RELIABILITY OF FUNCTION POINTS MEASUREMENT: A FIELD EXPERIMENT

Chris F. Kemerer

October 1990

CISR WP No. 216
Sloan WP No. 3193-90-MSA
RELIABILITY OF FUNCTION POINTS MEASUREMENT: A FIELD EXPERIMENT

Chris F. Kemerer

October 1990

CISR WP No. 216
Sloan WP No. 3193-90-MSA

©1990 Massachusetts Institute of Technology

Center for Information Systems Research
Sloan School of Management
Massachusetts Institute of Technology
Reliability of Function Points Measurement: A Field Experiment

Abstract

Despite the critical need for valid measurements of software size and complexity for the planning and control of software development, there exists a severe shortage of well-accepted measures, or metrics. One promising candidate has been Function Points (FPs), a relatively technology-independent metric originally developed by Allan Albrecht of IBM for use in software cost estimation, and now also used in software project evaluation. One barrier to wider acceptance of FPs has been possible concern that the metric may have low reliability. The very limited research that has been done in this area on individual programs has only been able to suggest a degree of agreement between two raters measuring the same program as +/- 30%, and inter-method reliability across different methods of counting has remained untested.

The current research consisted of a large scale field experiment involving over 100 FP measurements of actual medium-sized software applications. Measures of both inter-rater and inter-method reliability were developed and estimated for this sample. The results showed that the FP counts from pairs of raters using the standard method differed on average by +/- 10.78%, and that the correlation across the two methods tested was as high as .95 for the data in this sample. These results suggest that FPs are much more reliable than previously suspected, and therefore wider acceptance and greater adoption of FPs as a software metric should be encouraged.


Research support from the International Function Point Users Group (IFPUG) and the MIT Center for Information Systems Research is gratefully acknowledged. Helpful comments on the original research design and/or earlier drafts of this paper were received from A. Albrecht, N. Campbell, J. Cooprider, B. Dreger, J. Henderson, R. Jeffery, C. Jones, W. Orlikowski, D. Reifer, H. Rubin, E. Rudolph, W. Rumpf, C. Symons, and N. Venkataraman. Provision of the data was made possible in large part due to the efforts of A. Belden and B. Porter, and the organizations that contributed data to the study. Special thanks are also due my research assistant, M. Connolley.
I. INTRODUCTION

Software engineering management encompasses two major functions, planning and control, both of which require the capability to accurately and reliably measure the software being delivered. Planning of software development projects emphasizes estimation of the size of the delivered system in order that appropriate budgets and schedules can be agreed upon. Without valid size estimates, this process is likely to be highly inaccurate, leading to software that is delivered late and over-budget. Control of software development requires a means to measure progress on the project and to perform after-the-fact evaluations of the project in order, for example, to evaluate the effectiveness of the tools and techniques employed on the project to improve productivity.

Unfortunately, as current practice often demonstrates, both of these activities are typically not well performed, in part because of the lack of well-accepted measures, or metrics. Software size has traditionally been measured by the number of source lines of code (SLOC) delivered in the final system. This metric has been criticized in both its planning and control applications. In planning, the task of estimating the final SLOC count for a proposed system has been shown to be difficult to do accurately in actual practice [Low and Jeffery, 1990]. And in control, SLOC measures for evaluating productivity have weaknesses as well, in particular, the problem of comparing systems written in different languages [Jones, 1986].

Against this background, an alternative software size metric was developed by Allan Albrecht of IBM [Albrecht, 1979] [Albrecht and Gaffney, 1983]. This metric, which he termed "function points" (hereafter FPs), is designed to size a system in terms of its delivered functionality, measured in terms of such entities as the numbers of inputs, outputs, and files. Albrecht argued that these entities would be much easier to estimate than SLOC early in the software project lifecycle, and would be generally more meaningful to non-programmers. In addition, for evaluation purposes, they would avoid the difficulties involved in comparing SLOC counts for systems written in different languages.

FPs have proven to be a broadly popular metric with both practitioners and academic researchers. Dreger estimates that some 500 major corporations world-wide are using FPs [Dreger, 1989], and, in a survey by the Quality Assurance Institute, FPs were found to be regarded as the best available productivity metric [Perry, 1986]. They have also been widely used by researchers in such applications as cost estimation [Kemerer, 1987], software development productivity evaluation [Behrens, 1983] [Rudolph, 1983], software maintenance productivity evaluation [Banker, Datar

---

1Readers unfamiliar with FPs are referred to Appendix A for an overview of FP definitions and calculations.
Despite their wide use by researchers and their growing acceptance in practice, FPs are not without criticism. The main criticism revolves around the alleged low inter-rater reliability of FP counts, that is, whether two individuals performing a FP count for the same system would generate the same result. Barry Boehm, a leading researcher in the software estimation and modeling area, has described the definitions of function types as “ambiguous” [Boehm, 1987]. And, the author of a leading software engineering textbook describes FPs as follows:

“The function-point metric, like LOC, is relatively controversial... Opponents claim that the method requires some ‘sleight of hand’ in that computation is based on subjective, rather than objective, data...” [Pressman, 1987, p. 94]

This perception of FPs as being unreliable has undoubtedly slowed their acceptance as a metric, as both practitioners and researchers may feel that in order to ensure sufficient measurement reliability either a) a single individual would be required to count all systems, or b) multiple raters should be used for all systems and their counts averaged to approximate the ‘true’ value. Both of these options are unattractive in terms of either decreased flexibility or increased cost.

A second, related concern has developed more recently, due in part to FPs’ growing popularity. A number of researchers and consultants have developed variations on the original method developed by Albrecht [Rubin, 1983] [Symons, 1988] [Desharnais, 1988] [Jones, 1988] [Dreger, 1989]. A possible concern with these variants is that counts using these methods may differ from counts using the original method [Verner et al., 1989] [Ratcliff and Rollo, 1990]. Jones has compiled a list consisting of fourteen named variations, and suggests that the values obtained using these variations might differ by as much as +/- 50% from the original Albrecht method [Jones, 1989b]. If true, this lack of inter-method reliability poses several practical problems. From a planning perspective, one problem would be that for organizations adopting a method other than the Albrecht standard, the data they collect may not be consistent with that used in the development and calibration of a number of estimation models, e.g., see [Albrecht and Gaffney, 1983] and [Kemerer, 1987]. If the organization’s data were not consistent with this previous work, then the parameters of those models would no longer be directly useable by the organization. This would then force the collection of a large, internal dataset before FPs could be used to aid in cost and schedule estimation, which would involve considerable extra delay and expense. A second problem would be that for organizations that had previously adopted the Albrecht standard, and desired to switch to another variation, the switch might render previously developed models and heuristics less accurate.
From a control perspective, organizations using a variant method would have difficulty in comparing their ex post FP productivity rates to those of other organizations. For organizations that switched methods, the new data might be sufficiently inconsistent as to render trend analysis meaningless. Therefore, the possibility of significant variations across methods poses a number of practical concerns.

This study addresses these questions of FP measurement reliability through a carefully designed field experiment involving a total of 111 different counts of a number of real systems. Multiple raters and two methods were used to repeatedly count the systems, whose average size was 433 FPs. Briefly, the results of the study were that the FP counts from pairs of raters using the standard method differed on average by approximately +/- 10%, and that the correlation across the two methods was as high as .95 for the data in this sample. These results suggest that FPs are much more reliable than previously suspected, and therefore wider acceptance and greater adoption of FPs as a software metric should be encouraged.

The remainder of this paper is organized as follows. Section II outlines the research design, and summarizes relevant previous research in this area. Section III describes the data collection procedure and summarizes the contents of the dataset. Results of the main research questions are presented in section IV, with some additional results presented in section V. Concluding remarks are provided in the final section.

II. RESEARCH DESIGN

Introduction

Despite both the widespread use of FPs and the attendant criticism of their suspected lack of reliability, supra, there has been almost no research on either the inter-rater question or the inter-method question. Perhaps the first attempt at investigating the inter-rater reliability question was made by members of the IBM GUIDE Productivity Project Group, the results of which are described by Rudolph as follows:

"In a pilot experiment conducted in February 1983 by members of the GUIDE Productivity Project Group ... about 20 individuals judged independently the function point value of a system, using the requirement specifications. Values within the range +/- 30% of the average judgement were observed ... The difference resulted largely from differing interpretation of the requirement specification. This should be the upper limit of the error range of the function point technique. Programs available in source code or with detailed design specification should have an error of less than +/- 10% in their function point assessment. With a detailed description of the system there is not much room for different interpretations." [Rudolph, 1983, p. 6]
Aside from this description, no other research seems to have been documented, up until very recently. In January of 1990 a study by Low and Jeffery was published, which is the first widely available, well-documented study of this question [Low and Jeffery, 1990].

The Low & Jeffery Study
Low and Jeffery's research focused on one of the issues relevant to the current research, inter-rater reliability of FP counts [Low and Jeffery, 1990]. Their research methodology was an experiment using professional systems developers as subjects, with the unit of analysis being a set of program level specifications. Two sets of program specifications were used, both of which having been pre-tested with student subjects. For the inter-rater reliability question, 22 systems development professionals who counted FPs as part of their employment in 7 Australian organizations were used, as were an additional 20 inexperienced raters who were given training in the then current Albrecht standard. Each of the experienced raters used his or her organization's own variation on the Albrecht standard [Jeffery, 1990]. With respect to the inter-rater reliability research question Low and Jeffery found that the consistency of FP counts "appears to be within the 30 percent reported by Rudolph" within organizations, i.e., using the same method [Low and Jeffery, 1990, p. 71].

Design of the Study
Given the Low and Jeffery research, a deliberate decision was made at the beginning of the current research to select an approach that would complement their work by a) addressing the inter-rater reliability question using a different design and by b) directly focusing on the inter-method reliability question. The current work is designed to strengthen the understanding of the reliability of FP measurement, building upon the base started by Low and Jeffery.

The main area of overlap is the question of inter-rater reliability. Low and Jeffery chose a small group experiment, with each subject's identical task being to count the FPs implied from the two program specifications. Due to this design choice, the researchers were limited to choosing relatively small tasks, with the mean FP size of each program being 58 and 40 FPs, respectively. A possible concern with this design would be the external validity of the results obtained from the experiment in relation to real world systems. Typical medium sized application systems are generally an order of magnitude larger than the programs counted in the Low and Jeffrey experiment [Emrick, 1988] [Topper, 1990]2. Readers whose intuition is that FPs are relatively unreliable might argue that the unknown true reliability numbers are less than those estimated in the

---

2In addition to the references cited, an informal survey of a number of FP experts was highly consistent on this point. Their independent answers were 300 [Rudolph, 1989], 320 [Jones, 1986] and 500 [Albrecht, 1989].
experiment, since presumably it is easier to understand, and therefore count correctly, a small problem than a larger one. On the other hand, readers whose intuition is that the unknown true reliability numbers are better than those estimated in the experiment might argue that the experiment may have underestimated the true reliability since a simple error, such as omitting one file, would have a larger percentage impact on a small total than a large one. Finally, a third opinion might argue that both effects are present, but that they cancel each other out, and therefore the experimental estimates are likely to be highly representative of the reliability of counts of actual systems. Given these competing arguments, validation of the results on larger systems is clearly indicated. Therefore, one parameter for the research design was to test inter-rater reliability using actual average sized application systems.

A second research design question suggested by the Low and Jeffery results, but not explicitly tested by them, is the question of inter-method reliability. Reliability of FP counts was greater within organizations than across them, a result attributed by Low and Jeffery to possible variations in the methods used [Jeffery, 1990]. As discussed earlier, Jones has also suggested the possibility of large differences across methods [Jones, 1989b]. Given the growing proliferation of variant methods this question is also highly relevant to the overall question of FP reliability.

The goal of estimating actual medium-sized application systems requires a large investment of effort on the part of the organizations and individuals participating in the research. Therefore, this constrained the test of inter-method reliability to a maximum of two methods to assure sufficient sample size to permit statistical analysis. The two methods chosen were 1) the International Function Point Users Group (IFPUG) standard Release 3.0, which is the latest release of the original Albrecht method, [Sprouls, 1990] and 2) the Entity-Relationship approach developed by Desharnais [Desharnais, 1988]. The choice of the IFPUG 3.0-Albrecht Standard method (hereafter the “Standard method”) was relatively obvious, as it is the single most-widely adopted approach in current use, due in no small part to its adoption by the over 200 member organization IFPUG. Therefore, there is great practical interest in knowing the inter-rater reliability of this method.

The choice of a second method was less clear cut, as there are a number of competing variations. Choice of the Entity-Relationship (hereafter “E-R”) method was suggested by a second concern often raised by practitioners. In addition to possible concerns about reliability, a second explanation for the reluctance to adopt FPs as a software metric is the perception that FPs are relatively expensive to collect, given the current reliance on labor-intensive methods. Currently, there is no fully automated FP counting system, in contrast to many such systems for the
competing metric, SLOC. Therefore, many organizations have adopted SLOC not due to a belief in greater benefits, but due to the expectation of lower costs in collection. Given this concern, it would be highly desirable for there to be a fully automated FP collection system, and vendors are currently at work attempting to develop such systems. One of the necessary preconditions for such a system is that the design-level data necessary to count FPs be available in an automated format. One promising first step toward developing such a system is the notion of recasting the original FP definitions in terms of the Entity-Relationship model originally proposed by Chen, and now perhaps the most widely used data modeling approach [Teorey, 1990]. Many of the Computer Aided Software Engineering (CASE) tools that support data modeling explicitly support the Entity-Relationship approach, and therefore a FP method based on E-R modeling seems to be a highly promising step towards the total automation of FP collection. Therefore, for all of the reasons stated above, the second method chosen was the E-R approach.\(^3\)

In order to accommodate the two main research questions, inter-rater reliability and inter-method reliability, the research design depicted in Figure 1 was developed, and executed for each system in the dataset.

\(^3\)Readers interested in the E-R approach are referred to [Deshamais, 1988]. However, a brief overview and an example is provided in Appendix B.
For each system \( i \) to be counted, four independent raters from that participating organization were assigned, two of them to the Standard method, and two of them to the E-R method. (Selection of raters and their organizations is further described in Section III, "Data Collection," below.) These raters were identified only as Raters A and B (Standard method) and Raters C and D (E-R method) as shown in Figure 1.

The definition of reliability used in this research is that of Carmines and Zeller, who define reliability as concerning

"the extent to which an experiment, test, or any measuring procedure yields the same results on repeated trials... This tendency toward consistency found in repeated measurements of the same phenomenon is referred to as reliability" [Carmines and Zeller, 1979, pp. 11-12].

Allowing for standard assumptions about independent and unbiased error terms, the reliability of two parallel measures, \( x \) and \( x' \), can be shown to be represented by the simple statistic, \( \rho_{xx'} \) [Carmines and Zeller, 1979]. Therefore, for the design depicted in Figure 1, the appropriate statistics are\(^4\):

\[
\begin{align*}
\rho(FP_{Ai} FP_{Bi}) &= \text{inter-rater reliability for Standard method for System } i \\
\rho(FP_{Ci} FP_{Di}) &= \text{inter-rater reliability for E-R method for System } i \\
\rho(FP_{1i} FP_{2i}) &= \text{inter-method reliability for Standard (1) and E-R (2) methods for System } i.
\end{align*}
\]

While this design neatly addresses both major research questions, it is a very expensive design from a data collection perspective. Collection of FP counts for one medium-sized system was estimated to require 4 work-hours on the part of each rater\(^5\). Therefore, the total data collection cost for each system, \( i \), was estimated at 16 work-hours, or 2 work-days per system. A less expensive alternative would have been to only use 2 raters, each of whom would use one method and then re-count using the second method, randomized for possible ordering effects. Unfortunately, this alternative design would suffer from a relativity bias, whereby raters would tend to remember the answer from their first count, and thus such a design would be likely to produce artificially high correlations [Carmines and Zeller, 1979, ch. 4]. Therefore, the more expensive design was chosen, with the foreknowledge that this would likely limit the number of organizations willing and able to participate, and therefore limit the sample size.

\(^4\)In order to make the subscripts more legible, the customary notation \( \rho_{xx'} \) will be replaced with the parenthetical notation \( \rho(x \times') \).

\(^5\)For future reference of other researchers wishing to duplicate this analysis, actual reported effort averaged 4.45 hours per system.
III. DATA COLLECTION

The pool of raters all came from organizations that are members of the International Function Point Users Group (IFPUG), although only a small fraction of the raters are active IFPUG members. The organizations represent a cross-section of US, Canadian, and UK firms, both public and private, and are largely concentrated in either the Manufacturing or the Finance, Insurance & Real Estate sectors. Per the research agreement, their actual identities will not be revealed. The first step in the data collection procedure was to send a letter to a contact person at each organization explaining the research and inviting participation. The contacts were told that each system would require four independent counts, at an estimated effort of 4 hours per count. Based upon this mailing, 63 organizations expressed interest in the research, and were sent a packet of research materials. The contacts were told to select recently developed medium sized applications, defined as those that required from 1 to 6 work-years of effort to develop. After a follow-up letter, and, in some cases, follow-up telephone call(s), usable data were ultimately received from 27 organizations, for a final response rate of 43%. Given the significant effort investment required to participate, this is believed to be a high response rate, as the only direct benefit promised to the participants was a report comparing their data with the overall averages. The applications chosen are almost entirely interactive MIS-type systems, with the majority supporting either Accounting/Finance or Manufacturing-type applications.

Experimental Controls

A number of precautions were taken to protect against threats to validity, the most prominent being the need to ensure that the four counts were done independently. First, in the instructions to the site contact the need for independent counts was repeatedly stressed. Second, the packet of research materials contained four separate data collection forms, each uniquely labeled “A”, “B”, “C”, and “D” for immediate distribution to the four raters. Third, 4 FP manuals were included, 2 of the Standard method (labeled “Method I”) and 2 of the E-R method (labeled “Method II”). While increasing the reproduction and mailing costs of the research, it was felt that this was an important step to reduce the possibility of inadvertent collusion through the sharing of manuals across raters, where the first rater might make marginal notes or otherwise give clues to a second reader as to the first rater’s count. Fourth and finally, 4 individual envelopes, pre-stamped and pre-addressed to the researcher, were enclosed so that immediately upon completion of the task the rater could place the data collection sheet into the envelope and mail it to the research team in order that no post-count collation by the site contact would be required. Again, this added some extra cost and expense to the research, but was deemed to be an important additional safeguard. Copies of all of these research materials are available in [Connolly, 1990] for other researchers to examine and use if desired to replicate the study.
One additional cost to the research of these precautions to assure independence was that the decentralized approach led to the result that not all four counts were received from all of the sites. Table 1 summarizes the number of sets of data for which analysis of at least one of the research questions was possible.

<table>
<thead>
<tr>
<th>Counts Received:</th>
<th>Systems</th>
<th>Observations</th>
<th>Research Question</th>
</tr>
</thead>
<tbody>
<tr>
<td>A ∧ B</td>
<td>27</td>
<td>54</td>
<td>Standard method Inter-rater reliability</td>
</tr>
<tr>
<td>C ∧ D</td>
<td>21</td>
<td>42</td>
<td>E-R method Inter-rater reliability</td>
</tr>
<tr>
<td>A ∧ B ∧ C ∧ D</td>
<td>17</td>
<td>68</td>
<td>Inter-method reliability (&quot;Quadset&quot;)</td>
</tr>
<tr>
<td>A ∨ B ∧ C ∨ D</td>
<td>26</td>
<td>90</td>
<td>Inter-method reliability (&quot;Fullset&quot;)</td>
</tr>
</tbody>
</table>

Table 1: Summary of primary data collected

In Table 1, the first column shows the type of data. The row labeled “A ∧ B” indicates that data from both the “A” and “B” rater were received. Since both of these raters used the Standard method, the inter-rater reliability for this method can be assessed using these data. The second row is similar, except that it applies to the E-R method. The third row refers to systems for which all four counts were received, and can be used as originally designed to measure inter-method reliability. This set will be referred to as the “Quadset” to indicate that all four counts were present. The fourth row refers to systems for which at least one “A” or “B” count exists and at least one “C” or “D” count exists. These data can also be used to test inter-method reliability, and will be referred to as the “Fullset”. The “Fullset” naturally includes all of the systems in the “Quadset”.

These counts reflect the data after the removal of five systems’ data that was deemed unusable for purposes of the study. Data for two systems were not used as only one count for each system (an “A” in one case and a “D” in the other) were received, and therefore no comparison of any kind could be made. Data for two other systems, one an average of 3,590 FPs, and the other of 2,294 FPs, approximately 9.1 and 5.3 standard deviations above the mean for the inter-rater sample respectively, were also excluded, on the grounds that they reflected large size systems rather than the medium size (1-6 work years) systems requested. Finally, data for a fifth system for which independence of the raters was in doubt were also excluded.

---

6It should be noted that the correlations of the counts for two of these three latter systems were extremely high, and their exclusion in the interests of conservatism has the effect of reducing the overall reliability measures for the dataset.
Table 2 summarizes the data collected in the current research with that of the previous study of inter-rater reliability:

<table>
<thead>
<tr>
<th>Study</th>
<th>Number of Organizations</th>
<th>Total Number of Counts</th>
<th>Unit of Analysis</th>
<th>Countries of Origin</th>
<th>Mean FP Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low &amp; Jeffery</td>
<td>7+</td>
<td>88</td>
<td>Program</td>
<td>Australia</td>
<td>49</td>
</tr>
<tr>
<td>Kemerer</td>
<td>27</td>
<td>111</td>
<td>Application System</td>
<td>US, Canada, UK</td>
<td>450</td>
</tr>
</tbody>
</table>

Table 2: Comparison with Low & Jeffery Data Inter-rater data

Check of Random Assignment

Given that the four raters were assigned to one of the two methods by the site contact, one possible concern might be that the final assignment may have been biased in some way. For example, if raters “A” and “B” had greater FP experience, on average, than raters “C” and “D”, then any comparison of methods would be simultaneously testing the methods hypothesis and a hidden experience hypothesis [Low and Jeffery, 1990]. Given the number of field sites involved, assignment of raters could not be rigorously controlled a priori, other than through the instructions given to the site contact. Therefore, ex post tests of independent variables that could be postulated to have some effect were done, and the results of these tests are presented in Tables 3 and 4 below:

<table>
<thead>
<tr>
<th>Experience Type:</th>
<th>“A” Raters Mean or %</th>
<th>“B” Raters Mean or %</th>
<th>“C” Raters Mean or %</th>
<th>“D” Raters Mean or %</th>
<th>ANOVA F-test, by Rater</th>
<th>ANOVA F-test, by Method</th>
<th>Scheffe Test, α = .10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Systems Development</td>
<td>11.3 yrs.</td>
<td>9.7 yrs.</td>
<td>10.9 yrs.</td>
<td>11.2 yrs.</td>
<td>F=.21 (p=.89)</td>
<td>F=.08 (p=.77)</td>
<td>Negative, all cases</td>
</tr>
<tr>
<td>Function Points</td>
<td>1.3 yrs.</td>
<td>1.5 yrs.</td>
<td>1.7 yrs.</td>
<td>1.7 yrs.</td>
<td>F=.41 (p=.75)</td>
<td>F=.96 (p=.33)</td>
<td>Negative, all cases</td>
</tr>
<tr>
<td>This Application System</td>
<td>6%</td>
<td>19%</td>
<td>15%</td>
<td>13%</td>
<td>F=.76 (p=.52)</td>
<td>F=.08 (p=.78)</td>
<td>Negative, all cases</td>
</tr>
</tbody>
</table>

Table 3: Check of Rater & Method Assignment Randomness, Experience

As shown in the table, the average overall experience of the raters, in terms of their systems development experience, their experience in counting FPs, and the percentage of raters who were involved with the development or maintenance of the system being counted, was relatively consistent across all four groups. The results of one-way ANOVA tests for both rater differences and method differences (where “A” and “B” represent the Standard method, and “C” and “D” represent the E-R method), did not support rejecting the null hypothesis of zero difference between the mean levels of experience. In addition, the Scheffe multiple comparison procedure was run on
the full raters nested within methods model, with the same result that no statistically significant difference was detectable at even the α=.10 level for any of the possible individual (e.g., A vs B, A vs C, A vs D, B vs C...) cases [Scheffe, 1959]. Therefore, later tests of possible methods effects on FP count data will be assumed to have come from randomly assigned raters with respect to relevant experience.

In addition to experience levels, another factor that might be hypothesized to affect FP measurement reliability might be the system source materials with which the rater has to work. As suggested by Rudolph, three levels of such materials might be available: I) requirements analysis phase documentation, II) external design phase documentation (e.g., hardcopy of screen designs, reports, file layouts, etc.), and III) the completed system, which could include access to the actual source code [Rudolph, 1983]. Each of the raters contributing data to this study was asked which of these levels of source materials he or she had access to in order to develop the FP count. The majority of all raters used design documentation ("level II"). However, some had access only to level I documentation, and some had access to the full completed system, as indicated in Table 4. In order to assure that this mixture of source materials level was unbiased with respect to the assigned raters and their respective methods, ANOVA analysis as per Table 3 was done, and the results of this analysis are shown in Table 4.

<table>
<thead>
<tr>
<th>Source Materials Type:</th>
<th>&quot;A&quot; Raters %</th>
<th>&quot;B&quot; Raters %</th>
<th>&quot;C&quot; Raters %</th>
<th>&quot;D&quot; Raters %</th>
<th>ANOVA F-test, by Rater</th>
<th>ANOVA F-test, by Method</th>
<th>Scheffe Test, α = .10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Requirements Analysis Documentation (level I)</td>
<td>11%</td>
<td>6%</td>
<td>14%</td>
<td>14%</td>
<td>F= .37 (p=.78)</td>
<td>F= .82 (p=.37)</td>
<td>Negative, all cases</td>
</tr>
<tr>
<td>Detailed Design documentation (level II)</td>
<td>68%</td>
<td>66%</td>
<td>64%</td>
<td>67%</td>
<td>F= .03 (p=.99)</td>
<td>F= .03 (p=.87)</td>
<td>Negative, all cases</td>
</tr>
<tr>
<td>Completed system (level III)</td>
<td>21%</td>
<td>28%</td>
<td>23%</td>
<td>19%</td>
<td>F= .22 (p=.88)</td>
<td>F= .23 (p=.63)</td>
<td>Negative, all cases</td>
</tr>
</tbody>
</table>

Table 4: Check of Rater & Method Assignment Randomness, Materials

Similar to the results for experience levels, it appears that access to source materials was sufficiently similar for each rater group as to rule this out as a probable source of bias. Therefore, later tests of possible methods effects on FP count data will be assumed to have come from randomly assigned raters with respect to source material.
IV. MAIN RESEARCH RESULTS

A. Introduction

1. Statistical Test Power Selection

An important research parameter to be chosen in building specific research hypotheses from the general research questions is the tradeoff between possible Type I errors (mistakenly rejecting a true null hypothesis) and Type II errors (accepting a false null hypothesis). Typically, the null hypothesis represents “business as usual”, and is uninteresting from a practical point of view [Cohen, 1977]. That is to say managers would be moved to change their actions only if the null hypothesis were rejected. For example, a test is suggested to see whether use of a new tool improves performance. The null hypothesis is that the performance of the group using the new tool will not differ from that of the group that does not have the new tool. If the null hypothesis is not rejected, nothing changes. However, if the null hypothesis is rejected, the implication (assuming the tool users’ performance was significantly better, not significantly worse) is that the tool had a positive effect. Given this type of null hypothesis, the researcher is most concerned about Type I errors, since a Type I error would imply that managers would adopt a tool, presumably at some extra expense, that provided no benefit. In order to guard against Type I errors, a high confidence level (1 - α) is chosen. This has the effect of increasing the likelihood of Type II errors, or, stated differently, reducing the power of the test (1 - β).

The current research differs from the canonical case described above in that the null hypothesis is “substantial” rather than “uninteresting”. That is to say that there are actions managers can take if the null hypothesis is believed to be true. For example, if the null hypothesis H0: FPa = FPb is believed to be true, then managers’ confidence in using FPs as a software metric is increased, and some organizations that may currently not be using FPs will be encouraged to adopt them. Researchers might also have greater willingness to adopt FPs, as they may interpret these results to mean that inter-rater reliability is high, and therefore only single counts of systems may be necessary, and that all systems may not have to be counted by the same rater. Similarly, if the null hypothesis H0: FP1 = FP2 is believed to be true, then organizations who had previously adopted the Standard method may no longer have reason to be reluctant to adopt the E-R method.

Therefore, given that the null hypotheses are substantial, and that a Type II error is of greater than usual concern, α will be set to .1 (90% confidence level) in order to reduce the probability of a Type II error and to raise the power of the test to detect differences across raters and across methods. Compared to a typical practice of setting α = .05, this slightly raises the possibility of a Type I error of claiming a difference where none exists, but this balance is believed to be appropriate for the research questions being considered.


2. Magnitude of Variation Calculation
The above discussion applied to the question of statistical tests, those tests used to determine whether the notion of no difference can be rejected, based upon the statistical significance of the results. For practitioners, data that would be additionally useful is the estimated magnitude of the differences, if any. For example, one method might produce counts that are always higher, and this consistency could allow the rejection of the null hypothesis, but the magnitude of the difference might be so small as to be of no practical interest. This question is of particular relevance in the case of using FPs for project estimating, as the size estimate generated by FPs is a good, but imperfect predictor of the final costs of a project. Therefore, small differences in FP counts would be unlikely to be managerially relevant, given the “noise” involved in transforming these size figures into dollar costs.

As the reliability question has been the subject of very little research, there does not seem to be a standard method in this literature for representing the magnitude of variation calculations. In the more common problem of representing the magnitude of estimation errors versus actual errors, the standard approach is to calculate the Magnitude of Relative Error (MRE), as follows:

\[ \text{MRE} = \left| \frac{x - \hat{x}}{x} \right| \]

where \( x \) = the actual value, and \( \hat{x} \) = the estimate. This proportional weighting scales the absolute error to reflect the size of the error in percentage terms, and the absolute value signs protect against positive and negative errors cancelling each other out in the event that an average of a set of MREs is taken [Conte, Dunsmore and Shen, 1986].

In the current case, there is no value for \( x \), only two estimates, \( \hat{x} \) and \( \hat{y} \). Therefore, some variation of the standard MRE formula will be required. A reasonable alternative is to substitute the average value for the actual value in the MRE formula, renaming it the Average Relative Error (ARE):

\[ \text{ARE} = \left| \frac{\hat{x} \hat{y} - \hat{x}}{\hat{x} \hat{y}} \right| \]

where \( \hat{x} \hat{y} = \frac{\hat{x} + \hat{y}}{2} \).

The ARE will be shown for each comparison, to convey a sense of the practical magnitudes of the differences as a complement to the other measures that show their relative statistical significance.
B. Inter-rater reliability results

1. Standard method \[ H_0: \overline{FP}_A = \overline{FP}_B \]

Based on the research design described earlier, the average value for the “A” raters was 436.36, and for the “B” raters it was 464.02, with \( n = 27 \). The results of a test of inter-rater reliability for the standard method yielded a Pearson correlation coefficient of \( \rho = .80 \), (\( p=.0001 \)), suggesting a strong correlation between FP counts of two raters using the standard method. The results of a paired \( t \)-test of the null hypothesis that the difference between the means is equal to 0 was only -0.61 (\( p=.55 \)), indicating no support for rejecting the null hypothesis. The power of this test for revealing the presence of a large difference, assuming it were to exist, is approximately 90\% [Cohen, 1977, Table 2.3.6]\(^7\). Therefore, based on these results, there is clearly no statistical support for assuming the counts are significantly different. But, also of interest is the average magnitude of these differences. The average ARE is equal to 10.78\%, a difference which compares quite favorably to the approximately 30\% differences reported by Rudolph, and Low and Jeffrey [Rudolph, 1983, p. 6] [Low and Jeffery, 1990, p. 71]. This suggests that, at least for the Standard method, inter-rater reliability of multiple FP raters using the same standard is high, with an average difference that is likely to be quite acceptable for the types of estimation and other tasks to which FPs are commonly applied. Further discussion of the conclusions from these results will be presented below, after presentation of the results of the tests of the other hypotheses.

2. Entity-Relationship method \[ H_0: \overline{FP}_C = \overline{FP}_D \]

The same set of tests was run for the two sets of raters using the E-R method, *mutatis mutandis*. For an \( n = 21 \), values of \( \overline{FP}_C \) and \( \overline{FP}_D \) were 476.33 and 411.00 respectively. Note that these values are not directly comparable to the values for \( \overline{FP}_A \) and \( \overline{FP}_B \), as they come from different samples. The reliability measure is \( \rho(\overline{FP}_C, \overline{FP}_D) = .74 \) (\( p=.0001 \)), not quite as high as for the Standard method, but nearly as strong a correlation. The results of an equivalent \( t \)-test yielded a value of 1.15 (\( p=.26 \)), again indicating less reliability than the Standard method, but still well below the level where the null hypothesis of no difference might be rejected. The power of this test is approximately 82\%. The value of ARE was 18.13\%, also not as good as that for the Standard method, but still clearly better than the oft-quoted 30\% figure.

In general, from all of these tests it can be concluded that, on average, the reliability of FP counts obtained with the E-R method, though more reliable than previous speculation, are currently not as reliable as those obtained using the Standard method, at least as reflected by the data from this

\(^7\)All later power estimates are also from this source, *loc. cit.*
experiment. Some suggestions as to why this might be the case will be presented in the discussion of results section below.

C. Inter-method reliability results

1. Quadset analysis \((n = 17)\)

The test of inter-method reliability is a test of the null hypothesis:

\[ H_0: \overline{FP}_1 = \overline{FP}_2 \]

where \(\overline{FP}_1 = \sum_{i=1}^{n} \frac{FP_{Ai} + FP_{Bi}}{2} \) and \(\overline{FP}_2 = \sum_{i=1}^{n} \frac{FP_{Ci} + FP_{Di}}{2} \)

At issue here is whether FP raters using two variant FP methods will produce highly similar (reliable) results, in this particular case the two methods being the Standard method and the E-R method. In the interests of conservatism, the first set of analyses uses only the 17 systems for which all four counts, A, B, C, and D, were obtained. This is to guard against the event, however unlikely, that the partial response systems were somehow different. The values for \(\overline{FP}_1\) and \(\overline{FP}_2\) were 417.63 and 412.92, respectively, and yielded a \(\rho(\overline{FP}_{1i}; \overline{FP}_{2i}) = .95\) \((p=.0001)\). The \(t\)-test of the null hypothesis of no difference resulted in a value of \(.18\) \((p=.86)\), providing no support for rejecting the hypothesis of equal means. The ARE for this set was 8.48%. These results clearly speak to a very high inter-method reliability. However, the conservative approach of only using the Quadset data yielded a smaller sample size, thus reducing the power of the statistical tests. For example, the relative power of this \(t\)-test is 74%. To increase the power of the test in order to ensure that the results obtained above were not simply the result of the smaller sample, the next step replicates the analysis using the Fullset data, those for which at least one count from the Rater A and B method and at least one count from the Rater C and D method were available.

2. Fullset analysis \((n = 26)\)

The results from the Fullset analysis are somewhat less strong than the very high values reported for the Quadset, but they also show high correlation, and since the Fullset test has greater power to detect differences, should they exist, greater confidence can be placed in the result of no difference. The values of \(\overline{FP}_1\) and \(\overline{FP}_2\) were 403.39 and 363.04, respectively, and yielded a \(\rho(\overline{FP}_{1i}; \overline{FP}_{2i}) = .84\) \((p=.0001)\). The \(t\)-test of the null hypothesis was 1.25 \((p=.22)\), with a power of 89%. The ARE was 12.17%. Thus, it is still appropriate not to reject the null hypothesis of no difference across these two methods, and, based on the Fullset analysis, not rejecting the null hypothesis can be done with increased confidence.
D. Analysis and discussion of research results

All of the primary research results are summarized in Table 5 below:

<table>
<thead>
<tr>
<th></th>
<th>Inter-rater, Standard method</th>
<th>Inter-rater, E-R method</th>
<th>Inter-method, Quadset</th>
<th>Inter-method, Fullset</th>
</tr>
</thead>
<tbody>
<tr>
<td>n (Systems, Counts)</td>
<td>27, 54</td>
<td>21, 42</td>
<td>17, 68</td>
<td>26, 90</td>
</tr>
<tr>
<td>$p_{xy} : p$</td>
<td>.80 (.0001)</td>
<td>.74 (.0001)</td>
<td>.95 (.0001)</td>
<td>.84 (.0001)</td>
</tr>
<tr>
<td>paired t-test; p</td>
<td>-.61 (.55)</td>
<td>1.15 (.26)</td>
<td>.18 (.86)</td>
<td>1.25 (.22)</td>
</tr>
<tr>
<td>$1 - \beta$ (power)</td>
<td>.90</td>
<td>.82</td>
<td>.74</td>
<td>.89</td>
</tr>
<tr>
<td>Avg. Relative Error</td>
<td>10.78%</td>
<td>18.13%</td>
<td>8.48%</td>
<td>12.17%</td>
</tr>
</tbody>
</table>

Table 5: Summary of Reliability Statistics

Based upon these data, the inter-rater reliability of FP measurements using the Standard method can be treated by managers and researchers as relatively high, with an average approximate error of +/- 10.78%. This is well under previous reports of +/- 30% error on counts developed from small program specifications, and is similar to that predicted by Rudolph for raters using the same method and having access to detailed design specifications or completed systems [Rudolph, 1983]. Two possible factors are at work here: 1) the use of actual medium-sized systems, and 2) the use of detailed design documents or completed systems rather than requirements analysis documents only.

The influence of access to detailed design documents rather than requirements analysis specifications is difficult to isolate in this study, since so few raters had access to only requirements analysis documents. However, if counts using requirement analysis documents only were truly less reliable than those with level II or later documentation, then exclusion of the counts based on requirements analysis documentation only should improve the average reliability of the remainder of the sample. The results of this sensitivity analysis for the replication of the A vs B inter-rater reliability calculation minus the 3 pairs (of the original 27) using only requirements analysis documents generated a $p = .79$, and an ARE of 11.54 %, which, being somewhat worse rather than better than those obtained using all of the data, do not tend to support an argument based upon access to more detailed source documents. Rather, it is suggested that the higher reliability score reflects the use of actual medium-sized systems, where small errors are less important, on a percentage basis, than they would be on counts of small programs. It is also suggested that the results obtained from these data are more likely to reflect the experience of the use of FPs in practice, since they were obtained from counts of actual systems. Therefore, in assessing the reliability of FP counts in practice, based on the data from this study and in the absence of additional information, managers can assume an average variation across raters of approximately +/- 10.78%.
The inter-rater error for the E-R method, while almost 50% better than that suggested by previous authors, was still almost twice that of the Standard method. There are a number of possible explanations for this difference. The first, and easiest to check is whether the slightly different samples used in the analysis of the two methods (the 27 systems used by the Standard method and the 21 systems used by the E-R method) may have influenced the results. To check this possibility, both sets of analyses were re-run, using only the Quadset of 17 systems for which all four counts were available. This sub-analysis generated an ARE= 10.37% and a \( \rho = .79 \) for the Standard method, and an ARE=16.41% and a \( \rho = .73 \) for the E-R method, so it appears as if the difference cannot simply be attributed to a sampling difference.

More likely explanations stem from the fact that the E-R approach, while perhaps the most common data modeling approach in current use, is still sufficiently unfamiliar as to cause errors. Of the raters contributing data to this study, 23% of the C and D raters reported having no prior experience or training in E-R modeling, and thus were relying solely upon the manual provided. Thus, the comparison of the Standard and E-R methods results shows the combined effects of both the methods themselves, and their supporting manuals. Therefore, the possibility of the test materials, rather than the method per se, being the cause of the increased variation cannot be ruled out by this study\(^8\).

The inter-method results are the first documented study of this phenomenon, and thus provide a baseline for future studies. The variation across the two methods (ARE=12.17%) is similar to that obtained across raters, and thus does not appear to be a major source of error for these two methods. Of course, these results cannot necessarily be extended to pairwise comparisons of two other FP method variations, or even of one of the current methods and a third method. Determination of whether this result represents typical, better, or worse effects of counting variations must await further validation. However, as a practical manner, the results should be encouraging to researchers or vendors who might automate the E-R method within a tool, thus at least partially addressing both the reliability concerns and the data collection costs. The results also suggest that organizations choosing to adopt the E-R method, although at some risk of likely lower

\(^8\)An additional hypothesis has been suggested by Allan Albrecht. He notes that the E-R approach is a user functional view of the system, a view that is typically captured in the requirements analysis documentation, but sometimes does not appear in the detailed design documentation. To the degree that this is true, and to the degree that counters in this study used the detailed design documentation to the exclusion of using the requirements analysis documents, this may have hindered use of the E-R method[Albrecht, 1990]. A similar possibility suggested by some other readers is that the application system’s documentation used may not have contained E-R diagrams, thus creating an additional intermediate step in the counting process for those raters using the E-R method, which could have contributed to a greater number of errors and hence a wider variance.
inter-rater reliability, are likely to generate FP counts that are sufficiently similar from counts obtained with the Standard method so as to be a viable alternative. In particular, an analysis of the Quadset data revealed a mean FP count of 417.63 for the Standard method and 412.92 for the E-R method, indistinguishable for both statistical and practical purposes.

V. CONCLUDING REMARKS

If software development is to fully establish itself as an engineering discipline, then it must adopt and adhere to the standards of such disciplines. A critical distinction between software engineering and other, more well-established branches of engineering is the clear shortage of well-accepted measures of software. Without such measures, the managerial tasks of planning and controlling software development and maintenance will remain stagnant in a ‘craft’-type mode, whereby greater skill is acquired only through greater experience, and such experience cannot be easily communicated to the next project for study, adoption, and further improvement. With such measures, software projects can be quantitatively described, and the managerial methods and tools used on the projects to improve productivity and quality can be evaluated. These evaluations will help the discipline grow and mature, as progress is made at adopting those innovations that work well, and discarding or revising those that do not.

Currently, the only widely available software metric that has the potential to fill this role in the near future is Function Points. The current research has shown that, contrary to some speculation and to the limited prior research, the inter-rater and inter-method reliability of FP measurement are sufficiently high that their reliability should not pose a practical barrier to their continued and further adoption.

The collection effort for FP data in this research averaged approximately 1 work-hour per 100 FPs, and can be expected to be indicative of the costs to collect data in actual practice, since the data used in this research were real world systems. For large systems this amount of effort is non-trivial, and may account for the relative paucity of prior research on these questions. Clearly, further efforts directed towards developing aids to greater automation of FP data collection should continue to be pursued. However, even the current cost is small relative to the large sums spent on software development and maintenance in total, and managers should consider the time spent on FP collection and analysis as an investment in process improvement of their software development capability. Such investments are also indicative of true engineering disciplines, and there is increasing evidence of these types of investments in leading edge software firms in the United
States and in Japan [Cusumano and Kemerer, 1990]. Managers wishing to quantitatively improve their software development and maintenance capabilities should adopt or extend software measurement capabilities within their organizations. Based upon the current research, FPs seem to offer a reliable yardstick with which to implement this capability.  

Research support from the International Function Point Users Group (IFPUG) and the MIT Center for Information Systems Research is gratefully acknowledged. Helpful comments on the original research design and/or earlier drafts of this paper were received from A. Albrecht, N. Campbell, J. Cooprider, B. Dreger, J. Henderson, R. Jeffery, C. Jones, W. Orlikowski, D. Reifer, H. Rubin, E. Rudolph, W. Rumpf, G. Sosa, C. Symons, and N. Venkatraman. Provision of the data was made possible in large part due to the efforts of A. Belden and B. Porter, and the organizations that contributed data to the study. Special thanks are also due my research assistant, M. Connolley.
A. FUNCTION POINTS CALCULATION APPENDIX

Readers interested in learning how to calculate Function Points are referred to one of the fully documented methods, such as the IFPUG Standard, Release 3.0 [Sprouls, 1990]. The following is a minimal description only. Calculation of Function Points begins with counting five components of the proposed or implemented system, namely the number of external inputs (e.g., transaction types), external outputs (e.g., report types), logical internal files (files as the user might conceive of them, not physical files), external interface files (files accessed by the application but not maintained, i.e., updated by it), and external inquiries (types of on-line inquiries supported). Their complexity is classified as being relatively low, average, or high, according to a set of standards that define complexity in terms of objective guidelines. Table A.1 is an example of such a guideline, in this case the table used to assess the relative complexity of External Outputs, such as reports:

<table>
<thead>
<tr>
<th>1 - 5 Data Element Types</th>
<th>6-19 Data Element Types</th>
<th>20+ Data Element Types</th>
</tr>
</thead>
<tbody>
<tr>
<td>0-1 File Types Referenced</td>
<td>Low</td>
<td>Low</td>
</tr>
<tr>
<td>2-3 File Types Referenced</td>
<td>Low</td>
<td>Average</td>
</tr>
<tr>
<td>4+ File Types Referenced</td>
<td>Average</td>
<td>High</td>
</tr>
</tbody>
</table>

Table A.1: Complexity Assignment for External Outputs [Sprouls, 1990]

To use this table in counting the number of FPs in an application, a report would first be classified as an External Output. By determining the number of unique files used to generate the report ("File Type Referenced"), and the number of fields on the report ("Data Element Types"), it can be classified as a relatively Low, Average, or High complexity External Output. After making such determinations for each of the five component types, the number of each component type present is placed into its assigned cell next to its weight in the matrix shown in Table A.2. Then, the total number of function counts (FCs) is computed as shown in equation (1).
<table>
<thead>
<tr>
<th></th>
<th>Low</th>
<th>Average</th>
<th>High</th>
</tr>
</thead>
<tbody>
<tr>
<td>External Input</td>
<td>x3</td>
<td>x4</td>
<td>x6</td>
</tr>
<tr>
<td>External Output</td>
<td>x4</td>
<td>x5</td>
<td>x7</td>
</tr>
<tr>
<td>Logical Internal File</td>
<td>x7</td>
<td>x10</td>
<td>x15</td>
</tr>
<tr>
<td>External Interface File</td>
<td>x5</td>
<td>x7</td>
<td>x10</td>
</tr>
<tr>
<td>External Inquiry</td>
<td>x3</td>
<td>x4</td>
<td>x6</td>
</tr>
</tbody>
</table>

Table A.2: Function Count Weighting Factors

\[(1) \text{FC} = \sum_{i=1}^{5} \sum_{j=1}^{3} w_{ij}x_{ij}\]

where \(w_{ij}\) = weight for row \(i\), column \(j\) and \(x_{ij}\) = value in cell \(i, j\).

The second step involves assessing the impact of fourteen general system characteristics that are rated on a scale from 0 to 5 in terms of their likely effect for the system being counted. These characteristics are: 1) data communications, 2) distributed functions, 3) performance, 4) heavily used configuration, 5) transaction rate, 6) on-line data entry, 7) end user efficiency, 8) on-line update, 9) complex processing, 10) reusability, 11) installation ease, 12) operational ease, 13) multiple sites, and 14) facilitates change. These values are then summed and modified to compute the Value Adjustment Factor, or VAF:

\[(2) \text{VAF} = .65 + .01 \sum_{i=1}^{14} c_i\]

where \(c_i\) = value for general system characteristic \(i\), for \(0 <= c_i <= 5\).

Finally, the two values are multiplied to create the number of Function Points (FP):

\[(3) \text{FP} = \text{FC} \times \text{VAF}.\]
B. ENTITY-RELATIONSHIP APPROACH SUMMARY APPENDIX

The following material is excerpted directly from the materials used by Raters C and D in the experiment, and highlights the general approach taken in the E-R approach to FP counting. Readers interested in further details regarding the experimental materials should see [Connolley, 1990], and for further detail regarding the E-R approach see [Desharnais, 1988].

"This methodology's definition of function point counting is based on the use of logical models as the basis of the counting process. The two primary models which are to be used are the "Data-Entity-Relationship" model and the "Data Flow Diagram." These two model types come in a variety of forms, but generally have the same characteristics related to Function Point counting irrespective of their form. The following applies to these two models as they are applied in the balance of this document.

**Data Entity Relationship Model (DER).** This model typically shows the relationships between the various data entities which are used in a particular system. It typically contains "Data Entities" and "Relationships", as the objects of interest to the user or the systems analyst. In the use of the DER model, we standardize on the use of the "Third Normal Form" of the model, which eliminates repeating groups of data, and functional and transitive relationships. ... Data Entity Relationship models will be used to identify Internal Entities (corresponding to Logical Internal Files) and External Entities (corresponding to Logical External Interfaces).

**Data Flow Diagrams (DFD).** These models typically show the flow of data through a particular system. They show the data entering from the user or other source, the data entities which are used, and the destination of the information out of the system. The boundaries of the system are generally clearly identified, as are the processes which are used. This model is frequently called a "Process" model. The level of detail of this model which is useful is the level which identifies a single (or small number) of individual business transactions. These transactions are a result of the decomposition of the higher level data flows typically at the system level, and then at the function and sub-function level. Data Flow Diagrams will be used to identify the three types of transactions which are counted in Function Point Analysis (External Inputs, External Outputs and Inquiries)."

The following is an example of the documentation provided to count one of the five function types, Internal Logical Files.
Internal Logical Files

"Definition. Internal entity types are counted as Albrecht's internal file types. An entity-type is internal if the application built by the measured project allows users to create, delete, modify and/or read an implementation of the entity-type. The users must have asked for this facility and be aware of it. All attributes of the entity-type, elements that are not foreign keys, are counted. We also count the number of relation types that the entity-type has. The complexity is determined by counting the number of elements and the number of relationships:

<table>
<thead>
<tr>
<th>1 Relationship or other Entity Type</th>
<th>1 - 19 Data Attribute Types in the Entity</th>
<th>20-50 Data Attribute Types in the Entity</th>
<th>51+ Data Attribute Types in the Entity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low</td>
<td>Low</td>
<td>Average</td>
<td></td>
</tr>
<tr>
<td>Low</td>
<td>Average</td>
<td>High</td>
<td></td>
</tr>
<tr>
<td>Average</td>
<td>High</td>
<td>High</td>
<td></td>
</tr>
</tbody>
</table>

Table B.1: Complexity Assignment for Internal Logical Files, E-R Method

Guidance

- Entities updated by application are counted as logical internal files.
- Complexity is based on the number of relationships in which the entity participates as well as the number of DET's.
- When considering an Entity-Relationship chart, be sure to consider the real needs of the application. For instance, frequently attributes required are attributes of the relationship rather than the entities, thus requiring a concatenated key to satisfy the requirement. The related entities may or may not be required as separate USER VIEWS."
BIBLIOGRAPHY


Boehm, B., "Improving Software Productivity", Computer, 43-57, (September 1987).


