

AD-753 814

THE EVALUATION OF RESEARCH SUPPORTED
BY A MISSION-ORIENTED AGENCY-STRUCTURED
APPROACHES. VOLUME I: ON MEASURES OF
RESEARCH EFFECTIVENESS

Roger C. Molander, et al

Institute for Defense Analyses

Prepared for:

Defense Advanced Research Projects Agency

December 1971

DISTRIBUTED BY:

NTIS

National Technical Information Service
U. S. DEPARTMENT OF COMMERCE
5285 Port Royal Road, Springfield Va. 22151

PAPER P-753

THE EVALUATION OF RESEARCH SUPPORTED BY A
MISSION-ORIENTED AGENCY--STRUCTURED APPROACHES

in two volumes

Volume I: On Measures of Research Effectiveness

Roger C. Molander
A. Fenner Milton
Philip A. Selwyn

December 1971

Report number
NATIONAL TECHNICAL
INFORMATION SERVICE
U.S. Department of Commerce
NIST-71-22751

DDC
RECEIVED
JAN 16 1973
B



INSTITUTE FOR DEFENSE ANALYSES
SCIENCE AND TECHNOLOGY DIVISION

DISTRIBUTION STATEMENT A

Approved for public release;
Distribution unlimited

IDA Log No. HQ 71-12639
Copy 83 of 95 copies

54

AD753814

UNCLASSIFIED

Security Classification

DOCUMENT CONTROL DATA - R & D		
(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)		
1. ORIGINATING ACTIVITY (Corporate author) INSTITUTE FOR DEFENSE ANALYSES 400 Army-Navy Drive Arlington, Virginia 22202		2a. REPORT SECURITY CLASSIFICATION UNCLASSIFIED
		2b. GROUP --
3. REPORT TITLE The Evaluation of Research Supported by a Mission-Oriented Agency--Structured Approaches - Volume I: On Measures of Effectiveness		
4. DESCRIPTIVE NOTES (Type of report and inclusive dates) Paper P-753 - December 1971 (published December 1972)		
5. AUTHOR(S) (First name, middle initial, last name) Roger C. Molander, A. Fenner Milton, Philip A. Selwyn		
6. REPORT DATE December 1971	7a. TOTAL NO OF PAGES 46	7b. NO OF REFS 5
8a. CONTRACT OR GRANT NO. DAHC15 67 C 0011	8b. ORIGINATOR'S REPORT NUMBER(S) P-753	
8c. PROJECT NO. DARPA Assignment 20	8d. OTHER REPORT NO(S) (Any other numbers that may be assigned this report) None	
10. DISTRIBUTION STATEMENT Approved for public release; distribution unlimited.		
11. SUPPLEMENTARY NOTES None	12. SPONSORING MILITARY ACTIVITY Defense Advanced Research Projects Agency Arlington, Virginia 22209	
13. ABSTRACT The feasibility of developing structured approaches to prospective and retrospective evaluation of basic research supported by a mission-oriented agency, specifically the Advanced Research Projects Agency, is examined. It is argued that the existence of an applied technological mission for the agency, combined with the unreliability of assessing the potential applied impact of individual research projects, calls for an approach to research evaluation which (1) measures scientific disciplines and subdisciplines according to their relevance to the agency's mission and (2) measures individual research projects within these categories against a yardstick of scientific excellence. Potential structured research methods to accomplish these tasks are considered. However, no successful measures for relevance assessment or for prospective evaluation of individual projects, as cited above, were identified. Structured retrospective evaluation measures for project evaluation are feasible, but their associated problems are too severe to warrant relying on them for more than the purpose of identifying individual projects for more thorough but more subjective review. A number of research-on-research projects that can increase our knowledge of the research process are discussed, but in the absence of a conceptual breakthrough, the likelihood of further work of that type significantly improving the prospects for structured methods is too low to recommend their support.		

DD FORM 1473
1 NOV 66

UNCLASSIFIED
Security Classification

ia

Security Classification

UNCLASSIFIED
Security Classification

Security Classification

PAPER P-753

**THE EVALUATION OF RESEARCH SUPPORTED BY A
MISSION-ORIENTED AGENCY--STRUCTURED APPROACHES**

in two volumes

Volume I: On Measures of Research Effectiveness

Roger C. Molander
A. Fenner Milton
Philip A. Selwyn

December 1971



**INSTITUTE FOR DEFENSE ANALYSES
SCIENCE AND TECHNOLOGY DIVISION
400 Army-Navy Drive, Arlington, Virginia 22202**

**Contract DAHC15 67 C 0011
DARPA Assignment 20**

PREFACE

This report is divided into two volumes. The first is a summary volume reporting the results of a broad look at measures of research effectiveness and the second describes a detailed examination of one approach to purely quantitative methods of evaluating research.

Volume I - On Measures of Research Effectiveness

Volume II - A Preliminary Look at Quantitative Methods
for Evaluating Research

ABSTRACT

The feasibility of developing structured approaches to prospective and retrospective evaluation of basic research supported by a mission-oriented agency, specifically the Advanced Research Projects Agency, is examined. It is argued that the existence of an applied technological mission for the agency, combined with the unreliability of assessing the potential applied impact of individual research projects, calls for an approach to research evaluation which (1) measures scientific disciplines and subdisciplines according to their relevance to the agency's mission and (2) measures individual research projects within these categories against a yardstick of scientific excellence. Potential structured research methods to accomplish these tasks are considered. However, no successful measures for relevance assessment or for prospective evaluation of individual projects, as cited above, were identified. Structured retrospective evaluation measures for project evaluation are feasible, but their associated problems are too severe to warrant relying on them for more than the purpose of identifying individual projects for more thorough but more subjective review. A number of research-on-research projects that can increase our knowledge of the research process are discussed, but in the absence of a conceptual breakthrough, the likelihood of further work of that type significantly improving the prospects for structured methods is too low to recommend their support.

CONTENTS

I. Study Description	1
II. Summary of Conclusions and Recommendations	4
III. Introduction	9
IV. The Research Endeavor	12
V. The Selection and Evaluation of Disciplines and Subdisciplines--The Relevance Problem	15
VI. The Selection and Evaluation of Individual Research Projects--The Problem of Assessing Scientific Excellence	20
A. Research Proposals	20
B. Research Progress	21
C. Research Results	22
VII. Research Evaluation Models and Techniques	24
A. The NIH Research Grant Proposal Evaluation Process	24
B. The Abt Associates Study	26
C. The IDA Study	27
D. Other Studies of Interest	30
References	32
Appendix A--DoD RDT&E Categories	34
Appendix B--The NIH Research Proposal Evaluation Process	36
Appendix C--Abt Associates Study on Evaluation of Basic Research	40
Appendix D--Computer Horizons Work on Indexing Usage	44

FIGURE

1. Study Outline	5
------------------	---

TABLES

1. Productivity Measures	29
2. Comparison of Project Rank Orderings Obtained Using Different Productivity Measures	29

I. STUDY DESCRIPTION

This study examines the feasibility of developing meaningful structured* approaches for evaluating research** in the particular context of research supported by a mission-oriented government agency. Particular emphasis is given to basic research supported at universities by the Advanced Research Projects Agency (ARPA) of the Department of Defense (DoD) and the formulation of general research evaluation methods, i.e., methods that have broad applicability and do not require substantial modification between cases. Both the prospective evaluation of research proposals and the retrospective evaluation of completed projects are considered.

The investigative program included:

1. In-depth discussions with ARPA-sponsored university researchers, ARPA program managers, the current ARPA director, and other individuals in and out of government with interest and experience in research evaluation.
2. Examination of the available literature on research evaluation methods. (One study by Abt Associates was identified as particularly pertinent.)

* We choose to use the term "structured" rather than "quantitative" since we consider research evaluation methods that range from the quantification of subjective value judgments (e.g., rank ordering of projects by peer groups) to quantitative methods that are completely devoid of any subjective judgments (e.g., counting published papers). The more objective methods are of particular interest in the present study.

** For this study, the term "research" refers to those studies commensurate with the basic definition of DoD 6.1 Research as given in Appendix A. These are studies whose objective is to advance the state of knowledge in a particular scientific area.

3. Examination and evaluation of all measures which the authors could identify as being potentially useful in the structured evaluation of proposed, on-going, and completed research projects.
4. Effort at formulating a highly structured research evaluation method for completed research projects using real data obtained from an ARPA-sponsored university laboratory (described in detail in Volume II of this report).
5. Meeting of individuals (representing the universities, industry, and government) familiar with research evaluation problems to discuss the preliminary conclusions of this study as well as the general state of the art in structured approaches to research evaluation.
6. Analysis of the problems encountered in assessing the relevance of research supported by a mission-oriented agency.

It should be noted that in the course of the study the authors could identify no structured methods of evaluating research which they could recommend for implementation (other than simple modifications of peer evaluation). Furthermore, despite numerous possibilities, they could identify no "research-on-research" projects whose successful completion would significantly improve the potential utility of any identifiable structured evaluation method. Since the conclusions of this study are not positive with regard to structured methods, this report is not intended to provide an implementation rationale. It is hoped that this report will be of value by serving to:

1. Describe the state of the art in structured approaches to research evaluation.
2. Document the analysis leading to the study's inconclusive findings.
3. Set forth certain conclusions and recommendations concerning nonstructured evaluation of research supported by a mission-oriented agency and the possible partial contribution structured methods could make to that process.

4. Warn those individuals, who in the future might look to more structured approaches as a promising means of evaluating research, of the difficulties involved.

II. SUMMARY OF CONCLUSIONS AND RECOMMENDATIONS

The authors' principal conclusions and corresponding recommendations are given below. A flow chart that details the analysis leading to these conclusions and recommendations is shown in Fig. 1.

Conclusion: Prospective relevance* can probably be adequately judged on a subdiscipline level if extensive technical knowledge and a keen appreciation of possible future defense problems are available.

Conclusion: For individual research projects, prospective relevance cannot reasonably be predicted. No general structured research evaluation method was identified that would directly reflect the most important objective of ARPA (i.e., to support research that will eventually be useful to the defense community) and that could feasibly be implemented on a project-by-project basis.

The inherent statistical character of basic research success when judged on the basis of eventual applicability argues strongly against the logic of attempting to predict the outcome and importance of individual basic research projects. Although a retrospective investigation to determine which research projects have been the most useful should be possible, at least in principle, a number of factors (e.g., the long time period that would require examination, the vague and often circuitous routing of scientific information, and the poor source documentation characteristic of applied endeavors) generate a situation where this can be done only by laborious case studies that are simply not amenable to general structured approaches.

* In this study relevant research areas/projects are defined as those whose results impact on applied problems, in this case, in weapons systems technology.

STUDY OBJECTIVE: To examine the feasibility of developing structured approaches to prospective

respective evaluation of ARPA-supported basic research

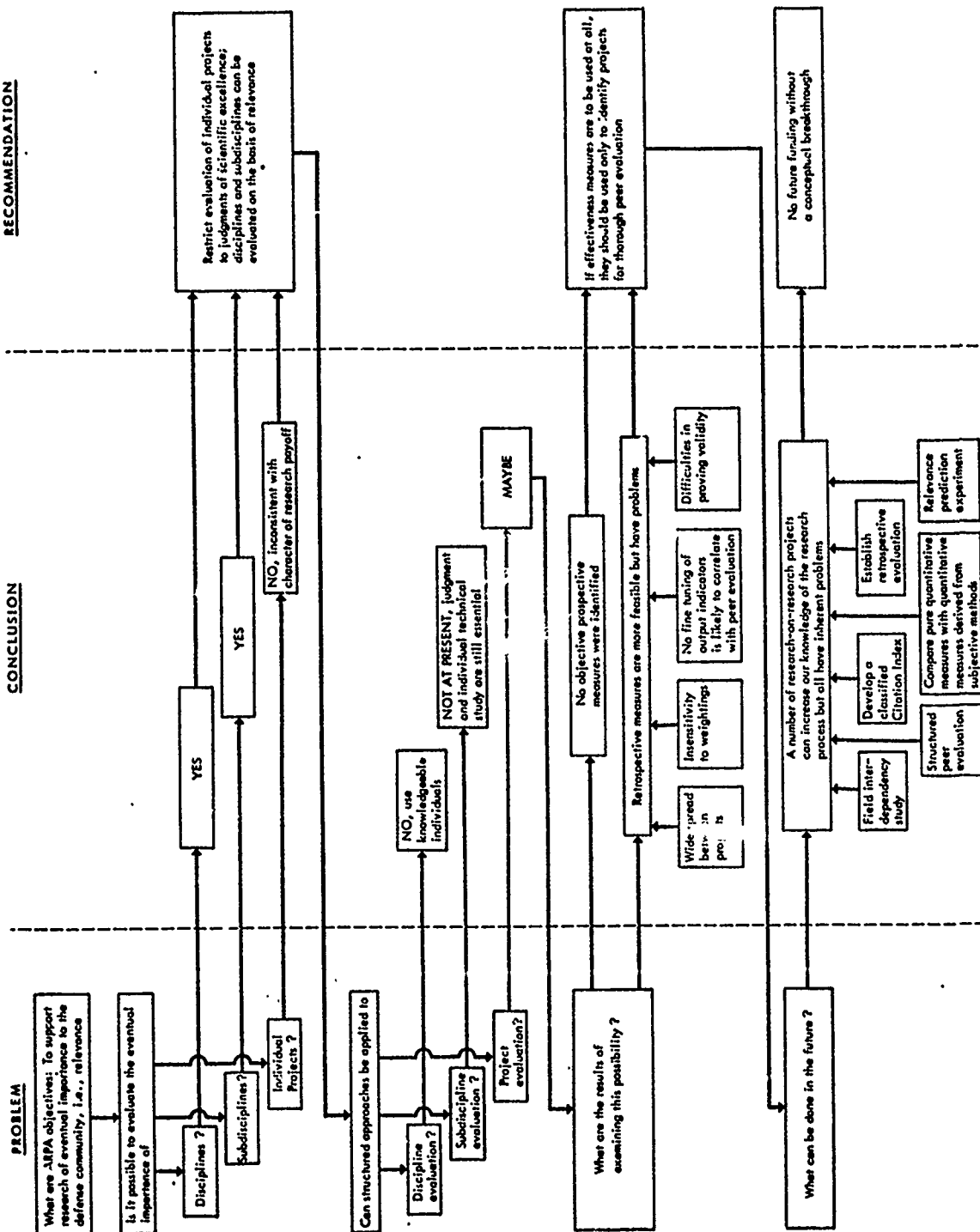


FIGURE 1. Study Outline

Recommendation 1: The evaluation of research supported by a mission-oriented agency should consist of:

1. The determination of those scientific disciplines and sub-disciplines that are particularly relevant to the agency's problem, and
2. Within relevant subdisciplines, individual research projects should be evaluated on the basis of judgment of scientific excellence.

* * * * *

Conclusion: The determination of relevant disciplines and, at least at present, subdisciplines does not lend itself to general structured approaches other than group consensus techniques.

Because of the requirement for detailed examination of research opportunities and potential applications, discipline and subdiscipline relevance assessment can probably be done only by knowledgeable individuals on a case-by-case basis.

Conclusion: No objective quantitative indices identified in this study can reasonably hope to prospectively measure the most important scientific characteristics of a proposed research project. These characteristics are (1) scientific value of a proposed research result if successfully attained, and (2) the likelihood of the investigator successfully completing the project if it were funded.

Conclusion: Several identifiable measures appear to be of possible utility in the retrospective evaluation of individual research projects. Among these measures the number of publications and their utilization (as reflected in the Science Citation Index) are of particular interest.

An IDA analysis of individual research projects yielded indices of effectiveness which (1) showed a large spread (a factor of 10) between individual research projects, and (2) produced a rank ordering of projects that was relatively insensitive to the exact weights given to various components of the indices. An Abt Associates study exhibited a similar insensitivity to weighting factors.

Conclusion: Even sophisticated output indicators are not likely to correlate with exhaustive retrospective peer evaluation.

This conclusion derives from the Abt Associates study comparing qualitative indices with peer evaluation and is predicated on the validity of that study. The projects selected for the Abt study were subjected to exhaustive peer evaluation, both prospectively and retrospectively (by different groups). The rank orderings as to scientific quality so obtained did not correlate with the rank ordering derived from any of the quantitative indices (prospective and retrospective) although the prospective and retrospective peer evaluations did correlate. Very sophisticated output indicators involving the actual use of the output information by subsequent investigators were examined.

Conclusion: All novel research evaluation methods suffer from the difficulty of finding a means of proving or disproving their validity. In the absence of other established methods, comparison with exhaustive peer evaluation seems the best means of testing validity.

Recommendation 2: Structured effectiveness measures (e.g., those based on the number and quality of publications) should be utilized only as a guide to selecting projects for further examination (i.e., more exhaustive peer evaluation). They should not be used alone as a basis for funding decisions.

* * * * *

Conclusion: It is possible to identify a number of research-on-research projects that could contribute to the general understanding of the research process and the evaluation of research. These include:

1. A relevance predictability experiment that would examine the level (discipline, subdiscipline, etc.) at which relevance judgments can be made, the appropriate time frame for such judgments, and the type of individuals best able to make such judgments.

2. Analysis of interdependence of fields (e.g., using literature citations; see Computer Horizons study described in Appendix D) that could identify the scientific disciplines or subdisciplines upon which identifiable areas of DoD interest are dependent.
3. The development of a Citation Index for classified literature.
4. The further structuring of peer-evaluation processes by quantifying subjective judgments of several characteristics of individual research efforts.
5. The comparison of purely quantitative evaluation methods with properly quantified subjective value judgments such as those obtained from a peer group.
6. The establishment of a formal retrospective review process.

Unfortunately, it appears doubtful that the successful completion of any of the above projects would significantly improve the potential validity of any identifiable structured evaluation method.

Conclusion: It is not likely that the steady funding of research-on-research will significantly improve the prospects for structured research evaluation methods. This conclusion applies to the relatively basic research funded by a mission-oriented government agency.

Recommendation 3: Future funding of research directed to the achievement of structured research evaluation methods should be conditional on a conceptual breakthrough.

* * * * *

III. INTRODUCTION

During the last few years, defense research programs have experienced increasing Congressional budgetary pressure. In the Defense Department this has led to a strong desire on the part of program managers to increase the productivity of available research funds (as far as their eventual contribution to the achievement of DoD goals is concerned) and a companion desire to evaluate research productivity in a more defensible manner by making the whole research process more subject to objective analysis and review.

One of the review processes frequently considered is structured evaluation. Interest in structured evaluation techniques is, of course, motivated by a lack of complete satisfaction with the more traditional methods. Traditionally, research proposals have been evaluated either by executive judgment or by peer evaluation. Executive judgment has always been hard to defend and peer evaluation can suffer from subjective bias and from misunderstandings concerning agency objectives. In any case, the traditional university reviewer will have difficulty implementing objectives concerned with value or relevance. Explicit retrospective evaluation of completed research projects has traditionally been ignored.

As indicated in the Study Description (Section I), this effort has been principally concerned with the problems encountered in the evaluation of ARPA-supported research. Nevertheless, it is anticipated that the findings will have general applicability to research (in particular, university research) supported by other offices within DoD and by other mission-oriented government agencies, such as the National Institutes of Health (NIH) and the Environmental Protection

Agency. The results will also be applicable in part to research supported by the National Science Foundation (NSF), although perhaps to a lesser degree in view of the NSF emphasis on non-mission-oriented research.

The ARPA mission may be broadly defined as the support of research in rapidly expanding scientific areas of potential DoD significance. The ARPA research evaluation problem can be broken into two separate issues or problems.

1. Determination of which disciplines (or subdisciplines) to support and the fraction of the ARPA research budget to allocate to each.
2. Determination of which research projects to fund within a specified discipline (or subdiscipline) and the evaluation of the progress or results of these research projects.

The first problem is essentially one of assessing the degree to which various scientific disciplines are relevant to recognized DoD problem areas, i.e., systems needs. The second problem may also include relevance considerations, although evaluation of the comparative scientific merit of individual research efforts should be the dominant concern; the likelihood of being able to measure the worth of an individual project meaningfully against the agency mission is too small to consider on a routine basis.

The reader will find that in many instances the material that follows is not purely analytical in nature, but instead includes judgments and considered opinions derived by the authors in their pursuit of a workable structured research evaluation method. Even though these comments support the contention of many individuals within the scientific community that research evaluation is not amenable to analytical methods, it is felt that their inclusion here can be of assistance to investigators who pursue this difficult subject in the future.

Section IV describes what is meant by a research activity and lists the various ways such an activity can contribute to applied

goals. Section V describes the problem of determining the degree of relevance of scientific disciplines and subdisciplines to DoD systems needs, and Section VI describes potential measures for evaluating individual research projects. Section VII describes structured evaluation methods that have been subjected to experiment, including a brief description of an IDA experiment that is described in detail in Volume II of this report.

IV. THE RESEARCH ENDEAVOR

For the purposes of this study, the term "research" refers specifically to research as defined under Category 6.1 activities in the Department of Defense (Appendix A provides complete definitions of the various research and development terms, i.e., "6.1--6.6," in common use within DoD.). These activities run the spectrum from what has been traditionally called "basic research" to more applied activities, but they all have the common denominator of generating new scientific knowledge. The distinguishing characteristic between this research and that traditionally supported by the National Science Foundation (NSF) is not the nature of the work but rather a restriction in scope just to fields of interest to DoD. In contrast, the term "development" refers to activities that generally involve the design, construction, and testing of prototype materials, devices, and equipment for practical systems. The outcome of development activities may be uncertain, but ordinarily the motivation for such activities is clear since potential applications usually have been identified. In the case of research, however, there is generally much more uncertainty. Occasionally a specific requirement for new knowledge may have been generated by a potential development. But normally the outcome of the more basic forms of research is so uncertain that it is virtually impossible to predict where (i.e., in which development program) the results of a proposed research project will prove useful--or even whether they will be useful at all. Yet it is beyond dispute that in the last 40 years, many forms of development have relied on an understanding of the physical universe that has been built up by basic research activities. In recognition of this fact, there is a widely held impression that a small but healthy research activity is essential to an innovative

development program (i.e., the more we understand, the more we can build). Accepting this argument, the difficult questions involve how much and what kinds of research are necessary to fuel an innovative development program.

Research can contribute to development in a number of ways. Research may uncover a wholly new phenomenon that can open up a development activity that did not exist before. Research can answer questions concerning the performance of existing equipment or devices in the development stage and thereby point the way to improvements. This diversity of possible types of contribution hinders efforts to measure relevance. A research activity may also aid a development program indirectly through synergistic effects that result from the interpersonal exchange that occurs when research and development activities are carried out in close proximity. In addition, researchers frequently become developers, and apparently very good ones. Such effects are often quoted by industry as the prime motivation for supporting small-scale basic research activities in a development laboratory. As Alvin Weinberg has stated, "In a large multidisciplinary applied laboratory, basic scientists keep their technological colleagues honest. They are the eyes through which the institution keeps in touch with the rest of the world of science." (Reference 1.) Unfortunately, in the university community (where most DoD-supported research takes place) there is usually very little interaction between basic research and applied work. Thus the synergistic product of research and development activity is probably absent for most of the research of principal interest to this study, i.e., university research. Besides, there remains a question whether meaningful quantitative measures could be found for this influence.

The direct product of a basic research activity (i.e., new knowledge) usually appears in publications that can be identified. Unfortunately (from the point of view of a mission-oriented agency), the publication of new knowledge is only the first step toward achieving a practical goal. In the short run, basic research tends to generate queries for more basic research. The information transfer

to a real development program can be very circuitous and often occurs only after a large number of basic studies (which feed off each other) has advanced the state of knowledge in a particular field. Our quantitative knowledge of this process depends upon case studies, such as HINDSIGHT (Ref. 2) and TRACES (Ref. 3), which have shown that the time period between the generation of new knowledge and the appearance of a new system or product whose development depended upon that knowledge is often several tens of years. It is also clear that from the point of view of generating a breakthrough that will lead directly to system or product improvements, basic research is indeed a high-risk enterprise. Only a very small portion of basic research projects pays off in that way, even though in some cases the payoff can be very large.

V. THE SELECTION AND EVALUATION OF DISCIPLINES AND SUBDISCIPLINES--THE RELEVANCE PROBLEM

As noted in the Introduction, a major concern in increasing the effectiveness of ARPA-supported research is the determination of those scientific disciplines or subdisciplines that are relevant to recognized DoD problem areas. In this regard, it should be emphasized that ARPA is concerned with relevance not to a weapon system per se but rather to weapon system technology.

Research can impact on current problem areas. ARPA's principal concern, however, is with anticipated problem areas, i.e., the response to future threats. An example of the former might be research on weather prediction, or perhaps more specifically tropospheric winds where improved knowledge of such phenomena could result in increased missile accuracy for existing strategic systems. Clearly, the identification of those scientific areas that are relevant to problems of this nature is a relatively straightforward endeavor (and, in fact, research in these areas might be characterized as applied research). However, defining the relevant research areas becomes progressively more difficult as one contends with future threats for which the system response is progressively less well defined. Yet it is on these problem areas that research, and particularly basic research, can have the greatest impact. For example, the viability of the sea-based deterrent may be jeopardized at some time in the future so that research on a wide spectrum of underwater detection phenomenology is currently warranted. Such research could profoundly affect U.S. SSBN operational procedures and deployment decisions as well as dictate the need for new and improved surveillance systems even though it is impossible at this time to predict the precise nature of these effects. The discussion that follows is directed to the identification of scientific areas relevant to future problems of such a nature.

The relevance assessment problem is made particularly difficult by the considerable time lapse between the initiation of research and its eventual appearance in a new or improved military system. This time lapse is an inescapable characteristic of the RDT&E process--a natural result of the need for new ideas and concepts to go through a long progression of research, development, test and evaluation. Although the actual length of time required for this process is not well known and is probably variable, one might infer from the results of TRACES (Ref. 3) that for most basic research it is generally in excess of 15 to 20 years. One might be led to conclude that a valid relevance assessment is dependent upon the degree to which one can accurately predict the important military systems almost two decades into the future. Such a demand for prescience is not only excessive, but, in fact, illogical and, in practice, no such demand is made. As implied in the previous discussion, it is not the prediction of the precise system which will respond to a particular threat that is asked for, but the identification of the threat and concomitantly the identification of a spectrum of possible responses to the threat. It is this effort that yields the relevant scientific areas to be explored. Thus in funding research, one should be interested in the identification of concepts that will reach the (6.2) exploratory development or (6.3) advanced development stages, recognizing that some concepts may never progress beyond these stages for reasons relating to cost, competitive effectiveness, or even geopolitics. However, this does not mean that the research that contributed to such development projects, can be considered irrelevant or unsuccessful. Feasibility knowledge gained in development programs contributes directly to a lessening of a military risk. Since all systems in the development stages are by definition candidates for eventual deployment¹, it is illogical to decide that the research which contributed to one such system is "more relevant" than that which contributed to another. Acceptance of the idea that research need not contribute to a deployed system in order to be considered successful and relevant results in a much less severe demand on prescience and perceptivity. Thus value for a conceptual system should provide as much justification for a research effect as value for an operational system.

There still remains the inherent difficulty in predicting scientific success and the resulting technological development, i.e., the difficulties in predicting which research areas will actually pay off. In this regard, an important question is the level (discipline, subdiscipline, or project) at which an accurate assessment of promising research areas can be made. Because individual research projects have a high failure rate if judged on the basis of technological payoff, it is generally inappropriate to evaluate such projects on a relevance-prediction basis. Relevance criteria should be employed to determine those scientific areas to receive funding (and their budget levels), but within these areas projects should be funded on the basis of their scientific merit alone. Consequently, it is desirable that a judgment concerning promising research areas be made at the narrowest, statistically meaningful level above the project level. For example, one would like to be able to say with confidence that composites and thin-film semiconductors are relevant subdisciplines rather than simply saying that materials science or solid-state physics is a relevant discipline. At the present time within ARPA, the choices in question appear to be made at the discipline level (e.g., materials sciences, computer science, atmospheric physics, etc.) through budget allocations. Within each discipline there appears to be an informal shopping list of relevant subdisciplines (with no formal budget allocation or comparative relevance judgment) within which research projects are funded on the basis of scientific merit. The choices of which disciplines and subdisciplines to support are in many respects conjectural and their defense must rely on arguments that are inherently qualitative. It becomes of interest to determine whether there is an objective way to validate or invalidate this decision-making process, or whether the process can be structured or quantified in a general way so that the resulting choices will be more objective or superior.

In the absence of established methods for predicting the future relevance (DoD or otherwise) of a spectrum of individual scientific disciplines or subdisciplines, consideration was given to the formulation of a "predictability experiment" that would attempt to determine:

1. The extent to which valid assessment of the comparative relevance of different research areas can be made and the level of detail (discipline, subdiscipline, etc.) at which such judgment can be meaningful.
2. The type of individuals (scientists, DoD program managers, system designers, etc.) best able to make a relevance assessment.
3. The required time period for conclusive determination of relevance.

The experiment envisioned would solicit from individuals drawn from groups, such as those mentioned above, a rank ordering of the relative importance of the various subdisciplines of solid-state physics. The subdisciplines might be further grouped according to mature and immature areas under the assumption that predictability might be much easier in the case of more mature areas. Similarly, it may be necessary to specify the time frame for which the respondents are asked to make the relevance judgment. The resultant rank orderings might then be subjected to a Delphi approach to obtain a consensus within each of the groups or across groups. At the specified times in the future, perhaps 5, 10, and 15 years later, the previously obtained rank ordering of subdisciplines and DoD perception of subdiscipline importance (as reflected in budget allocations) could be compared with the then current perceptions of relative importance in order to assess changes and to determine which, if any, of the groups polled was capable of an accurate relevance prediction. This information could then be used to modify the current budget allocation process. This experiment has several disadvantages, the most severe of which is the time span required for significant results. It may be as long as 10 or 15 years in the future before a valid relevance assessment can be made. In addition, the experiment would not be a controlled one since interim funding allocations could markedly affect the relative development of the different subdisciplines. A rigidly controlled experiment would require even funding over all the candidate subdisciplines. Thus extreme care would have to be exercised in the

interpretation of results. As a consequence, it would be difficult to justify generalizing the results obtained for this one research area, whether they are conclusive or inconclusive, to other areas.

As indicated above, there are no established methods that could lead to near-term modification of current practices of assessing the future relevance of different research areas. However, one might be less ambitious and simply attempt to show how certain research areas have been relevant to certain exploratory or advanced development activities. This might be done by analyzing the scientific disciplines or subdisciplines upon which identifiable areas of DoD interest are dependent through an analysis of literature citations. This approach has been pursued by Computer Horizons (Ref. 4) in showing the dependence of special education on research in psychology. The disadvantage in such an approach for ARPA-supported research is that it may be necessary to go to the classified literature before conclusive relevance arguments can be made. The absence of a classified citation index would severely inhibit this process. Because of this problem, some consideration was given to the difficulties involved in compiling a citation index for classified literature; such a compilation would be very costly and time-consuming. In addition, the classified literature does not have the strong tradition of documenting its sources, such as exists in the open literature, so that one could not be assured that reference lists are really complete. Similarly, some information transfer in the classified community takes place through informal channels such as memoranda that neither contain references nor are they often referenced.

It can be seen from the above discussion that the inherent difficulty in the relevance question is the inescapable requirement for a prediction of either future defense problem areas or the development of scientific research in specific areas. Because of such a requirement, the relevance assessment effort lends itself very poorly to quantitative methods and must rely on subjective value judgments by knowledgeable individuals.

VI. THE SELECTION AND EVALUATION OF INDIVIDUAL RESEARCH PROJECTS--THE PROBLEM OF ASSESSING SCIENTIFIC EXCELLENCE

Organizations that support large-scale research programs have traditionally relied upon the judgment of knowledgeable individuals to determine whether a given research project has the potential for success or, in retrospect, whether the project has been successful. Since this evaluation process, generally characterized by the term "peer evaluation" (although at times involving only line management) involves the subjective judgment of human beings, it is subject to personal prejudices, lack of total understanding, incompatibility of objectives between the evaluators and the supporting agency, and other human failings, as well as being costly and time-consuming. As a consequence, it is hardly surprising that there would be considerable interest in being able to utilize meaningful structured approaches to research evaluation--approaches that would eliminate the human judgment factor as much as possible--to minimize these problems. The discussion that follows explores this problem in the context of evaluating the scientific excellence of (1) research proposals, (2) research progress, and (3) research results.

A. RESEARCH PROPOSALS

An objective judgment on the scientific merit of the proposed research is clearly desirable. Traditionally, such a judgment has been obtained by a peer-group assessment or management judgment with the assessment frequently expressed either qualitatively or through the attribution of a score (e.g., 1-5 or excellent-poor) to the research proposal. One can conceive of injecting further refinement into the scientific merit evaluation by seeking a quantitative measure of the specific goals of the proposed research, for example, by examining the improvement in accuracy of a physical constant or a

physical relationship that would result if the research were successful (this approach was taken in the Abt Associates study, Ref. 4). However, such a measure would still require a judgment of the importance of such an advance in knowledge.

As a necessary concern in pursuing the above factors, one is also interested in the probability of successful completion of the proposed research. Such a probability assessment is frequently a component of quantitative evaluation methods proposed for industrial R&D programs. In most cases, however, rather than attempt an overall assessment of probability of success, it is more common to provide an assessment of the competence of the investigator(s). A quantitative measure of competence could be obtained from the investigator's productivity as reflected in the number or quality of his publications. (Since this essentially constitutes retrospective evaluation of research results, further discussion is reserved for a later section.) The quality of the facilities available to the investigator might also affect the probability of success as well as the cost of the research. Such a measure could probably be obtained only in a very qualitative fashion from an individual who is familiar with the investigator's institution and appreciative of the difficulty of the problem.

One might also have a peripheral interest in the enhancement of the investigator's capabilities (especially for younger researchers) so that an assessment of potential for scientific growth of the investigator would be of interest. It is difficult to conceive of any measure that would reflect this quality other than a judgment obtained from an individual familiar with both the researcher and the area of research.

B. RESEARCH PROGRESS

Evaluation of research progress could conceivably be obtained by a comparison of actual progress with a previously established sequence of intermediate events leading toward a final attainment of the research project objectives. Not all research efforts would lend themselves to the establishment of milestones although it would be

advantageous when feasible, particularly for research projects that span a period of several years. Such procedures are not uncommon in the evaluation of progress on industrial R&D projects. The actual evaluation process would require an establishment of acceptable time lags between the planned and achieved progress as reflected in achievement of the milestones.

C. RESEARCH RESULTS

It can be assumed that the significant results of basic research projects will be contained in the publications directly attributable to the project. Thus an evaluation of research results can focus on an evaluation of project publications. Only in rare cases do the supporting agencies perform a direct evaluation of the output of a research project (although as indicated previously, the evaluation of past performance is used as an input to the evaluation of new research proposals). Instead, there is a tacit assumption that research results will be reviewed and evaluated in the scientific community as reflected in publication of results in prestigious journals and the enhancement of the stature of the investigator(s). Thus, in most cases, the only formal review given most research results is that received by the resulting papers when they are submitted to scientific journals. However, this situation does offer an opportunity for a direct evaluation of the research results through a methodology that considers the quantity and the quality of the publications produced by the research project. (Quality in this case can include an evaluation of both scientific merit and relevance.)

The level of analysis that can be performed on publications in the open literature is certainly time-dependent. The first opportunity for evaluation probably occurs 3 to 12 months after the completion of a piece of work when the resulting publication appears in a journal. The crudest level of analysis one can accomplish in this (or a longer) time frame is to simply count the number of publications. However, it has been suggested that a judgment of quality can be obtained from the prestige of the particular journal in which an article is published.

Preliminary efforts in this direction (Computer Horizons study) have shown that there appears to be a definite ordering of journal quality, although the weighting problem has not as yet been solved in a convincing fashion. Another indication of publication quality is embodied in the utilization of research results by other investigators. A measure of this characteristic can be obtained from the frequency with which publications are cited, as detailed in the Science Citation Index. Because of the time lag between performing research and publishing results, it appears that there is a time lag of at least two to three years before such a measure could become useful. Again one has a weighting problem in assessing the relative importance of numbers of citations.

VII. RESEARCH EVALUATION MODELS AND TECHNIQUES

As noted previously, this analysis is principally directed to structured approaches to research evaluation. Such approaches run the complete spectrum from structured peer-evaluation techniques to purely quantitative methods of evaluating research. The material that follows briefly describes the NIH research grant proposal evaluation process (a highly structured peer-evaluation process), the purely quantitative approach investigated by Abt Associates, and the quantitative methods investigated by the IDA study group (and reported in detail in Volume II of this report).

A. THE NIH RESEARCH GRANT PROPOSAL EVALUATION PROCESS

The process employed by NIH in evaluating research grant proposals is of particular interest because it is the most highly structured research evaluation process employed in organizations with large-scale research programs. The primary responsibility for evaluating applications for NIH research funds rests, not with the individual Institutes, but with the NIH Division of Research Grants (DRG). This division essentially provides the administrative framework within which peer groups are convened in "Study Sections" organized according to scientific disciplines and medical specialties to thoroughly evaluate all health-related research proposals received by NIH.

Following an initial screening to remove those applications which are not health-related, all applications are thoroughly evaluated on the basis of scientific merit by the Study Sections. With the exception of an Executive Secretary, all Study Section members are from outside NIH, although other NIH staff members are present at the Study Section meetings in a non-voting capacity. Those applications which are approved are given a numerical "priority score" (the average of scores

produced by the individual peer-group members) and then transmitted to a specific program area within one of the NIH institutes. (It should be noted that approval does not imply the granting of funds, but only that the application is a potential recipient of funds.) A research program is then generated in each program area by rigidly allocating a specified budget to the candidate research applications (which have been received from several different Study Sections) in order of priority score. The resulting research programs are reviewed by the National Advisory Council of each Institute which may make small changes derived primarily from priority considerations.

A particularly noteworthy characteristic of the above process is the division of responsibility between assessing relevance and assessing scientific merit. The Study Sections do the scientific merit evaluation while the relevance assessment is embodied in the budgets of the individual program areas and in the option of the Advisory Councils to modify allocation decisions through a perception of priorities.

The availability of the "priority scores" described above offers a good opportunity for comparing this prospective evaluation scheme with a retrospective evaluation of completed research projects. However, a comparison of this type has not as yet been attempted.

One might conceive of structuring a peer-evaluation process even further by quantifying subjective judgments of specific characteristics, such as scientific merit of the proposal, capability of the investigator, etc. Such a procedure might reveal those characteristics which dominate the probability of success and the degree to which peer groups can perform valid prospective evaluations. However, there is currently no test bed within ARPA for such an experiment, i.e., none of the research funded by ARPA is subjected to a peer-evaluation process which would lend itself to such an experiment. (NIH would clearly be an excellent test bed for such an experiment.)

B. THE ABT ASSOCIATES STUDY

The Abt Associates study (Ref. 5) was directed to the feasibility of quantitative methods for prospectively and retrospectively evaluating basic research projects by a non-mission-oriented support agency (NSF). Prospective measures were developed for individual research proposals and focused on the characteristics of the fundamental scientific "Relation" which the researcher proposed to investigate. These measures included: (1) a measure of the number of dependent variables involved in the relation under consideration, the range over which these variables were to be determined (experimentally or theoretically), and the precision with which the parameters of interest were to be determined, (2) a measure of what materials the relation applied to, and (3) a measure of how fundamental a relation was.

The retrospective measures developed focused on the utilization of the research results of individual projects (as put forth in actual publications) by subsequent investigators in the same field. The Science Citation Index was employed to find a first generation of publications which had used (not merely cited) distinct scientific results from project publications (source papers) as inputs to their own research. The analysis was also carried to a second generation of publications which used the outputs from the first generation. Three distinct measures of merit (or indices) were developed for individual source papers. These measures included: (1) a measure of the number of legitimate users of the source paper, (2) a measure of the number of "new queries" generated by the source publication, and (3) a measure of the rapidity with which the results of the source paper spread through the resulting network of publications.

The quantitative measures which were developed were tested by applying them to a sample of completed research projects in the field of solid-state physics and comparing the results with peer evaluation both of the original proposals and of the outputs of the completed research projects. These evaluations were performed by two distinct groups of NSF-selected judges. The important results of that comparison were as follows:

1. Peer evaluation of proposals showed a significant positive correlation ($r = 0.60$)* with the ranking of the completed papers by the same method.
2. There was no significant correlation ($r < 0.40$), either positive or negative, between the rank orderings by the NSF method and any of the quantitative methods.
3. Rankings based on a simple counting of citations correlated highly ($r = 0.96$) with a much more complex retrospective method of evaluating the utilization of research results, thus implying that a simple citation counting method may give results as good as more complex methods (which were poor).

The second result is of principal importance: even sophisticated retrospective measures based on the Citation Index did not correlate with peer evaluation. Since this study represents the most comprehensive effort to date to compare a sophisticated quantitative scheme with an independent evaluation method (albeit one which has its own shortcomings), the outlook for quantitative methods of evaluating the scientific merit of research is clearly discouraging.

C. THE IDA STUDY

An IDA study (reported in detail in Volume II of this report) examined the utility of purely quantitative measures of evaluating research. Consistent with the arguments made previously on the difficulty of obtaining quantitative measures which reflected the relevance of a particular project to DoD problems, the effort focused on deriving measures for the scientific excellence of individual research projects. Candidate productivity measures were derived and

*For the sample size (10), a correlation coefficient (r) of 0.4 is significant at the 0.05 level and a value of 0.6 is significant at the 0.01 level.

applied on a project-by-project basis to a four-year sample of research projects at a Materials Science Interdisciplinary Laboratory (IDL).*

The productivity measures examined consisted principally of simple combinations of various input parameters (number of full-time equivalent faculty, number of graduate students, project budget, etc.) and output parameters (articles, books, etc.). For example, the manpower input to a project was summed over all the major contributors to the project with appropriate weightings based on estimates of the relative worth of each manpower type (e.g., in one case the relative worth of professors, post doc's and graduate students was estimated at 4:3:1, respectively, per unit time). Similarly, the total project output was summed over all visible products of the project, again with appropriate weightings (e.g., in one case the relative worth of articles in reviewed journals, articles in unreviewed journals, books, and chapters in books was estimated at 2:1:5:2, respectively). By dividing output by manpower input or total budget, a possible measure of relative project quality is obtained. Another measure of project quality was obtained by attempting to rate the quality of journal articles on the basis of number of citations as obtained from the Science Citation Index. Application of productivity measures such as these to individual research projects resulted in a considerable spread between projects. However, comparisons between projects were made principally on a rank ordering basis rather than on the basis of the absolute value of the productivity measures. Tables 1 and 2 summarize the results for a sample of 23 individual research projects.

*The Materials Science IDL program consists of 12 university laboratories with a total annual operating budget of about \$40 million. About 35% of this operating budget was supplied by ARPA until 1971 when the bulk of ARPA's IDL financial responsibility was transferred to NSF. The IDL from which data were taken had an average annual budget of \$5.2 million during the four years studied. The ARPA portion of this budget (about 32%) was divided among about 25 individual research projects.

TABLE 1. PRODUCTIVITY MEASURES

<u>Measure</u>	<u>Description</u>	<u>Range</u>
α_{FTE}	Publication output divided by manpower input assuming faculty input equaled time charged to the project	0.00-3.65 mean = 1.13
$\alpha_{\frac{1}{2}}$	Publication output divided by manpower input assuming faculty input at half-time for each project	0.00-2.76 mean = 0.84
$\alpha_{\$}$	Publication output divided by total project budget	0.000-0.295 mean = 0.094
$\alpha_{1:1}$	Publication output plus Ph.D and M.S. output (assuming one paper in reviewed journals equals one Ph.D) divided by total project budget	0.011-0.418 mean = 0.118
$\alpha_{FTE/SCI}$	Same as α_{FTE} only publication output modified by Science Citation Index weightings	0.00-8.72 mean = 2.84
α^*_{FTE}	Same as α_{FTE} only faculty:post doc:grad student productivity weighting assumed 6:3:1 rather than 4:3:1	0.00-5.26 mean = 1.41
$\alpha_{3:1}$	Same as $\alpha_{1:1}$ only assuming one Ph.D equals three journal articles	0.027-0.659 mean = 0.173

TABLE 2. COMPARISON OF PROJECT RANK ORDERINGS OBTAINED USING DIFFERENT PRODUCTIVITY MEASURES. SPEARMAN RANK ORDER CORRELATION COEFFICIENTS. (N = 23 PROJECTS)

	α_{FTE}	$\alpha_{\frac{1}{2}}$	$\alpha_{\$}$	$\alpha_{1:1}$	$\alpha_{FTE/SCI}$	α^*_{FTE}	$\alpha_{3:1}$
α_{FTE}							
$\alpha_{\frac{1}{2}}$	0.83						
$\alpha_{\$}$	0.84	0.79					
$\alpha_{1:1}$	-	-	0.93				
$\alpha_{FTE/SCI}$	0.97	-	-	-			
α^*_{FTE}	0.97	-	-	-	-		
$\alpha_{3:1}$	-	-	0.82	-	-	-	

It was found that the resulting rank orderings were relatively insensitive to various weightings applied to the different types of manpower inputs and research outputs in the form of publications. Citation indexing was only found to be useful 2-3 years after the completion of a piece of work and, even then, did not significantly alter the rank orderings found by simply counting the number of publications. The absence of an independent method of judging product quality (no peer-group evaluation was available as in the Abt study) left open the questions of the validity and applicability of the measures derived. However, it seemed clear that projects at the extremes of the productivity spectrum could be isolated in this manner.

The IDA study also applied simple productivity measures on an overall laboratory basis to the twelve ARPA Materials Sciences IDL's in an attempt to test the feasibility of comparing large-scale research programs on a purely quantitative basis. Since the effort was restricted to gross parameters, such as number of publications, number of graduate students, etc., the detailed examination of quantitative measures applied at the project level was not possible. Attempts to correlate level of funding with laboratory productivity measures similar to those discussed above were essentially unsuccessful. A rough independent measure of laboratory quality was available in the percentage change in laboratory budgets subsequent to the four-year period for which average measures were derived. A rank ordering of laboratories on this basis produced no correlation with any of the proposed productivity measures.

D. OTHER STUDIES OF INTEREST

Computer Horizons has undertaken a study for NSF which is directed, in part, to an examination of the feasibility of evaluating completed research on the basis of the journals in which the resultant papers are published and which employs the Science Citation Index as the source of this information. (The effort also includes an examination of the flow of information between disciplines as well as between basic and applied areas.) Such an evaluation technique

offers a 2-3 year advantage over counting citations themselves since it is not necessary to wait for statistically significant citations to appear. Although the effort is more concerned with evaluations of large programs or university departments, there is some possibility of extending the results to individual research projects. At the present time, the study has shown that publications within a given field, e.g., physics or mathematics, tend to group in quality with usually one "super" journal, 4-6 very important journals, about 10 important journals, and then the balance where order is less clearly defined. If this grouping can be shown to be statistically significant and stable, a range of values assigned to the four different types of publications (e.g., 10-5-2-1) could conceivably be used as an average measure of quality. A more detailed description of this study is contained in Appendix D.

REFERENCES

1. Alvin Weinberg, Science, 167, 144, 9 January 1970.
2. Project HINDSIGHT, Final Report, Task I, 1 July 1967.
3. Illinois Institute of Technology Research Institute, Technology in Retrospect and Critical Events in Science (TRACES), Contract NSF-C535, 15 December 1968.
4. Francis Narin, "Analysis of Research Journals and Related Research Structure in Special Education," Computer Horizons, Inc., February 1971.
5. Abt Associates, Inc., A Comparative Study of the Prospective and Retrospective Approaches to the Evaluation of Proposed Basic Research, July 1969.

APPENDIX A--DoD RDT&E Categories	34
APPENDIX B--The NIH Research Proposal Evaluation Process	36
APPENDIX C--Abt Associates Study on Evaluation of Basic Research	40
APPENDIX D--Computer Horizons Work on Citation Indexing Usage	44

APPENDIX A

DOD RDT&E CATEGORIES

The Department of Defense RDT&E (Research, Development, Test and Evaluation) Program is structured as follows:

6.1. Research - Includes all effort directed toward increased knowledge of natural phenomena and environment and efforts directed toward the solution of problems in the physical, behavioral and social sciences that have no clear direct military application. It would, thus, by definition, include all basic research and, in addition, that applied research directed toward the expansion of knowledge in various scientific areas. It does not include efforts directed to prove the feasibility of solutions of problems of immediate military importance or time-oriented investigations and developments. The Research elements are further characterized by using level of effort as the principal program control.

6.2. Exploratory Development - Includes all effort directed toward the solution of specific military problems, short of major development projects. This type of effort may vary from fairly fundamental applied research to quite sophisticated bread-board hardware, study, programming and planning efforts. It would thus include studies, investigations and minor development effort. The dominant characteristic of this category of effort is that it be pointed toward specific military problem areas with a view toward developing and evaluating the feasibility and practicability of proposed solutions and determining their parameters. Program control of the Exploratory Development element will normally be exercised by general level of effort.

6.3. Advanced Development - Include all projects which have moved into the development of hardware for experimental or operational test. It is characterized by line item projects and program control is exercised on a project basis. A further descriptive characteristic lies in the design of such items being directed toward hardware for test or experimentation as opposed to items designed toward hardware for test or experimentation as opposed to items designed and engineered for eventual Service use. Examples are VTOL Aircraft, ARTEMIS, Experimental Hydrofoil, X-15, and Aerospace Plane Components.

6.4. Engineering Development - Include those development programs being engineered for Service use but which have not yet been approved for procurement or operation. For example: MAULER, TYPHON, B-70. This area is characterized by major line item projects and program control will be exercised by review of individual projects.

6.5. Management and Support - Include research and development effort directed toward support of installations or operations required for general research and development use. Included would be test ranges, military construction, maintenance support of laboratories, operations and maintenance of test aircraft and ships. Costs of laboratory personnel, either in-house or contract-operated, would be assigned to appropriate projects or as a line item in the Research, Exploratory Development, or Advanced Development Programs areas, as appropriate. Military Construction costs directly related to a major development program will be included in the appropriate element.

6.6. Operational System Development - Includes research and development effort directed toward development, engineering and test of systems, support programs, vehicles and weapons that have been approved for production and Service employment. This area is included for convenience in considering all RDT&E projects. All items in this area are major line item projects which appear as RDT&E Costs of Weapons Systems Elements in other Programs. Program control will thus be exercised by review of the individual research and development effort in each Weapon System Element.

Categories are further subdivided into elements and aggregations. The R&D program element is the smallest subdivision of the R&D Program considered in this system. Each element will consist of RDT&E projects in the same budget activity. It may consist of a number of projects in a related field as in the Research and Exploratory Development categories or it may be a single major project. In the Advanced Development and Engineering Development categories it may be desirable to group a number of related elements under a descriptive title. Such groupings are called aggregations; e.g., in the Army Engineering Developments, the elements dealing with communications are grouped into a Communications Aggregation.

Table A-1 summarizes the budget allocations in these categories for FY 70-71.

TABLE A-1. ESTIMATED DISTRIBUTION OF DOD RDT&E (DOLLARS IN MILLIONS)

CATEGORY	FY 1970		FY 1971	
6.1 Research	\$ 368.5	5.0%	\$ 369.6	5.0%
6.2 Exploratory Development	857.0	11.5%	897.4	12.2%
6.3 Advanced Development	938.7	12.6%	1,112.7	15.3%
6.4 Engineering Development	1,021.8	13.7%	1,395.9	18.9%
6.5 Management & Support	1,205.0	16.2%	1,167.5	15.9%
- Emergency Fund	75.0	1.0%	50.0	0.7%
6.6 Operational Systems Development	2,972.9	40.0%	2,352.5	32.0%
TOTAL	\$7,438.9	100.0%	\$7,345.6	100.0%

APPENDIX B

THE NIH RESEARCH PROPOSAL EVALUATION PROCESS

The primary responsibility for evaluating applications for NIH research funds rests, not with the individual Institutes, but with the NIH Division of Research Grants (DRG). This division essentially provides the administrative framework within which peer groups are convened in "Study Sections" organized according to scientific disciplines and medical specialties to thoroughly evaluate all health-relevant research proposals received by NIH.

The review process is initiated in the DRG Referral Office (see Fig. B-1) where all applications are screened and those which are not health-related (about 1%) are immediately rejected. Each application is then assigned to the appropriate Study Section* according to the nature of the proposed research. Although the Study Sections are not organized along Institute lines, in many cases the great majority of the proposals reviewed by a Study Section are appropriate only to a single Institute. Approximately two months prior to a Study Section meeting (there are three meetings a year), all proposals to be considered at the upcoming meeting are sent to Study Section members for perusal. In addition, each proposal is given a thorough review by one to three Study Section members (designated by the Executive Secretary of the Study Section, a DRG staff member) depending on the workload of the group.

* Each Study Section (currently there are 46) consists of 10 to 15 distinguished scientists from universities and other public and non-profit institutions who essentially serve as consultants to NIH. Members are appointed by the Director of NIH to serve overlapping four-year terms.

As a first step, the Study Sections evaluate the applications on the basis of scientific merit. This evaluation includes the significance of the proposed project, the qualifications of the investigator, the proposed methodology, and the facilities available at the investigator's institution. With these criteria, the Study Section recommends approval or disapproval (or deferral if the group finds some inadequacy that might be remedied) of each application (see Fig. B-1). Approval does not imply the granting of funds, but only that the application is a potential recipient of funds. Those proposals that have been approved are then assigned a priority score by soliciting a rating (on a 1-5 basis, secret ballot) from each Study Section member and averaging the ratings for the Study Section as a whole. The Executive Secretary of the Study Section prepares a summary recommendation for each proposal which, along with the priority score, represents the Study Section evaluation of the proposal. All approved applications are then forwarded to the appropriate Institute.

The above procedure is followed for each of the Study Sections. As noted previously, all the proposals from a single Study Section do not go to the same Institute, much less to the same Program Area within an Institute. Since some Study Sections tend to score higher or lower than others (on an average), in order to give equitable consideration to all proposals, the priority scores from individual Study Sections are normalized to some average score.

Within each Institute, the approved proposals are transmitted to the appropriate Program Area. Each Program Area has an anticipated research budget and essentially allocates that budget to the approved proposals it receives, rigidly funding proposals from the top down on the basis of their priority scores. The results constitute the recommended research programs for each Program Area.

The recommendations of the individual Program Areas are then subjected to review by the appropriate National Advisory Council for the Institute that will do the funding (see Fig. B-1). The Councils consist of 12 members of which at least one-half are experts in the field

while the others are leaders in public affairs. The Councils rely heavily on the technical evaluation of the Study Section, i.e., summaries and priority scores, and defer on only 2 to 3% of the applications recommended for funding by the different Program Areas. The Council considers projects from the perspective of the Institute as a whole which implies a consideration of high priority areas. To accommodate this sense of priorities, the Council may increase the priority scores of applications in certain areas by an appropriate amount. This results in a reordering of the applications and accounts for the bulk of the 2 to 3% of changes. There are also other peripheral considerations, such as granting funds to promising young scientists or to certain universities, which are sometimes treated in a like manner.

Final decisions officially rest with the Surgeon General, although he seldom differs with the decision of the Council. He also cannot award a grant without recommendation from the Council. Pay lists go to the Councils and according to priority score and adequacy of budget, the director of the Institute or division that will fund the grant notifies the investigator.

If an application has been disapproved, the investigator may ask why and resubmit his proposal in an altered or amended form. Sixty-five percent of all applications for grants are approved, of which approximately 50% are funded. Grants are usually for a period of three years with a current maximum of seven years.

The procedure described above is employed for all undirected research supported by NIH (\$445,000,000 in FY 70). A different procedure is employed for directed research and development programs. In contrast to the "grants" given for undirected research, "contracts" are issued for directed research and development programs.

APPENDIX C

ABT ASSOCIATES STUDY ON EVALUATION OF BASIC RESEARCH

The Abt Associates study was directed to the feasibility of quantitative methods for prospectively and retrospectively evaluating basic research projects. Emphasis was given to developing measures of relative effectiveness in terms that are directly relevant to the mission of the sponsoring agency. The sponsoring agency of interest in this case was the National Science Foundation whose primary mission was described as "the advancement of science."

PROSPECTIVE MEASURES

Prospective measures were developed for individual research proposals and focused on the characteristics of the fundamental scientific "relation" which the researcher proposed to investigate. Three distinct measures were developed for individual proposals.

1. Range-Precision, which was a measure of the number of dependent variables involved in the relation under consideration, the range over which these variables were to be determined (experimentally or theoretically), and the precision with which the parameters of interest were to be determined.
2. Region of Applicability, which was a measure of what materials the relation applied to, and
3. Depth, which was a measure of how fundamental a relation was.

In addition, a measure of the agreement between the experimental and theoretical expressions of the same relation was developed although it was eventually included as part of the Depth measure.

RETROSPECTIVE MEASURES

The retrospective measures developed focused on the utilization of the research results of individual projects (as put forth in actual publications) by subsequent investigators in the same field. The Science Citation Index was employed to find "first-generation receptors" which had used the "outputs" of the "source papers" as "inputs" to their own research. The outputs of the source papers were distinct scientific research results (perhaps 2 to 3 per source paper) as determined by a knowledgeable individual in the field. The analysis was also carried to "second-generation receptors" which used the outputs from the first-generation receptors (but did not directly use the outputs from the source paper). First-generation receptors generally employed 3 to 8 inputs, one or more of which might have come from the source paper. A similar statement holds for second-generation papers using inputs from first-generation papers (or for that matter for any research paper).

Three distinct measures of merit (or indices) were developed for individual publications (source papers).

1. A Utility Index, which measures the number of legitimate users of the source paper.
2. A Fertility Index, which measures the number of "new queries" generated by the source publication, as reflected in the number of distinct "outputs" from the first-generation receptors (weighted against the fraction of "inputs" which came from the source paper).
3. A Diffusion Index, which measures the rapidity with which the results of the source paper spread through the resulting network of first- and second-generation receptors as reflected in the nodes (source paper plus its receptors) minus the number of generations (counted as three when there are two generations of receptors) all divided by the number of content links (input-output links) between nodes.

COMPARISON OF PROSPECTIVE, RETROSPECTIVE, AND PEER EVALUATION

The quantitative measures that were developed were tested by applying them to a sample of ten completed research projects in the field of solid-state physics and comparing the results with peer evaluation by NSF-selected judges of the original proposals and (by a separate group of judges) of the outputs of the completed research projects. The important results of that comparison were as follows:

1. Rank ordering of proposals by the subjective scale (5 grades from poor to excellent) employed by NSF-selected judges had a significant positive correlation with the ranking of the completed papers by the same method.
2. There was no significant correlation, either positive or negative, between the rank orderings by the NSF method and any of the quantitative methods.
3. The only significant correlation between prospective and retrospective rankings by the indicators was a negative correlation between Range-Precision of a proposal and average Diffusion per paper generated.
4. Rankings obtained by all three of the retrospective measures are positively correlated. It was also found that rankings based on a simple counting of citations (first-generation receptors) correlated positively with the more complex retrospective methods, thus implying that a simple citation counting method may give results as good as the more complex methods.
5. Rankings obtained by the prospective methods of Range-Precision and Region of Applicability were positively correlated.

In view of the results described above, the authors were somewhat discouraged about the future of purely quantitative methods. Their pessimism is, of course, based in part on the assumption that the NSF peer-group ratings were an accurate appraisal of the relative quality of the projects which is not necessarily true. It was recognized, however, that the peer-group evaluations could, in part, have reflected evaluations of the scientific stature of the investigators

rather than the quality of the results (or the proposal) and, in fact, it was suggested that some measure of investigator competence be included in future indices.

The authors' final recommendation was to go back to peer-group evaluations as the basis for judging projects, but to include such evaluations in a formal system which includes explicit consideration of the other, more administrative factors (presumably this means cost, etc.) that influence the process of choosing from among alternative programs of basic research.

APPENDIX D

COMPUTER HORIZONS WORK ON CITATION INDEXING USAGE

DESCRIPTION

Computer Horizons (CH), under NSF sponsorship, has been exploring the possibility of generating importance and utilization measures by citation indexing of 250 journals in the Physical Sciences. The approaches being employed are to:

1. Develop one- and two-step models by which each journal is surveyed to determine the first and second other journal which it references the most.
2. Determine how often a given journal cites references in other journals compared to how often it is cited by other journals.

One can then establish a hierarchy of physics journals in which one orders, in a branched tree, the journals according to the comparative magnitude of the percentage of one journal's references to another with the reverse percentage. CH has observed that in almost every field there is one unique hierarchy in which all journals can be placed with a minimum of conflict.

The sum total of these kinds of measures appears to provide (in CH's opinion) a gross breakdown of journals according to their importance. In each field there appears to be one "Super" journal (Physical Review in physics), three or four "Very Important" journals, a larger group of "Important" journals, and then the rest. These judgments are fraught with problems insofar as their use is concerned as will be discussed below.

Another measure of interest is the Dispersion measure, i.e., the number of journals necessary to encompass 50% of the references for

the given journal. For example, Dispersion for Physical Review is 6, Astrophysics is 5, Journal of Applied Physics is 21, Journal of Geophysical Research is 26. The significance of this measure is that it reflects the existence of a well-ordered body of knowledge in a given field or perhaps of an establishment in the field. Generally, high quality journals are concentrated, lesser journals are more dispersed. But interdisciplinary journals cannot be measured against this criterion.

These approaches are also being used to study university appearances in the literature and agency support. A study of 20 universities produced the surprising result that despite numerical differences in publications, if point quality ratings are assigned to journals, the resulting average points/article varies little between schools.

POSSIBLE VALUE AND LIMITATIONS OF THIS WORK

Through citation indexing work of this nature, one can determine the information flow pattern from field to field. For instance, CH has shown that the special education literature is dependent on the psychology literature so that if one were interested in the long-range support of special education, one would be in trouble if psychology funding dried up. In the context of mission-oriented research, if an agency could determine certain "obviously relevant fields" and then demonstrate the dependence of these fields on more basic science, they could very possibly serve to justify the funding of the more basic science.

There is the hope of using such citation techniques to evaluate the quality of research output. But one must be careful in the application of such an approach. Perhaps at the level of a laboratory or a department this may have validity, but variations associated with why an individual researcher publishes in a given journal would be comparable to the statistical noise. At the very least, sub-disciplines must be analyzed.

The determination of quality ratings and weightings is somewhat subjective. At present, there are normalization problems which are

not adequately considered in CH's work which invariably cause small journals to come out as being no better than "Important" and usually merely "other." Also, there should be some method to determine whether the rating is deemed reasonable by workers in the field.