AD-753 815

THE EVALUATION OF RESEARCH SUPPORTED BY A MISSION-ORIENTED AGENCY-STRUCTURED APPROACHES. VOLUME II: A PRELIMINARY LOOK AT QUANTITATIVE METHODS FOR EVALUATING RESEARCH

Roger C. Molander

Institute for Defense Analyses

Prepared for:

Defense Advanced Research Projects Agency

December 1971

PAPER P-753

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

in two volumes

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

Roger C. Molander

December 1971

IDA
INSTITUTE FOR DEFENSE ANALYSES
SCIENCE AND TECHNOLOGY DIVISION

IDA Log No. HQ 71-12640
Copy 75 of 95 copies

41

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 II: A Preliminary Look at Quantitative Methods for Evaluating Research

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

6. REPORT DATE December 1971	7a. TOTAL NO OF PAGES 33	7b. NO OF REFS 1
8a. CONTRACT OR GRANT NO DAHC15 67 C 0011 b. PROJECT NO DARPA Assignment 20	9a. ORIGINATOR'S REPORT NUMBER(S) P-753 9b. 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

Because of ~~dissatisfaction with current~~ methods of evaluating research, ~~in particular,~~ peer evaluation, an attempt is made to find quantitative methods that might provide a straightforward means of comparing research projects and research laboratories. The effort was partly unsuccessful in that no quantitative measures were found that would reflect the principal objectives of the agency of interest. These objectives relate to ~~a desire~~ for knowledge applicable ~~to~~ current or projected DoD problem areas and education of the defense R&D community, with respect to developments in the basic research community. However, quantitative measures were found that offered the potential for measuring scientific excellence. Without an independent measure of scientific excellence, however, it was impossible to determine whether these measures provided a valid tool for measuring project or laboratory quality.

DD FORM 1 NOV 65 1473

UNCLASSIFIED
Security Classification

UNCLASSIFIED

Security Classification

14 KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
Research evaluation						
Basic research evaluation						
Quantitative measures of research effectiveness						
Structured methods						

UNCLASSIFIED

Security Classification

PAPER P-753

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

in two volumes

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

Roger C. Molander

December 1971

IDA

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

Contract DAHC15 67 C 0011
DARPA Assignment 20

1c

ACKNOWLEDGMENT

The author wishes to acknowledge the helpful comments of Drs. Philip Selwyn and A. Fenner Milton in the preparation of this Volume.

ABSTRACT

Because of dissatisfaction with current methods of evaluating research, in particular, peer evaluation, an attempt is made to find quantitative methods that might provide a straightforward means of comparing research projects and research laboratories. The effort was partly unsuccessful in that no quantitative measures were found that would reflect the principal objectives of the agency of interest. These objectives relate to a desire for knowledge applicable to current or projected DoD problem areas and education of the defense R&D community with respect to developments in the basic research community. However, quantitative measures were found that offered the potential for measuring scientific excellence. Without an independent measure of scientific excellence, however, it was impossible to determine whether these measures provided a valid tool for measuring project or laboratory quality.

CONTENTS

I. INTRODUCTION

Organizations that support large-scale research programs have traditionally relied upon the judgment of knowledgeable individuals to determine whether a given research effort has the potential for success or, in retrospect, whether the effort has been successful. This evaluation process, generally characterized by the term "peer evaluation," involves the subjective judgment of human beings, and is subject to personal prejudices, lack of total understanding, and other comparable human failings--as well as being tremendously costly and time-consuming. In recognition of these shortcomings, interest has arisen in the utility of much more structured approaches to research evaluation--approaches that would eliminate the human judgment factor as much as possible. Attention has thus been directed to the potential utility for research evaluation of those characteristics (e.g., papers published, patents, budget, etc.) of a research program that can be quantified in a purely objective fashion. This effort has been characterized by the term "quantitative methods for evaluating research." As implied by the title, this Volume represents a preliminary look at the feasibility of such methods, in particular, for evaluating that research supported by ARPA, with due consideration for ARPA's role as part of a mission-oriented agency and the time urgency of the evaluations it must make.

The term "research evaluation" carries with it the implication of evaluation against some perceived set of objectives. However, a precise definition of objectives is not always readily available, as was the case at the inception of this study. Furthermore, as objectives were defined, it became clear that the objectives prevalent within ARPA were not completely consistent with those existing in parts of the research community which received ARPA support. This lack of uniformity of objectives gives rise to significant problems in performing evaluations. The evaluation of a given piece of work would probably be different if one were interested in advancing the state of knowledge in a particular area of physics versus being interested in the development of military technology. This dilemma is common to efforts to evaluate research supported by a mission-oriented agency, particularly when that research is performed at a university.

In pursuing this study, the search for meaningful quantitative measures was not restricted to those that would reflect the particular objectives emphasized by ARPA, although it was recognized that such measures would be of principal interest. In fact, as will be discussed later, the effort to find quantitative measures that would reflect in a meaningful manner the principal objectives put forth by ARPA was singularly unsuccessful. Instead, it has been found that the only quantitative measures with potential utility are those that may reflect scientific excellence (it should be pointed out that this objective is generally given principal emphasis by the university research community).

There is interest in evaluating research in both a prospective and a retrospective sense. Clearly the latter is easier since one has available the results of the research effort. This analysis will focus on the retrospective evaluation, in part because of the logic of attacking easier problems first and in part because it is felt that only if quantitative methods can be shown to be feasible for retrospective studies is there any opportunity of their being used for prospective evaluations. In addition, prospective quantitative methods may never be shown to be feasible--in which case past performance as measured by retrospective studies would weigh heavily in making funding decisions.

II. EVALUATION BY PROJECT

A. MEASURES OF PROJECT QUALITY

We seek a measure of the relative quality of individual research projects and hypothesize that there exists an appropriate combination of those parameters available for individual projects which provides such a measure. The logical procedure for testing this hypothesis would be to formulate "candidate" combinations of the available parameters and then to employ established principles of statistical inference to determine how well these candidate combinations correlate with an independent method of determining project "quality" whose validity is well established. The principal difficulty encountered in this procedure is the absence of a universally accepted, independent method of measuring project quality against which the candidate measures can be correlated. The source of this difficulty is in part the lack of agreement on what characterizes a high-quality project, where this lack of agreement derives principally from different perceptions of project objectives. This problem is particularly acute when the traditional method of judging project quality, peer evaluation, is employed. There is simply no guarantee that the peer group embraces the same set of objectives as the supporting agency. Still further, even if objectives and, in turn, characteristics of "quality" could be agreed upon, not all objectives, e.g., "relevance to military technology," can be interpreted in a sufficiently structured fashion to be of real utility in objectively evaluating hypotheses. This is discussed in more detail in the material that follows.

An examination of research program objectives revealed that ARPA's principal interest in supporting basic research is the "development of new scientific knowledge applicable to current or projected DoD problem areas." However, in examining the quantitative data available for individual research projects, it is difficult, if not impossible, to find data that would reflect this particular objective in any meaningful manner. Project output is almost exclusively publications in scientific journals that have no identifiable connection with "DoD problem areas." One might conceive of employing the Science Citation Index and looking for instances where the project publications are cited in journals with a defense orientation. However, even the citations of these papers (as found in the Science Citation Index) occur almost exclusively in similar non-DoD-oriented publications. This is not surprising in view of the nature of the work. As previous studies such as TRACES (Ref. 1) have shown, there is generally a long time lag before basic research developments are utilized in technological applications. Thus this study has been unable to find any purely quantitative measures that would be appropriate for judging whether a particular research project has the potential for developing new scientific knowledge with potential DoD application.

A second major ARPA objective is the "education of the defense R&D community" with respect to developments in the basic research community. Again one is bankrupt for meaningful quantitative measures that would reflect this objective. One could not simply count number of days of consulting for DoD laboratories and contractors and assume that such a number reflects the transfer of information from Category 6.1 research areas to Category 6.2 areas and above. Even though such contact is a potential vehicle for such information transfer, it is not unique (simply publishing also makes developments available) and provides no assurance that information transfer does take place.

On the other hand, publications in the open literature do offer promise of reflecting scientific excellence, an objective which is given great emphasis in the university research community but is not explicitly emphasized by ARPA. For example, publications in most scientific journals have received critical reviews from peer groups whose principal criterion is scientific excellence. Hence one might hypothesize that the number of such publications, weighed against project manpower or budget, is a candidate measure of project quality, but only from the standpoint of scientific excellence. The difficulties of confirming or denying such hypotheses are discussed in detail in the material that follows.

In summary, the inability to find meaningful quantitative measures for ARPA's principal objectives restricts this analysis to the utility of quantitative measures in evaluating scientific excellence.

B. UTILIZATION OF INFORMATION CONTAINED IN ANNUAL TECHNICAL REPORTS

In an attempt to determine the level of analysis that might be performed at the project level, a sample laboratory (the University of Illinois Materials Research Laboratory--U of I MRL) was first examined using information available from the annual technical reports (ATRs). It was found that these ATRs contained the following relevant information at the project level:

1. Number of participating faculty
2. Number of post doctoral fellows
3. Number of graduate students (broken down into those on fellowship and those given salary support by the project)
4. Number of M.S. and Ph.D. degrees awarded
5. A listing of publications including books.

The following relevant information was not available either in the ATRs or in the administrative reports that are also prepared annually:

1. Number of Full-Time Equivalent (FTE) faculty (neither by project nor total for the laboratory).
2. Project funding. The ATRs contained no cost information whatever, much less a breakdown by project of funds allocated to personnel, equipment procurement, etc.

As might be expected the ATRs made no attempt to specifically identify those efforts that were principally theoretical as opposed to those that were principally experimental. Arguments for the advisability of making this distinction rest principally on the fact that theoretical efforts, because of their lower operating costs, are likely to appear much more productive on a per dollar basis. Isolation of students and post doctoral fellows on fellowship is also desirable since they are available at zero personnel cost to the project in contrast to those individuals who receive salary support from the project. The interest in personnel and equipment procurement cost are also obvious because in the absence of such a cost breakdown one sacrifices any opportunity of fairly judging return on funds allocated to equipment. There is also an interest in identifying thesis-related publications which stems from the view that one measure of the merit of a thesis is the fact that it led to a publication in a good journal.

It was concluded that a comparison of individual projects using the information contained in the ATRs could do little more than isolate some of those projects at the extremes of the spectrum, i.e., those which were very productive and those which were very unproductive based on the papers produced and the project manpower. Even this judgment could only be viewed as a temporary one, subject to confirmation by a more detailed scrutiny of the isolated projects using information from some other source.

C. UTILIZATION OF DETAILED INFORMATION OBTAINED FROM A LABORATORY

In view of the shortcomings of the analysis (based purely on information contained in annual technical reports), more detailed information on individual projects was obtained from the University of Illinois MRL. Table 1 is a sample of the information form which was completed for each project. The instruction sheet for completing the forms is found in the Appendix.

As noted in a previous section, this analysis, by default, is focusing on the possible use of quantitative methods for evaluating the scientific excellence of individual research projects. Because no independent method (such as peer evaluation) has been employed to

TABLE 1. PROJECT INFORMATION FORM

Project No. ____________________ Project Supervisor ____________________

Project Title ____

____

	Academic Year			
	65-66	66-67	67-68	68-69
I. Faculty				
A. Number of Full-Time Equivalent Man-Years				
B. Total Number of Participating Faculty				
II. Post Doctoral Fellows				
A. Total Number				
B. Number Receiving Salary Support				
III. Graduate Students				
A. Total Number				
B. Number Receiving Salary Support				
IV. Other Students Receiving Salary Support from ARPA Contract				
V. Theses Completed				
A. MS				
1. Experimental				
2. Theoretical				
B. Ph.D.				
1. Experimental				
2. Theoretical				
VI. Thesis-Related Publications in Reviewed Journals				
1. Experimental				
2. Theoretical				
VII. Other Publications in Reviewed Journals				
1. Experimental				
2. Theoretical				
VIII. Other Publications				
1. Articles in Unreviewed Publications				
2. Books				
3. Articles in Books				
IX. Total Personnel Salary Support ($1000's)				
X. Equipment Procurement ($1000's)				
XI. Total Project Funding ($1000's)				

judge the scientific excellence of the MRL projects, there will be no way to confirm or deny the validity of the measures to be proposed. However, it will be possible to judge their sensitivity to arbitrary weightings, and to determine the correlation between different proposed measures.

1. Productivity as a Function of Manpower Input

Examination of the detailed information forms revealed one source of potential difficulty: the unexpectedly low values for faculty FTEs. As shown in Fig. 1, the average number of faculty FTEs per project per year is about 0.22. This is surprisingly low since the average faculty member probably devotes about half his time to research. The discrepancy results principally from the fact that not all of the time a faculty member performs on research is charged to the research contract.* This results in part from university policy with respect to source of funds for faculty salaries. The magnitude of the discrepancy between the actual contribution and the amount charged to the research contract varies from school to school. For example, at Harvard all of a tenured faculty member's academic year salary is paid by the university. Such a practice is consistent with an effort to ensure that support of faculty members is not heavily dependent on research contracts. This situation obviously creates considerable difficulty in properly weighting faculty contribution to research projects. Should one (a) use FTEs charged to the project, (b) assume all faculty members spend roughly the same fraction of their time on their research projects, or (c) ask faculty members to estimate their actual contribution in FTEs? The latter would probably give the best estimate although its use on a regular basis would create a tendency to bias estimates in the direction most favorable to the individual performing the estimate. In the analysis that follows, calculations will be made for two separate cases; one using the given figures for FTEs

*Although a faculty member might have other research contracts, this was generally not of significance for the research projects that were examined.

charged to the project and one using an assumed average contribution (half-time) for all faculty members. The half-time estimate is based on an assumption of full-time participation on a research project during the summer and about one-third time participation during the school year.

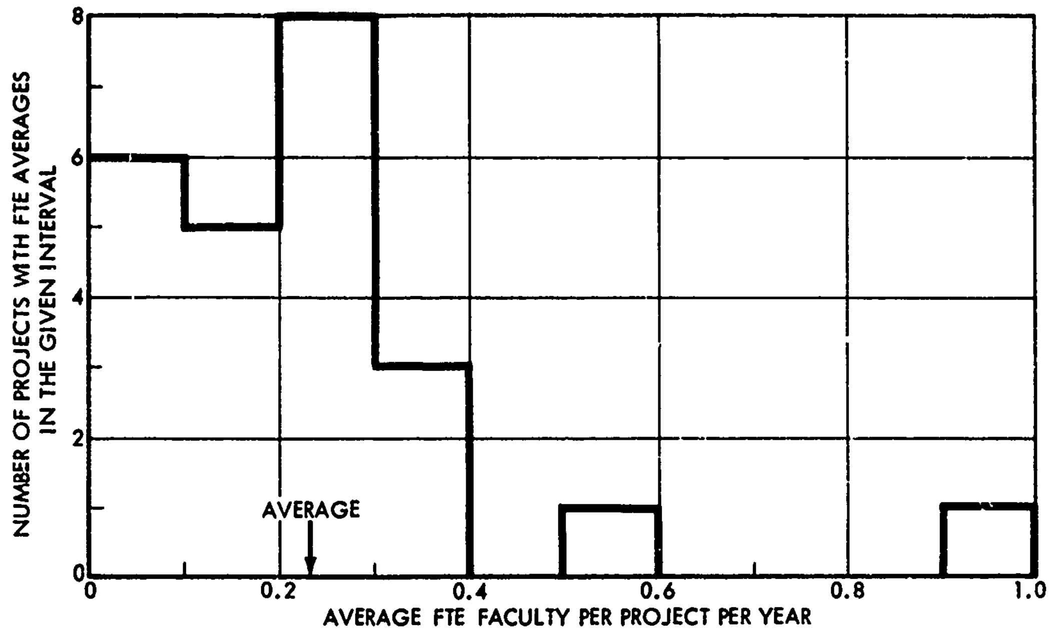

FIGURE 1. Faculty FTEs per Project (Four-Year Averages)

In order to properly weight manpower input, it is necessary to make judgments on the relative productivity of faculty, post doctoral students and graduate students. One index of the total effective manpower input to a project from n different sources could be given by the quantity

$$M \equiv \sum_{i=1}^{n} \beta_i F_i \tag{1}$$

where the weighting factors, β_i, reflect the relative productivity of type i manpower and F_i is the number of FTE man-years of type i manpower. The relative productivity of faculty, post doctoral fellows, and graduate students is difficult to determine but is probably roughly in the ratio 4:3:1. Explicit FTE information was not obtained for post doctoral students and graduate students since, for university research projects, it is reasonable to expect that post doctoral fellows spend almost all of their time on the project while graduate students spend an average of about 75 percent of their time (ranging from full-time for doctoral students working on their theses to about one-third to one-half time for first- and second-year graduate students). Thus, the F_i's for post doctoral and graduate students become

$$F_{\text{Post Doc.}} = N_{\text{Post Doc.}}$$

$$F_{\text{Grad. Stud.}} = \frac{3}{4} N_{\text{Grad. Stud.}}$$

where N_i simply represents the total number of participants of type i. Thus for the case where the given FTE figures will be used, Eq. 1 becomes

$$M_{\text{FTE}} = 4\ F_{\text{Faculty}} + 3\ N_{\text{Post Doc.}} + \frac{3}{4} N_{\text{Grad. Stud.}}$$

while for the case where all faculty are assumed to spend half-time on the project, Eq. 1 becomes

$$M_{\frac{1}{2}} = 2\ N_{\text{Faculty}} + 3\ N_{\text{Post Doc.}} + \frac{3}{4} N_{\text{Grad. Stud.}}$$

To compare manpower input with output it is presumably necessary to weight various types of publications. For example, an article in a reviewed journal should receive more credit than a publication in a conference proceedings. Similarly, an article in a publication such as Physical Review is presumably worth more credit than an article in

Physical Review Letters. (The merit of articles might also be judged using the Citation Index. As discussed later in the analysis, there is at least a two-year lag before such a measure is meaningful.) However, rather than attempt to weigh the merits of individual journals, proceedings, etc. (an effort which would generate considerable controversy with dubious payoff) for the purposes of measuring publication output against manpower input, publications will simply be grouped into articles in reviewed journals,* articles in unreviewed publications, books, and articles in books. It is recognized that weighting of these categories is extremely difficult. In this analysis these four categories will be weighted in the ratios

$$N_{r.j.} : N_{u.p.} : N_{b.} : N_{a.b.} = 2:1:5:2$$

where $N_{r.j.}$ = number of articles in reviewed journals, etc. A measure of the total publication output from a given project is given by

$$P_1 = 2\,N_{r.j.} + N_{u.p.} + 5\,N_{b.} + 2\,N_{a.b.}$$

Assuming productivity is reflected in the ratio

$$\alpha = \frac{\text{Publication output}}{\text{Manpower input}} = \frac{P_1}{M}$$

Table 2 presents average values of $\alpha_{FTE} = P_1/M_{FTE}$ and $\alpha_{\frac{1}{2}} = P_1/M_{\frac{1}{2}}$ (and the appropriate rank orderings) based on the four-year period from 65-66 to 68-69. Only those projects with annual budgets in excess of $15,000 were considered. (Table 2 also shows productivity as a function of total funding which is discussed in the next section.)

---

*It is recognized that counting numbers of publications also ignores the fact that some research efforts lend themselves to many short publications while others are constrained to fewer, more detailed publications.

TABLE 2. PRODUCTIVITY COMPARISONS

Project Identification Number	$\alpha_{FTE} = \frac{P_1}{M_{FTE}}$		$\alpha_{\frac{1}{2}} = \frac{P_1}{M_{\frac{1}{2}}}$		$\alpha_{\$} = \frac{P_1}{\$TP}$	
	α_{FTE}	Rank	$\alpha_{\frac{1}{2}}$	Rank	$\alpha_{\$}$	Rank
18	3.65	1	2.76	1	0.295	2
15	3.02	2	1.28	2	0.195	5
13	1.81	3	1.27	3	0.269	4
4	1.67	4	1.20	4	0.352	1
3	1.50	5	0.84	13	0.285	3
10	1.33	6	0.57	16	0.118	11
6	1.22	7	0.99	8	0.085	15
19	1.17	8	1.06	6	0.103	12
33	1.17	9	1.00	7	0.175	6
16	1.16	10	0.95	9	0.138	7
5	1.16	11	1.19	5	0.125	10
1	1.13	12	0.84	12	0.075	16
29	1.06	13	0.94	10	0.135	8
14	1.03	14	0.87	11	0.130	9
34	0.99	15	0.53	18	0.068	17
32	0.93	16	0.71	15	0.065	18
20	0.64	17	0.50	19	0.043	20
8	0.62	18	0.40	20	0.089	14
24	0.57	19	0.54	17	0.094	13
30	0.53	20	0.33	21	0.027	21
26	0.38	21	0.75	14	0.048	19
25	0.31	22	0.23	22	0.021	22
31	0.00	23	0.00	23	0.000	23

Correlation Coefficients (rank ordering):

Quantities Being Compared	Correlation Coefficient,* r_1
$\alpha_{FTE}:\alpha_{\frac{1}{2}}$	0.83
$\alpha_{FTE}:\alpha_{\$}$	0.84
$\alpha_{\frac{1}{2}}:\alpha_{\$}$	0.79

*For 23 samples, $r_1 = 0.496$ is significant at 0.01 level.

As can readily be seen from Table 2, despite some individual differences, there is close overall correlation (Spearman rank-ordering correlation coefficient ≡ $r_1 = 0.83$) between the values of α_{FTE} and $\alpha_{\frac{1}{2}}$. Since α_{FTE} and $\alpha_{\frac{1}{2}}$ differ only in the calculation of a faculty contribution to manpower input, one might conclude that the inability to accurately fix faculty FTE contribution may not be as serious a shortcoming as had been feared. Nevertheless, the arguments still hold regarding the weakness of faculty contribution based on FTEs charged to the project and constant contribution (half-time) for all participating faculty. The best available method would be to solicit from each faculty member an estimate of his actual FTE contribution; this could then be compared with his known teaching load, consulting, and other activities to eliminate "cheating" as much as possible.

To determine the sensitivity of the results to the assumed faculty:post doctoral student:graduate student productivity ratios, α_{FTE} and $\alpha_{\frac{1}{2}}$ were recalculated assuming a 6:3:1 ratio rather than 4:3:1. There was very close correlation between the α's obtained using the two different ratios (e.g., correlation coefficient $r_1 = 0.97$ for values of α_{FTE} calculated from the two ratios). Thus the correlation between α_{FTE} and $\alpha_{\frac{1}{2}}$ was again very close ($r_1 = 0.81$). One can conclude that calculations of relative productivity are relatively insensitive (within reason, of course) to choice of productivity ratios between faculty, post doctoral students, and graduate students.

The raw data show that the dominant output (in terms of publications) is papers published in reviewed journals. As a consequence, the results (i.e., rank orderings of Table 2) are almost completely insensitive to the relative weighting of articles in reviewed journals, books, etc.

2. Productivity as a Function of Funding Level

Productivity measured on the basis of manpower input, as was done in the previous section, sacrifices any opportunity of judging how the dollars allocated to a project are actually spent. This section will attempt to compare productivity of the various projects where productivity will be determined by comparing output and funding level for each project.

As noted previously, the major difficulties in using funding level as the project input is taking proper account of experimental versus theoretical efforts. Figure 2 shows the distribution of the ratio of Personnel Salary Support to Total Project Funding (Table 1) for the projects being considered. The projects are broken into those that are principally theoretical and those that are principally experimental on the basis of the project supervisor's classification of the project publications into those categories (Table 1). In the projects considered the choice was easily made. The difference in experimental and theoretical projects is clearly shown in Fig. 2. It would appear advisable to consider experimental and theoretical projects separately when comparing projects. There are only two theoretical projects for the laboratory under consideration, and this analysis will focus on those projects that are principally experimental.

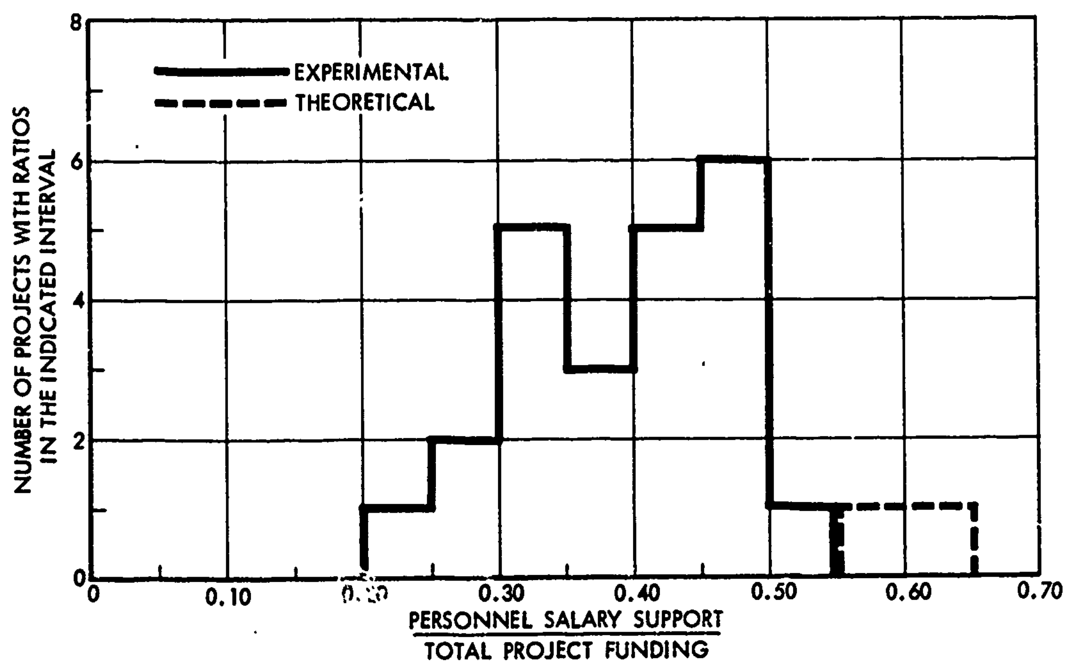

FIGURE 2. Distribution of Salary Support/Project Funding Ratios

One method of defining productivity on a per dollar basis is obtained by dividing output measured in terms of publications (P_1, as previously defined) by total project funding (TPF). The results, $\alpha_\$ = P_1/TPF$, were shown in Table 2. The correlation between productivity measured on a manpower input basis and productivity measured on a per dollar basis is quite good (correlation coefficient $r_1 = 0.83$ for $\alpha_{FTE}:\alpha_\$$ and 0.79 for $\alpha_{\frac{1}{2}}:\alpha_\$$).

One might also consider the production of skilled researchers as a valid output of a project. Such output is, in part, reflected in counting publications for those instances where a Ph.D. or M.S. thesis results in a publication. As noted previously, whether a thesis gets published is a measure of the quality of the work. Thus one could compute output on the basis of adding publications and degrees granted (with proper weighting) with the understanding that the overlap rewards those theses that produced publications. Assuming a Ph.D. is worth 2 M.S.'s, the output in terms of trained researchers can be written

$$P_2 = 2\,N_{Ph.D.} + N_{M.S.}$$

where $N_{Ph.D.}$ and $N_{M.S.}$ are the number of degrees granted (see Table 1). It is extremely difficult to weight trained researchers versus publications. The judgment is certain to be based on a perception of objectives. Two weightings will be used in this analysis. A weighting of $P_2:P_1$ in the ratio 1:1 assumes that the value of each Ph.D. produced is equal to that of each paper in a reviewed journal and thus is weighted toward publications as the principal program objective. A $P_2:P_1$ ratio of 3:1 is also used, which favors the production of trained researchers. Productivity is thus defined by

$$\alpha_{1:1} = \frac{P_1 + P_2}{TPF}\ ;\ \alpha_{3:1} = \frac{P_1 + 3\,P_2}{TPF}$$

The results are shown in Table 3 along with the results for $\alpha_\$$. The correlation between $\alpha_\$$ and $\alpha_{1:1}$ is very good ($r_1 = 0.93$) and still good between $\alpha_\$$ and $\alpha_{3:1}$ ($r_1 = 0.82$). Table 3 shows there are strong individual differences for only a few projects (e.g., projects 10 and 8).

TABLE 3. PRODUCTIVITY AS A FUNCTION OF FUNDING LEVEL

Project Identification Number	$\alpha_\$ = \frac{P_1}{\$TP}$ $\alpha_\$$	Rank	$\alpha_{1:1} = \frac{P_1+P_2}{\$TP}$ $\alpha_{1:1}$	Rank	$\alpha_{3:1} = \frac{P_1+3P_2}{\$TP}$ $\alpha_{3:1}$	Rank
4	0.352	1	0.418	1	0.549	4
18	0.295	2	0.416	2	0.659	1
3*	0.285	3	0.302	5	0.336	6
13	0.269	4	0.319	3	0.420	5
15	0.195	5	0.317	4	0.561	3
33	0.175	6	0.227	7	0.330	7
16	0.138	7	0.181	9	0.266	9
29	0.135	8	0.182	8	0.278	8
14	0.130	9	0.166	11	0.237	11
5	0.125	10	0.169	10	0.256	10
10	0.118	11	0.118	13	0.118	17
19	0.103	12	0.145	12	0.230	12
24*	0.094	13	0.094	17	0.094	18
8	0.089	14	0.267	6	0.622	2
6	0.085	15	0.108	14	0.154	16
1	0.075	16	0.081	19	0.094	19
34	0.068	17	0.102	15	0.169	15
32	0.065	18	0.101	16	0.173	13
26	0.048	19	0.048	20	0.048	20
20	0.043	20	0.086	18	0.172	14
29	0.027	21	0.027	21	0.027	23
25	0.021	22	0.027	22	0.037	21
31	0.000	23	0.011	23	0.032	22

Correlation Coefficients (rank ordering):

Quantities Being Compared	Correlation Coefficient, r_1
$\alpha_\$: \alpha_{1:1}$	0.93
$\alpha_\$: \alpha_{3:1}$	0.82

*Theoretical Project.

3. Utilization of the Science Citation Index

The Science Citation Index, an annual listing of the citations of journal articles, was considered as a potential aid in judging the merit of individual publications. To explore this question, a sampling of publications of the University of Illinois Materials Research Laboratory has been checked in the Citation Index. The sampling includes those publications supported wholly or in part by ARPA for 20 different projects (as listed in the 1965-66, 1966-67, 1967-68 and 1968-69 ATRs). Only articles in reviewed journals have been included. In addition, in those instances where a paper is cited by one of the original authors or (where identifiable) another member of the same research project, the citation has been excluded. Table 4 summarizes the data obtained from the 1965-1969 volumes of the Science Citation Index.

Assuming the number of citations is a valid measure of the quality of a publication, one is led to attempt to weight individual publications consistent with such a measure. The relative weightings, W, shown in Table 5 have been assumed in this analysis. (There is obviously some arbitrariness in the choice of weighting factors and no attempt will be made to defend the precise values chosen.) Since there has not been sufficient time for a significant number of citations to appear for those publications listed in the 1968-69 ATRs, two different weightings will be applied to those publications. The first assumes that all publications receive a weighting of three (which is the average weighting for the previous years). The second assumes that performance is approximately constant and assigns a weighting to 1968-69 publications consistent with the average weighting obtained by other publications from that project in the previous three years.

Since the average weighting is almost exactly three, the equation used previously for output can be rewritten as

$$P_1' = \frac{2}{3} W_{r.j.} + N_{u.p.} + 5\ N_{b.} + 2\ N_{a.b.}$$

where $W_{r.j.}$ represents the sum of weightings over all the project's publications.

TABLE 4. NUMBER OF CITATIONS FOR INDIVIDUAL PUBLICATIONS

Project Identification Number	1965-66*	1966-67	1967-68	1968-69
1	17	-	2,6,3	(2)**
3	0,1,17,0,3 11,62,24	6,1,3,48,11,11 9,158,25,14	0,5,7,1 0,0,4	(4)
4	7	12,5	7,2	(9)
5	0,1	1	2,2	(5)
6	12,29	1,8	5	(2)
13	18,18,0,13	10	-	(11)
14	1,0	5,2,3,6	-	(1)
15	23,1	0	1,0	(1)
16	0,0,48	8,17,38,2	3,10,16	(2)
18	4,1,10,6, 2,11,13,11	1,5,4,0	4,1,1,2,1	(7)
19	24,13,8,29	1,15	4,6,3	(5)
20	26	17	5	(2)
24	-	9	3,7,1	(6)
25	-	-	-	(2)
26	-	-	2,5	(1)
29	2,23,7,4	0,20,6,0 2,2,2	0,0,2,0	(8)
30	-	-	0	(1)
31	-	-	-	(0)
32	29,9,7	1	0	(4)
33	-	7	3,0,1,6, 12,0,0	(8)

*Year of ATR.

**Number in parentheses is number of publications.

NOTE: Numbers given represent number of citations found in the 1966-69 volumes of the Science Citation Index for publications listed in the ATRs indicated. Citations by the author(s) or other members of the research project are excluded.

TABLE 5. WEIGHTINGS OF INDIVIDUAL PUBLICATIONS

Weighting, W	Range of Citations which Receive the Given Weighting		
	1965-66	1966-67	1967-68
1	0	0	0
2	1-4	1-3	1-2
3	5-10	4-8	3-5
4	11-20	9-15	6-10
5	20-∞	15-∞	10-∞

Values of $\alpha'_{FTE} = P'_1/M_{FTE}$ have been calculated for the two different treatments of 1968-69 publications. In Table 6 the resulting rank ordering of projects is compared with previous results for $\alpha_{FTE} = P_1/M_{FTE}$ where no attempt was made to weight individual publications. Clearly there is very close correlation between the results ($r_1 = 0.97$ for α_{FTE} and α'_{FTE} when W = 3 and $r_1 = 0.93$ when W = the project average for the previous three years). Since 20 projects and 223 individual articles were included in the calculation, considerable reliability can be attributed to the results. It thus appears that within the restrictions of the weighting method (i.e., the results could not be expected to correlate as well if a broader range of weightings were used) applying citation index weightings does not produce a marked difference in the quantitative evaluation of projects as compared to simply counting number of publications.

There are some weaknesses in the use of the Citation Index. Most importantly, checking the submission dates for articles shows that about two years separate submission and the appearance of significant citations. Thus the Citation Index would only appear to be a potential tool for judging performance up to about two years prior to the evaluation. However, if performance is assumed to be relatively constant (at least on a two-year time scale), the effect of this time lag might not be nearly as important as it appears.

TABLE 6. PRODUCTIVITY RANK ORDERINGS WITH CITATION INDEX WEIGHTINGS

Project Identification Number	α_{FTE}	$\alpha_{FTE}\|_{W=3}$	$\alpha_{FTE}\|_{W=3\text{-year average}}$
18	1	1	1
15	2	2	2
13	3	3	3
4	4	4	4
3	5	5	5
6	6	7	7
19	7	6	6
33	8	10	11
16	9	9	9
5	10	11	15
1	11	8	8
29	12	15	14
14	13	14	13
32	14	12	12
20	15	13	10
24	16	16	16
30	17	17	19
26	18	18	17
25	19	19	18
31	20	20	20

Correlation Coefficients (rank ordering):

Quantities Being Compared	Correlation Coefficient,* r_1
$\alpha_{FTE}:\alpha_{FTE\|W=3}$	0.97
$\alpha_{FTE}:\alpha_{FTE\|W=3\text{-yr average}}$	0.93
$\alpha_{FTE\|W=3}:\alpha_{FTE\|W=3\text{-yr average}}$	0.97

*For 20 samples, $r_1 = 0.534$ is significant at the 0.01 level.

Another weakness of the Citation Index is the incestuousness of some scientific endeavors and the tendency to show favoritism to the work of laboratory co-workers. It would be extremely difficult to filter out such effects in any systematic fashion.

There is also a problem with publications that have several sponsors where one wishes to isolate the work attributable to a single sponsor--as was desired in this analysis. In some instances, project supervisors indicate that all project publications are co-sponsored by all project sponsors while in other cases an attempt is made to distinguish between project sponsors for individual publications. Projects that follow the first practice will appear much more productive under an analysis, such as that pursued in this paper, where ARPA sponsorship alone is considered.

III. EVALUATION BY INSTITUTE

A limited amount of quantitative information has been compiled for the Inter-Disciplinary Laboratories that are funded by the ARPA Materials Sciences Office. Annual figures (for about a ten-year period up to the present) are available for each of the laboratories for the following parameters:

1. Number of Ph.D. degrees
2. Number of papers published
3. Number of graduate students
4. Number of faculty (number of full-time equivalent faculty in some cases)
5. Level and source of support: University, ARPA, non-ARPA DoD, and non-DoD government (AEC, NASA, etc.).

Because the number of Ph.D.'s, number of papers, etc., is not broken down into those supported by ARPA, those supported by University funds, etc., the discussion will center on the overall institute performance rather than attempt to isolate the performance on projects supported wholly or in part by ARPA. The discussion that follows is an attempt to determine what level of analysis can be performed with information at this level of sophistication. A summary of the information available for the last four years is given in Table 7. Part of the annual support at eight of the institutes is building use charges. In comparing institutes, the building use charges will be subtracted from the total support figures.

A performance analysis based on information of the type given in Table 7 suffers from several particular weaknesses. These include:

TABLE 7. SUMMARY OF LABORATORY DATA: FOUR-YEAR TOTALS FOR THE YEARS 65-66 TO 68-69

	Ph.D.'s	Papers	Graduate Students	Faculty	Total Support, $\$10^5$	% ARPA	Building Use Charges, $\$10^5$
MIT	278	1572	1697	387*	300.6	26	12.0
Cornell	151	876	790	237	212.8	47	16.0
Illinois	98	618	564	223	210.5	32	12.3
Chicago	87	714	634	147	189.5	19	--
Pennsylvania	115	527	626	173	188.1	44	9.6
Stanford	177	704	1070	221	183.8	30	4.2
Brown	87	889	827	318	143.2	39	5.3
Northwestern	113	702	625	165	113.9	48	7.6
Purdue	109	438	600	222	88.5	32	--
Maryland	62	368	434*	157*	69.3	33	--
Harvard	44	242	259	82	52.4	56	2.9
North Carolina	62	393	430	113	47.4	43	--
Overall Totals	1383	8043	8556	2445	1800		72.1

*Originally given in FTEs.

1. The listing of number of papers published gives no indication whether this is simply a count of papers published in reviewed journals or includes proceedings and other unreviewed publications. Unless the criterion for inclusion was the same at all institutes (doubtful), comparisons based on these numbers will be inconclusive.

2. The number of Ph.D.'s granted is not broken down into those which are principally theoretical as compared to those which are principally experimental. It appears that one Ph.D. is produced for every $100,000 to $200,000 of annual laboratory support, thus the importance of counting the production of theoretical Ph.D.'s (probably each requires only about $10,000 of annual support) is obvious. Otherwise an effort that is strongly theoretical will appear much more productive than an effort that is primarily experimental. A similar argument holds for judgments based on the number of papers published where it is unclear what fraction of the papers published results from theoretical studies.

3. The number of graduate students listed gives no indication of whether it is the total number of students who have worked on laboratory projects during the preceding year, the average number who were working at any one time, or the number of full-time equivalent (FTE) graduate students. Neither is it indicated if those students on fellowship are included or only those receiving salary support. It will be assumed that the number given is the total number of students who participated regardless of their source of support.

4. There is no mention of post doctoral students who are generally full-time researchers and very productive compared to a graduate student who has not yet completed his doctoral research.

5. Except in two instances there was no indication if the number of faculty listed was the total number of participating faculty

or the number of FTE faculty. For several laboratories, the number appeared to be consistent with total number of participating faculty as obtained from annual reports. Therefore it was assumed that, unless otherwise stated, the number given is total number of participating faculty. Where the number of FTE faculty was given it was converted to approximate total number of participating faculty using information from annual reports.

There did not appear to be marked annual variations in the proportionality between parameters for individual laboratories except in number of Ph.D.'s. This is presumably a result of the relatively small number of Ph.D.'s produced annually. To improve the statistics the individual laboratories were compared on the basis of laboratory performance over a four-year period ending with the 1968-69 academic year. The results of this comparison are presented in Tables 8, 9, and 10.

For each laboratory, Table 8 gives the percentage variation of the four-year per dollar averages of the parameters of interest from the overall average of all the laboratories. Note in Table 8 that the five institutes receiving the lowest annual support had above average outputs (Ph.D.'s and papers) on a per dollar basis. In a parallel manner, of the five institutes receiving the highest annual support, all but number one (MIT) had below average outputs on a per dollar basis. The negative correlation coefficients (computed using the deviation score method) given in Table 7 confirm the implications of the raw data. One might conclude from this that smaller laboratories are inherently more efficient than larger ones. However, such conclusions could hardly be supported in the absence of other information. For example, the differences might be rooted in the fact that large laboratories can support larger, more expensive equipment with the result that a greater fraction of the support at these laboratories goes to equipment. Hence the smaller laboratories would tend to undertake more limited experiments or perhaps be more theoretically oriented. An argument has already been made that theoretical efforts will appear

TABLE 8. DEVIATIONS FROM OVERALL AVERAGES--FOUR-YEAR PERIOD

Institute	Ph.D.'s $$	Papers $$	Grad. Student $$	Faculty $$
1. MIT*	+19%	+17%	+19%	- 5%
2. Cornell	- 8%	- 8%	-22%	-18%
3. Illinois	-39%	-34%	-43%	-22%
4. Chicago	-40%	-16%	-29%	-43%
5. Pennsylvania	-21%	-37%	-30%	-32%
6. Stanford	+25%	-11%	+23%	-12%
7. Brown	-21%	+39%	+22%	+63%
8. Northwestern	+29%	+38%	+15%	+ 7%
9. Purdue	+60%	+11%	+43%	+85%
10. Maryland	+16%	+19%	+32%	+67%
11. Harvard	+ 9%	+ 3%	+ 4%	+15%
12. No. Carolina	+70%	+84%	+91%	+75%
Overall Averages	$\frac{0.77}{\$10^5}$	$\frac{4.47}{\$10^5}$	$\frac{4.75}{\$10^5}$	$\frac{1.36}{\$10^5}$

	Correlation Coefficients: Deviation Score	Correlation Coefficients: Rank Ordering
$$ and Ph.D.'s/$$:	- 0.50	- 0.42
$$ and Papers/$$:	- 0.52	- 0.55
$$ and Grad. Students/$$:	- 0.57	- 0.62
$$ and Faculty/$$:	- 0.70	- 0.75

*Institutes are listed in order of level of support from highest to lowest. Thus MIT receives the largest amount of money.

more productive on a per dollar basis. Support for the argument that the larger laboratories have more extensive experimental facilities can be found in the faculty per dollar column in Table 8. If the level of effort at a laboratory is roughly proportional to faculty participation, then one would expect faculty FTEs per dollars to be lower at those laboratories (here hypothesized to be the larger laboratories) which have more extensive experimental programs. As shown in Table 8, this is, in fact, the case. The top six laboratories in level of support are all below average in faculty/$$ while the bottom six are all above average. The high negative correlation coefficient for $$ and faculty/$$ confirms these conclusions.

TABLE 9. INSTITUTE PRODUCTIVITY: DEVIATIONS FROM OVERALL AVERAGE
(No. of Papers)/(No. of Grad. Students + 2 x No. of Faculty)

1.	MIT	+ 7%
2.	Cornell	+15%
3.	Illinois	+25%
4.	Chicago	+28%
5.	Pennsylvania	-10%
6.	Stanford	-18%
7.	Brown	+ 2%
8.	Northwestern	+23%
9.	Purdue	-17%
10.	Maryland	-18%
11.	Harvard	- 5%
12.	North Carolina	0%

Since figures are not available for the fraction of support going to equipment procurement, one might consider judging output merely on the basis of manpower input, recognizing that in so doing one abandons

TABLE 10. DEVIATIONS FROM OVERALL AVERAGES--FOUR-YEAR PERIOD

	Papers / Grad. Students	Papers / Faculty	Grad. Students / Faculty	Ph.D.'s / Faculty	Grad. Students / Ph.D.'s
1. MIT	- 1%	+23%	+25%	+26%	- 2%
2. Cornell	+17%	+12%	- 5%	-67%	-16%
3. Illinois	+16%	-16%	-28%	-23%	- 6%
4. Chicago	+20%	+48%	+24%	+ 4%	+18%
5. Pennsylvania	-11%	- 8%	+ 3%	+16%	-13%
6. Stanford	-30%	- 3%	+37%	+40%	-23%
7. Brown	+14%	-15%	-26%	-53%	+53%
8. Northwestern	+19%	+29%	+ 8%	+19%	-11%
9. Purdue	-22%	-40%	-23%	-14%	-11%
10. Maryland	-10%	-29%	-21%	-32%	+13%
11. Harvard	- 1%	-10%	-10%	- 5%	- 5%
12. North Carolina	- 3%	+ 6%	+ 9%	- 4%	+11%
Overall Averages	0.94	3.29	3.50	0.57	6.2

any attempt to evaluate the efficiency with which one uses dollars spent for equipment. Manpower input is generally made up of graduate students, faculty, and post doctoral students. However, as mentioned previously, the latter are not included in the data available. Ideally, one would be able to weight the different types of manpower with respect to their productivity. In the absence of other information one might, for example, assume that productivity is roughly proportional to annual salary. This would give a faculty/graduate student productivity ratio of about 4:1 (on a full-time basis) which is probably not unreasonable. If a graduate student spends roughly twice as much time as a faculty member on the project, the resultant productivity ratio is 2:1 based on number of participants. A comparison of laboratory productivity on this basis is shown in Table 9. The correlation coefficient between the level of laboratory support and laboratory productivity, as defined in this manner, is 0.38 which implies that less than 15% ($0.38^2 = 0.145$) of the variation in productivity as defined above can be accounted for by the level of support.

The manpower weighting question pursued above could be bypassed if the input ratio of the different types of manpower for any effort or laboratory is roughly constant. This might be expected in the case of graduate students and faculty since one faculty member can efficiently guide the research of only a limited number of students. However, for the data given (shown in the third column of Table 10), the manpower ratio varies between laboratories. As stated previously this may be a result of different methods of calculating number of graduate students and number of faculty. The other columns of Table 10 show the comparison of outputs when the two types of manpower are not combined. The last column is inverted compared to the others and provides a rough identification of the relative length of time necessary to obtain a Ph.D. at the various universities.

IV. SUMMARY AND CONCLUSIONS

A. EVALUATION BY PROJECT

No quantitative measures were found that would reflect the principal ARPA objectivės in supporting basic research, i.e., "development of new scientific knowledge applicable to current or projected DoD problem areas" or "education of the defense R&D community" with respect to developments in the basic research community. However, simple quantitative measures were developed that offer the potential of reflecting scientific excellence by ranking individual projects on a productivity basis. It was shown that the resulting rank orderings are, on the whole, relatively insensitive to weightings applied to the various types of manpower inputs and product outputs. Because of the absence of an independent method of judging project quality (no peer evaluation of individual projects was available), it was not possible to confirm or deny that the measures developed provide valid reflections of project quality. Although they do provide a quantitative description of project productivity, it is particularly difficult to obtain an accurate appraisal of the quality of individual publications. The Science Citation Index offers a potential vehicle for such a judgment (by counting the number of times a paper is cited) but suffers from at least a two- to three-year lag time--which is presumably too long to be of utility to a decision-maker who is interested in a researcher's current or very recent productivity. Another inherent weakness in counting number of publications is the pressure to publish in the academic community, particularly among individuals whose reputation is not so well established and who are seeking tenure or advancement in academic standing.

It is difficult to make judgments as to the utility of the rank orderings obtained. At the very least, they allow the isolation of those projects at the extremes of the productivity spectrum. However, the assumptions and simplifications which attend the analysis argue strongly against a rigid interpretation of the rank orderings as being absolutely representative of the relative merit of the individual projects.

No quantitative methods were found that offered any potential of judging projects in a prospective sense, i.e., in judging the relative merit of individual research proposals and their potential for success. Since this is what the decision-maker really wants to know, it is a significant shortcoming of quantitative methods.

B. EVALUATION BY LABORATORY

The laboratory comparison performed in this paper suffered from the crudeness of the available data. Attempts to correlate level of funding with laboratory productivity were essentially unsuccessful. Although there was a spectrum of productivities, the crudeness of the available data does not allow conclusions as to the relative merit of the laboratories.

C. FURTHER STUDIES

Further studies of quantitative methods of evaluating research might attempt a correlation between peer evaluations of completed research projects and the quantitative measures proposed in this paper to see if the latter do reflect scientific excellence as assessed by the peer group (with its inherent weaknesses). As noted previously, the data available from individual research projects could be improved if a better assessment of actual FTE manpower contributions were available.

REFERENCE

1. Technology in Retrospect and Critical Events in Science (TRACES), Illinois Institute of Technology Research Institute, Contract NSF-C535, 15 December 1968.

APPENDIX

INSTRUCTIONS FOR USE OF PROJECT INFORMATION FORM

It is recognized that all of the information requested on the /enclosed/ forms may not be immediately available for all projects. However, if the information can be obtained with a reasonable effort, it would be appreciated if that effort could be made.

Note that each form is for a specific project, using descriptions consistent with those given in the Annual Technical Reports. All efforts active in 1968-69 have been included.

The specific instructions which follow are numbered according to the corresponding blanks on the information forms.

I. One full-time equivalent (FTE) man year corresponds to nine months during the regular academic year and two months during the summer. If n_9 is the number of FTE's during the academic year and n_2 the number of FTE's during the summer, the effective number of FTE's for the entire year is

$$n = \frac{9n_2 + 2n_9}{11}$$

This is the number which should appear under IA.

Include visiting faculty as well as regular faculty in this category.

II. Distinguish between the total number of post doctoral fellows and those receiving salary support from the project. Presumably the difference represents the number of individuals on fellowship.

III. Include students pursuing both M.S. and Ph.D. degrees.

IV. Include nonacademic professional staff members in this category.

V. Count theses completed during the appropriate fiscal year (July 1 to June 30).

VI. and VII. Distinguish between those publications in reviewed journals which were a direct result of theses and those which were not. Presumably the individual to whom the thesis was awarded would appear as major author for those publications in VI.

VIII. List articles in categories shown.

IX. Indicate the amount of salary support given to the faculty, post doctoral fellows, students, and other research staff members. Include salary support only. Do not count accompanying overhead. Thus one FTE faculty member would presumably receive salary support of about $15,000-$20,000 while graduate students would receive $2500-$3500.

X. Indicate the dollar expenditure for equipment procurement.

XI. Indicate total project funding.