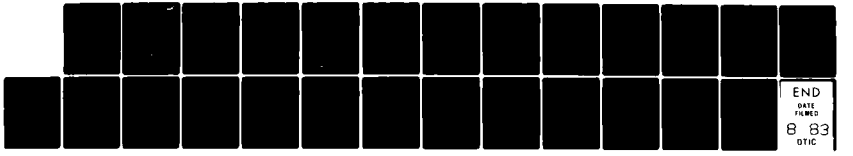
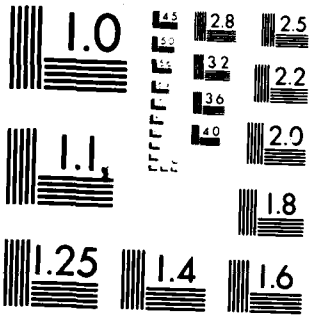
AD-A131 332 A METHODOLOGY FOR COLLECTING VALID SOFTWARE ENGINEERING 1/1
DATA(U) NAVAL RESEARCH LAB WASHINGTON DC
V R BASILI ET AL. 12 JUL 83 NRL-8679

UNCLASSIFIED F/G 5/1 NL





NRL Report 8679

A Methodology For Collecting Valid Software Engineering Data

VICTOR R. BASILI

University of Maryland

and

DAVID M. WEISS

Computer Science and Systems Branch
Information Technology Division

July 12, 1983

NAVAL RESEARCH LABORATORY
Washington, D.C.

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER NRL Report 8679	2. GOVT ACCESSION NO. AD-A131 332	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) A METHODOLOGY FOR COLLECTING VALID SOFTWARE ENGINEERING DATA		5. TYPE OF REPORT & PERIOD COVERED Interim report on a continuing NRL problem. 6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) Victor R. Basili* and David M. Weiss		8. CONTRACT OR GRANT NUMBER(s)
9. PERFORMING ORGANIZATION NAME AND ADDRESS Naval Research Laboratory Washington, DC 20375		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS 61153N, RR0140941 75-0199-0-3
11. CONTROLLING OFFICE NAME AND ADDRESS Office of Naval Research Washington, DC 20360		12. REPORT DATE July 12, 1983 13. NUMBER OF PAGES 24
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)		15. SECURITY CLASS. (of this report) UNCLASSIFIED 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

*University of Maryland

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Software engineering
Software development
Data collection

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

An effective data collection method for evaluating software development methodologies and for studying the software development process is described. The method uses goal-directed data collection to evaluate methodologies with respect to the claims made for them. Such claims are used as a basis for defining the goals of the data collection, establishing a list of questions of interest to be answered by data analysis, defining a set of data categorization schemes, and designing a data collection form.

(Continued)

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE
S/N 0102-014-6601

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

20 ABSTRACT (Continued)

The data to be collected are based on the changes made to the software during development and are obtained when the changes are made. To insure accuracy of the data, validation is performed concurrently with software development and data collection. Validation is based on interviews with those people supplying the data. Results from using the methodology show that data validation is a necessary part of change data collection. Without it, as much as 50% of the data may be erroneous.

Feasibility of the data collection methodology was demonstrated by applying it to five different projects in two different environments (other NRL Reports). The application showed that the methodology was both feasible and useful.

CONTENTS

A METHODOLOGY FOR COLLECTING VALID SOFTWARE ENGINEERING DATA

INTRODUCTION

According to the mythology of computer science, the first computer program ever written contained an error. Error detection and error correction are now considered to be the major cost factors in software development [1-3]. Much current and recent research is devoted to finding ways of preventing software errors. This research includes areas such as requirements definition [4], automatic and semiautomatic program generation [5,6], functional specification [7], abstract specification [8-11], procedural specification [12], code specification [13-15], verification [16-18], coding techniques [19-24], error detection [25], testing [26,27], and language design [16,28-31].

One result of this research is that techniques claimed to be effective for preventing errors are in abundance. Unfortunately, there have been few attempts at experimental verification of such claims. The purpose of this report is to show how to obtain valid data that may be used both to learn more about the software development process and to evaluate software development methodologies in a production environment. Previous [15] and companion papers [32] present the data and evaluation results. The methodology described in this report was developed as part of studies conducted by the Naval Research Laboratory and by NASA's Software Engineering Laboratory (SEL) [33].

SOFTWARE ENGINEERING EXPERIMENTATION

The course of action in most sciences when faced with a question of opinion is to obtain experimental verification. Software engineering disputes are not usually settled that way. Data from experiments exist, but rarely apply to the question to be settled. There are a number of reasons for this state of affairs. Probably the two most important are the number of potentially confounding factors involved in software studies and the expense of attempting to do controlled studies in an industrial environment involving medium or large-scale systems.

Rather than attempting controlled studies, we have devised a method for conducting accurate causal analyses in production environments. Causal analyses are efforts to discover the causes of errors and the reasons that changes are made to software. Such analyses are designed to provide some insight into the software development and maintenance processes, to help confirm or reject claims made for different methodologies, and to lead to better techniques for prevention, detection, and correction of errors. Relatively few examples of this kind of study exist in the literature; some examples are [34,35,4,15,36].

Most of the work in this area has centered on collecting data on errors committed by programmers during or after the coding phase of software development, or the recoding phase of maintenance. Analysis of the data consists of grouping the errors according to some classification scheme, such as by symptom or by cause, and calculating distributions. Results obtained are quite sensitive to the error classification scheme(s) selected.

As an example, Endres [35] studied the errors made in preparing a new release of the DOS/VS operating system. Two of his conclusions were that 1. interfaces between modules* were not a major

Manuscript approved December 15, 1982.

*Endres' modules correspond to subroutines.

source of error and that 2. nearly half of the errors originated in misunderstandings either of the problem to be solved or potential solutions, and were not susceptible to improvements in programming techniques. Operating system developers might well conclude from this study that better system specifications and more refined models of system behavior would result in a decreased incidence of errors during development.

Such conclusions are tempting, but have several drawbacks. First, there is but a single study to support these conclusions; more data are needed on operating system developments that show the same error patterns before a firm conclusion may be drawn.

A second drawback is that Endres did not report the cost to correct errors in different categories; the total effort expended in detecting and correcting errors caused by problem misunderstandings may have been small compared to the total development effort.

A third drawback is that the errors were classified after development ended, and without contact with the programmers. There is no way to estimate the accuracy of the data reported.

Causal analyses published in the literature suffer from a number of deficiencies. Many studies, such as [37] seem to be based on unverified data; i.e., programmer responses on data collection forms were not discussed with the programmers. Validation of error categorization by an independent validator is also apparently rare. (An exception is [38]). Studies of several projects using many data points often either use too many error categories to permit useful analysis, or do not give sufficient early thought to the categorization scheme to permit useful conclusions to be drawn. Examples of these problems, and methods for avoiding them, will be discussed in later sections.

To provide useful data, a data collection methodology must display certain attributes. Since much of the data of interest for real projects are collected during the test phase, complete analysis of the data must await project completion. Although it is important that data collection and validation proceed concurrently with development, the final analysis must be done from a historical viewpoint, after the project ends.

Developers can provide data as they make changes during development. In a reasonably well-controlled software development environment, documentation and code are placed under some form of configuration control before being released for use by others than the author. Changes are defined as alterations to baselined design, code, or documentation.

A key factor in the data gathering process is validation of the data as they become available. Such validity checks result in corrections to the data that cannot be captured at later times owing to the nature of human memory [39]. Timeliness of both data collection and data validation is quite important to the accuracy of the analysis.

Careful validation means that the data to be collected must be carefully specified, so that those supplying data, those validating data, and those performing the analyses will have a consistent view of the data collected. This is especially important for the purposes of those wishing to repeat studies in both the same and different environments.

Careful specification of the data requires the data collectors to have a clear idea of the goals of the study. Specifying goals is itself an important issue, since, without goals, one runs the risk of collecting unrelated, meaningless data.

To obtain insight into the software development process, the data collectors need to know the kinds of errors committed and the kinds of changes made. To identify troublesome issues, the effort

needed to make each change is necessary. For greatest usefulness, one would like to study projects from software production environments involving teams of programmers.

We may summarize the preceding as the following six criteria:

1. the data must contain information permitting identification of the types of errors and changes made;
2. the data must include the cost of making changes and correcting errors;
3. data to be collected must be defined as a result of clear specification of the goals of the study;
4. data should include studies of projects from production environments, involving teams of programmers;
5. data analysis should be historical, but data must be collected and validated concurrently with development; and
6. data classification schemes to be used must be carefully specified for the sake of repeatability of the study in the same and different environments.

SCHEMA FOR THE INVESTIGATIVE METHODOLOGY

Our data collection methodology is goal oriented. It starts with a set of goals to be satisfied, uses these to generate a set of questions to be answered, and then proceeds step-by-step through the design and implementation of a data collection and validation mechanism. Analysis of the data yields answers to the questions of interest and may also yield a new set of questions. The procedure relies heavily on an interactive data validation process; those supplying the data are interviewed for validation purposes concurrently with the software development process. The methodology has been used in two different environments to study five software projects developed by groups with different backgrounds using very different software development methodologies. In both environments it yielded answers to most questions of interest and some insight into the development methodologies used.

The projects studied vary widely with respect to factors such as application, size, development team, methodology, hardware, and support software. Nonetheless, the same basic data collection methodology was applicable everywhere. The schema used has six basic steps, listed in the following, with considerable feedback and iteration occurring at several different places.

1. Establish the goals of the data collection

We divide goals into two categories: those that may be used to evaluate a particular software development methodology relative to the claims made for it, and those that are common to all methodologies to be studied.

As an example, a goal of a particular methodology, such as information hiding [40], might be to develop software that is easy to change. The corresponding data collection goal is to evaluate the success of the developers in meeting this goal, i.e., evaluate the ease with which the software can be changed. Goals in this category may be of more interest to those who are involved in developing or testing a particular methodology, and must be defined cooperatively with them.

A goal that is of interest regardless of the methodology being used is to characterize changes in ways that permit comparisons across projects and environments. Such goals may interest software

engineers, programmers, managers, and others more than goals that are specific to the success or failure of a particular methodology.

Consequences of omitting goals

Without goals, one is likely to obtain data in which either incomplete patterns or no patterns are discernible. As an example, one goal of an early study [15] was to characterize errors. During data analysis, it became desirable to discover the fraction of errors that were the result of changes made to the software for some reason other than to correct an error. Unfortunately, none of the goals of the study were related to this type of change, and there were no such data available.

2. Develop a list of questions of interest

Once the goals of the study have been established, they may be used to develop a list of questions to be answered by the study. Questions of interest define data parameters and categorizations that permit quantitative analysis of the data. In general, each goal will result in the generation of several different questions of interest. As an example, if the goal is to characterize changes, some corresponding questions of interest are: "What is the distribution of changes according to the reason for the change?" "What is the distribution of changes across system components?" "What is the distribution of effort to design changes?"

As a second example, if the goal is to evaluate the ease with which software can be changed, we may identify questions of interest such as: "Is it clear where a change has to be made in the software?" "Are changes confined to single modules?" "What was the average effort involved in making a change?"

Questions of interest form a bridge between subjectively determined goals of the study and the quantitative measures to be used in the study. They permit the investigators to determine the quantities that need to be measured and the aspects of the goals that can be measured. As an example, if one is attempting to discover how a design document is being used, one might collect data that show how the document was being used when the need for a change to it was discovered. This may be the only aspect of the document's use that is measurable.

Goals for which questions of interest cannot be formulated and goals that cannot be satisfied because adequate measures cannot be defined may be discarded. Once formulated, questions can be evaluated to determine if they completely cover their associated goals and if they define quantitative measures. Finally, questions of interest have the desirable property of forcing the investigators to consider the data analyses to be performed before any data are collected.

Consequences of omitting questions of interest

Without questions of interest, there may be no quantitative basis for satisfying the goals of the study. Data distributions that are needed for evaluation purposes, such as the distribution of effort involved in making changes, may have to be constructed in an ad hoc way, and may be incomplete or inaccurate.

3. Establish data categories

Once the questions of interest have been established, categorization schemes for the changes and errors to be examined may be constructed. Each question generally induces a categorization scheme. If one question is, "What was the distribution of changes according to the reason for the change?" one will want to classify changes according to the reason they are made. A simple categorization scheme of this sort is *error corrections* as opposed to *nonerror corrections* (hereafter called *modifications*).

Each of these categories may be further subcategorized according to reason. As an example, modifications could be subdivided into those modifications resulting from requirements changes, those resulting from a change in the development support environment (e.g., compiler change), planned enhancements, optimizations, and others.

Such a categorization permits characterization of the changes with respect to the stability of the development environment, with respect to different kinds of development activities, etc. When matched with another categorization such as the difficulty of making changes, this scheme also reveals which changes are the most difficult to make.

Each categorization scheme should be complete and consistent, i.e., every change should fit exactly one of the subcategories of the scheme. To insure completeness, the category "Other" is usually added as a subcategory. Where some changes are not suited to the scheme, the subcategory "Not Applicable" may be used. As an example, if the scheme includes subcategories for different levels of effort in isolating error causes, then errors for which the cause need not be isolated (e.g., clerical errors noticed when reading code) belong in the "Not Applicable" subcategory.

Consequences of not defining data categories before collecting data

Omitting the data categorization schemes may result in data that cannot later be identified as fitting any particular categorization. Each change then tends to define its own category, and the result is an overwhelming multiplicity of data categories, with little data in each category.

4. Design and test data collection form

To provide a permanent copy of the data and to reinforce the programmers' memories, a data collection form is used. Form design was one of the trickiest parts of the studies conducted, primarily because forms represent a compromise among conflicting objectives. Typical conflicts are the desire to collect a complete, detailed set of data that may be used to answer a wide range of questions of interest, and the need to minimize the time and effort involved in supplying the data. Satisfying the former leads to large, detailed forms that require much time to fill out. The latter requires a short form organized so that the person supplying the data need only check off boxes.

Including the data suppliers in the form design process is quite beneficial. Complaints by those who must use the form are resolved early (i.e., before data collection begins), the form may be tailored to the needs of the data suppliers (e.g., for use in configuration management), and the data suppliers feel they are a useful part of the data collection process.

The forms must be constructed so that the data they contain can be used to answer the questions of interest. Several design iterations and test periods are generally needed before a satisfactory design is found.

Our principal goals in form design were to produce a form that:

1. fit on one piece of paper,
2. could be used in several different programming environments, and
3. permitted the programmer some flexibility in describing the change.

Figures 1a and 1b show the last version of the form used for the SEL studies. (An earlier version of the form was significantly modified as a result of experience gained in the data collection and analysis processes.) The first sections of the form request textual descriptions of the change and the reason it was made. Following sections contain questions and check-off tables that reflect various categorization schemes.

NUMBER _____

CHANGE REPORT FORM

PROJECT NAME _____ CURRENT DATE _____

SECTION A - IDENTIFICATION

REASON: Why was the change made? _____

DESCRIPTION: What change was made? _____

EFFECT: What components (or documents) are changed? (Include version) _____

EFFORT: What additional components (or documents) were examined in determining what change was needed? _____

	(Month	Day	Year)
Need for change determined on			
Change started on			

What was the effort in person time required to understand and implement the change?

____1 hour or less, ____1 hour to 1 day, ____1 day to 3 days, ____more than 3 days

SECTION B - TYPE OF CHANGE (How is this change best characterized?)

- ☐ Error correction
- ☐ Planned enhancement
- ☐ Implementation of requirements change
- ☐ Improvement of clarity, maintainability, or documentation
- ☐ Improvement of user services
- ☐ Insertion/deletion of debug code
- ☐ Optimization of time/space/accuracy
- ☐ Adaptation to environment change
- ☐ Other (Explain in E)

Was more than one component affected by the change? Yes_____ No_____

FOR ERROR CORRECTIONS ONLY

SECTION C - TYPE OF ERROR (How is this error best characterized?)

- ☐ Requirements incorrect or misinterpreted
- ☐ Functional specifications incorrect or misinterpreted
- ☐ Design error, involving several components
- ☐ Error in the design or implementation of a single component
- ☐ Misunderstanding of external environment, except language
- ☐ Error in use of programming language/compiler
- ☐ Clerical error
- ☐ Other (Explain in E)

FOR DESIGN OR IMPLEMENTATION ERRORS ONLY

If the error was in design or implementation:

The error was a mistaken assumption about the value or structure of data _____

The error was a mistake in control logic or computation of an expression _____

580-2 (6/78)

Fig. 1a – SEL change report form (front)

FOR ERROR CORRECTIONS ONLY

SECTION D - VALIDATION AND REPAIR

What activities were used to validate the program, detect the error, and find its cause?

	Activities Used for Program Validation	Activities Successful in Detecting Error Symptoms	Activities Tried to Find Cause	Activities Successful in Finding Cause
Pre-acceptance test runs				
Acceptance testing				
Post-acceptance use				
Inspection of output				
Code reading by programmer				
Code reading by other person				
Talks with other programmers				
Special debug code				
System error messages				
Project specific error messages				
Reading documentation				
Trace				
Dump				
Cross-reference/attribute list				
Proof technique				
Other (Explain in E)				

What was the time used to isolate the cause?

____one hour or less, ____one hour to one day, ____more than one day, ____never found

If never found, was a workaround used?_____Yes_____No (Explain in E)

Was this error related to a previous change?

____Yes (Change Report #/Date_____) ____No ____Can't tell

When did the error enter the system?

____requirements ____functional specs ____design ____coding and test ____other ____can't tell

SECTION E - ADDITIONAL INFORMATION

Please give any information that may be helpful in categorizing the error or change, and understanding its cause and its ramifications.

Name:_____ Authorized:_____ Date:_____

SEN-2 (6/78)

Fig. 1b — SEL change report form (back)

As an example, a categorization of time to design changes is requested in the first question following the description of the change. The completer of the form is given the choice of four categories (1 hour or less, 1 hour to 1 day, 1 day to 3 days, and more than 3 days) that cover all possibilities for design time.

Consequences of not using a data collection form

Without a data collection form, it is necessary to rely on the developers' memories and on perusal of early versions of design documentation and code to identify and categorize the changes made. This approach leads to incomplete, inaccurate data.

5. Collect and validate data

Data are collected by requiring those people who are making software changes to complete a change report form for each change made, as soon as the the change is completed. Validation consists of checking the forms for correctness, consistency, and completeness. As part of the validation process, in cases where such checks reveal problems, the people who filled out the forms are interviewed. Both collection and validation are concurrent with software development; the shorter the lag between programmers completing forms and being interviewed concerning those forms, the more accurate the data.

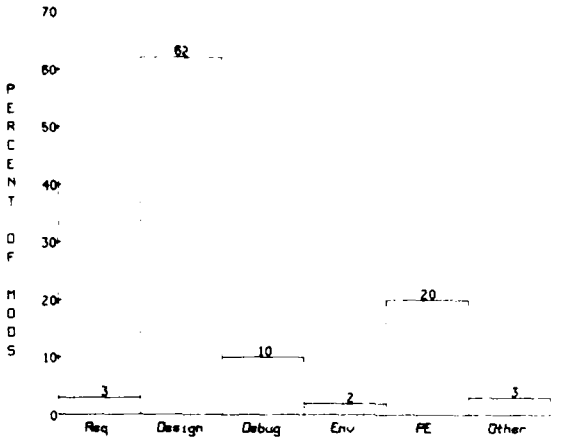
Perhaps the most significant problem during data collection and validation is insuring that the data are complete, i.e., that every change has been described on a form. The better controlled the development process, the easier this is to do. At each stage of the process where configuration control is imposed, change data may be collected. Where projects that we have studied use formal configuration control, we have integrated the configuration control procedures and the data collection procedures, using the same forms for both, and taking advantage of configuration control procedures for validation purposes. Since all changes must be reviewed by a configuration control board in such cases, we are guaranteed capture of all changes, i.e., that our data are complete. Furthermore, the data collection overhead is absorbed into the configuration control overhead and is not visible as a separate source of irritation to the developers.

Consequences of omitting validation

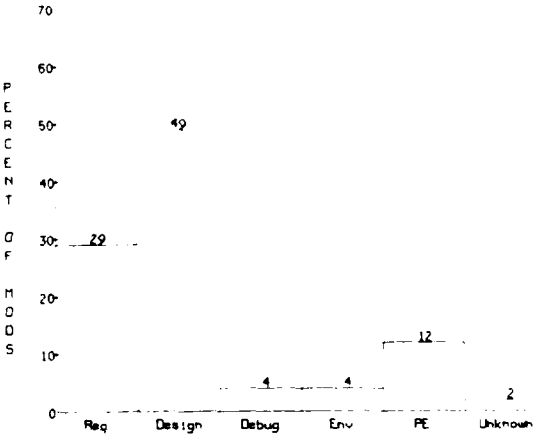
One result of concurrent development, data collection, and data validation is that the accuracy of the collection process may be quantified. Accuracy may be calculated by observing the number of mistakes made in completing data collection forms. One may then compare, for any data category, prevalidation distributions with post-validation distributions. We call such an analysis a validation analysis. The validation analysis of the SEL data shows that it is possible for inaccuracies on the order of 50% to be introduced by omitting validation. To emphasize the consequences of omitting the validation procedures, we present some of the results of the validation analysis of the SEL data in the next section.

6. Analyze data

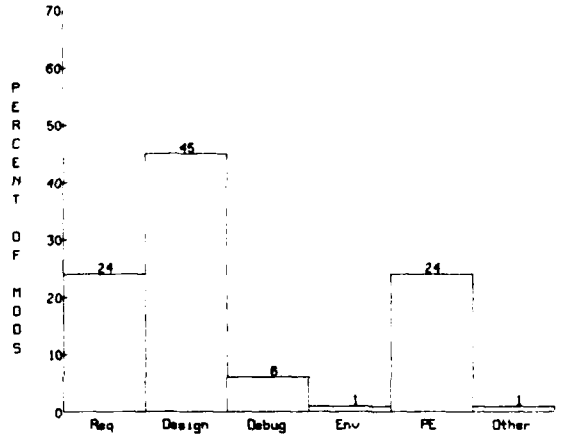
Data are analyzed by calculating the parameters and distributions needed to answer the questions of interest. As an example, to answer the question "What was the distribution of changes according to the reason for the change?" a distribution such as that shown in Fig. 2 might be computed from the data.



SEL1



SEL2



SEL3

Key to Figure

Design	Modifications caused by changes in design
Debug	*Modifications to insert or delete debug code*
Env	Modifications caused by changes in the hardware or software environment
PE	Planned Enhancements
Req	Modifications caused by changes in requirements or functional specifications
Unknown	Causes of these modifications are not known

Fig. 2 — Sources of modification

Application of the Schema

Applying the schema requires iterating among the steps several times. Defining the goals and establishing the questions of interest are tightly coupled, as are establishing the questions of interest, designing and testing the form(s), and collecting and validating the data. Many of the considerations involved in implementing and integrating the steps of the schema have been omitted here so that the reader may have an overview of the process. The complete set of goals, questions of interest, and data categorizations for the SEL projects are shown in a companion report [32].

Support Procedures and Facilities

In addition to the activities directly involved in the data collection effort, there are a number of support activities and facilities required. Included as support activities are testing the forms, collection and validation procedures, training the programmers, selecting a data base system to permit easy analysis of the data, encoding and entering data into the data base, and developing analysis programs.

DETAILS OF SEL DATA COLLECTION AND VALIDATION

In the SEL environment, program libraries were used to support and control software development. A full-time librarian was assigned to support SEL projects. All project library changes were routed through the librarian. In general, we define a change to be an alteration to baselined design, code, or documentation. For SEL purposes, only changes to code, and documentation contained in the code, were studied. The program libraries provided a convenient mechanism for identifying changes. A programmer causing a library change was required to complete a change report form (Fig. 1).

The data presented here are drawn from studies of three different SEL projects, denoted SEL1, SEL2, and SEL3. The processing procedures were as follows.

1. Programmers were required to complete change report forms for all changes made to library routines.

2. Programs were kept in the project library during the entire test phase.

3. After a change was made a completed change report form describing the change was submitted. The form was first informally reviewed by the project leader. It was then sent to the SEL library staff to be logged and a unique identifier assigned to it.

4. The change analyst reviewed the form and noted any inconsistencies, omissions, or possible miscategorizations. Any questions the analyst had were resolved in an interview with the programmer. (Occasionally the project leader or system designer was consulted rather than the individual programmer.)

5. The change analyst revised the form as indicated by the results of the programmer interview, and returned it to the library staff for further processing. Revisions often involved cases where several changes were reported on one form. In these cases, the analyst insured that there was only one change reported per form; this often involved filling out new forms. Forms created in this way are known as *generated* forms.

   (Changes were considered to be different if they were made for different reasons, if they were the result of different events, or if they were made at substantially different times

(e.g., several weeks apart). As an example, two different requirements amendments would result in two different change reports, even if the changes were made at the same time in the same subroutine.)

Occasionally, one change was reported on several different forms. The forms were then merged into one form, again to insure one and only one change per form. Forms created in this way are known as *combined* forms.

6. The library staff encoded the form for entry into the (automated) SEL data base. A preliminary, automated check of the form was made via a set of data base support programs. This check, mostly syntactic, ensured that the proper kinds of values were encoded into the proper fields, e.g., that an alphabetic character was not entered where an integer was required.

7. The encoded data were entered into the SEL data base.

8. The data were analyzed by a set of programs that computed the necessary distributions to answer the questions of interest.

Many of the reported SEL changes were error corrections. We define an error to be a discrepancy between a specification and its implementation. Although it was not always possible to identify the exact location of an error, it was always possible to identify exactly each error correction. As a result, we generally use the term error to mean error correction.

For data validation purposes, the most important parts of the data collection procedure are the review by the change analyst, and the associated programmer interview to resolve uncertainties about the data.

The SEL validation procedures afforded a good chance to discover whether validation was really necessary; it was possible to count the number of miscategorizations of changes and associated misinformation. These counts were obtained by counting the number of times each question on the form was incorrectly answered.

An example is misclassifications of errors as clerical errors. (Clerical errors were defined as errors that occur in the mechanical translation of an item from one format to another, e.g., from one coding sheet to another, or from one medium to another, e.g., coding sheets to cards.) For one of the SEL projects, 46 errors originally classified as clerical were actually errors of other types. (One of these consisted of the programmer forgetting to include several lines of code in a subroutine. Rather than clerical, this was classified as an error in the design or implementation of a single component of the system.) Initially, this project reported 238 changes, so we may say that about 19% of the original reports were misclassified as clerical errors.

The SEL validation process was not suited for verfiying the completeness of the reported data. We cannot tell from the validation studies how many changes were never reported. This weakness can be eliminated by integrating the data collection with stronger configuration control procedures.

Validation Differences Among SEL Projects

As experience was gained in collecting, validating, and analyzing data for the SEL projects, the quality of the data improved significantly, and the validation procedures changed slightly. For SEL1 and SEL2, completed forms were examined and programmers interviewed by a change analyst within a few weeks (typically 3 to 6 weeks) of the time the forms were completed. For project SEL2, the task leader (lead programmer for the project) examined each form before the change analysts saw it.

Project SEL3 was not monitored as closely as SEL1 and SEL2. The task leader, who was the same as for SEL2, by then understood the data categorization schemes quite well and again examined the forms before sending them to the SEL. The forms themselves were redesigned to be simpler but still capture nearly all the same data. Finally, several of the programmers were the same as on project SEL2 and were experienced in completing the forms.

Estimating Inaccuracies in the Data

Although there is no completely objective way to quantify the inaccuracy in the validated data, we believe it to be no more than 5% for SEL1 and SEL2. By this we mean that no more than 5% of the changes and errors are misclassified in any of the data collection categories. For the major categories, such as whether a change is an error or modification, the type of change, and the type of error, the inaccuracy is probably no more than 3%.

For SEL3, we attempted to quantify the results of the validation procedures more carefully. After validation, forms were categorized according to our confidence in their accuracy. We used four categories:

1. Those forms for which we had no doubt concerning the accuracy of the data. Forms in this cateogry were estimated to have no more than a 1% chance of inaccuracy.

2. Those forms for which there was little doubt about the accuracy of the data. Forms in this category were estimated to have at most a 10% chance of an inaccuracy.

3. Those forms for which there was some uncertaincy about the accuracy, with an estimated inaccuracy rate of no more than 30%.

4. Those forms for which there was considerable uncertaincy about the accuracy, with an estimated inaccuracy rate of about 50%.

Applying the inaccuracy rates to the number of forms in each category gave us an estimated inaccuracy of at most 3% in the validated forms for SEL3.

Prevalent Mistakes in Completing Forms

Clear patterns of mistakes and misclassifications in completing forms became evident during validation. As an example, programmers on projects SEL1 and SEL2 frequently included more than one change on one form. Often this was a result of the programmers sending the changes to the library as a group.

Comparative Validation Results

Figure 3 provides an overview of the results of the validation process for the 3 SEL projects. The percentage of original forms that had to be corrected as a result of the validation process is shown. As an example, 32% of the originally completed change report forms for SEL3 were corrected as a result of validation. The percentages are based on the number of original forms reported (since some forms were generated, and some combined, the number of changes reported after validation is different than the number reported before validation). Figure 4 shows the fraction of generated forms expressed as a percentage of total validated forms.

Figure 3 shows that pre-validation SEL3 forms were significantly more accurate than the pre-validation SEL1 or SEL2 forms. When the generated and combined forms are also considered, the pre-validation SEL3 data appear to be considerably better then the pre-validation data for either of the other projects. We believe the reasons for this are the improved design of the form, and the familiarity

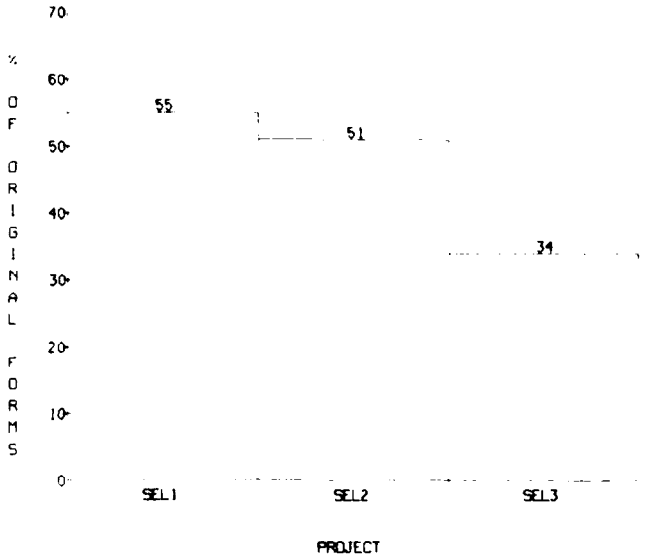


Fig. 3 — Corrected forms

of the task leader and programmers with the data collection process. (Generated forms are shown in Fig. 4. Combined forms for all of the projects represented a very small fraction of the total validated forms.)

These (overall) results show that careful validation, including programmer interviews, is essential to the accuracy of any study involving change data. Furthermore, it appears that with well-designed forms, and programmer training, there is improvement with time in the accuracy of the data one can obtain. We do not believe that it will ever be possible to dispense entirely with programmer interviews, however.

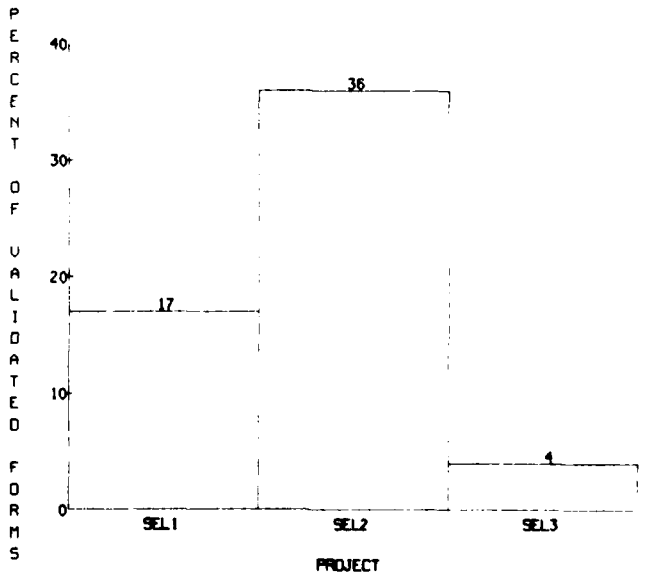


Fig. 4 — Generated forms

Erroneous Classifications

Table 1 shows misclassifications of errors as modifications and modifications as errors. As an example, for SEL1, 14% of the original forms were classified as modifications, but were actually errors. Without the validation process, considerable inaccuracy would have been introduced into the initial categorization of changes as modifications or errors.

Table 2 shows a sampling of other kinds of classification errors that could contribute significantly to inaccuracy in the data. All involve classification of an error into the wrong subcategory. The first row shows errors that were classified by the programmer as clerical, but were later reclassified as a result of the validation process into another category. For SEL1, significant inaccuracy (19%) would be introduced by omitting the validation process.

Table 3 is similar to Table 2, but shows misclassifications involving modifications. The first row shows modifications that were classified by the programmer as requirements or specifications changes, but were reclassified as a result of validation.

Table 1—Erroneous Modification and Error Classifications
(Percent of Original Forms)

	SEL1	SEL2	SEL3
Modifications classified as errors	1%	5%	less than 1%
Errors classified as modifications	14%	5%	2%

Table 2—Typical Error Type Misclassifications
(Percent of Original Forms)

Original Classification	SEL1	SEL2	SEL3
Clerical Error	19%	7%	6%
(Use of) Programming Language	0%	5%	3%
Incorrect or Misinterpreted Requirements		0%	less than 1%
Design Error		8%	1%

Table 3—Erroneous Modification Classifications
(Percent of Original Forms)

	SEL1	SEL3
Requirements or specification change	1%	less than 1%
Design change	8%	1%
Optimization	8%	less than 1%
Other	3%	less than 1%

Variation in Misclassification

Data on misclassifications of change and error type subcategories, such as shown in Table 2, tend to vary considerably among both projects and subcategories. (Misclasssification of clerical errors as shown in Table 2 is a good example.) Most likely this is because the misclassifications represent biases in the judgements of the programmers. It became clear during the validation process that certain programmers tended toward particular misclassifications.

The consistency between projects SEL2 and SEL3 in Table 2 probably occurs because both projects had the same task leader, who screened all forms before sending them to the SEL for validation.

Conclusions Concerning Validation

The preceding sections have shown that the validation process, particularly the programmer interviews, are a necessary part of the data collection methodology. Inaccuracies on the order of 50% may be introduced without this form of validation. Furthermore, it appears that with appropriate form design and programmer experience in completing forms, the inaccuracy rate may be substantially reduced, although it is doubtful that it can be reduced to the level where programmer interviews may be omitted from the validation procedures.

A second significant conclusion is that the analysis performed as part of the validation process may be used to guide the data collection project; the analysis results show what data can be reliably and practically collected, and what data cannot be. Data collection goals, questions of interest, and data collection forms may have to be revised accordingly.

RECOMMENDATIONS FOR DATA COLLECTORS

We believe we now have sufficient experience with change data collection to be able to apply it successfully in a wide variety of environments. Although we have been able to make comparisons between the data collected in the two environments we have studied [41], we would like to make comparisons with a wider variety of environments. Such comparisons will only be possible if more data become available. To encourage the establishment of more data collection projects, we feel it is important to describe a successful data collection methodology, as we have done in the preceding sections, to point out the pitfalls involved, and to suggest ways of avoiding those pitfalls.

Procedural Lessons Learned

Problems encountered in various procedural aspects of the studies were the most difficult to overcome. Perhaps the most important are the following.

1. Clearly understanding the working environment and specifying the data collection procedures were a key part of conducting the investigation. Misunderstanding by the programmer of the circumstances that require him/her to file a change report form will prejudice the entire effort. Prevention of such misunderstandings can be accomplished partly by training procedures and good forms design, but feedback to the development staff, i.e., those filling out the data collection forms, must not be omitted.

2. Similarly, misunderstanding by the change analyst of the circumstances that required a change to be made will result in misclassifications and erroneous analyses. Our SEL data collection was helped by the use of a change analyst who had previously worked in the NASA environment and understood the application and the development procedures used.

3. Timely data validation through interviews with those responsible for reporting errors and changes was vital, especially during the first few projects to use the forms. Without such validation procedures, data will be severely biased, and the developers will not get the feedback to correct the procedures they are using for reporting data.

4. Minimizing the overhead imposed on the people who were required to complete change reports was an important factor in obtaining complete and accurate data. Increased overhead brought increased reluctance to supply and discuss data. In projects where data

collection has been integrated with configuration control, the visible data collection and validation overhead is significantly decreased, and is no longer an important factor in obtaining complete data. Because configuration control procedures for the SEL environment were informal, we believe we did not capture all SEL changes.

5. In cases where an automated data base is used, data consistency and accuracy checks at or immediately prior to analysis are vital. Errors in encoding data for entry into the data base will otherwise bias the data.

Nonprocedural Lessons Learned

In addition to the procedural problems involved in designing and implementing a data collection study, we found several other pitfalls that could have strongly affected our results and their interpretation. They are listed in the following.

1. Perhaps the most significant of these pitfalls was the danger of interpreting the results without attempting to understand factors in the environment that might affect the data. As an example, we found a surprisingly small percentage of interface errors on all of the SEL projects. This result was surprising since interfaces are an often-cited source of errors. There was also other evidence in the data that the software was quite amenable to change. In trying to understand these results, we discussed them with the principal designer of the SEL projects (all of which had the same application). It was clear from the discussion that as a result of their experience with the application, the designers had learned what changes to expect to their systems, organized the design so that the expected changes would be easy to make, and then re-used the design from one project to the next. Rather than misinterpreting the data to mean that interfaces were not a significant software problem, we were led to a better understanding of the environment we were studying.

2. A second pitfall was underestimating the resources needed to validate and analyze the data. Understanding the change reports well enough to conduct meaningful, efficient programmer interviews for validation purposes initially consumed considerable amounts of the change analysts' time. Verifying that the data base was internally consistent, complete, and consistent with the paper copies of reports, was a continuing source of frustration and sink for time and effort.

3. A third potential pitfall in data collection is the sensitivity of the data. Programmers and designers sometimes need to be convinced that error data will not be used against them. This did not seem to be a significant problem on the projects studied for a variety of reasons, including management support, processing of the error data by people independent of the project, identifying error reports in the analysis process by number rather than name, informing newly hired project personnel that completion of error reports was considered part of their job, and high project morale. Furthermore, project management did not need error data to evaluate performance.

4. One problem for which there is no simple solution is the Hawthorne (or observer) effect [42]. When project personnel become aware that an aspect of their behavior is being monitored, their behavior will change. If error monitoring is a continuous, long-term activity that is part of the normal scheme of software development, not associated with evaluation of programmer performance, this effect may become insignificant. We believe this was the case with the projects studied.

5. The sensitivity of error data is enhanced in an environment where development is done on contract. Contractors may feel that such data are proprietary. Rules for data collection may have to be contractually specified.

Avoiding Data Collection Pitfalls

In the foregoing sections a number of potential pitfalls in the data collection process have been described. The following list includes suggestions that help avoid some of these pitfalls.

1. Select change analysts who are familiar with the environment, application, project, and development team.

2. Establish the goals of the data collection methodology and define the questions of interest before attempting any data collection. Establishing goals and defining questions should be an iterative process performed in concert with the developers. The developers' interests are then served as well as the data collector's.

3. For initial data collection efforts, keep the set of data collection goals small. Both the volume of data and the time consumed in gathering, validating, and analyzing it will be unexpectedly large.

4. Design the data collection form so that it may be used for configuration control, so that it is tailored to the project(s) being studied, so that the data may be used for comparison purposes, and so that those filling out the forms understand the terminology used. Conduct training sessions in filling out forms for newcomers.

5. Integrate data collection and validation procedures into the configuration control process. Data completeness and accuracy are thereby improved, data collection is unobtrusive, and collection and validation become a part of the normal development procedures. In cases where configuration control is not used or is informal, allocate considerable time to programmer interviews, and, if possible, documentation search and code reading.

6. Automate as much of the data analysis process as possible.

Limitations

It has been previously noted that the main limitation of using a goal-directed data collection approach in a production software environment is the inability to isolate the effects of single factors. For a variety of reasons, controlled experiments that may be used to test hypotheses concerning the effects of single factors do not seem practical. Neither can one expect to use the change data from goal-directed data collection to test such hypotheses.

A second major limitation is that lost data cannot be accurately recaptured. The data collected as a result of these studies represent five years of data collection. During that time there was considerable and continuing consideration given to the appropriate goals and questions of interest. Nonetheless, as data were analyzed, it became clear that there was information that was never requested but that would have been useful. An example is the length of time each error remained in the system. Programmers correcting their own errors, which was the usual case, can supply this data easily at the time they correct the error. Our attempts to discover error entry and removal times after the end of development were fruitless. (Error entry times were particularly difficult to discover.) Given such data, one could isolate errors that were not easily susceptible to detection. This type of example underscores the need for careful planning prior to the start of data collection.

Recommendations That May Be Provided to the Software Developer

The nature of the data collection methodology and the environments in which it can be used do not generally permit isolation of the effects of particular factors on the software development process. The results cannot be used to suggest that controlling a particular factor in the development process will reduce the quantity or cost of particular kinds of errors. We have found that the patterns found in the data do suggest that certain approaches, when applied in the environment studied, will improve the development process.

As an example, in the SEL environment neither external problems, such as requirements changes, nor global problems, such as interface design and specification, were significant. Furthermore, the development environment was quite stable. Most problems were associated with the individual programmer. The data show that in the SEL environment it would clearly pay to impose more control on the process of composing individual routines. Since detecting and correcting most errors was apparently quite easy in the overwhelming majority of cases, more attention should be paid to preventing errors from entering the code initially.

CONCLUSIONS CONCERNING DATA COLLECTION FOR METHODOLOGY EVALUATION PURPOSES

The data collection schema presented has been applied to five different projects in two different environments. We have been able to draw the following conclusions as a result of designing and implementing the data collection processes.

1. In all cases, it has been possible to collect data concurrently with the software development process in a software production environment.

2. Data collection may be used to evaluate the application of a particular software development methodology, or simply to learn more about the software development process. In the former case, the better defined the methodology, the more precisely the goals of the data collection may be stated.

3. The better controlled the development process, the more accurate and complete the data.

4. For all projects studied, it has been necessary to validate the data, including interviews with the project developers.

5. As patterns are discerned in the data collected, new questions of interest emerge. These questions may not be answerable with the available data, and may require establishing new goals and questions of interest.

The difficulties involved in conducting large scale controlled software engineering experiments have prevented evaluations of software development methodologies in the environments where they are often claimed to work best. As a result, software engineers must depend on less formal techniques that can be used in real working environments to establish long-term trends. We view change analysis as one such technique and feel that more techniques, and many more results obtained by applying such techniques, are needed.

ACKNOWLEDGMENTS

Without the many people who took time to fill out forms and submit to being interviewed, this research would be impossible. All the NASA/GSFC contractor-supplied programmers and in-house

NASA programmers were helpful in this regard. Particularly helpful were Jean Grondalski and Dr. Gerald Page. Sam DePriest was particularly adept at organizing, storing, and retrieving large numbers of change report forms.

Support for a research project involving data collection in a production environment must come from many sources. These sources include project management, the programmers supplying the data, those maintaining the data base, those assisting in data analysis, and those providing technical review and guidance. A few of the people providing such support were Dr. John Gannon, Dr. Richard Meltzer, Frank McGarry, Dr. Gerald Page, Dr. David Parnas, Dr. John Shore, and Dr. Marvin Zelkowitz.

Research that relies on the analysis of large quantities of data must also rely on human (and computer) efforts to analyze that data. Part of the necessary computer support was provided by the University of Maryland Computer Center. Other, significant, support came from Kathy Kragh, Tamara Lewis, Joanne Glazer-Weiss, Joshua Weiss, and Shana Weiss.

Deserving of special mention is Frank McGarry, who had sufficient foresight and confidence to sponsor much of this work and to offer his projects for study.

REFERENCES

1. B. Boehm and A. Haile, "Information processing/Data Automation Implications Of Air Force Command and Control Requirements in the 1980's (CCIP-85)," Space and Missile Systems Organization, Los Angeles, SAMSO-TR-72-200-**11** (February 1972).

2. B. Boehm "Software and Its Impact: A Quantitative Assessment," Datamation **19**(5), 48-59 (May 1973).

3. R. Wolverton, "The Cost of Developing Large Scale Software," IEEE Trans. Computers **23**(6), 615-636 (1974).

4. T. Bell, D. Bixler, and M. Dyer, "An Extendable Approach to Computer-Aided Software Requirments Engineering," IEEE Trans. Software Engineering **SE-3**(1), 49-60 (January 1977).

5. A. Ambler, D. Good, J. Browne, W. Burger, R. Cohen, C. Hoch, and R. Wells, "GYPSY: A Language for Specification and Implementation of Verifiable Programs," Proc. of the ACM Conference on Language Design for Reliable Software, 1-10 (March 1977).

6. Z. Manna and R. Waldinger, "Synthesis: Dreams => Programs," IEEE Trans. Software Engineering **SE-5**(4), 294-329 (July 1979).

7. K. Heninger, "Specifying Software Requirements for Complex Systems: New Techniques and Their Application," IEEE Trans. Software Engineering **SE-6**, 2-13 (January 1980).

8. D.L. Parnas, "A Technique for Software Module Specification with Examples," Comm. ACM **15**(5), 330-336 (May 1972).

9. J. Guttag, "The Specification and Application to Programming of Abstract Data Types," CSRG-59, University of Toronto Dept. of Computer Science Computer Systems Research Group (1975).

10. J. Guttag, "Abstract Data Types and the Development of Data Structures," Comm. ACM **20**, 396-404 (June 1976).

11. B. Liskov and S. Zilles, "Specification Techniques for Data Abstractions," IEEE Trans. Software Engineering **SE-1**(1), 7-19 (March 1975).

12. R. Linger, H. Mills, and B. Witt, *Structured Programming Theory and Practice*, Addison-Wesley, Reading (1979).

13. S. Caine and E. Gordon, "PDL—A tool for software design," Proc. Nat. Computer Conf., 271-276 (1975).

14. H. Elovitz, "An Experiment in Software Engineering: The Architecture Research Facility as a Case Study," Proc. Fourth Intnl. Conf. Software Engineering, 145-152 (1979).

15. D. Weiss, "Evaluating Software Development by Error Analysis: The Data from the Architecture Research Facility," J. Systems and Software **1**, 57-70 (1979).

16. E.W. Dijkstra, *A Discipline of Programming*, Prentice-Hall, Englewood Cliffs (1976).

17. R.W. Floyd, "Assigning Meanings to Programs," Proc. Symposium in Applied Mathematics **XIX**, 19-32 American Mathematical Society (1967).

18. C.A.R. Hoare, "An Axiomatic Basis for Computer Programming," Comm. ACM **12**(10), 576-580 (October 1969).

19. F. Baker, "Chief Programmer Team Management of Production Programming," IBM Systems Journal **11**(1), 56-73 (1972).

20. E.W. Dijkstra, "Notes on Structured Programming," in *Structured Programming*, Academic Press, London (1972).

21. D.E. Knuth, "Structured Programming with Go To Statements," Computing Surveys **6**(4), 261-301 (December 1974).

22. H. Mills, "Chief Programmer Teams: Principles and Procedures," FSC 71-5108, IBM Federal Systems Division (1971).

23. H. Mills, "Mathematical Foundations for Structured Programming," FSC 72-6012, IBM Federal Systems Division (1972).

24. N. Wirth, "Program Development by Stepwise Refinement," Comm. ACM **14**(4), 221-227 (April 1971).

25. E. Satterthwaite, "Debugging Tools for High-Level Languages," Software—Practice and Experience **2**(3), 197-217 (July-September 1972).

26. W. Howden, "Theoretical and Empirical Studies of Program Testing," Proc. Thrid Intntl. Conf. Software Engineering, 305-311 (May 1978).

27. J. Goodenough and S. Gerhart, "Toward a theory of test data selection," Proc. Intntl. Conf. Reliable Software, 493-510 (1975).

28. J. Gannon, "Language Design to Enhance Programming Reliability," CSRG-47, University of Toronto Dept. of Computer Science Computer Systems Research Group (1975).

29. J. Gannon and J. Horning, "Language Design for Programming Reliability," IEEE Trans. Software Eng. **SE-1**(2), 179-191 (June 1975).

30. C.A.R. Hoare and N. Wirth, "An Axiomatic Definition of the Programming Language Pascal," Acta Informatica **2**, 335-355 (1973).

31. K. Jensen and N. Wirth, *PASCAL: User Manual and Report Second Edition*, Springer-Verlag, New York (1974).

32. V. Basili and D. Weiss, "Evaluating Software Development by Analysis of Changes: The Data From the Software Engineering Laboratory," *NRL Report in preparation.*

33. V. Basili, M. Zelkowitz, F. McGarry, et. al., "The Software Engineering Laboratory," Report TR-535, University of Maryland (May 1977).

34. B. Boehm, "An Experiment in Small-Scale Application Software Engineering," Report TRW-SS-80-01, TRW (1980).

35. A. Endres, "Analysis and Causes of Errors in Systems Programs," Proc. Intntl. Conf. Reliable Software, 327-336 (1975).

36. V. Basili and D. Weiss, "Evaluation of a Software Requirements Document by Analysis of Change Data," Proc. Fifth Intntl. Conf. Software Engineering, 314-323 (March 1981).

37. G. Craig, W. Hetrick, M. Lipow, T. Thayer, and J. Yoxtheimer, "Software Reliability Study," Report RADC-TR-74-250, Rome Air Development Center (October 1974).

38. E. Youngs, "Error-Proneness in Programming," Ph.D. Thesis, University of North Carolina, Dept. of Psychology, Chapel Hill (1970).

39. G. Miller, "The Magical Number Seven, Plus or Minus Two: Some Limits on our Capacity for Processing Information," The Psychological Review **63**(2), 81-97 (March 1956).

40. D.L. Parnas, "On the Criteria to be used in Decomposing Systems into Modules," Comm. ACM **15**(12), 1053-1058 (December 1972).

41. D. Weiss, "A Comparison of Errors in Different Software—Development Environments," NRL Report 8598, July 9, 1982.

42. J. Brown, *The Social Psychology of Industry*, Penguin Books, Baltimore (1954).


DATE
ILME