, albeit less dramatic, examples. Newspapers and magazines last summer told the story of a coed who earned her bachelor's degree at the University of Illinois in a single year. Students who finish college in less than four years have been known to all of us. Lehman has piled evidence on evidence of the early age at which many great accomplishments in science, literature, and the arts have been made. Professor Pressey's excellent work at the Ohio State University constitutes a model of how to bring a larger number of able students into productive careers at a younger than usual age.

There is a common element in Professor Pressey's work, in the quickly-earned degree of the Illinois coed, in Hamilton's youthful attainment of professorial status: in each case a college faculty recognized and nurtured the outstanding ability of a young man or woman whose interests, earlier attainments, background, and financial resources had brought him into a college in which his ability could come under the observation of interested faculty members. It is important to identify such students and to provide them with opportunities for the full development of their abilities. But we must also remember that there are other highly capable young men and women who do not go to college and who do not come under the eye of an interested faculty

member. Because they lack the money or more frequently because they lack academic interest they go to work instead of to college. They constitute a portion of the nation's intellectual resources which we are not now utilizing as effectively as we might.

Before taking up this [illegible] greater detail, let me give a basis for being concerned about [illegible]mal utilization of the nation's intellectual resources. The usual justification consists of a comparison between anticipated supply and anticipated demand for scientists, engineers, doctors, school teachers, or professional men and women in other fields. These supply-demand comparisons are based upon rather crude analyses of the manpower market—crude because our techniques for estimating manpower demand are considerably more primitive than are our techniques for estimating a commodity demand. They result in such estimates as that in 1960 the nation will be short 40,000 doctors, will need so and so many more engineers or school teachers than are in sight, or will have an adequate supply of foresters or lawyers. Despite their limitations, such analyses have value; for a few years into the future they can be of aid in the counseling of students and they provide guide lines for the planning of educational programs, personnel policies, and manpower utilization programs.

But the supply and demand analysis is not the only justification for feeling a serious concern over the problem of how we utilize our intellectual resources:

"A society which permits a significant portion of its members to work at levels below their capabilities is failing to achieve its full potential strength. The ability of a society to progress, the ability to better the goals for which it strives, and the skill and wisdom with which it meets its challenges are likely to be the decisive factors in determining its fate . . . the wisdom of the society is the wisdom of its members. Thus a society can attain its full potential only when each of its members is enabled to contribute as fully as his individual abilities permit.

"Judged by this standard the United States has failed to reach the strength which it might attain, for it wastes the abilities of many of its most capable sons and daughters and thereby loses the contributions they might have made. Our potential is greater than our achievement; progress in science and the arts might go on at a faster pace than it does; the moral strength of the country might be greater than it is.

"This standard is fundamentally different from the supply and demand concept . . . Instead of asking *How many jobs are there for scientists?* it asks *How much scientific talent is not being utilized?* Instead of asking *How many additional social scientists and humanists*

could be employed? it asks *How much important and useful knowledge about humanity which we might learn are we failing to learn?*"[1]

Several kinds of information are required to answer the question of how much talent is being wasted in the United States. We need to know what happens to bright youngsters who go to college, what happens to equally bright ones who stay out of college, and how many there are in each group. A full examination of these questions would require agreement upon what we mean by good utilization of talent and agreement upon the levels and methods of measuring the ability we are talking about. I will take up these points later in a published account, but let me skip over them here and give only the highlights of a recently completed study.

In cooperation with the school and college officials of Illinois, Minnesota, and Rochester, New York, my colleagues and I in the Commission on Human Resources and Advanced Training studied a group of superior high school graduates of some 20 years ago in order to find out what they were doing and how much they were earning in 1954. We analyzed the differences in their occupations and incomes in terms of the ability they had shown in high school and in terms of the amount of education they had obtained after graduating from high school. The 8,000 or more subjects for whom we had fairly full records came from large and small high schools located in urban, suburban, and rural areas in three regions of the United States.

Ability, whether measured by score on a standardized test or by rank in high school graduating class, was related to adult occupation. Those who in the 1930s had been far above average as high school students were in 1954 more likely to be engaged in professional occupations than were those who were only slightly superior to the average high school graduate. Moreover, the occupational differences remained when we corrected for differences in amount of post-high school education. Higher-level positions were associated with higher ability, among those who had gone to technical schools, among those who had entered college but had dropped out before graduation, and among those who had received college degrees.

Similar differences in earnings were found. Those who had finished high school in the top 20 per cent of their classes received higher incomes in 1954 than did those who finished closer to the high school average. And that also was true within each of several educationally homogeneous sub-groups.

[1] Wolfle, Dael. *America's resources of specialized talent.* New York: Harper and Brothers, 1954. P. 137.

Because family background might well have contaminated the analysis, we took a look at what had happened to the sons of professional men, the sons of laborers, the sons of farmers, and the sons of men in other occupational groups. In general, family background showed a small influence on adult earnings and positions. For example, the sons of professional men were receiving higher incomes than were the sons of laborers, whether the comparison was made among college graduates, college dropouts, or those who had never attended college.

The differences in occupations and incomes which were associated with differences in ability, and those which were associated with differences in family background, were much smaller than were the occupational and earning differentials associated with differences in education. At any ability level, as ability had been manifest in high school, and for any background, as background was indicated by the father's occupation, college paid off. The superior high school graduates who had spent some time in college were better off, occupationally and financially, in 1954 than were those who had not, and those who had graduated from college were better off than were those who had entered college but had not remained long enough to earn a degree. Within the restricted range of ability with which we dealt it meant a few hundred dollars a year higher income to be near the top of the high school ability distribution than to be near the middle. It meant a few hundred dollars a year more to have the background of a professional home than to come from a laboring family background. But possession of a college degree meant two to three thousand dollars a year higher income than was received by otherwise comparable classmates who went directly to work after graduating from high school. Moreover—and this point I want to stress—the income differential was greatest for those of greatest ability. If income can be accepted as a rough measure of what a man is worth to society, then it seems that a man of moderately superior qualifications can earn more and contribute more to society with a college education than he can without it; a man of very superior qualifications can earn and contribute much more.

These findings lead back to the distinction I drew earlier between able young men and women who go to college and equally able ones who do not. Both groups are part of the nation's intellectual resources. If we are seriously concerned—as I think we must be—about how the nation uses its intellectual resources, we must interest ourselves in both of these groups, the ones who go to college and the ones who do not. The two groups are similar in that they include people of high ability,

but the most effective techniques of capitalizing on their abilities are somewhat different, and herein lies a problem of educational and social planning.

On the one hand we have able young men and women who are academically motivated. Because of family interests, educational success, the aspirations of their fellows, or for other reasons, they go to college and sometimes to graduate or professional school. The primary problem for them is to see that they get the best possible education. The primary danger is that they will be lost in the mass of less capable high school and college classmates, will be offered a less stimulating educational fare than would be best for them, and will, as a result, waste a portion of their time and effort and not develop as far as they might with a more appropriate education. There is a growing recognition of the special needs of these students. The programs described by Dr. Pressey and Dr. Cornog are specifically designed to enable them to make better use of their educational years than would be the case if they followed the route of the average student.

The other group includes those who are bright but are not academically motivated. There are substantial numbers of such persons. At an intelligence level one standard deviation above the population average about one man and woman in four currently graduates from college. At a level two standard deviations above the population average, about one in two graduates from college. The number going through college could be increased by offering more scholarships. But financial help is not the only remedy necessary. Interest is more often lacking than money. Encouragement, guidance, and an elementary and secondary school experience that will challenge their abilities are at least as necessary as are scholarships. To bring about these improvements will, obviously, require changes in attitude on the part of many people and will require action at the local community level. The task is one for the entire educational system, and for the society that supports and benefits from that system.

In conclusion I would like to make one final observation. It is this: such evidence as we have indicates that it is worthwhile to give greatest attention to those of highest ability. With limited resources, I would concentrate on the top one or two or five per cent; with more generous resources I would extend the effort over the top five or ten or twenty per cent. There are tens of thousands of young men and women who are or could be above average college students; there are quite a few with very superior ability; there are precious few William Hamiltons. The widely ramifying contributions which the very ablest can make are

so important that every reasonable effort should be made to identify and educate all of them. The nation needs the scientific progress, technological achievement, and moral strength that can come from the minds of its ablest sons and daughters. Those of us who can help to identify the youngsters of highest potential, who can point the way to overcoming the obstacles which now keep some of this talent from coming to full development, and who can help to develop the kinds of elementary, secondary, and collegiate education which is best adapted to the training of highly able youth share much of the responsibility for determining the future progress of society.

DISCUSSION

PARTICIPANTS

RALPH BERDIE, HERBERT S. CONRAD, EDWARD E. CURETON, WALTER N. DUROST, DAEL WOLFLE

DR. CONRAD: Are these people whom Dr. Wolfle was speaking about those who graduated from high school twenty years ago?

DR. WOLFLE: They graduated between 1920 and 1938.

DR. CONRAD: Do you generalize from that generation to the present one?

DR. WOLFLE: In general principles, yes. In terms of amounts and details, no. For example, the earning differential would change as you studied a large fraction of the population or studied graduates of a different time period.

Appendix

Participants—1954 Invitational Conference on Testing Problems

ADKINS, Dorothy C., University of North Carolina
AFFLERBACH, Janet, Professional Examination Service
AHMANN, J. Stanley, Cornell University
ALMAN, John E., Boston University
ALLEN, Kathryn M., Department of Education, Schenectady
ALLEN, Margaret E., Board of Education, Portland, Maine
ALLISON, Roger, Educational Testing Service
ALT, Pauline M., Teachers College of Connecticut
ANDERHALTER, Oliver F., St. Louis University
ANDERSON, Rosa G., Psychological Corporation
ANDERSON, Roy N., North Carolina State College
ANDERSON, T. W., Columbia University
ANDERSON, Mrs. T. W., New York City
ANGELL, George W., Jr., Educational Testing Service
ANGOFF, William H., Educational Testing Service
ANTHONY, William, Maryland State Department of Education
APPEL, Valentine, Greenwich, Connecticut
ARMSTRONG, Fred, Lehigh University
ARONOW, Miriam S., New York City Board of Education
ARSENIAN, Seth, Springfield College, Massachusetts
BALLISTY, Al, Personnel Department, Philadelphia
BANNON, Charles J., Crosby High School, Waterbury, Connecticut
BARDACK, Herbert D., New York State Department of Civil Service
BARNES, Paul J., World Book Company
BARTELME, Phyllis F., Blythedale, New York City
BARTNICK, Robert, Educational Testing Service
BATTISON, Mrs. C., Remedial Education Center, Washington, D. C.
BAYROFF, A. G., Personnel Research Branch, AGO
BEDARD, Joseph A., Public Schools, New Britain, Connecticut
BEMENT, Dorothy M., Northampton School for Girls
BENDA, Harold, New Jersey State Department of Education
BENNETT, George K., Psychological Corporation
BENNETT, Ralph, New York City
BENSON, Arthur L., Educational Testing Service
BENT, Alma, State Teachers College, New Paltz, New York
BENTZ, Jon, Sears Roebuck & Company
BERDIE, Ralph F., University of Minnesota
BERG, Joel, University of Connecticut
BERGER, Bernard, Municipal Civil Service Commission, New York City
BERGESEN, B. E., Personnel Press, Incorporated
BLAUL, R. Elizabeth, Highland Park High School, Illinois
BLOOM, Benjamin, University of Chicago
BOASI, Veronica, Department of Personnel, New York City
BOGAR, Jack, Board of Education, Richmond, Virginia
BOGOSIAN, Dorothy, Queens College
BOLDT, R. F., Educational Testing Service
BOLLENBACHER, Joan, Cincinnati Public Schools
BONNER, Hubert, Columbia University
BRACE, Susan E., Archdiocesan Vocational Service, New York City
BRANDT, Hyman, American Occupational Therapy Association
BRANSFORD, Thomas L., New York State Department of Civil Service

BRAY, Douglas W., Columbia University
BRIDGMAN, Donald S., American Telephone and Telegraph Company
BRISTOW, William H., Bureau of Curriculum Research, New York City
BRODERICK, J., Lawrence, YMCA, New York City
BROGDEN, Hubert E., Personnel Research Branch, AGO
BROLYER, Cecil R., New York State Department of Civil Service
BROOKS, Richard B., College of William and Mary
BROWN, Frederick S., Great Neck Public Schools, New York
BRYAN, Ned, Rutgers University
BRYAN, Miriam M., Silver Burdett Company
BUCKTON, LaVerne, Brooklyn College
BUEL, William D., Temple University
BURDOCK, Eugene, Carnegie Corporation of New York
BURKE, James M., Darien Public Schools, Connecticut
BURKE, Paul J., Bell Telephone Company
BURNHAM, Paul S., Yale University
BUROS, Oscar K., Rutgers University
BYRNE, Lois A., Temple University
BYRNE, Richard Hill, University of Maryland
CADWELL, Dorothy, Civil Service Commission, Canada
CAMPBELL, Donald W., Board of Education, Newark, New Jersey
CAPPS, Marian P., South Carolina State College
CARLSON, C. Ray, Air University, Maxwell AFB
CARLSON, Harold S., Upsala College
CARLSON, J. Spencer, University of Oregon
CARROLL, John B., Harvard University
CELLIERS, Peter J., *Pathfinder . . . The Town Journal*
CHAUNCEY, Henry, Educational Testing Service
CHURCHILL, Ruth, Antioch College
CLARK, Willis W., California Test Bureau
COBB, William E., Pennsylvania State University
COCKLIN, John H., Temple University
COFFMAN, William E., Educational Testing Service
COHEN, Philip S., State Teachers College, Montclair, New Jersey
COLEMAN, William, University of Tennessee
CONRAD, Herbert S., U. S. Office of Education
CONWAY, C. B., Department of Education, Victoria, British Columbia
COPELAND, Herman A., Atlantic Refining Company
CORNEHLSEN, John H., Tufts College
CORNOG, William H., Central High School, Philadelphia
COX, Henry M., University of Nebraska
CRANE, Percy F., University of Maine
CRAVEN, Ethel C., Polytechnic Institute of Brooklyn
CRAWFORD, Barbara, Educational Testing Service
CRAWFORD, J. R., University of Maine
CRISSY, W. J. E., Personnel Development, Inc.
CRONBACH, Lee J., University of Illinois
CRUTCHFIELD, Richard S., University of California
CUMMING, Mary, Boston Public Schools
CURETON, Edward E., University of Tennessee
CURETON, Louise W., Board of Examiners in Psychology, Tennessee
CYNAMON, Manuel, Brooklyn College
DAHNKE, Harold, Michigan State College
DAILEY, John T., Bureau of Naval Personnel
DALY, Alice T., New York State Department of Education
DAMRIN, Dora E., Educational Testing Service
DAVIDOFF, M. D., U. S. Civil Service Commission
DAVIDSON, Helen H., City College of New York

DAVIS, Fred B., Hunter College
DAVISON, Hugh M., Pennsylvania State University
DEES, Bowen C., National Science Foundation
DEGAN, James W., Massachusetts Institute of Technology
DEVINNEY, Leland C., Rockefeller Foundation
DIAMOND, Lorraine K., Teachers College, Columbia University
DICK, G. W., International Business Machines
DIEDERICH, Paul B., Educational Testing Service
DION, Robert, California Test Bureau
DOBBIN, John E., Educational Testing Service
DOPPELT, Jerome E., Psychological Corporation
DRAGOSITZ, Anna, Educational Testing Service
DRAKE, L. E., University of Wisconsin
DRUZ, William, New Jersey Civil Service Department, Trenton
DUKER, Sam., Brooklyn College
DUNN, Frances E., Brown University
DURAN, June C., California Test Bureau
DUROST, Walter N., Test Service and Advisement Center, New Hampshire
DWYER, Paul S., University of Michigan
DYER, Henry S., Educational Testing Service
EBEL, Robert L., State University of Iowa
EDRINGTON, T. C., Department of Defense
EDWARDS, Robert E., Rensselaer Polytechnic Institute
ENGELHART, Max D., Chicago City Junior College
EPSTEIN, Bertram, City College of New York
ESTAVAN, Donald, Educational Testing Service
EWERS, Dorothea, Bloom Township High School, Chicago Heights
FAN, C. T., Educational Testing Service
FARNUM, Henry M., Human Engineering Devices
FARR, George C., International Business Machines
FEINBERG, Mortimer R., Bernard M. Baruch School and Research Institute of America
FELDT, Leonard S., State University of Iowa
FENDRICK, Paul, Western Electric Company
FERGUSON, William C., World Book Company
FERRIS, F. L., Jr., Educational Testing Service
FIELDS, Mrs. Carlyle, Croydon Hall, Atlantic Highlands, New Jersey
FIFER, Gordon, Test Research Service, Inc.
FINDLEY, Warren G., Educational Testing Service
FINK, August A., Jr., Columbia University
FINKLE, Robert B., Metropolitan Life Insurance Company
FIRTH, Louise, Educational Testing Service
FLANAGAN, John C., American Institute for Research
FLETCHER, Frank M., Ohio State University
FOLGER, John, Southern Regional Education Board
FORLANO, George, Board of Education, New York City
FORLANO, Mrs. George, Board of Education, New York City
FORRESTER, Gertrude, West Side High School, Newark
Fox, William H., Indiana University
FREDERIKSEN, Norman, Educational Testing Service
FREEMAN, Paul M., Educational Testing Service
FRENCH, Benjamin J., New York State Department of Civil Service
FRENCH, John W., Educational Testing Service
FRIEDENBERG, Edgar Z., Brooklyn College

FRIEDMAN, Sidney, Bureau of Naval Personnel
FULTON, Ailsa W., American and Foreign Teachers' Agency
FULTON, Renee J., Board of Education, New York City
FURST, Edward J., University of Michigan
GALLAGHER, Henrietta L., Educational Testing Service
GARDNER, Eric F., Syracuse University
GAWOSKI, Roman S., Yonkers, New York
GERBERICH, J. R., University of Connecticut
GIDDINGS, Frank, Springfield Trade High School, Massachusetts
GLASER, Robert, American Institute for Research
GLASS, Albert A., The Signal School, Fort Monmouth
GODDARD, W. A., International Business Machines
GODSHALK, Fred I., Educational Testing Service
GOLDSTEIN, Leo S., Teachers College, Columbia University
GOODMAN, Joan, World Book Company
GOODMAN, Samuel M., Board of Education, New York City
GORDON, Mary Alice, Macy's
GRIMM, Elaine R., Professional Examination Service
GREEN, Bert F., Massachusetts Institute of Technology
GROVES, Kenneth J., Air University, Maxwell AFB
GUTHRIE, George M., Pennsylvania State University
GULLIKSEN, Harold, Educational Testing Service
HAERTEN, Heinz, Studienstiftung, Bad Godesberg, Germany
HAGEN, Elizabeth, Teachers College, Columbia University
HAGMAN, Elmer R., Board of Education, Greenwich, Connecticut
HALPERN, Joseph, New York State Department of Civil Service
HARMON, Lindsey B., National Research Council
HARTER, R. K., American Telephone and Telegraph Company
HARVEY, Philip R., University of Connecticut
HASTINGS, J. Thomas, University of Illinois
HAYWARD, Priscilla, Educational Testing Service
HEATON, Kenneth L., Richardson Bellows, Henry and Company
HEIL, Louis M., Brooklyn College
HELMICK, John S., Educational Testing Service
HELMSTADTER, G. C., Educational Testing Service
HEMPHILL, John K., Ohio State University
HILL, Walker H., Michigan State College
HILLS, John R., Educational Testing Service
HITTINGER, William F., Pennsylvania State College
HOBERMAN, Solomon, Department of Personnel, New York City
HOLLEY, Clifford S., Personnel Department, Philadelphia
HOLLIS, William H., New York City
HOLLISTER, John S., Educational Testing Service
HOLMES, A. F., Royal Canadian Air Force
HOLMES, Therese, Metropolitan Life Insurance Company
HOROWITZ, Milton W., Queens College
HORTON, Clark W., Dartmouth
HUDDLESTON, Edith, Educational Testing Service
HUNSICKER, Paul, University of Michigan
HUNTER, Genevieve, P., Archdiocesan Vocational Service
HURLBURT, Allan S., North Carolina State Department of Public Instruction
HYLLA, Erich, Hochschule Internationale Paedagogische Forschung, Germany
JANOWSKY, John, Personnel Department, Philadelphia

JAENICKE, Walter H., West Virginia University
JASPEN, Nathan, National League for Nursing
JEFFREY, W. E., Barnard College, Columbia University
JOHNSON, A. Pemberton, Educational Testing Service
JOHNSON, Lewis W., Personnel Department, Philadelphia
JORDAN, Arthur M., University of North Carolina
KABACK, Goldie R., City College of New York
KALIN, Robert, Educational Testing Service
KARON, Bertram P., Educational Testing Service
KEATS, J. A., Educational Testing Service
KIDD, John W., Michigan State College
KIGER, Joseph C., American Council on Education
KIMBALL, Elizabeth G., Educational Testing Service
KING, Richard G., College Entrance Examination Board
KIPNIS, David, New York University
KIRKPATRICK, Forrest H., Bethany College
KOLKEBECK, Robert F., Educational Testing Service
KOSMERL, Alice F., Washington, D. C.
KOSTICK, M. M., State Teachers College, Boston
KRATHWOHL, David R., University of Illinois
KUBIS, Joseph F., Fordham University
KUDER, G. Frederic, Duke Universtiy
KURTZ, Albert K, University of Florida
KUSHNER, Rose E., City College of New York
KUTCHER, Charlotte, American Public Health Association
LANGMUIR, C. R., Psychological Corporation
LANNHOLM, G. V., Educational Testing Service
LAYTON, Wilbur L., University of Minnesota
LEACH, Kent W., University of Michigan
LEAVY, Sylvia, Queens College
LEBOLD, William K., Purdue University
LEE, Marilyn C., Science Research Associates
LENNON, Roger T., World Book Company
LEV, Joseph, New York State Department of Civil Service
LEVINE, Richard, Educational Testing Service
LEVY, Charlotte, Chunky Chocolates
LIMBURG, Charles C., U. S. Air Force
LINDQUIST, E. F., State University of Iowa
LOHMAN, Maurice A., New York State Department of Education
LOOS, Gordon M., Educational Testing Service
LORCH, Vera, Great Neck, New York
LORD, Frederic, Educational Testing Service
LORD, Shirley, Educational Testing Service
LORGE, Irving, Teachers College, Columbia University
LORR, Maurice, Veterans Administration
LUND, Kenneth W., Public Schools, Chicago
LUSK, Louis T., Norwalk, Connecticut
LUTZ, Orpha M., State Teachers College, Montclair, New Jersey
MCARTHUR, Charles, Harvard University
MCCABE, Frank J., Metropolitan Life Insurance, New York City
MCCALL, W. C., University of South Carolina
MCCAMBRIDGE, Barbara, Educational Testing Service
MCCANN, Forbes E., Personnel Department, Philadelphia
MCCORD, Richard B., Personnel Department, Philadelphia
MCEWAN, Dorothy M., New York State Department of Civil Service
MCGILL, William J., Massachusetts Institute of Technology

McGINNIES, Elliott M., University of Maryland
McINTOSH, Vergil M., Air University, Maxwell AFB
McLAUGHLIN, Eugenia, New York State Department of Civil Service
McNAMARA, W. J., International Business Machines
MacPHAIL, A. H., Brown University
McQUITTY, John V., University of Florida
MACALUSO, Charles J., U. S. Naval Examining Center
MACHI, Vincent S., Columbia University
MANUEL, Herschel T., University of Texas
MARRIOTT, John C., World Book Company
MARSTON, Helen M., Educational Testing Service
MASIA, Bertram B., New York University
MATHEWS, Chester O., Ohio Weslayan University
MATHEWSON, Robert H., Division of Teacher Education, New York City
MAXSON, Georgia, Educational Testing Service
MEDLEY, Donald M., Division of Teacher Education, New York City
MELVILLE, S. Donald, Educational Testing Service
MERENDA, Peter F., U. S. Naval Examining Center
MERRY, Robert W., Harvard University
METZ, Elliott, Queens College
MICHAEL, William B., University of Southern California
MICHAEL, S. R., Educational Testing Service
MICHELI, Gene, Metropolitan Life Insurance Company
MILES, Matthew B., Teachers College, Columbia University
MILL, Cyril R., Public Schools, Richmond
MILLETT, Esther, Westover School, Middlebury
MITCHELL, Blythe C., World Book Company
MITZEL, Harold E., Division of Teacher Education, New York City
MOLLENKOPF, William G., Educational Testing Service
MORGAN, Antonia Bell, Aptitude Associates Incorporated
MORGAN, H. H., Psychological Corporation
MORRIS, John B., University of Minnesota
MORRIS, Nancy, Educational Testing Service
MORRISON, Alexander W., Polytechnic Institute, Brooklyn
MORRISON, J. Cayce, Puerto Rican Study
MORTON, Anton S., Educational Testing Service
MOSELY, Russell, Wisconsin State Department of Public Instruction
MUNGER, A. M., Standard Oil Company, New Jersey
MURRAY, John E., Special Devices Center, ONR
MYERS, Robert L., Temple University
NELSON, Kenneth D., Division of Research, Albany
NELSON, M. J., Iowa State Teachers College
NEVIN, Margaret, Educational Testing Service
NILL, Kathryn F., Rinehart & Company
NOLL, Victor H., Michigan State College
NORTH, Robert D., Educational Records Bureau
NORTON, Dee W., Air Force Personnel & Training Research Center, Lackland AFB
OLSEN, Marjorie, Educational Testing Service
ORLEANS, Beatrice S., Bureau of Ships, Navy Department
ORLEANS, Joseph B., George Washington High School, New York City
PACE, C. Robert, Syracuse University
PALMER, Orville, Educational Testing Service
PASHALIAN, Siroon, Queens College
PEARSON, Richard, Educational Testing Service