

# To Evaluate the Function Point Analysis: A Case Study

**TSOI Ho-Leung**

Software Quality Institute  
Griffith University  
Email:tsoihl@yahoo.com

## **Abstract**

*Nowadays Function Point Analysis (FPA) model is widely used by academic researchers and growing acceptance in practice. Low and Jeffery claim that experience of software development will be a factor to influence function point estimates. However, the study had concentrated on identification the variation in function point estimates between experience and inexperience programmers but lack of further investigation the constitution. . This paper presents an empirical study that is designed to strengthen the understanding of the reliability of function point measurement.*

*Keywords: Function Point Analysis (FPA) model, software estimation*

## **1 INTRODUCTION**

Function Point Analysis (FPA) model has been proposed in 1979 by Allan Albrecht of IBM [Albrecht79] and has been further refined by Albrecht and Gaffney in 1983 [Albrecht83]. The basic concept of FPA is quite simple and is based on the number of “Functions” the software has to be developed. All such “Functions” are related to the types of data which the software uses and generates. As FP counts appear easy to calculate, Albrecht and many researchers believe that FPA model can be understood

and evaluated by relatively non-technical project team members [Albrecht84, Kemerer87, Symons88, Sumner99].

There are only a limited number of well-documented studies of this issue [Jeffery90]. Low and Jeffery claim that experience of software development will be a factor to influence function point estimates. However, the study had concentrated on identification the variation in function point estimates between experience and inexperience programmers but lack of further investigation the constitution of the variation. Low's findings suffer from the disadvantage that they consider the FPA as a single entity. In fact, FPA is made up of two parts, Unadjusted Function Point (UFP) and Technical Complexity Factor (TCF), which are not same functionality and assessment method. This paper presents an empirical study which is designed to strengthen the understanding of the reliability of function point measurement, building on the base started by Low and Jeffery. In short, this paper presents an experiment which aims at

- Are UFP counts and Complexity Factors equally contribution to the variation for the technical experience subjects?
- Do technical and non-technical experience subjects use same amount of time to find out the function point counting?

To study these problems, an experiment, involving 21 subjects, was conducted. They have been divided into two different groups, namely technical and non-technical. Each subject has to apply FPA to estimate the size of a real-life application. As part of this experiment, we have developed some intensive guidelines and support for a FPA counting process.

Section 2 of this paper describes the concept of the FPA model. Section 3 introduces the research background and limitations of the experiment. Section 4 presents the results of the experiment at the final phase of counting process. The result shows that FPA can be evaluated by non-technical team members but their FPA counts are significant different and the efficiency is very low. Besides, the interrater reliability of FP counts for relatively non-technical group is very low. Moreover, the results show that most the technical experience subjects face the problem of rating the TCF. Finally, in section 5, the summary and future study directions are discussed.

## 2 FUNCTION POINT ANALYSIS MODEL

Function Point Analysis (FPA) was introduced by Albrecht in 1979 and has been developed further since then [Albrecht79, Albrecht83]. This metric first measure of functionality involved by counting the number of unique input types, output files, logical files, and external queries included by software program. These counts are then weighted depending upon difficulty and further modified by 14 “complexity factors” defined by Albrecht. It makes that these metrics can capture the magnitude and complexity of the analysis and design task of various projects.

In summary, The FPA model is shown as below:

$$\text{Unadjusted Function Point} = \sum_x F_x(C_x)$$

$$\text{Total degree of influence (TDI)} = \sum D_j$$

$$\text{The value adjustment factor (VAF)} = (0.01 * \text{TDI} + 0.65)$$

where

$D_j$ : 14 general information processing adjustments

$F_x(C_x)$ : raw function point measure for the  $x$ th component

FPA = Unadjusted Function Point \* The degree of influence (DI)

$$= \sum_x F_x(C_x) * (0.01 * \text{TDI} + 0.65)$$

$$= \sum_x F_x(C_x) * \text{VAF}$$

The calculation processing for  $F_x(C_x)$  takes place according to table 1 and 2, which show the component types, the items whose counts determine the raw function points of components, the class intervals for those item counts, and the weights for low(L), average(A) and high(H) function point levels for each component type. Readers interested in learning how to calculate Function Points are suggested to refer the IFPUG Standard [IFPUG94].

## 3 MOTIVATION

FPA model has been chosen for investigation because it is a popular software sizing model in the information systems (IS) community. Dreger claims that around 500 major cooperates worldwide are using FPA model [Dreger89]. Nowadays FPA is widely used by academic researchers and is growing in acceptance in practice. Thus more understanding of the way to apply this model is important and necessary.

Low and Jeffery claim that experience of software development will be a factor to influence function point estimates. However, the study had concentrated on identification the variation in function point estimates between experience and inexperience programmers but lack of further investigation the constitution. Low's findings suffer from the disadvantage that they consider the FPA as a single entity. In fact, FPA is made up of two parts which are not same functionality and assessment requirement. The main objective of this study is to advance the understanding raising from Low and Jeffery. For the variation between technical and non-technical subjects, one of the following issues may be the case of the constitution

- i. variation is mainly caused by the rating of UFC
- ii. variation is mainly caused by the rating the TCF
- iii. variation is mainly caused by both the UFC and TCF

For the case 1 and 2, corrective actions just focus on the usage of either UFC or TCF. For the last case, however, it shows that the overall quality of the FPA method is not great and need to be re-considered whether it is advisable to be used as an estimation tool. This paper presents an empirical study which is mainly designed to clarify this point.

#### **4. RESEARCH BACKGROUND AND LIMITATIONS**

Software sizing is a complex human endeavor and FPA is the most proper used software sizing model which quantifies the size and complexity of a software system in terms of the functions that the system will be produced. Most previous research studies are focus on the accuracy and reliability of the FPA model [Kemerer87], but lack of

discussion the constitution of variation and reasons causing this problem. The authors believe that no such kind of investigation can be completed without an empirical investigation.

In general, investigating the reasons causing the low interrater reliability should consider the actual estimation from large-scale software projects. However, there are a lot of factors will affect the estimation process in "large-scale" projects. Most of them, such as time constraints, political, etc., are difficult to control. Any one of the uncontrolled factors can mask the effects of the factors under study. In addition, one serious problem faced by researchers today is the lack of record-keeping mechanisms for collecting on-going development data during the project. The seriousness of this problem has been emphasized by many researchers [Jones86].

Conduct controlled experiments involving non-trivial programming tasks can be considered as another method to investigate of the estimation process. Our research approach has been to conduct controlled experiments involving a real-life programming task to student programmers. It is important to aware the benefits and limitations of such empirical studies conducted at semi-real-life environments. On the negative side, generalization from such studies should be quite limited. Because of some artificial controls, results derived under controlled conditions may not be totally reflection the real-life situation. Fortunately, the objective of this experiment is to identify the difference between technical and non-technical experience people in using FPA model. Thus academic students are suitable to be employed in this experiment. Moreover, on the positive side, rigorous controls allow researchers to investigate thoroughly the effects of experimentally-manipulated

factors and identify possible cause-and-effect relationships among them.

Most researchers agree that results obtained from well-controlled experiments can be used to gain a better understanding of the isolated factor being tested. In addition, the insights obtained from these studies can form a firm basis for studying different kind of projects in real-life environments. Therefore we believed that an in-house empirical study is an essential first step before further investigation. This can provide more evidence to support later real world application.

#### **4.1 RESEARCH METHODOLOGY**

The exploratory study was undertaken within the Computer Science Department of a medium-sized private post secondary Christian college, Zeta (a pseudonym). Zeta was established in 1986 and was a non-profit-making school registered with the Education Department of Hong Kong. The College observes the Christian Principle of “not to be served, but to serve” in providing quality education services to students and working youths. The Computer Science department has 7 teaching staff graduating from various Computing majors, such as Computer Science, Computer Engineering, and Information Systems.

The participants in this experiment were fifteen full-time computer science students enrolled in a higher diploma course at Zeta. They were computer science major and had previously taken several programming and software engineering courses. All these students are classified as non-working subjects. Beside, six people with several years programming or analysis experience (max. 12 years) are classified as working subjects. They are working as analysis or programmers in MIS department associated with Commerce and Industry. In this experiment, the number of non-technical

groups is larger than technical groups. We believe that the difference for non-technical groups is quite higher than technical groups. Thus large number of non-technical groups will be more clear to illustrate the weakness points, such as low interrater reliability.

It is important to note that all subjects in this experiment lack of prior experience in using the FPA estimating technique but were assigned to use Albrecht Function Point (FPA) to evaluate the size of a real software system. This evaluation is based on a real commercial software development tender. Moreover, the FPA manual [IFPUG94] is provided and they have been given long enough time, around four weeks, to complete this evaluation so that the experimental results would be meaningful. Several precautions were taken to protect against threats to validity, the most prominent being the need to ensure that all subjects fully understand the system requirement specification and the model used. To understand of applying FPA estimating technique, several intensive training sessions were held prior the experiment. Besides, sufficient interview sessions were arranged to answer the problems from estimators during experiment.

#### **4.2 DEVELOPMENT PROCESS:**

It was a very important for the subjects in this experiment not only to find out the size of the application and the estimation attribute, but also to follow the prescribed experimental procedures in estimation process. They were strongly requested to do their best to follow these guidelines even if the procedures were not their favorite estimation style. The estimation process in this experiment was considered as a two-stage procedure:

(i) *Counting Stage*

During this stage, all participants were requested and supposed to read the program specification carefully until they had a very good understanding what was required of this program. Several open seminars were offered in order to let them more understanding the FPA estimating technique. That is, in this stage, they should identify the basic functionality components of this program. At the end of this stage, they count the raw unadjusted numbers, such as Unadjusted Function Points counts, used by FPA model. Such results were referred to as an “error-free” version.

(ii) *Evaluation Stage*

During this stage, all participants were requested to evaluate the adjustment factors base on good understanding the programming specification and rating definitions. To investigate the subjectivity issue, confidence levels of the final rated values are recorded in their reports.

All participants were asked to keep an accurate record and detailed explanation in their report. It ensures all counting process will be controlled.

### 4.3 COMPARISON OF TECHNICAL AND NON-TECHNICAL FUNCTION POINT COUNTERS

All participants were classified as either technical or non-technical counters to estimate the number of function points for the application. The findings are presents in this section.

***FINDING 1: There exists a variation in function point estimates between technical and non-technical estimators.***

***True.*** The results presented in Table 3 provide general support this conclusion. The average function point counts for technical and non-technical subjects are 43.05 and 101.29, respectively. The difference between these two subjects is 58.24 FPCs and more than one time to those of the technical subject. This large difference adds validity to this study's findings. Moreover, the average function point counts of the technical subject is close to the actual FP counts (39.36 FPCs, where UFP=41 and TCF=0.96).

Table 3: Function Point Counts Result of this Experiment

	No. of participants	Unadjusted Function Points	14 System Characteristics	Estimated No. of Function Points (Mean)	Std. Dev.	Std. Dev. /Mean	Range	Time Used (Hours)
Non-technical	15	102.86	34.9	101.29	31.43	31.04%	49 -148	30.7
Technical	6	42.33	36	43.05	7.63	17.95%	30 - 48.8	17.5

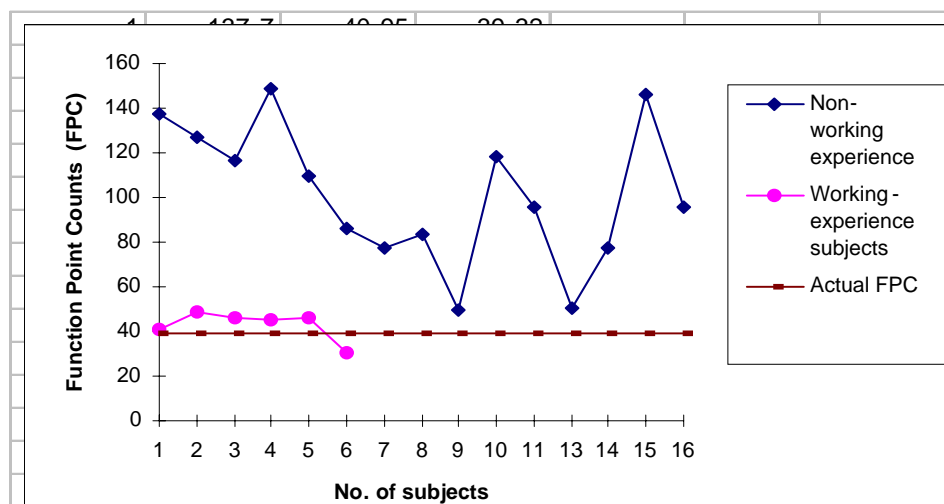


Figure 1: Function Point Counts for Technical and Non-Technical Groups

In order to analyze the variation between technical and non-technical participants, *t*-test approach is adopted in this paper. Let  $\mu_1$  and  $\mu_2$  be the means of the function point counts of the non-technical and technical estimators, respectively. Two hypotheses were established for the statistical analysis:

$$H_0 : \mu_1 - \mu_2 = 0 \quad \text{and}$$

$$H_1 : \mu_1 - \mu_2 > 0$$

Some statistical measurements for the two groups of characteristics are given as below:

$\mu_1 = 101.29$  and the standard derivation  $\sigma_1 = 31.43$  for the non-technical estimators  
 $\mu_2 = 43.05$  and the standard derivation  $\sigma_2 = 7.63$  for the technical estimators

The realization of statistics *t* is 12.54. For a significant level of  $\alpha = 0.05$  and the degree of freedom  $DI = 19$ , we have  $t_{19}(0.95) = 1.7291$ . Hence we conclude that the difference between the two groups of characteristic is significant and reject  $H_0$ .

There appears to be a significant difference in function point estimates between technical and non-technical estimators. In other word, working experience background influences function point estimates.

Another interesting finding is that the FP counts for all non-technical subjects are greater than the actual value. The authors investigate the designs for the non-technical subjects and discuss this matter with them. Two possible explanations are shown as below:

- Because of lack of working experience, non-technical subjects wrongly interpret the specifications, some extra unnecessary function modules are included.
- It may be the reason that the non-technical groups haven't sufficient working background to make the judgment, they prefer to choose a pessimistic rating for the TCF.

**FINDING 2: The fluctuation of FP counts for non-technical estimators is high while for technical estimators is low.**

**True.** The major disadvantage of the non-technical estimators is the fact of high level of fluctuation of function point counting. In other words, two individuals performing an FP count for the same system would generate a significant different result. Table I summarize the counting statistics collected, it shows that the range of function point counts for non-technical estimators is between 48 to 153 and the relative variation (Std.Dev. / Mean) for the non-technical and technical subjects are 31.04% and 16.67%, respectively. Obviously, we can say that function point counts can be estimated by non-technical estimator but the relative variation is quite high. Thus we can not fully rely on their estimates. However, the estimates from technical estimators are relatively small fluctuation.

**FINDING 3: Non-technical estimators have to spend more time than technical estimators in estimates.**

**True.** On average, the average time spent for doing the estimation was 30.7 hrs and 17.5 hrs for the non-technical and technical estimators, respectively. Obviously, the efficient for non-technical estimators is quite lower than technical estimators. One could speculate that there are many possible reasons including insufficient working experience, poor understanding the specification, etc.

**FINDING 4: Are UFP counts and Complexity Factors equally contribution to the variation for the technical experience subjects?**

**False.** This conclusion can be drawn on the basis of the remarkable difference between the  $\overline{MRE}$  for the UFP and TCF. This again underlines the need for re-consideration the definition of the fourteen system characteristics.

Because only technical experience subjective may be the “qualified” users of the FPA method, this paper just focus on the constitution of the variation for this group. To answer the above problem, Average Magnitude of Relative Error ( $\overline{MRE}$ ) will be applied.  $\overline{MRE}$  is one of the most common method to measure the accuracy of estimation. It is suggested and used by number of researchers [Thebaut83, Kemerer87, Conte88], and can be calculated by used following formula:

$$MRE_i = \frac{|E_i - \tilde{E}|}{E_i} \quad \text{and}$$

$$\overline{MRE} = \frac{1}{n} \sum MRE_i$$

where

$E_i$ : the  $i$  th predicted values by technical experience groups

$\tilde{E}$ : the actual value

In this paper,  $\overline{MRE}$  is the average absolute relative error between the actual FPCs and the technical experience group. Therefore, the lower the value of  $\overline{MRE}$ , the closer the estimation to the actual FPCs. Since errors of opposite sign do not cancel one another, this evaluation should adequately represent the average performance of each group. Table 4 summarises the results of this analysis.

Table 4:  $\overline{MRE}$  for Technical Groups

	UFP	$MRE_i$	TCF	$MRE_j$	FP	$MRE_k$
	44	7.32%	1.04	8.33%	45.76	16.26%
	44	7.32%	1.04	8.33%	45.76	16.26%
	43	4.88%	1.12	16.67%	48.16	22.36%
	43	4.88%	1.06	10.42%	45.58	15.80%
	42	2.44%	0.97	1.04%	40.74	3.51%
	38	7.32%	0.85	11.46%	32.3	17.94%
<i>Mean</i>	42.33		1.01		43.05	
<i>MRE</i>		5.69%		9.38%		15.35%

Where  $MRE_i$ : the MRE of the Unadjusted Function Point (UFP) Counts  
 $MRE_j$ : the MRE of the 14 Technical Complexity Factors (TCF)  
 $MRE_k$ : the MRE of the Function Point (FP) Counts  
 (Actual UFP = 41, TCF = 0.96, FPC = 39.36)

The results presented in Table 4 provide general support this conclusion. The average  $\overline{MRE}$  for UFP and TCF are 5.69% and 9.38%, respectively. The difference between these two elements is 3.69%. It implies that the  $\overline{MRE}$  of TCF is around 50% more than UFP. The possible explanations, without going into detail, for above differences is that-

*Reason 1:* These 14 cost drivers are used to adjust effort predictions. however, most research works in FPA are focused on validate and usage the definition of UFC, leaving the rating TCF unexplored. There are short of discussions how to improve the assessment of 14 general system characteristics that are rated on a scale from 0 to 5.

*Reason 2.* For all 14 system characteristics, some of them have been rated based on subjective judgment. For example, Transaction Rate, one of the system characteristics, is not well-defined. This characteristic can vary between installations depending on two main factors, such as the machine capacity and the machine loading

factor as stated for the 'Performance' characteristic. No quantification is stated for any of the items listed under this characteristic and introduces subjectivity into this characteristic.

It may be the reason why the scores of 14 system characteristics on average for technical groups is higher than the actual value (higher than 9.38%). It is the main constitution of the overall variation.

To confirm the subjectivity issue, an informal interview session was carried out, in which four questions were asked, after the experiment.

- (i) Is subjectivity the major problem in rating the FP counts?'
- (ii) Do you feel difficult in rating the UFP counts?'
- (iii) Do you feel difficult in rating the 14 system characteristics?'
- (iv) Is fuzzy statement suitable being used to express your judgement in more natural terms?'

The first question is asked relating to the FP counts and is a borderline result. When asked "Is the subjectivity the major problem in rating the FP counts?" the responding 'yes' for non-technical and technical experience subjects are 100% (15/15) and 83.3% (5/6). The first point questions the overall feeling of using the FP counts. The result confirms that subjectivity is a major problem for this estimating tool. It would seem to suggest that

the reliability of FP count need to be further improved.

The second point ('Do you feel difficult in rating the UFP counts?') questions the difficulty in rating the UFP. According to this study, 81.3% (13/15) non-technical subjects would say 'yes' while just 50% (3/6) non-technical subjects replied 'yes'. This result shows that the subjectivity problem in rating the UFP counts to non-technical subjects is much greater than technical subjects. It may be the reason that technical subjects more understand the user requirements and the way in rating the UFP counts.

The third point ("Do you feel difficult in rating the 14 system characteristics?") questions the difficulty in rating the 14 system characteristics. According to this study, it is one of the most support issues from the sample. The responding 'yes' for technical and non-technical subjects are 100% (15/15) and 83.3% (5/6), respectively. The vast majority of respondents feel difficulty in rating the 14 system characteristics. In summary, the result from the interview shows that FP estimators feel problems in determining the FP counts, especially the 14 system characteristics.

To tackle the subjectivity problem in rating the 14 system characteristics, fuzzy concept is introduced. It is used to calibrate the human factors to ensure that estimating model can be used in a more realistic way. Fuzzy statement allows estimators to express their judgement in more natural terms which represent more information about the expert judgement. When asked "*Is fuzzy statement suitable being used to express your judgement in more natural terms?*", the responding "Yes" for non-technical and technical experience subjects are 93.3% (14/15) and 83.3% (5/6), respectively. No matter having working experience or not, the vast majority of respondents see fuzzy statement as a useful tool to express their judgement in more natural terms.

## 5 SUMMARY AND CONCLUSIONS

The first finding of the study is that software development experience influences function point estimates. This can be understood by comparing the results of the technical and non-technical groups. Although the statistics were based on the analysis of a small number of examples and may not constitute any strong recommendation for management to use FP technique, the experiment can be regarded as a very useful reference to correct the general misunderstanding of using this model.

Many researchers believed that the processing complexity had no poor effect on the accuracy of the FPA. However, the result from this study shows that the processing complexity adjustment appear to affect the accuracy of the derived FPC and suffers of poor definition. We believe that on the reasons for this is that the definition for some system characteristics is not well defined. It reveals that this proposition may be incorrect and further research is required.

Another finding is that fuzzy concept can be applied with the FP counting process. Almost all FP counters see fuzzy statement as a useful tool to express their judgement in more natural terms. Although it is too early to make this proposition, but the results from this experiment so far are certainly encouraging.

## 6 ACKNOWLEDGMENTS

The authors would like to thank Mr Barrie Brandon and Miss Reddy Kwok for their invaluable support in doing this empirical study.

Table 5: Summary of the experiment

<i>There exists a variation in function point estimates between technical and non-technical estimators</i>	True
<i>The fluctuation of FP counts for non-technical estimators is high while for technical estimators is low.</i>	True
<i>Non-technical estimators have to spend more time than technical estimators in estimates</i>	True
<i>Are UFP counts and Complexity Factors equally contribution to the variation for the technical experience subjects?</i>	False

	non-technical	technical
	Yes	Yes
Is subjectivity the major problem in rating the FP counts?’	15	5
Do you feel difficult in rating the UFP counts?’	14	3
Do you feel difficult rating the 14 system characteristics?’	15	5
Is fuzzy concept suitable being used to express your judgement in more natural terms’	14	5

## REFERENCES

- [Albrecht79] A.J. Albrecht. “Measuring application development productivity”. In *Proceedings of the IBM Applications Development Joint SHARE/GUIDE/IBM Symposium*, CA, USA, pages 83-92, 1979.
- [Albrecht83] A.J. Albrecht and J.E. Gaffney. “Software function, source lines code, and development effort prediction: a software science validation”. *IEEE Transaction of Software Engineering*, 9(6):639-648, 1983.
- [Albrecht84] A.J. Albrecht. “AD/M productivity measurement and estimate validation”. *IBM Corporate Information Systems and Administration Guideline*, May, 1984.
- [Conte86] S. Conte, Dunsmore and H. Shen. *Software Engineering Metrics and Models*. Benjamin/Cummings Publishing Company, Menlo Park, CA, 1986.
- [Dreger89] J. B. Dreger. *Functional Point Analysis*, Prentice Hall, 1989.
- [IFPUG94] The International Function Point Users Group. *Function Point Counting Practices Manual*, Release 4.0, Westerville, Ohio, 1994.
- [Jeffery 90] D.R. Jeffery and C.G. Low. “Function Point in the Estimation and Evaluation of the Software Process”. *IEEE Transaction of Software Engineering*, 16(1):64-71, 1990.
- [Jones86] C. Jones. *Programming Productivity*, McGraw-Hill, N.Y. 1986.
- [Kemerer87] C. Kemerer. “An Empirical Validation of Software Cost Estimation Models”. *Communication of the ACM*, 30(5):416-429, 1987
- [Symons88] C. Symons. “Function Point Analysis: Difficulties and Improvements”. *IEEE Transaction of Software Engineering*, 14(1):2-11, 1988
- [Summer99] Mary Sumner. “Critical Success Factors in Enterprise Wide Informaion Management System Projects”. *Proceeding of the ACM SIGCPR*, New Orlean, LA, USA, 1999.
- [Thebaut93] S.M. Thebaut. “Model evaluation in software metrics research”, in *Proceedings of the 15th Symposium on Computer Science and Statistics*, 1993.

