DEVELOPMENT OF A RECORD FORM FOR EVALUATING RESEARCH THROUGH THE REPORT

The fifth in a series of reports dealing with the evaluation and measurement of research performance.

Prepared under the sponsorship of the
Personnel and Training Branch
Psychological Sciences Division
Office of Naval Research

January 1954

AMERICAN INSTITUTE FOR RESEARCH

Pittsburgh, Pennsylvania

DEVELOPMENT OF A RECORD FORM FOR EVALUATING RESEARCH THROUGH THE REPORT

The fifth in a series of reports dealing with the evaluation and measurement of research performance.

Report prepared under the sponsorship of the

Personnel and Training Branch
Psychological Sciences Division
Office of Naval Research

by

James W. Altman, Research Associate

and

Mahlon V. Taylor, Jr., Project Director

Advisory Staff
John C. Flanagan, Director of Research
Elmer D. West Marion F. Shaycoft

Research Adviser
Paul Horst

Project Designation No. NR 153-146

January 1954

AMERICAN INSTITUTE FOR RESEARCH

AIR, 1500

CONTENTS

ACKNOWLEDGMENTS

The work described in this report was carried out under contract with the Office of Naval Research. Dr. Denzel D. Smith, Head of the Personnel and Training Branch, Psychological Sciences Division, served as monitor for the Office of Naval Research and provided valuable advice and assistance to the project. Work was conducted under the general cognizance of Dr. Howard E. Page as Head of the Psychological Sciences Division. Drs. Ralph M. Hogan and J. W. MacMillan monitored early stages of the project.

A special acknowledgment is due Naval Research Reserve Units 4-3 and 9-6 for their participation in the study. The assistance of Commander Joseph A. Wise and Mr. Louis J. Sparvero in obtaining the cooperation of these groups was very helpful.

Without the generous cooperation of more than 300 physicists, chemists, and engineers it would have been impossible to carry out the work of this project. Limitations of space preclude listing each individual, but the contribution of each is gratefully acknowledged.

Special thanks are also due to Dr. George H. Young, Director of Research, Mellon Institute, and the members of the Mellon Institute staff who contributed a substantial proportion of the evaluations obtained in this study.

The assistance of a number of persons during the formative stages of the project deserves particular mention. Among those who participated in these early phases are:

Dr. J. Alfred Berger
Head, Department of Metallurgy
University of Pittsburgh

Mr. Holbrook G. Botset
Head, Department of Petroleum Eng.
University of Pittsburgh

Dr. Harry W. Braun
Assistant Professor of Psychology
University of Pittsburgh

Dr. Patrick Conley
Manager, Ultrasonics Section
Westinghouse Research Laboratories

Dr. H. C. Corben
Professor of Physics
Carnegie Institute of Technology

Dr. James Coull
Professor and Head, Department of
Chemical Engineering
University of Pittsburgh

Dr. Alfred G. Dietze
Associate Professor of Psychology
Michigan State University

Dr. Malcolm F. Dull
Associate Professor of Chemistry
University of Pittsburgh

Dr. Max Garbuny
Research Physicist of Electronics Dept.
Westinghouse Research Laboratories

Dr. Bertram K. Hovey
Associate Professor of Electrical Eng.
University of Pittsburgh

Dr. Jack Matthews
Associate Professor of Psychology
University of Pittsburgh

Dr. George E. Owen
Assistant Professor of Physics
University of Pittsburgh

Dr. Frederick D. Rossini
Head, Dept. of Chemistry
Carnegie Institute of Technology

Mr. Philip E. Rush
Professor of Electrical Engineering
University of Pittsburgh

Mr. Leo A. Shinn
Chemist
Office of Naval Research

Dr. P. M. Stehle
Associate Professor of Physics
University of Pittsburgh

Dr. James R. Tomlinson
Group Leader
Gulf Research and Development Co.

Dr. William E. Wallace
Associate Research Professor of Chemistry
University of Pittsburgh

Mr. Theodore A. Werkenthin
Bureau of Ships
Navy Department

The project was carried out under the general direction of Dr. John C. Flanagan, Director of Research for the American Institute for Research. Dr. Elmer D. West and Miss Marion F. Shaycoft provided valuable advice throughout the project.

Other members of the American Institute for Research staff who contributed to the project include: Joanne P. Folley, Robert Glaser, Clifford P. Halin, Eugene Heide, Barbara Lupfer, George L. Murphy, and Beatrice M. Shriver.

Dr. Paul Horst has served as research adviser to the project staff throughout the project and many of the basic ideas incorporated in the design are his. Early phases of the project were under the direction of Mrs. Mary H. Weislogel. The authors of this report have depended in large measure upon her phase reports for a description of early phases of the study.

Finally, special thanks are due to Mrs. Anne M. Canzian who has had the responsibility for most of the administrative details connected with the project.

James W. Altman
Mahlon V. Taylor, Jr.

January 1954

ABSTRACT

This report describes the development of a procedure for evaluating research through the report as part of a comprehensive research program for the study of potentiality for and successful performance in research work. The program is sponsored by the Personnel and Training Branch, Psychological Sciences Division, Office of Naval Research. The total program of research includes the following steps:

1. Determination of the critical requirements for successful participation in research and engineering work

2. Development of an aptitude test for the selection of research personnel

3. Development of tests to measure proficiency in specific areas of scientific work

4. Development of procedures for evaluating the job performances of research personnel

5. Determination of the predictive value of the tests developed in steps (2) and (3) using the procedures developed in step (4) to obtain evaluations of personnel for comparison with test predictions.

The first three of the above steps have been completed. Preliminary work on step 4 has previously been reported. The present report deals with additional work on that step.

A preliminary version of the Record Form for Evaluating Research Through the Report was constructed directly from the first five areas of the Observational Record Form for Research Personnel which was developed in a previous study. The areas included on the form were:

I. Formulating Problems and Hypotheses
II. Planning and Designing the Investigation
III. Conducting the Investigation
IV. Interpreting Research Resul
V. Preparing Reports

The preliminary form and instructions for its use were revised on the basis of results from several small scale tryouts.

Chemists, physicists, and engineers who were listed in American Men of Science as fellows, officers, or past officers of their respective professional societies were contacted by mail and each requested to select one report of especially effective and one report of relatively mediocre research. The selected reports were then evaluated by other workers in the field using a trial Record Form

composed of 114 items descriptive of effective and 114 items descriptive of ineffective research performance. The trial form was revised on the basis of results from the evaluations and the comments of evaluators concerning the form. Fifty items describing effective performance were selected for the revised form. No ineffective items were selected since, as a group, they were less used and did not seem to function as well as the effective items.

The method of developing the revised Record Form for Evaluating Research Through the Report should maximize its value as an aid to careful evaluation. However, its usefulness for evaluation has not been fully determined. A study for this purpose is recommended.

The estimate of inter-evaluator reliability obtained in this study for the 50 items selected for the revised form was low. The estimates of intra-evaluator reliability were rather high. Although the estimates may be somewhat in error due to the effect of selecting the items from a large number of other items, the evidence obtained supports a conclusion that standards for evaluating research through the report, using items describing effective research performance, vary greatly from one evaluator to another. It is therefore recommended that item responses not be combined into a total score except in certain special circumstances. Individual items should be used by the evaluator as an aid to arriving at a careful over-all judgment about the value of the research. It is suggested that methods for reducing the variation among evaluators' standards may be a fruitful area for further research.

Chapter I

THE PROBLEM

Introduction

The impact upon modern society of research and the many complex technologies to which it is wedded hardly needs mention. We are continually reminded by the daily press, popular magazines, radio, and television that research and technics are playing an increasingly important role in our daily lives. There are many who decry the effect of scientific advance upon our way of life and there are perhaps still more who laud it as contributor and defender. But no one detracts from its importance.

Effective communication among research workers is an essential and time honored aspect of research. For without effective communication each research worker must spend so much time unearthing findings which have already been developed elsewhere that scientific progress is greatly impeded. There are many channels through which research workers communicate -- including professional meetings, postgraduate seminars, and the publication of basic reference books, but it seems safe to say that the single most important mode for communicating research knowledge is publication in professional journals. It is therefore surprising that the effectiveness of this communication, that is, the amount of information conveyed to different readers by scientific reports, has not been investigated. A survey of the literature* concerning scientific manpower revealed no systematic studies of this problem.

The main purpose of the present investigation was to develop a new instrument which might be useful to persons responsible for evaluating research through its report. However, much of the information obtained in this study should be relevant to an assessment of the agreement among research workers concerning research evaluated through the report. Such an assessment should provide valuable knowledge concerning the effectiveness with which written reports communicate to research workers. For an important function of scientific report writing is to communicate to the qualified reader the contribution which has been made by the reported research. Unless there is agreement among qualified readers concerning this contribution it is apparent that scientific communication is not functioning optimally. Many of the data gathered in the subject research project are relevant to this problem. There is a real need for procedures and instruments which will help to increase the effectiveness of research performance. There seems to be little likelihood that the demands for results of governmental research operations will be reduced in the foreseeable future. Rather, the need for new research information seems to be steadily increasing. On the other hand, it appears that governmental

---

* Weislogel, M. H. and Altman, J. W. Abstracts of Literature Concerning Scientific Manpower. Pittsburgh: American Institute for Research, 1952.

laboratories are reaching the saturation point in the number of professional research workers they will be able to absorb. As the numbers of professional workers begin to level off, while the demand for results continues to mount, there are two major ways of preventing a widening of the breach between the Government's imperative need for effective research and the ability of the laboratories to produce research.

The first way to increased research effectiveness is through careful evaluation and development of currently employed research workers. This includes matching the right man with the right job, letting the man know his own strengths and weaknesses in an objective and constructive manner, and rewarding outstandingly effective performance with recognition and increased responsibility. The misplaced or unrecognized worker is likely to seek new employment, with a resulting waste of the time and money that went into recruiting, orienting, and training him. Even if he remains on the job, his achievement will fall below capacity.

A second way to improved productivity in research is through careful selection and placement of new employees and selection of the most promising candidates for advanced training. Performance on standardized tests and records of past research performance are useful sources of information. School grades are only a special, and relatively indirect, index of past research performance which may best be used in conjunction with scores on carefully constructed standardized tests and more direct measures of previous research performance. A careful appraisal of this information should improve identification of applicants with superior abilities and of those who are not qualified for professional research work. It should also provide an estimate of the level at which applicants should begin work and the level they should ultimately attain. This would improve research performance in at least three ways: 1) it would start workers with superior talents on creative research early in their career; 2) it would place satisfactory, but not outstanding, applicants at a job level where work would be sufficiently difficult to present a challenge but not so difficult as to be discouraging; and 3) it would reject employees lacking sufficient potentiality before considerable money and effort were expended on them.

A Long-range Personnel Research Program

Since January 1948 the Office of Naval Research has sponsored a comprehensive program of the American Institute for Research dealing with personnel research problems of scientific and engineering workers. The reports of work already accomplished have been published.* The total program includes the following steps:

1. Determination of the critical requirements for successful participation in research and engineering work

---

* See the bibliography of these reports on page 40.

2. Development of an aptitude test for the selection of research personnel

3. Development of tests to measure proficiency in specific areas of scientific work

4. Development of procedures for evaluating the job performances of research personnel

5. Determination of the predictive value of the tests developed in steps (2) and (3) using the procedures developed in step (4) to obtain evaluations of personnel for comparison with test predictions.

The first three of the above steps have been completed. Preliminary work on step 4 has previously been reported. The present report deals with additional work on that step. A project is currently in progress which deals with step 5.

Functions of a Procedure for Evaluating Research Through the Report

A procedure for evaluating research through the report functions both in selection of new employees and in evaluating the performance of currently employed workers. It functions at a number of levels. Such a procedure might conveniently be used to evaluate theses or dissertations of prospective employees who are recently graduated from school. At a higher level it could be used to evaluate the published research reports of more experienced personnel. It might also be used to evaluate the performance of personnel engaged in more or less independent research since the performance of such persons is frequently little known except through the written report.

There are at least four advantages to evaluation of research through the report. The report:

1. is a permanent record of performance;

2. can be made available to several judges or raters;

3. can be evaluated by persons who do not know the author, thus decreasing personal bias;

4. can be evaluated in a single and relatively brief period, rather than requiring a series of observations of job performance.

Objectives of the Present Project

The general objective of the project was to develop a procedure which would be helpful to graduate and undergraduate thesis committees, research supervisors, potential employers, and others responsible for evaluating research through

the report. Specific steps of the project include the following:

1. Developing pre-field test forms of the procedure for small scale preliminary tryouts

2. Revising the preliminary forms to develop a trial form suitable for a large scale field test

3. Obtaining an independent measure of the effectiveness of the reports to be evaluated in the field test

4. Conducting a large scale field test to provide data suitable for the trial form of the procedure

5. Revising the trial form on the basis of statistical data and comments of evaluators obtained in the field test.

Chapter II

PLANNING THE RESEARCH

Formulating the Problem

In the spring of 1951 Dr. Paul Horst of the University of Washington was asked to study the problem of evaluating research through the report and to recommend a plan for developing a suitable evaluation form. His recommendations, in part, were:

1. Construct a check list of items describing critical behaviors which can be evaluated from the report. Horst pointed out that since the Observational Record Form for Research Personnel,* developed from a study of critical requirements for research workers,** is believed to represent the most accurate and comprehensive list of critical behaviors involved in success or failure of research personnel which is currently available, it should provide the basis for an experimental report evaluation form. He further suggested that the experimental form include only the first five main areas of the Observational Record Form, since the last three areas include critical behaviors which would probably not be reported in the research report. The first five areas, on the other hand, include behaviors which might be given in a research report. These five areas are listed below:

   I. Formulating Problems and Hypotheses
   II. Planning and Designing the Investigation
   III. Conducting the Investigation
   IV. Interpreting Research Results
   V. Preparing Reports

2. Have specialists from a variety of natural science and engineering fields and research functions select reports which are clearly examples of either especially effective or relatively mediocre research.

3. Have persons specializing in the appropriate field evaluate the research by indicating for each item on the experimental check list whether or not there is evidence in the report that the given critical behavior occurred.

---

* Weislogel, Mary H. Procedures for Evaluating Research Personnel with a Performance Record of Critical Incidents. Pittsburgh: American Institute for Research, 1950.

** Flanagan, John C. et al. Critical Requirements for Research Personnel. Pittsburgh: American Institute for Research, 1949.

4. Revise the experimental check list on the basis of statistical analysis of the item responses and other information provided by participants in the study.

Preliminary Tryout of the Experimental Check List

A preliminary Record Form for Evaluating Research Through the Report was made up directly from the Observational Record Form for Research Personnel. Thirteen individuals participated in the preliminary tryout of this form. Participants selected reports from their own field of specialization as effective or ineffective. Eight persons evaluated reports judged effective, and five evaluated reports judged ineffective. Four of those who evaluated effective reports were psychologists, two were chemists, and two were physicists. Three of those who evaluated ineffective reports were psychologists and two were physicists. Data obtained in the tryout were of two kinds, entries on the Record Form and comments of the respondents. A study of these data indicated:

1. the preliminary form was sufficiently comprehensive; no additional items needed to be added to the form.

2. where effective items were not balanced by an analogous ineffective item or an ineffective item was not balanced by an analogous effective item, participants found evaluation difficult.

3. certain changes in the instructions for use of the procedure were needed. It was also decided that the term "ineffective" was probably too strong for the less satisfactory research. The term "relatively mediocre" seemed more appropriate for reports to be selected in later stages of the project.

After several revisions and a tryout with six research psychologists the preliminary evaluation form was used by two physicists, two chemists, and two engineers. Interviews with these persons indicated that the preliminary form was satisfactory for large scale tryout but that additional changes in the instructions for use of the form were needed in the interest of brevity and clarity. The required changes were made in the instructions and they were judged to be satisfactory for large scale tryout.

Obtaining Reports for Evaluation

Several sources for reports of research judged to be either especially effective or relatively mediocre by specialists in a variety of natural science and engineering fields were investigated.

Doctoral dissertations selected by the candidates' major advisers would provide complete reports of original research which was the major responsibility of a single investigator. The major adviser seems to be in an especially good

position to judge the effectiveness of the individual research. However, an investigation of the university inter-library loan system indicated that the obtaining of a sufficiently large number of dissertations for the purposes of this investigation was not feasible.

Final reports of research for distribution within the producing laboratory or parent organization have many of the advantages of doctoral dissertations. Such reports are usually quite complete and can be judged by supervisory personnel under whom the research was conducted. Communication with cognizant persons in a research foundation and industrial laboratories indicated that such reports are usually highly confidential and not available for evaluation by persons not within the organization. The problem of the security of report content obtains to an even greater degree for most governmental laboratories. It was pointed out that reports of such research, i.e., of a non-confidential nature, are usually published either in scientific journals or as separate reports.

The published literature, particularly the scientific journals, proved to be the only source for a sufficiently large number of available reports. One limitation of such reports is the relatively condensed treatment which is required for publication in most journals This makes a comprehensive and detailed evaluation of the research through such reports more difficult than evaluation of more lengthy and complete research reports

Planning the Research Design

From the above considerations the following specific steps were planned.

1. Obtain selections of a large number of research reports in chemistry, physics and engineering which are judged by highly competent persons in the field to represent either especially effective or relatively mediocre research. These persons will be termed "selectors."

2. Have the selected reports evaluated by other workers in the field using the trial Record Form for Evaluating Research Through the Report These persons will be termed "evaluators." They will not be aware of the selectors' judgments concerning the effectiveness of the research they are evaluating through the report.

3. Revise the trial form on the basis of results from the evaluations and the comments of evaluators concerning the form. The primary method of revision will be the selection of items from the trial form which show the greatest amount of agreement between the selectors' judgment and the evaluators' evaluation of the effectiveness of the research being considered.

4. Have each of a number of reports evaluated independently by two persons. This will permit an estimation of the inter-evaluator reliability of the items selected for the revised version of the form. The item inter-evaluator reliability coefficients could also provide a basis for selecting items, although it was not planned that this be done.

Chapter III

CONDUCTING THE RESEARCH

Obtaining Selections of Reports

After preliminary investigations had been completed, the first step in conducting the research was to have highly competent persons in special areas of physics, chemistry, and engineering select a large number of reports as representing especially effective or relatively mediocre research. The judgment of these selectors was the criterion of research excellence with which results on the trial Record Form was compared. Each selector was requested to nominate one report of especially effective and one of relatively mediocre research.

The questionnaire method seemed to be the most feasible way of obtaining report selections from a large number of persons in widely scattered geographical areas. Biographical data contained in American Men of Science provided a convenient source for compiling a mailing list.

Requests for selections of reported research were sent to 50 person whose names were "starred"* in the 1944 edition and to 511 persons listed in the 1949 edition as fellows, officers, or past officers of a professional society in their main field of interest. Approximately equal numbers of chemists, physicists and engineers were contacted. Two follow-up letters were sent to all but a small number of persons contacted early in the study. In preliminary mailings various modifications of the original request and follow-up letters** were tried out, in order to explore the possibility of increasing the proportion of returns from later mailings.

In reply to 561 mail requests there were 133 or 24 per cent, returns. Each completed return represents two report selections, one of especially effective and one of relatively mediocre research. Differences in proportion of returns in answer to various modifications of the request and follow-up letters were not appreciable.

Returns were reviewed, as they came in from selectors, to determine which of the various fields were represented by the reports, and whether the references could be readily located in university or technical libraries. There were 254 available and usable research reports selected, 90 in chemistry, 81 in physics, and 83 in engineering.

---

* Starred men are those mentioned most frequently by their colleagues as outstanding men of science. Starring was discontinued in the 1949 edition.

** Copies of these materials are shown in Technical Appendix A.

Obtaining Evaluations of Research Through Reports

The second step in conducting the research was to have a group of evaluators read the reports chosen by selectors in the previous step and use the trial Record Form in evaluating the research through the report. Minimum requirements for evaluators were set as follows: for physics and chemistry, a doctor's degree or a master's degree plus two years' experience in research or teaching; and for engineering, a master's degree or a bachelor's degree plus three years' experience in research or teaching.

Prospective participants were contacted by mail in various universities and research organizations. Each potential evaluator was requested to indicate on a data form his willingness to participate in the study and certain information concerning his training and experience. From the information given on the data form it was usually possible to provide each evaluator with a report in his general area of specialization.

Two reports were usually referred to each evaluator, but never more than two. Usually, each evaluator was sent one report selected as an example of outstandingly effective and one report selected as an example of relatively mediocre research. A few evaluators evaluated either two "especially effective" or "relatively mediocre" research studies and a few evaluated only one study. The evaluator, of course, did not know on what basis the reports had been selected.

Two hundred and two reports were evaluated; 116 were evaluated only once, while each of the remaining 86 was independently evaluated by two persons. These independent evaluations were obtained to permit an estimation of the inter-evaluator reliability of the procedure.

Chapter IV

RESEARCH RESULTS

Evaluations of Research Through Reports

The method of obtaining evaluations was discussed in the previous chapter. An effort was made to obtain approximately equal numbers of evaluations from the fields of chemistry, physics, and engineering and to include evaluations of research from most of the more important specializations within each field. This effort was made because it was thought that a Record Form revised on the basis of results from a sample including representative reports from the three disciplines would have wider applicability than if revised on the basis of results from a more restricted sample. The number of evaluations obtained for each specialization is shown in Table I. The dual evaluations of the 43 especially effective and 43 relatively mediocre reports were used to obtain an estimate of the inter-evaluator reliability of the revised Record Form. The trial Record Form and instructions for its use are presented on pages 11-19.

Table I

Reports Evaluated in Physics, Chemistry, and Engineering

	Evaluated by One Person		Evaluated by Two Persons		Total Reports Evaluated		Total Number of Evaluations	
	E*	M**	E	M	E	M	E	M
Chemistry	19	21	15	15	34	36	49	51
Inorganic-Analytical	3	5	4	4	7	9	11	13
Physical	4	5	6	4	10	9	16	13
Organic	8	11	5	5	13	16	18	21
Biochemistry	4	0	0	2	4	2	4	4
Physics	23	20	12	14	35	34	47	48
Atomic-Nuclear	9	5	5	5	14	10	19	15
Electricity	7	7	5	6	12	13	17	19
Mechanics	1	2	1	0	2	2	3	2
Spectroscopy	3	2	1	3	4	5	5	8
Meteorology	0	1	0	0	0	1	0	1
Sound	3	3	0	0	3	3	3	3
Engineering	16	17	16	14	32	31	48	45
Chemical	5	4	4	2	9	6	13	8
Civil	0	1	0	1	0	2	0	3
Electrical	2	5	4	3	6	8	10	11
Mechanical	3	2	2	2	5	4	7	6
Metallurgical	3	3	0	0	3	3	3	3
Petroleum	1	1	0	0	1	1	1	1
Aeronautical	0	0	5	5	5	5	10	10
Ceramic	2	1	0	0	2	1	2	1
Textile	0	0	1	1	1	1	2	2
All Fields	58	58	43	43	101	101	144	144

* E indicates reports selected as representing especially effective research.
** M indicates reports selected as representing relatively mediocre research.

INSTRUCTIONS FOR EVALUATORS

1. Evaluate the research by filling in the Record Form for Evaluating Research Through the Report. It is organized as an outline with major and sub-headings with numbered categories below them. Each category is composed of an effective item to the left and an ineffective item to the right. There is a blank space beside each item.

   Place one of the following symbols in each blank space:

   NA = This item could not have occurred because such activity is not applicable to this research.

   ? = This item might have occurred, but the reader cannot tell from this report.

   0 = This item did not occur.

   X = This item occurred, but was not at all important to conducting this research well or poorly.

   √ = This item occurred and was of some importance to conducting this research well or poorly.

IMPORTANT: THE SPACE BESIDE EVERY ITEM SHOULD BE MARKED WITH SOME SYMBOL. THE TWO ITEMS FOR ANY CATEGORY WILL NOT NECESSARILY BE GIVEN THE SAME NOR NECESSARILY DIFFERENT SYMBOLS, SINCE EACH ITEM WILL BE EXAMINED SEPARATELY.

2. Complete the section on the last page of the Record Form which asks for your over-all judgment of the contribution of the research.
3. Circle the symbol for each item which in itself made this study of definite value (or reduced the value of this research appreciably) and had a noticeable effect upon your over-all judgment of the contribution of this research.
4. Of the items circled under 3 above, place an additional circle around the symbol for each item which in itself made this study an important contribution (or reduced the value of this research very significantly) and had a sizeable effect upon your over-all judgment of the importance of this research.
5. Report on the last page of the Record Form the time required to evaluate the research. This will not include time required to read the report.

EXAMPLES OF ENTRIES

1. The effective item is not applicable to this research; the reader cannot tell from the report whether the ineffective item occurred, which in itself reduced the value of this research appreciably.

   NA Effective item — Ineffective item (?)

2. The effective item occurred, but was not at all important to conducting the research well; the ineffective item occurred and was of some importance to conducting this research poorly.

   X Effective item — Ineffective item √

3. The effective item occurred and in itself made this study an important contribution; the ineffective item did not occur.

   (√) Effective item — Ineffective item 0

4. The effective item occurred and was of some importance to conducting the research well; the ineffective item occurred and in itself reduced the value of this research appreciably.

   √ Effective item — Ineffective item (√)

RECORD FORM FOR EVALUATING RESEARCH THROUGH THE REPORT

Title of Report_____

Author(s)_____ Evaluated by_____

I. FORMULATING PROBLEMS AND HYPOTHESES

Effective	A. Identifying and Exploring Problems	Ineffective
__1.	Investigated chance findings, unexpected results or difficulties encountered in work or mentioned significance of such findings.	Failed to investigate chance findings, unexpected results or difficulties encountered in work or failed to mention significance of such findings. __
__2.	Chose for investigation a problem for which solution was urgently needed.	Chose problem for which solution was not urgently needed although there were urgent problems in his research area. __
__3.	Suggested a new problem which could be studied with an already successful technique.	Allowed a successful technique to be dropped without further application to new problems. __
__4.	Proposed an entirely new problem or line of research.	Worked on a problem which had already been solved or proved unproductive. __
__5.	Used materials that had recently been made available to study previously unsolved problem.	Failed to use new methods or materials which were recently made available to study previously unsolved problems. __
__6.	Conducted preliminary investigation to see whether phenomena merited experimental study or to furnish essential basic data.	Failed to conduct preliminary investigation to see whether phenomena merited experimental study or to furnish essential data. __
	B. Defining the Problem	
__1.	Proposed investigation of basic factors and implications involved in the problem as well as its superficial aspects.	Proposed an investigation confined to superficial aspects of problem. __
__2.	Defined the problem and objectives of investigation.	Did not define problem or objectives of investigation. __
__3.	Gathered information on exact requirements, specifications, and goal of project.	Failed to obtain information needed to define the requirements, specifications, and goal of project. __
__4.	Proposed investigating only factors which could feasibly be studied under existing practical limitations.	Chose a problem which did not lend itself to investigation because of practical limitations. __
__5.	Covered both theoretical and experimental aspects of problem.	Ignored either theoretical or experimental aspects of problem. __

Effective	C. Setting Up Hypotheses	Ineffective
__1. Proposed hypothesis to direct research or to explain observed phenomena.	Proposed program of data collection undirected by any hypothesis.	__
__2. Proposed hypothesis in agreement with all known facts.	Proposed hypothesis contrary to known facts.	__
__3. Predicted phenomena by theoretical or mathematical analysis.	Failed to predict phenomena by theoretical or mathematical analysis when it was clearly possible to do this.	__
__4. Extended a theory to cover a broader range of problems.	Failed to extend theory to cover broader range of problems when possible.	__
__5. Reformulated a theory to improve its explanation of the facts.	Failed to reformulate a theory although this would clearly have improved its explanation of the facts.	__
__6. Explained observed phenomena through theory or analogous situation in the same field or a related field.	Failed to explain observed phenomena through theory or analogous situation in the same field or a related field which was clearly applicable.	__

II. PLANNING AND DESIGNING THE INVESTIGATION

A. Collecting Background Information

__1. Sought out information and ideas from existing literature, associates, or experts on problem before beginning work on project.	Did not consult those intimately concerned with problem, or investigate existing literature before beginning work on project.	__
__2. Included all relevant sources in surveying the literature or consulting experts.	Omitted an important source in surveying literature or consulting experts.	__
__3. Questioned the validity of material in the literature.	Took action based on unreliable information in the literature without checking.	__
__4. Obtained needed information from an uncommon source.	Consulted only the most common sources for needed information.	__
__5. Suggested that literature he had read in the past might apply to the current problem.	Ignored application of literature which he should have read in the past to the current problem.	__
__6. Performed experiments or gathered necessary information directly which was unavailable in usual sources.	Failed to perform experiments or gather necessary information directly which was unavailable in usual sources.	

B. Setting Up Assumptions

__1. Based research plan on assumptions which closely approximated actual conditions.	Used a research plan dependent on false assumptions or assumptions inapplicable to specific problem.	__
__2. Secured evidence of validity of assumptions.	Failed to secure evidence of validity of assumptions.	__
__3. Verified previous work before basing assumptions on it.	Based plan of investigation on opinion or previous work of others without question.	__

Effective	C. Identifying and Controlling Important Variables	Ineffective
__1.	Provided for control and systematic variation of all relevant variables.	Failed to provide for control and systematic variation of all relevant variables. __
__2.	Made provision for equated conditions in planning comparison tests.	Failed to make provision for equated conditions in planning comparison tests. __
__3.	Simulated actual conditions in a laboratory test.	Failed to simulate actual conditions in a laboratory test. __
__4.	Treated the various factors in accordance with their relative importance.	Failed to treat the various factors in accordance with their relative importance. __
__5.	Pointed out the significance of a factor overlooked or dismissed as trivial by others.	Failed to point out the significance of an important and obvious factor. __
	D. Developing Systematic and Inclusive Plans	
__1.	Included all relevant factors or phases in the investigation.	Omitted a relevant factor or included an irrelevant factor in the investigation. __
__2.	Included methods of integrating one factor or phase with others.	Considered one factor or phase in isolation from related phases. __
__3.	Pointed out the basic factors in a mass of information about the problem.	Failed to point out basic factors in a mass of information about the problem. __
__4.	Tried out various approaches to problem before choosing one.	Did not try out various approaches to the problem before choosing one. __
__5.	Studied each element of problem in proper sequence.	Studied elements of problem in illogical sequence. __
	E. Developing Plans for the Use of Equipment, Materials, or Techniques	
__1.	Used equipment, material, or techniques which met the requirements of the problem.	Used equipment, materials, or techniques not fitted to the requirements of the problem. __
__2.	Used simplified or substitute equipment, materials, or techniques, which met required standards and saved time or money.	Used equipment or materials more complex or expensive than necessary to produce results of required standards. __
__3.	Conducted pilot study to determine feasibility of proposed techniques, materials, or equipment.	Used a procedure that had never been tested. __
__4.	Used technique or equipment which would eliminate doubt of validity or accuracy of results.	Used technique or equipment which would leave doubt of validity or accuracy of results. __
__5.	Used the latest development of an appropriate equipment, technique, or material.	Ignored latest development of an appropriate equipment, technique, or material. __
__6.	Set up work for procedure in most efficient physical arrangement for handling details easily.	Set up work for procedure in an inefficient physical arrangement so details could not be handled easily. __

Effective	F. Anticipating Difficulties	Ineffective
__1. Made provision for an alternate approach or for handling difficulties which might arise at later stages.	Made no provision for handling difficulties which might arise at later stages.	__
__2. Included in plans internal or independent check on accuracy of data or method.	Failed to set up internal or independent check on accuracy of data or methods.	__
__3. Outlined probable consequences of various alternative approaches.	Failed to consider probable consequences of various alternative approaches.	__
__4. Took special precautions in planning to prevent damage to equipment.	Failed to take necessary precautions in planning to prevent damage to equipment.	__
	G. Determining the Number of Observations	
__1. Collected an appropriate quantity of data for the purpose of the investigation.	Collected data which were insufficient or considerably more than sufficient in quantity.	__

III. CONDUCTING THE INVESTIGATION

	A. Developing Methods, Materials, or Equipment	
__1. Devised an improved method, material, or equipment.	Developed a method, material, or equipment which was less effective or no better than existing one.	__
__2. Developed an entirely new and effective method, material, or equipment to fill a need.	Developed a new method, material, or equipment which did not meet assigned specifications or recognized need.	__
__3. Adapted available methods, materials, or equipment to meet requirements of new problem.	Failed to make proper adaptation of existing method, material, or equipment to permit its use in a specific problem.	__
__4. Showed experimentally the capabilities of method, material, or equipment he developed.	Failed to show experimentally the capabilities of method, material, or equipment he developed.	__
	B. Applying Methods and Techniques	
__1. Applied critical tests of equipment or material correctly.	Used equipment, material, or technique incorrectly.	__
__2. Used a technique, material, or equipment which solved problem or eliminated difficulty in the investigation.	Failed to use a technique, material, or equipment which would have solved problem or eliminated difficulty in the investigation.	__
__3. Tried even unlikely methods after obvious methods had failed.	Failed to try unlikely methods after obvious methods had failed.	__
__4. Demonstrated that material, technique, or equipment could be used for purposes other than original ones.	Failed to suggest that material, technique, or equipment could be used for purposes other than original ones.	__
__5. Used most accurate method of measurement that was available.	Estimated data when accurate methods of measurement were available.	__
__6. Applied all methods provided for in plans.	Failed to apply a method provided for in plans.	__

Effective	C. Modifying Planned Procedures	Ineffective
__1. Modified standards, methods, or work schedule to meet practical demands without reducing essential value of results.	Failed to modify standards, methods, or work schedule to meet practical demands or omitted an essential step or precaution for accuracy in modifying them.	__
__2. Adopted alternate procedures as soon as unforeseen negative conditions or difficulties were encountered.	Continued to follow old method or approach without change when evidence showed it had failed.	__
__3. Held up phases of work until results of earlier phase were available.	Began work on a phase of project without waiting for results of earlier phase.	__
__4. Accepted partial results or temporary solution to an urgent problem.	Refused to accept partial results or temporary solution to an urgent problem.	__
__5. Modified work to incorporate latest research findings.	Did not modify work to incorporate latest research findings.	__
__6. Instituted a change which prevented damage, incorrect performance, or inaccuracy.	Instituted a change causing damage, incorrect performance, or inaccuracy.	__
__7. Did not abandon or modify a method or device until sufficient evidence had been gathered.	Abandoned or modified a method or device before sufficient evidence had been gathered.	__
__8. Modified all of the materials or procedures giving trouble.	Modified only a part of the materials or procedures giving trouble.	__
	D. Applying Theory	
__1. Presented unique solution or technique developed by mathematical analysis.	Presented an erroneous mathematical solution.	__
__2. Transformed physical problem so that it could be solved by mathematical analysis.	Failed to transform physical problem so that it could be solved by mathematical analysis when such transformation was possible.	__
__3. Explained phenomenon by analyzing procedures used or data obtained.	Failed to explain phenomenon by analyzing procedures used or data obtained.	__
__4. Solved a problem by application or extension of textbook principles.	Failed to solve a problem requiring only direct application or elementary extension of textbook principles.	__
__5. Was able to provide answer to technical question, and gave no incorrect information.	Failed to provide answer to technical question or gave incorrect technical information.	__
__6. Correctly interpreted implications of fundamental theory in explaining application to a problem.	Omitted or misinterpreted implications of fundamental theory in explaining application to a problem.	__
	E. Attending To and Checking Details	
__1. Performed work which met standards for accuracy.	Performed work which contained errors and did not meet standards for accuracy.	__
__2. Gave proper proportion of time and attention to small details of procedure.	Gave disproportionate time and attention to small details of procedure.	__

Effective	F. Analyzing the Data	Ineffective
__1. Used the most efficient method for analyzing data.	Used inefficient method for analyzing data.	__
__2. Used data analysis method which was well suited to give required information.	Used a data analysis method which could not give required information.	__
__3. Completed only analyses necessary for data obtained.	Completed an analysis unnecessary for data obtained.	__
__4. Made all necessary mathematical analyses of data.	Failed to make the necessary mathematical analysis of data.	__

IV. INTERPRETING RESEARCH RESULTS

A. Evaluating Findings

__1. Produced conclusions or recommendations supported by data and interpreting results correctly.	Drew conclusions not supported by the data and interpreting results incorrectly.	__
__2. Drew conclusions in accordance with correct logical principles.	Presented a conclusion violating logical principles.	__
__3. Presented results of a check on validity of conclusions.	Did not present a check on validity of conclusions.	__
__4. Pointed out the limitations of data or method, conflicting elements, and conclusiveness of evidence.	Did not point out the limitations of data or method, conflicting elements, and conclusiveness of evidence.	__
__5. Presented logical explanation of unexpected results.	Reported unexpected results without logical explanation.	__
__6. Drew all conclusions from the data that were justifiable and showed solution to problem.	Failed to draw conclusions from data or to show solution to problem.	__
__7. Drew conclusions only from complete, adequate, and correct data.	Drew conclusions from incomplete, inadequate, or erroneous data.	__

B. Pointing Out Implications of Data

__1. Pointed out new and useful implications and possible extensions of work.	Failed to report implications and possible extensions of work or reported inapplicable ones.	__
__2. Worked out applications to other problems or fields.	Included inadequate discussion of applications which should have been discussed fully.	__

V. PREPARING REPORTS

A. Describing and Illustrating Work

__1. Included important details of procedure and results sufficient for checking or repetition of work.	Failed to include important details of procedure and results sufficient for checking or repetition of work.	__
__2. Used graphic, tabular, or pictorial material to clarify text.	Omitted necessary illustrative material.	__
__3. Defined all terms and symbols.	Presented ambiguous definition of terms or symbols or failed to define them.	__
__4. Used simple, direct language, concrete words, and correct English usage.	Made excessive use of complex sentence structure and unusual words or violated correct English usage.	__

Effective		Ineffective	
__	5. Kept statement of problem and conclusions brief enough for quick grasp.	Made statement of problem and conclusions unnecessarily long and involved.	__
__	6. Explained new material when first introduced.	Failed to explain or define new material.	__
__	7. Gave examples of practical applications or a simplified statement of complex theory.	Failed to give examples of practical applications or a simplified statement of complex theory.	__
__	8. Used a simple system of designation for items.	Used an unnecessarily complex system of designation for items.	__
__	9. Defined purpose of report or project explicitly.	Failed to define purpose of report or project explicitly.	__
__	10. Gave proper emphasis to major and unimportant findings.	Failed to emphasize major findings or overemphasized unimportant findings.	__
__	11. Limited treatment of elementary theory or well known materials to a brief discussion and gave only necessary detail in descriptions.	Discussed at length elementary theory or well known materials or gave excessive detail in descriptions.	__
__	12. Used only accurate labels on illustrative material, accurate references, and correct symbols.	Used inaccurate labels on illustrative material, inaccurate references, or incorrect symbols.	__
	B. Substantiating Procedures and Findings		
__	1. Described fully the basic principles involved.	Failed to give sufficient background or theory for full understanding.	__
__	2. Gave explicit statement of underlying assumptions and inferences.	Failed to state underlying assumptions.	__
__	3. Included detailed reasoning leading to conclusions presented.	Failed to present material on which conclusions were based.	__
__	4. Gave an especially complete, relevant bibliography.	Omitted necessary bibliography or failed to give full reference to related work area.	__
__	5. Presented proof for unusual theory.	Gave abstruse theory without presenting proof.	__
__	6. Gave derivations of all but very common equations or formulas.	Failed to give derivations of equations or formulas.	__
	C. Organizing the Report		
__	1. Gave problem and introductory material at the beginning of the report.	Failed to give plan and scope of the problem at the beginning of the report.	__
__	2. Summarized the important points.	Failed to bring out the main points.	__
__	3. Followed a logical outline.	Separated related sections or materials or jumped from one point to another.	__
__	4. Placed lengthy or detailed analysis or data in appendix.	Gave technical details in body of report.	__
__	5. Separated background material or discussion from presentation of method.	Mixed background material or discussion with presentation of method.	__
__	6. Placed references in appropriate location.	Placed references in an inappropriate location.	__

Effective		Ineffective
__7.	Presented figures or tables in order corresponding to text.	Presented figures or tables in order not corresponding to text. __
__8.	Used a logical order in tabulating data and presenting conclusions.	Used an illogical order in tabulating data or presenting conclusions. __

D. Using Appropriate Style in Presenting Report

__1.	Used a style adapted to probable readers.	Used an unduly informal style or style inappropriate for readers. __
__2.	Heightened interest and stimulated thought by skillful manner of presentation.	Presented material in an uninteresting and unstimulating way. __

* * * * *

We would like you to make an over-all judgment about the research you have just evaluated. In making this judgment it might be helpful to consider whether you think the time and money required for the project were well spent, whether the research made a significant contribution to knowledge in the field, whether findings of this study were substantiated by later research, and whether you would recommend that other persons read this report. It is especially important to keep in mind that negative results do not necessarily mean the research made no significant contribution to knowledge in the field. Taking these points into consideration, please check the one statement below which you consider to be most true. These statements should be considered definitions of five points on a continuum.

_____ This research made an especially significant contribution to knowledge in the field. I would strongly recommend that all persons in the field read this report.

_____ This research made a relatively significant contribution to knowledge in the field. I would recommend that interested persons in the field read this report.

_____ This research made a small but definite contribution to knowledge in the field. I would suggest that interested persons in the field might find this report of some value.

_____ This research contributed almost nothing to knowledge in the field. I would recommend to few persons in the field that they read this report.

_____ This research contributed nothing to knowledge in the field and probably has misled some workers in the field. I would never recommend that other persons in the field read this report.

> You should now circle important items as described under 3 and 4 of the instructions.

How much time did you spend completing the Record Form (including circling of important items, but not including time spent in reading the report)?_____

Comments: (e.g, important factors not covered or difficulties in using the procedure)

 (Use back of sheet if necessary)

Results on the Trial Record Form

The frequency with which each of the seven response symbols (defined in the Instructions for Evaluators) was used or a blank space was left was tabulated for each of the 228 items, 114 effective and 114 ineffective. This was done separately for the 144 evaluations of reports selected as examples of especially effective research and the 144 evaluations of reports selected as examples of relatively mediocre research. Tabulations for each item were also made separately for physics, chemistry, and engineering reports. Results for the three disciplines were sufficiently similar that the data were combined for further analysis.

Table II shows the relative frequency with which each symbol was used for all items for reports selected as representing especially effective and relatively mediocre research. Results are shown separately for effective and ineffective items.

Table II

Frequency With Which the Various Symbols Were Used on the Trial Record Form

	Effective Items				Ineffective Items			
	Effective Research		Mediocre Research		Effective Research		Mediocre Research	
Symbol*	Number	Percent	Number	Percent	Number	Percent	Number	Percent
ⓥ	399	2.4	224	1.4	38	0.2	83	.5
ⓥ	1099	6.7	946	5.8	117	0.7	190	1.2
✓	6978	42.5	6027	36.7	514	3.1	725	4.4
X	1288	7.8	1424	8.7	632	3.8	634	3.9
0	2763	16.8	3468	21.1	11516	70.1	10739	65.4
?	1604	9.8	1804	11.0	1503	9.2	1685	10.3
NA	2268	13.8	2498	15.2	1949	11.9	2223	13.5
Blank	17	0.1	25	.2	147	0.9	137	.8

* Definitions of Symbols

ⓥ = The item occurred and in itself made the study an important contribution (or reduced the value of the research very significantly).

ⓥ = The item occurred and in itself made the study of definite value (or reduced the value of the research appreciably).

✓ = The item occurred and was of some importance to conducting the research well or poorly.

X = The item occurred but was not at all important to conducting the research well or poorly.

0 = The item did not occur.

? = The item might have occurred, but the reader cannot tell from the report.

NA = The item could not have occurred because such activity is not applicable to the research.

Blank = The evaluator failed to place one of the above symbols in the space beside the item.

It may be seen from Table II that effective items were checked somewhat more frequently as having occurred for reports selected as representing especially effective research whereas ineffective items were checked somewhat more frequently as having occurred for reports selected as representing relatively mediocre research. It may also be noted that, for all reports, the ineffective items were checked much less frequently than the effective items. These same trends hold for the "check-circle" and "check-double circle" responses. (These symbols are explained in the footnote of Table II.)

On the last page of the trial Record Form evaluators were asked to make an over-all judgment about the research they had evaluated on the 228 items on the form. The results of these judgments are shown separately in Figure 1 for reports selected as especially effective and relatively mediocre.

It may be seen from Figure 1 that research which was selected as especially effective tended to be judged, through the report, to have made a greater contribution to knowledge in the field than research selected as relatively mediocre. There were sometimes, however, wide discrepancies between selectors' and evaluators' judgments. It may be noted, for example, that in eight cases where the research had been selected as relatively mediocre it was judged to have made an especially significant contribution to knowledge in the field and that in two cases where the research was selected as especially effective it was judged to have contributed nothing to knowledge in the field. The amount of agreement between selectors' and evaluators' judgments may be expressed by a point-biserial correlation coefficient of +.36 (N=287).

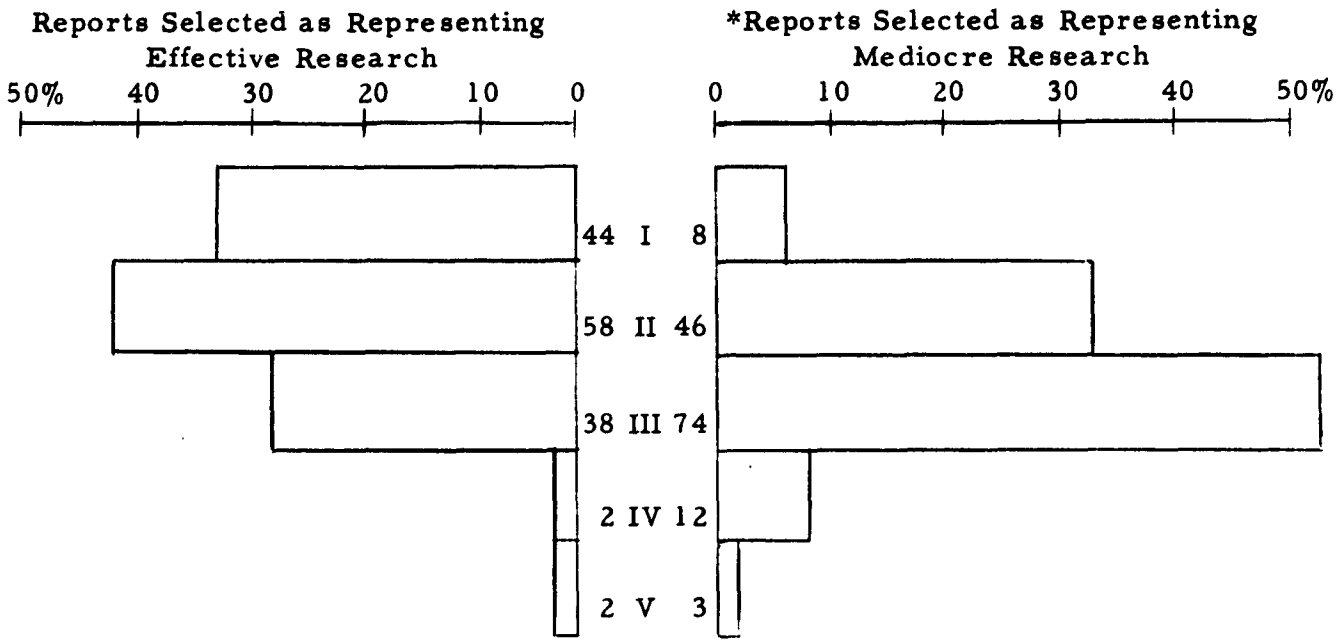
In Figure 1, as with Table II, it may be seen that there is a very strong tendency for reports of both effective and mediocre research to be evaluated more favorably than unfavorably. That is, effective items were checked as having occurred far more frequently than ineffective items (Table II) and far more research was judged to have made a significant contribution to knowledge in the field than was judged to have made little or no contribution. This is to be expected, since even the relatively mediocre research was, usually, of sufficient value to be accepted for publication in a professional journal. It could be expected, therefore, that the relatively mediocre research in this sample would not typify the least productive research conducted. It would rather be expected to typify some of the least productive research which is published in the literature. It seems likely that, on the average, the less productive research remains unpublished.

It has previously been mentioned that each of 86 research studies was independently evaluated by two persons. Of the pair of evaluations for each report one was randomly designated as evaluation A and the other as evaluation B. For purposes of the present analysis there were 85 pairs of evaluations since one evaluator failed to make an over-all judgment about the contribution of the research. Figure 2 shows the extent of agreement between evaluation A and evaluation B on over-all judgments about the contribution of the research.

Figure 1

Distribution of Over-all Judgments
About the Value of the Research

The Roman numerals represent judgments about the research as defined below. Arabic numerals indicate the number of evaluators making each judgment.



I This research made an especially significant contribution to knowledge in the field. I would strongly recommend that all persons in the field read this report.

II This research made a relatively significant contribution to knowledge in the field. I would recommend that interested persons in the field read this report.

III This research made a small but definite contribution to knowledge in the field. I would suggest that interested persons in the field might find this report of some value.

IV This research contributed almost nothing to knowledge in the field. I would recommend to few persons in the field that they read this report.

V This research contributed nothing to knowledge in the field and probably has misled some workers in the field. I would never recommend that other persons in the field read this report.

---

* One evaluator of a report representing mediocre research did not make an overall judgment making the total 143.

Figure 2

Agreement Between Independent Judgments
Of the Contribution Made by Research

Roman numerals indicate judgments as defined in Figure 1. Arabic numerals inside the small boxes indicate the number of reports with particular combinations of overall judgments by the A and B evaluators. Arabic numerals to the right of the large square indicate the number of reports given each type of judgment by B evaluators, and those under the large square indicate the number given each type of judgment by A evaluators. Frequencies inside the heavy boxes represent perfect agreement between evaluators. Those in light boxes indicate less than perfect agreement. The farther light boxes are from the nearest heavy black box, the greater the extent of disagreement. The frequency expected by chance in each box is indicated in parentheses in the lower right hand corner of the box.

Evaluation A

Evaluation B	V	IV	III	II	I	
I	1 (.3)	0 (.6)	0 (4.2)	4 (4.4)	7 (2.5)	12
II	0 (.8)	0 (1.7)	16 (12.7)	15 (13.1)	5 (7.6)	36
III	0 (.8)	2 (1.5)	13 (11.3)	11 (11.7)	6 (6.8)	32
IV	1 (.1)	2 (.2)	1 (1.8)	1 (1.8)	0 (1.1)	5
V	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0
	2	4	30	31	18	85 = total pairs of evaluations

It may be seen from Figure 2 that agreement in over-all judgment between two independent evaluations of research through the report is substantially better than chance, but is certainly far from perfect. The extent of agreement is roughly about the same as between evaluators' over-all judgment and selectors' designation as especially effective or relatively mediocre research. The data shown in Figure 2 yield a product-moment correlation coefficient of +.35. The point-biserial correlation between the over-all judgment with the selectors' effective-mediocre judgment was +.31 for the 85 A evaluators and +.27 for the 85 B evaluators.

Evaluators' Comments

About half of the evaluators made one comment or more about the trial Record Form. Each comment was listed separately and an attempt was made to classify all comments into relatively homogeneous areas. The main and sub-headings of classification are shown below:*

Evaluators' Comments Concerning the Trial Record Form for Evaluating Research Through the Report

(Number in parentheses after a given sub-heading indicates the number of evaluators making comments in this area.)

I. General Comments

- A. Record Form is well done. (5)
- B. Record Form is not well conceived. (2)
- C. Not completing form because of its length or complexity. (11)
- D. Evaluation procedure is subjective. (4)
- E. This type of evaluation requires very specialized knowledge. (3)
- F. Form is too long. (12)
- G. Procedure not applicable to the type of paper being evaluated. (49)

II. Evaluator's Task

- A. Instructions are not clear. (6)
- B. Form should be scored twice (tentatively the first time). (2)
- C. Not enough room is left on forms for author and title. (2)

III. The Items

- A. Items are not relevant to useful evaluation. (6)
- B. Designation of item as effective or ineffective is wrong or confusing. (9)
- C. It is difficult or impossible to answer certain of the items. (10)
- D. Statement of items is ambiguous or incomplete. (26)
- E. Items overlap. (5)

* This outline is reproduced in the technical appendices to this report with examples under each sub-area.

F. Procedure needs additional items. (7)
G. Items are in the wrong place. (1)

IV. The Symbol Scoring System

A. Scoring system should provide for degrees of effectiveness and ineffectiveness. (8)
B. Symbol system is ambiguous or inadequate. (5)
C. Additional symbols are needed. (6)
D. It is difficult to decide which symbol should be used. (11)

The large majority of critical comments seemed to stem from two related sources, the length and complexity of the trial form and the inappropriateness of the form for the type of report being evaluated. The most common single comment dealt with the point that published reports, and especially journal articles, must of necessity be greatly condensed. Many of the detailed questions posed by the trial Record Form concerning the research could not, consequently, be answered adequately from the report. This problem and the necessity for using published research were discussed in Chapter II. The length and complexity of the trial Record Form result from the developmental nature of the present research. A large number of items was included on the trial form to provide empirical evidence for selecting only the better items for the revised Record Form.

It is thought that both of the major points criticized on the trial form -- length and inappropriateness of many items for journal articles -- have been eliminated to a large extent in the revised form, since the aim of the revision was to select from the trial form only those items which functioned adequately in the present study. The revised Record Form is much shortened and composed only of the most cogent items from the trial Record Form. The comments and suggestions of evaluators, as well as the statistical and other considerations discussed in the following section, were carefully considered in revising the Record Form.

Item Analysis and Revision of the Trial Record Form

For purposes of the item analysis the eight possible responses to each item were grouped into two classes -- checks, including ✓, ⓥ, and ⓥ, and non-checks, including Blank, NA, ?, 0, and X. The number of times symbols in each class were used for each item was tabulated separately for research selected as especially effective and research selected as relatively mediocre. A phi (ϕ) coefficient of correlation was computed for each item between the type of selection (as especially effective or relatively mediocre) and the "check-non-check variable." The computation of a phi coefficient from these data for one effective item is illustrated in Table III.

Table III*

Relationship Between Selectors' Judgments of Research Effectiveness and Evaluators' Responses to Item I-A-2 (Effective)

	Number of reports		
	"Especially Effective"	"Relatively Mediocre"	Totals
Checked (✓, ⊘, and ⊘)	82 (a)	60 (b)	142 (a+b)
Not checked (Blank, NA, ?, 0, and X)	62 (c)	84 (d)	146 (c+d)
Totals	144 (a+c)	144 (b+d)	288 (a+b+c+d)

$$\phi = \frac{(ad) - (bc)}{\sqrt{(a+b)\ (c+d)\ (a+c)\ (b+d)}}$$

$$= \frac{(6888) - (3720)}{\sqrt{(142)\ (146)\ (144)\ (144)}} = +.15$$

The obtained values for ineffective items were reflected (plus changed to minus and minus changed to plus) so a positive phi coefficient would indicate an item, whether effective or ineffective, functioning in the desired manner. The distribution of phi coefficients obtained for the 114 effective items on the trial Record Form is shown in Figure 3. The distribution of phi coefficients obtained for the 114 ineffective items is shown in Figure 4.

Figures 3 and 4 show that the average phi coefficient for effective items and that for ineffective items both discriminate between reports of especially effective research and reports of relatively mediocre research. A comparison of Figures 3 and 4 reveals that effective items, in general discriminate to a greater degree than ineffective items.

---

* The greater the excess of checks for effective reports over those for mediocre reports the higher positive will be the phi coefficient. The greater the excess of checks for mediocre reports over those for effective reports the higher negative will be the phi coefficient. The limiting values of phi are plus and minus one.

Figure 3

Distribution of Validity (ϕ) Coefficients Obtained for the 114 Effective Items on the Trial Record Form

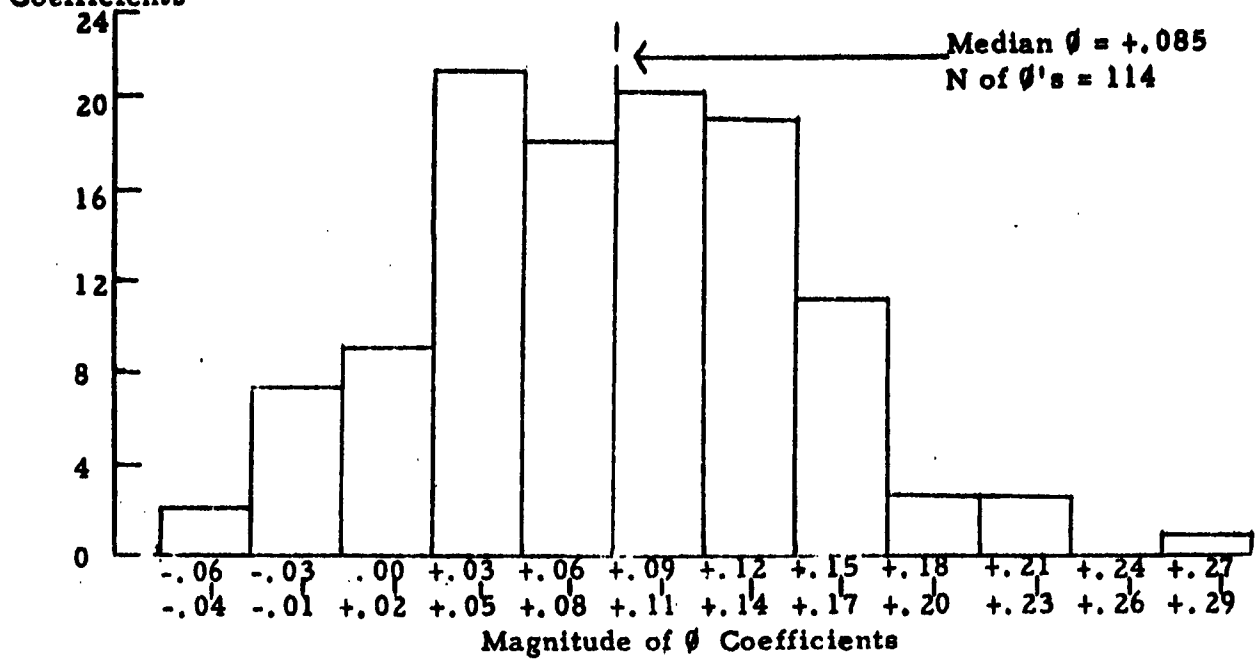
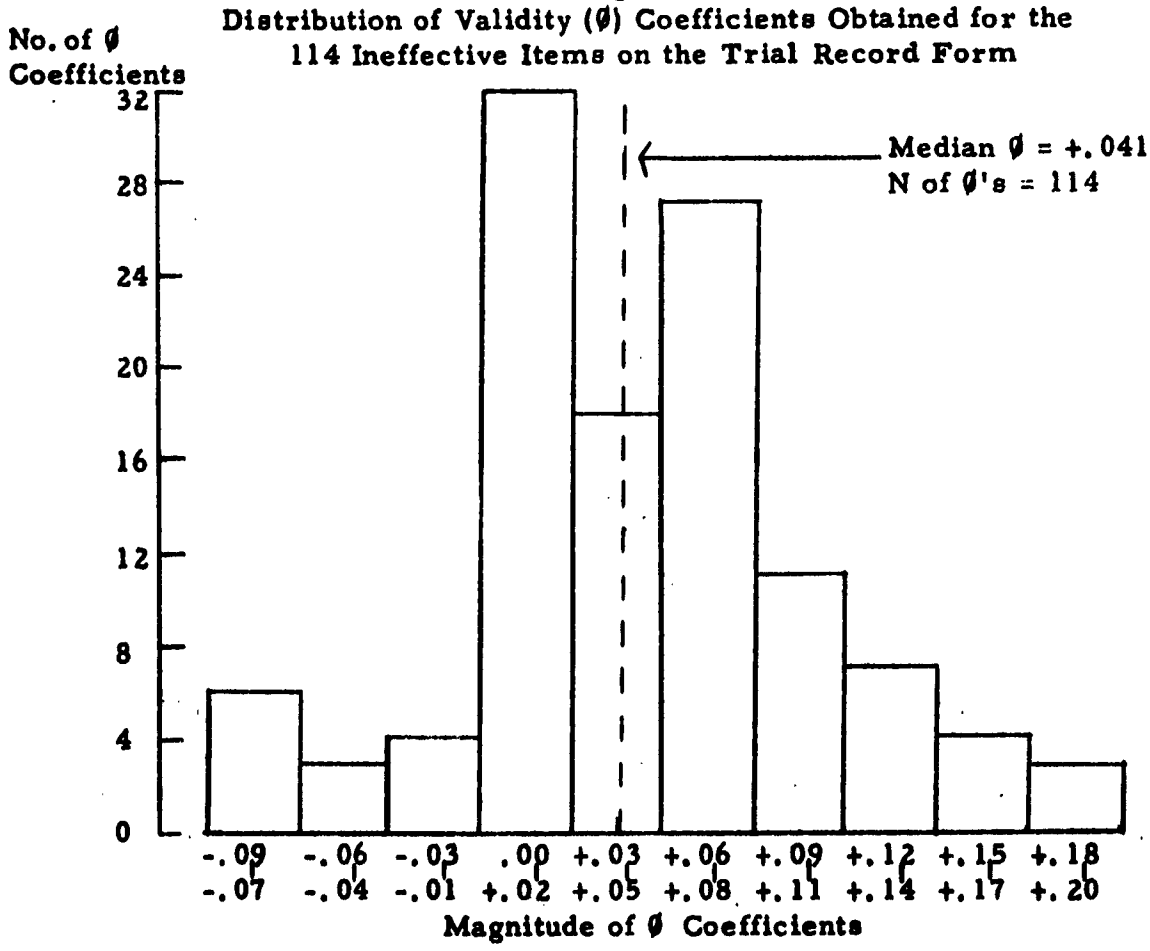


Figure 4

Distribution of Validity (ϕ) Coefficients Obtained for the 114 Ineffective Items on the Trial Record Form



The mean phi coefficients for the effective and for the ineffective items in each of the five main performance areas on the trial form were computed. The results are shown below:

Area	Mean Validity for Effective Items	Mean Validity for Ineffective Items*
I. Formulating Problems and Hypotheses	.11	.06
II. Planning and Designing the Investigation	.07	.05
III. Conducting the Investigation	.09	.03
IV. Interpreting Research Results	.09	.10
V. Preparing Reports	.08	.03

* Item validities reflected.

The mean validity coefficients for the various areas seem to be fairly similar. A further study of area scores and interrelationships would have been desirable, but was beyond the scope of the present study.

The obtained phi coefficients provided one basis for selecting items for the revised Record Form. Another consideration in selecting items was the extent to which each item correlated with other items selected for the revised Record Form. Other things being equal the higher the correlation of an item with other selected items the less will be its contribution to the validity of the selected items. Items with low positive or negative validity would not contribute much in any case, unless they were given special weights which did not seem justified in this study. Therefore, all items with a phi below +.08 were excluded from further analysis.

There were 65 effective and 34 ineffective items with phi's of +.08 or above. A score based on these items was obtained for each of the 288 trial Record Forms completed in the tryout of the form. The total score was obtained by counting each check (including ✓ Ⓥ, and ⓋⓋ) of an effective item as +1 and each check of an ineffective item as -1. The mean number of effective items checked was 35 and the mean number of ineffective items checked was two. Thus, the mean average total score was +33. The standard deviation for effective items was 12.7, for ineffective items was 3.7, and for the total score was 14.6. The product-moment correlation between the number of effective items checked and the number of ineffective items checked was - 39. The product-moment correlation between the number of effective items checked and the total score was +.96 and the product-moment correlation between the number of ineffective items checked and the total score was -.58.

The point biserial correlation of each item with the total score was obtained. This provided an estimate of the average overlap of each item entering into the total score with all other items in the total score. This correlation was spurious to a slight extent since each item for which the correlation was obtained was also included in the total score. The range of point biserial coefficients for the 65 effective

items with the total score was from +.11 to +.60 with a median of +.44. The range of point biserial coefficients for the 34 ineffective items with the total score was from -.11 to -.40 with a median of -.26.

The point biserial correlation between total score for the 99 items (65 effective and 34 ineffective) and the effective-mediocre criterion was obtained.* A weight was computed for each item, showing the relative contribution of the item in predicting the criterion in conjunction with the total score. This index was computed from the formula:**

$$\beta_{ci.t} = \frac{\phi_{ic} - (r_{ct})\ (r_{it})}{1 - (r_{it})^2}$$

where $\beta_{ci.t}$ is the beta weight for a given item (i) when the criterion is predicted by item and total score,

ϕ_{ic} is the phi coefficient of correlation*** between criterion and given item (i)

r_{ct} is the point biserial correlation between criterion and total score (constant for all items), and

r_{it} is the point biserial correlation*** between total score and a given item (i).

It may be seen from this formula that the weight for a given item depends upon both its correlation with the criterion and with a composite of other items. For effective items the greater the validity phi and the smaller the item-composite correlation the larger the beta weight becomes. For ineffective items this is also true, but since both validity phi and item-total score correlation are negative, the beta weight will also have a negative sign. That is, the size of the negative beta becomes larger as the negative validity coefficient increases and the item-total score correlation approaches zero. We would, of course, expect that the occurrence of an ineffective item of behavior should have negative weight in the evaluation of research.

---

* This coefficient was +.30, but this cannot be considered an estimate of the total score validity since individual items were selected for this score on the basis of their validities.

** This is an application of the usual formula for a beta weight in the three variable problem.

*** The signs of these coefficients were not reflected for purposes of this computation.

The obtained beta weight comments of evaluators the number of times the item was checked, and a review of each item in terms of whether it would logically be expected to function in the desired manner were all considered in selecting items for the revised Record Form. The first consideration was whether both effective and ineffective items should be selected for the revised form or whether it would be sufficient to use only one type of item. The relatively small number of times ineffective items were checked in comparison with effective items suggested that the ineffective items might make little additional contribution to the effective items.

It was found that the point biserial correlation between total score for the 65 effective items with a validity phi of +.08 or greater and the effective-mediocre criterion was increased only +.02 by adding the score for the 34 ineffective items to the score for effective items using the beta weights* for combining the scores. It was therefore concluded that it was probably not necessary to select ineffective items for the revised Record Form.

On the basis of the considerations mentioned above, 50 of the 65 trial Record Form effective items with a validity phi of +.08 or greater were selected for the revised Record Form. Four of the 50 items were revised slightly, largely on the basis of suggestions made by evaluators.

Reliability of the Revised Record Form

Since each of 86 reports was evaluated independently by two persons, it was possible to estimate the inter-evaluator reliability of a score obtained from the 50 selected items. The report on which one evaluator did not make an over-all judgment was not included in this analysis so results would be directly comparable with those for the inter-evaluator reliability of the over-all judgment. The product-moment correlation between the two independent evaluations of the 85 reports on the 50 selected items was +.16. This may be compared with the product-moment correlation coefficient of +.35 between two independent over-al judgments of the same reports by the same 85 pairs of evaluators.

The per cent of agreement between the evaluators on each of the 50 items was computed. The per cent of agreement expected by chance** was subtracted

---

* The beta weight for the effective score was +.25 and the beta weight for the ineffective score was -.12.

** The expected proportion of agreement for each item was computed from the formula $p_A p_B + q_A q_B$ expected proportion of agreement

where p_A is the proportion of A evaluators checking the item (one of each pair of evaluators evaluating a given report was randomly designated A, the other was designated B)
p_B is the proportion of B evaluators checking the item
q_A is 1.00 - p_A and
q_B is 1.00 - p_B.

from this obtained per cent of agreement to yield the per cent difference from chance agreement. The tally of results for the 50 items is shown below:

Per Cent Difference From Chance Agreement	Number of Items
+.18 to +.20	3
+.15 to +.17	3
+.12 to +.14	1
+.09 to +.11	6
+.06 to +.08	5
+.03 to +.05	16
.00 to +.02	12
-.03 to -.01	4
	$\sum$= 50

The extent of agreement between evaluators' over-all judgment and their responses may be indicated by a product moment correlation between their total score on the 50 items and their over-all judgment. For the 85 A evaluators this correlation was +.46 and for the 85 B evaluators it was +.36.

An estimate of the internal consistency of the composite of 50 items was obtained for both the A evaluators of the 85 reports and for the B evaluators. The Kuder-Richardson Formula (20) estimate for the A evaluators was .88 and for the B evaluators was .89. It may be seen from these estimates that the evaluators show very substantial self-consistency in checking the 50 items. This indicates that different items are measuring essentially the same thing and hence duplicating each other in the function measured. However, the very low inter-evaluator reliability would indicate that standards for checking items vary considerably from one evaluator to another.

The revised Record Form for Evaluating Research Through the Report is shown on pages 32-35.

RECORD FORM FOR EVALUATING RESEARCH THROUGH THE REPORT

Title of Report _____

Author(s) _____

Evaluated by _____

INSTRUCTIONS

1. Evaluate the research by completing this form. It is organized as an outline with headings and numbered items below them. Each item describes a specific area of effective research performance. There is a blank space beside each item.

   Place one of the following symbols in each blank space to indicate whether there is evidence in the report that effective performance occurred in the area described:

   NA = This item could not have occurred because such activity is not applicable to this research.
   ? = This item might have occurred, but the reader cannot tell from this report.
   0 = This item did not occur.
   X = This item occurred, but was not at all important to the effective conduct of this research.
   ✓ = This item occurred and was of some importance to the effective conduct of this research.

   IMPORTANT: THE SPACE BESIDE EVERY ITEM SHOULD BE MARKED WITH SOME SYMBOL.

2. Complete the section on the last page of the Record Form which asks for your over-all judgment of the contribution of the research.

3. Place a circle around the symbol for each item which in itself made this study an important contribution (or reduced the value of this research very significantly) and had a sizeable effect upon your over-all judgment of the importance of this research. Items marked ? or 0, as well as checked items, may be circled since inadequate reporting or lack of effective performance may significantly reduce the value of research.

EXAMPLES OF ENTRIES

1. ⓥ The item occurred and in itself made the study an important contribution.

2. (?) It cannot be told from the report whether the item occurred; this in itself reduced the value of the research very significantly.

3. (0) The item did not occur; this in itself reduced the value of the research very significantly.

I. FORMULATING PROBLEMS AND HYPOTHESES

_____ 1. Chose for investigation a problem for which solution would be a valuable contribution.
_____ 2. Proposed an entirely new problem or line of research.
_____ 3. Used materials that had recently been made available to study previously unsolved problem.
_____ 4. Conducted preliminary investigation to see whether phenomena merited experimental study or to furnish essential basic data.
_____ 5. Proposed investigation of basic factors and implications involved in the problem as well as its superficial aspects.
_____ 6. Gathered information on exact requirements specifications, and goal of project.
_____ 7. Covered both theoretical and experimental aspects of problem.
_____ 8. Proposed hypothesis to direct research or to explain observed phenomena.
_____ 9. Proposed hypothesis in agreement with all known facts.
_____ 10. Predicted phenomena by theoretical or mathematical analysis.
_____ 11. Explained observed phenomena through theory or analogous situation in the same field or a related field.

II. PLANNING AND DESIGNING THE INVESTIGATION

_____ 12. Included all relevant sources in surveying the literature or consulting experts.
_____ 13. Performed experiments or gathered necessary information directly which was not available in usual sources.
_____ 14. Based research plan on assumptions which closely approximated actual conditions.
_____ 15. Secured evidence of validity of assumptions.
_____ 16. Provided for control and systematic variation of all relevant variables.
_____ 17. Treated the various factors in accordance with their relative importance.
_____ 18. Included all relevant factors or phases in the investigation.
_____ 19. Included methods of integrating one factor or phase with others.
_____ 20. Pointed out the basic factors in a mass of information about the problem.
_____ 21. Tried out various approaches to problem before choosing one.
_____ 22. Made provision for an alternate approach or for handling difficulties which might arise at later stages.
_____ 23. Used equipment, material, or techniques which met the requirements of the problem.
_____ 24. Used technique or equipment which eliminated doubt of validity or accuracy of results.
_____ 25. Used the latest development of an appropriate equipment, technique, or material.

III. CONDUCTING THE INVESTIGATION

____26. Devised an improved method, material, or equipment.
____27. Developed an entirely new and effective method, material, or equipment to fill a need.
____28. Adapted available methods, materials, or equipment to meet requirements of new problem.
____29. Showed experimentally the capabilities of method, material, or equipment he developed.
____30. Used a technique, material, or equipment which solved problem or eliminated difficulty in the investigation.
____31. Modified work to incorporate latest research findings.
____32. Presented unique solution or technique developed by mathematical analysis.
____33. Transformed physical problem so that it could be solved by mathematical analysis.
____34. Correctly interpreted implications of fundamental theory in explaining application to a problem.
____35. Performed work which met standards for accuracy.
____36. Used data analysis method which was well suited to give required information.
____37. Completed only analyses necessary for data obtained.
____38. Made all necessary mathematical analyses of data.

IV. INTERPRETING RESEARCH RESULTS

____39. Presented results of a check on validity of conclusions.
____40. Drew all conclusions from the data that were justifiable.
____41. Drew conclusions only from complete, adequate, and correct data.
____42. Pointed out new and useful implications and possible extensions of work.

V. PREPARING REPORTS

____43. Used graphic, tabular, or pictorial material to clarify text.
____44. Explained new material when first introduced.
____45. Gave examples of practical applications or a simplified statement of complex theory.
____46. Followed a logical outline.
____47. Placed references in appropriate location.
____48. Presented figures or tables in order corresponding to text.
____49. Heightened interest and stimulated thought by skillful manner of presentation.
____50. Used a style adapted to probable readers.

We would like you to make an over-all judgment about the research you have just evaluated. In making this judgment it might be helpful to consider whether you think the time and money required for the project were well spent, whether the research made a significant contribution to knowledge in the field, whether findings of this study were substantiated by later research, and whether you would recommend that other persons read this report. It is especially important to keep in mind that negative results do not necessarily mean the research made no significant contribution to knowledge in the field. Taking these points into consideration, please check the one statement below which you consider to be most true. These statements should be considered definitions of five points on a continuum.

_____ This research made an especially significant contribution to knowledge in the field. I would strongly recommend that all persons in the field read this report.

_____ This research made a relatively significant contribution to knowledge in the field. I would recommend that interested persons in the field read this report.

_____ This research made a small but definite contribution to knowledge in the field. I would suggest that interested persons in the field might find this report of some value.

_____ This research contributed almost nothing to knowledge in the field. I would recommend to few persons in the field that they read this report.

_____ This research contributed nothing to knowledge in the field and probably has misled some workers in the field. I would never recommend that other persons in the field read this report.

You should now circle important items as described under 3 of the instructions.

Comments:

Chapter V

CONCLUSION

Conclusions and Recommendations

1. The Record Form for Evaluating Research Through the Report which has been developed provides an outline of important points for consideration in evaluating research through the report. The items tried out in this study were taken directly from critical behaviors suggested by a large number of senior research workers. Only those items which were found to be valid in this study were included on the form.

2. The Record Form was developed in a manner which should tend to maximize its validity, but the form has not been tried out. It is suggested that a full-scale tryout of the form for various research functions and scientific and engineering disciplines is desirable.

3. The comments of evaluators concerning the trial Record Form suggest that the estimate of inter-evaluator reliability obtained in this study for the 50 selected items may be low since the large number of items on the trial form probably tended to reduce the care with which individual items were answered. It is expected that evaluators using the shorter revised Record Form will be able to exert greater care in responding to individual items. Nevertheless, the evidence obtained indicates that scores on the 50 selected items are not very reliable from one evaluator to another. Although agreement between independent over-all judgments of the value of research is only moderate, it does appear to be greater than agreement between scores obtained from checking the occurrence of behaviors. It is entirely possible that consideration of a number of behavioral items prior to making the over-all judgment might improve the reliability and validity of the over-all judgment, but no data are available to test this possibility. It is therefore recommended that:

    a. checking of behaviors be used at this time only as an aid to arriving at a careful over-all judgment and that ordinarily no total score be obtained for comparison of various reports evaluated by different evaluators. There are, however, two circumstances in which the obtaining of total scores might be justified. They are:

        (1) when each evaluator completes a Record Form for a number of different reports, making it possible to convert each total score into a rank or standard score from the evaluator's distribution of evaluations, and

        (2) when each report is evaluated by a number of persons, making it possible to obtain an average of total scores for each report.

b. Whenever possible, each report should be evaluated by more than one person. The expected effect of such multiple evaluations upon the inter-evaluator reliability of the item total scores and over-all judgments, and the expected effect upon the validity of over-all judgments are indicated in Table IV. It should be pointed out that it is much less difficult to obtain multiple evaluations of research through the report than to obtain evaluations of research through direct observation, since the report can be evaluated by a number of competent persons at different times and in scattered geographical locations.

Table IV

Estimates of Reliability and Validity for Evaluations Obtained by Combining Varying Numbers of Individual Evaluations

	Obtained Reliability	Estimated Reliability*				Obtained Validity	Estimated Validity**			
	r_{xx}	n=3	n=5	n=10	n=20	r_{xy}	n=3	n=5	n=10	n=20
Over-all judgment	.35	.62	.73	.84	.92	.36	.48	.52	.56	.58
Total item score	.16	.36	.49	.66	.79	-	-	-	-	-

* From the Spearman-Brown formula

** From the following formula:

$$r_{(nx)y} = \frac{r_{xy}}{\sqrt{\frac{1 - r_{xx}}{n} + r_{xx}}}$$

where x is an individual evaluation,
r_{xy} is the obtained correlation between X(n=1) and Y,
r_{xx} is the obtained inter-evaluator reliability of X,
n is the number of evaluations of a given report,
nx is a combination of n independent evaluations, and
$r_{(nx)y}$ is the predicted correlation between a criterion, Y(selectors' judgment), and an average of n evaluations.

c. Methods for reducing the variation in evaluators' standards for checking behaviors should be investigated. It is suggested that both the effect of using different item forms and the possibility of obtaining area judgments in addition to the over-all judgment might be fruitfully investigated.

Significance of the Research Findings

One finding of this study seems to be of greater significance than all others. This is the consistently small amount of agreement between two independent evaluations of a given piece of research through the report. Agreement between selectors' and evaluators' over-all judgments concerning the research, between evaluators' individual item responses and the selectors' judgments, and between two evaluators' independent item responses for the same report all tend to be small.

We might suggest at least four possible explanations for this slight agreement.

1. There is no way in which research can be evaluated with satisfactory independent agreement among different qualified evaluators.

2. There are ways in which research may be evaluated with satisfactory independent agreement among different qualified evaluators but there can be no satisfactory method for evaluating research through the report.

3. There are methods for evaluating research through the report which would provide satisfactory independent agreement among qualified evaluators and these are known clearly to research workers in physics, chemistry, and engineering but they are not known to personnel research workers conducting this study.

4. There could be methods for evaluating research through the report which would provide satisfactory agreement among qualified persons evaluating the research independently; but the basic knowledge concerning the complex operations involved in satisfactory evaluations is not possessed either by personnel research workers or research workers in the fields of physics, chemistry, and engineering.

If we accept the first explanation above, we are forced to posit that the operations which, in sequence, constitute a piece of research are not amenable to rational-quantitative measurement. Otherwise, there is no reason why two equally competent measurers should not arrive at similar measures on the appropriate dimensions. If research is not amenable to measurement (evaluation) it then follows that it differs qualitatively from the many other human activities which have been successfully measured. It would certainly be agreed that research is a very complex activity, but this complexity should only make measurement more difficult, not impossible.

Acceptance of the second explanation would certainly be a major indictment of the way in which research reports are now written. It is commonly accepted that research reports are supposed to communicate why the investigation was undertaken, how it was conducted, what the results were, and what the investigator concluded from the results. If this information is not available in the printed reports, it would seem proper to question the value of mass publication of such reports. Of

course, as has previously been mentioned, not all information relevant to a given piece of research can be published, particularly in abbreviated treatments such as journal articles. It would seem then that certain aspects of research can be evaluated only through direct observation and not through the report. But this is no adequate justification for a belief that the aspects of research which are routinely reported in published articles are not subject to evaluation.

Our experience in this study does not lend support to the third explanation. Interviews with workers in various scientific and engineering fields during early phases of the study and comments of participating evaluators revealed wide variation in suggested methodology. There do not seem to be any standards by which different research workers consistently evaluate research through the report. Emphasis upon one or another aspect of research for evaluation seems to depend to a large extent upon the individual worker.

The last suggested explanation seems closest to adequate explanation as to why agreement among working researchers was so small. The dimensions on which a research report can be evaluated with a sufficient amount of agreement between independent evaluators have not yet been isolated and defined. That this is true should not be surprising for, as has been pointed out, research is an extremely complex activity. Since careful empirical studies of evaluation of research through the report have not been conducted, we cannot expect the answer to such a complex problem to come with little or no effort to all who desire it. We can expect that it will be necessary to study the problem intensively with the aid of rigorous logical and mathematical principles before anything approaching a satisfactory solution can be obtained.

It is little wonder, then, that standards for evaluating research through the report are now implicit, highly individual, and obscure. No doubt there are those who will be content to accept on faith that the many research reports to which each individual research worker is exposed are evaluated by him in a manner satisfactory for his own purposes, even though his evaluation is not in agreement with other workers in the field. It is the opinion of the present investigators, however, that an empirical demonstration by the rational-quantitative approach is preferable to unsupported intuitive judgment.

Certainly, the crucial position held in our society by research workers and the importance to their work of effective communication among them should compel careful consideration of the problem.

BIBLIOGRAPHY

1. Flanagan, John C. et al. Critical Requirements for Research Personnel. Pittsburgh: American Institute for Research, 1949.

2. Weislogel, Mary H. The Development of a Test for Selecting Research Personnel. Pittsburgh: American Institute for Research, 1950.

3. Weislogel, Mary H. Procedures for Evaluating Research Personnel with a Performance Record of Critical Incidents. Pittsburgh: American Institute for Research, 1950.

4. Weislogel, Mary H. The Development of Tests for Evaluating Research Proficiency in Physics and Chemistry. Pittsburgh: American Institute for Research, 1951.

5. Weislogel, Mary H. and Altman, James W. Abstracts of Literature Concerning Scientific Manpower. Pittsburgh: American Institute for Research, 1952.

6. Evaluating the Performance of Research Personnel: A Manual for Supervisors. Pittsburgh: American Institute for Research, 1951.

7. Critical Requirements for Research Personnel. Research Notes No. 2. Pittsburgh: American Institute for Research, 1949.

8. The Development of Tests of Aptitude and Proficiency. Research Notes No. 5. Pittsburgh: American Institute for Research, 1951.