An Evaluation of Functional Size Measurement Methods

Christian Quesada-López, Marcelo Jenkins

Center for ICT Research, University of Costa Rica, San Pedro, Costa Rica
{cristian.quesadalopez, marcelo.jenkins}@ucr.ac.cr

Abstract. Background: Software size is one of the key factors that has the potential to affect the effort of software projects. Providing accurate software size estimation is a complex task. A number of functional size measurement (FSM) methods have been proposed to quantify the size of software based on functional user requirements (user perspective). Function point analysis (FPA) was the first proposal for a FSM method and it is one of the most accepted FSM methods in the industry. Automated Function Point (AFP) method state the guidelines for automating FPA counting from software source code. **Objectives:** This paper reports on an experiment that compares FPA and AFP. The goal is to evaluate the measurement process on a range of performance and adoption properties such as accuracy, reproducibility, efficiency, perceived easy to use, usefulness, and intention to use. **Methods:** A controlled experiment was conducted to compare the two methods. Statistical analyses were conducted to find differences between the methods regarding performance and adoption properties. **Results:** The functional size results between the FPA and AFP methods were similar (MMRE 6-8%). Productivity rate was about the same reported for the industry (43.4 FPA/h, 37.8 AFP/h). There were no significant differences between the methods for functional size estimation, reproducibility, and accuracy. **Limitations:** This is an initial experiment of a work in progress. The limited sample size and nature of the subjects may influence the results. **Conclusions**: These results support the claim that AFP produces similar measurement results that FPA. The automation of the AFP method could produce more consistent measurement results in conformance with the FPA counting guidelines. An automated and quick FSM counting method will increase the adoption of this metric in industry. Further research is needed to conclude more on some perceived adoption properties.

Keywords: Function points, functional size measurement, Function Point Analysis FPA, Automated Function Points AFP, experimental procedure.

1 Introduction

Software estimation process is a key factor for software project success [1]. The complexity to provide accurate software size estimation and effort prediction models in software industry is well known. The need for accurate size estimates and effort predictions for projects is one of the most important issues in the software industry [2]. Software size measurement based on functional size has been studied for many years,

but many software organizations are still using expert judgment as their preferred estimation method, producing inaccurate estimations and severe schedule overruns in many of their projects [3]. Although software size measurement is an important part of the software development process [4, 5], several companies consider formal functional size estimation methods to be too complex and unpractical for their processes. Software size has proved to be one of the main effort-and-cost drivers [3, 8, 9, 10].

Functional size methods are used to measure the logical view of the software from the users' perspective by counting the amount of functionality to be delivered. These measures can be used for a variety of purposes, such as project estimation [4, 5, 6], quality assessment, benchmarking, and outsourcing contracts [5]. According to [7], functional size measurements are used for budgeting, tracking progress, negotiating modifications, sizing deliveries, estimating portfolio size, managing productivity, and managing defect density. Hence, functional size measures generate a variety of productivity, financial and quality indicators in different phases of the software development process [5].

The Function Point Analysis (FPA) counting manual is one of the mostly used functional size measurement methods in the software industry [5]. An automatic method of counting function points will increase the use of this technique, because automation reduces the cost of counting and the inconsistency of manual counts. An automated function point measurement can become a standard component of the software development and maintenance process. Besides, automatic counting could generate consistent and reliable historical project data for benchmarking. Finally, IT organizations whose manage many software projects can estimate the functional size of their application portfolio more accurately and usually within a short time frame [11]. A functional size estimation method based on input provided by artifacts such as design models and source code can help the process of regularly updating the baseline counts and taking into account changes made during application maintenance and during small application enhancement projects [12]. Recently, the Object Management Group (OMG) released the Automated Function Point (AFP) specification [11]. AFP provides a standard for automating function point measure according to the counting guidelines of the International Function Point User Group (IFPUG). According to OMG, this method ensures automation, consistency and verifiability. However, it is difficult to evaluate new proposals on a practical level due to the lack of rigorous empirical validation for new functional size measurement (FSM) methods. The absence of systematic evaluation could explain the low adoption rate of the new proposed FSM methods [13].

This paper reports on an experiment which compares FPA and AFP measurement process in terms of performance properties (accuracy, reproducibility, efficiency), and adoption properties (perceived easy to use, perceived usefulness and intention to use). This study was carried out at the University of Costa Rica with a group of 14 practitioners in a software metrics course. The experimental design follows the framework proposed by Wohlin et al. [14]. The structure of this paper is organized as follows. Section 2 provides the foundations about the compared function point methods. Section 3 presents the related work on empirical studies assessing FSM methods. Section 4 describes the experimental design process, and Section 5 presented and discusses the

results of the experiment. Section 6 presents the summary. Finally, Section 7 outlines the conclusions.

2 Functional Size Measurement

Functional Size Measurement (FSM) is defined as the process of measuring functional size. The ISO/IEC 14143-1 standard [7] defines the concepts related to FSM and describes the general principles for applying an FSM method. After the ISO/IEC 14143 standard series, several FSM methods have been proposed to quantify the software functional size based on functional user requirements, including COSMIC, IFPUG, Mk II, NESMA, and FiSMA.

2.1 Function Point Analysis

Many functional size measurement methods have been proposed to quantify the size of software based on functional user requirements (the user's perspective). Function point analysis (FPA) [8, 9] was the first proposal of a FSM. The International Function Point Users Group (IFPUG) FPA counting practice manual is one of the most used functional size measurement methods in the software industry [15]. ISO/IEC 20926:2009 standard [16] specifies the set of definitions, rules and steps for applying the IFPUG method. In FPA, user requirements are classified and counted in a set of basic functional size components (BFC). These elementary units are called data and transactional functions. They represent data and operations that are relevant to the users. FPA can be applied in early stages in the development process, and it is independent from technology-based influences [9]. FPA have been subject to a number of critiques: the reliability of FPA measurement [4], the BFCs have inter correlations with each other [6], the application and usefulness of the complexity adjustments [17]. FPA is prone to different interpretations by different subjects, hence a variation in the counts is expected. Besides, the counting method is slow and expensive [15]. OOther FSM methods have been proposed, but they also have some issues that have to be analyzed in order to create a reliable and consistent method [18].

2.2 Automated Function Points

The Automated Function Point (AFP) specification [11] provides a standard for automating function point measure according to the counting guidelines of the International Function Point User Group (IFPUG), release 4.3.1. This specification may differ from IFPUG counting practice manual at points where subjective judgments have to be replaced by the rules needed for automation [11], and it is applicable to the functional sizing of transaction-oriented software applications, and in particular those with data persistency. This method is the first standard that ensures the repeatability and consistency of the counting technique. Besides, this process ensures automation and veri-

fiability. The arrival of an automatic method of counting function points will most certainly increase the use of this technique because it reduces the cost of counting and reduces the inherent inconsistency of manual counts. Therefore, AFP measurement could become a standard component of the software development and maintenance process. Automatic counting could generate more consistent and reliable historical project data for benchmarking.

3 Related work

ISO standard series provides the basis against existing Functional Size Measurement (FSM) methods could be evaluated. Part 3 [19] describes the process for verification of a FSM method and establishes a framework for verifying the statements of an FSM method and for conducting tests requested by the verification about performance properties. This part aims at ensuring that the output from the verification is objective, impartial, consistent, repeatable, and reproducible. Jacquet and Abran [20, 21] suggest a process model for functional size measurement methods. The model details the steps from the design, its application, the analysis of its measurement results, and the exploitation of these results in subsequent prediction models, such as in quality and estimation models. Empirical validation relates to the second and third steps in the process model proposed for Jacquet and Abran [20, 21]. This process describes a step where the measurements results must be validated and verified. This evaluation validates the functional size of the measured application and verifies that the measurement rules are applied correctly. Evaluating the use of a FSM method allows the assessment of the degree of confidence in the measurement results and verifies whether the method satisfies its intended use and the user needs [22]. The evaluation of the FSM methods seeks objective evidence of the efficacy (effectiveness and efficiency) of a method in achieving its objectives and test the user response to a FSM method to try to predict its acceptance in practice [23]. Systematic evaluation of FSM method process could compare performance between proposed FSM methods [13].

Abrahao and Pastor proposed a method for evaluating FSM methods [22]. This proposal contains a rigorous empirical evaluation of the effectiveness. The proposal was based on ISO FSM standard part 3 [19], and the technology acceptance model (TAM) [24]. The evaluation model provides a range of performance-based and perceived-based variables: Performance-based (objective measures): How well are people able to use the FSM method? In addition, Perception-based (subjective measures): How effective do people believe the FSM method to be in achieving its objective?

Marin et al. [25] quantified precision by calculating repeatability and reproducibility of counts. It attempted to control all the factors that could affect the precision of the counts (knowledge of the measurement procedure, experience in using the measurement procedure). Several studies have evaluated measurement processes for different FSM methods through experiments following Abrahao and Pastor´s proposal [22]. For example, FPA and OOmFP [26] methods were compared in terms of reproducibility and accuracy [13]. FPA and OOmFPWeb [23] were evaluated on a range of performance-based and perception-based variables [27] [28].

Our work evaluates and compares FPA and AFP methods by conducting a controlled experiment. According to [23], the evaluation of the application of a FSM method should precede the validation of effort predictive models that are based on functional size measurement. To our knowledge, no academic empirical evaluations of this type on FPA and AFP methods have been published.

4 Experimental design

In this section, we describe the experimental design that follows the framework proposed by Wohlin et al. [14]. This paper reports on an experiment which compares FPA and AFP FSM measurement method process in terms of performance properties (accuracy, reproducibility, efficiency), and adoption properties (perceived easy to use, perceived usefulness and intention to use). The goal of the experiment is to evaluate and compare the measurement process on a range of performance and adoption properties. The objective written in GQM [29] form is:

Analyze FPA and AFP FSM method process
for the purpose of evaluating and comparing
with respect to performance and adoption properties
from the point of view of the researcher
in the context of a metrics course

4.1 Planning

4.1.1 Context selection

The context of the experiment is a software metrics course. This study was carried out at the University of Costa Rica with a group of 14 practitioners taking a graduate level metrics course. Practitioners applied the IFPUG Function Point Analysis (FPA) and OMG Automated Function Points (AFP) FSM methods to measure the same application. The sample application was a small web site for a fictional University.

4.1.2 Hypothesis formulation

This section states what is going to be evaluated in the experiment. In this study, we evaluate the OMG Automated Function Points (AFP) method process against IFPUG Function Point Analysis (FPA) method process. We state the hypothesis and define what measures are needed to evaluate them.

Hypothesis 0: test the variance between the measurement results.

- H_{0ufp}: AFP produces equal measurement results than FPA
- H_{1ufp}: AFP produces different measurement results than FPA

Measures needed: AFP and FPA functional size measurement for each subject.

Hypothesis 1: test the relationship between methods and reproducibility.

- H_{0rpd}: AFP produces equal consistent measurement results than FPA
- H_{1rpd}: AFP produces different consistent measurement results than FPA

Measures needed: AFP and FPA functional size measurement for each subject.

Hypothesis 2: test the relationship between methods and accuracy.

- H_{0acc}: AFP produces equal accurate measurement results than FPA
- H_{1acc}: AFP produces different accurate measurement results than FPA

Measures needed: AFP and FPA functional size measurement for each subject. AFP and FPA true value (by an expert) to compare.

Hypothesis 3: AFP is perceived as easy to use.

Measures needed: AFP perceived easy to use for each subject.

Hypothesis 4: AFP is perceived as useful.

Measures needed: AFP perceived usefulness for each subject.

Hypothesis 5: There is an intention to use AFP.

Measures needed: AFP intention to use for each subject.

The hypotheses mean that the following data needs to be collected. Metrics used in this study were based on [23]:

- Reproducibility (ratio scale): the closeness of the agreement between the results of successive measurement of the same product carried out under the same conditions. It refers to the use of the method on the same product and environment by different subjects. In order to evaluate the degree of variation in reproducibility, the statistic proposed in [13, 22, 30] was applied. This was calculated as the difference in absolute value between the count produced by a subject and the average count produced by the other subjects in the sample, relative to his average count for the same FSM method. Equation 1 describes the statistic:

$$\llbracket Rep \rrbracket\_i = |(AverageOthers - \llbracket Suject \rrbracket\_i)/AverageOthers| \quad (1)$$

- Accuracy (ratio scale): the closeness of the agreement between the result of a measurement and the true value. Magnitude of Relative Error (MRE) was used to evaluate accuracy results. The functional size calculated by an expert represented the "true value". Equation 2 describes the statistic:

$$\llbracket MRE \rrbracket\_i = |(TrueValue - \llbracket Suject \rrbracket\_i)/TrueValue| \quad (2)$$

- Measurement time (ratio scale): the time for the measurement process taken by each subject.
- Perceived easy to use (ordinal scale): the degree to which a person believes that using a particular method would be free of effort. This construct measure the perceptual judgment about the effort required to learn and use the FSM method. The items were formulated as a five-point Likert scale using an opposing statements question format. Perceived easy to use was measured using five items.
- Perceived usefulness (ordinal scale): the degree to which a person believes that a particular method will be effective in achieving its intended objectives. This construct measure the perceptual judgment about the effectiveness of the FSM method. The items were formulated as a five-point Likert scale using an opposing statements question format. Perceived usefulness was measured using five items.

- Intention to use (ordinal scale): the degree to which a person intends to use a particular method. This construct measure the perceptual judgment about the performance of the FSM method. The items were formulated as a five-point Likert scale using an opposing statements question format. Intention to use was measured using three items.

4.1.3 Variable selection

The independent variable is the FSM method used by subjects to size the web application: FPA or AFP. The dependent variables are functional size, measurement time, perceived easy to use, perceived usefulness, and the intention to use.

4.1.4 Selection of subjects

The subjects were chosen based on convenience. The subjects are professionals working on Costa Rican software companies with similar backgrounds in Computer Science. They were not experts in functional size measurement.

4.1.5 Experiment design

The definition, hypotheses and measures for the evaluation means that the design is one factor with two treatments. The factor is the FSM method and the treatments are FPA and AFP methods. The treatments correspond to the two levels of the independent variable: the use of AFP versus FPA sizing a web application. A between-subject design was conducted for the reason that the time for the experiment was limited. The initial 14 subjects were randomly assigned to two groups with the same number of subjects. Each group worked with a different FSM method and no blinds were applied (group 1: FPA method [n=7], group 2: AFP method [n=7]).

4.1.6 Instrumentation

The experiment included two tasks, the FSM process task and the post measurement survey. In the FSM process task, each subject used the methods rules to measure the same web application. Data were collected on a results sheet, data that were later used to evaluate performance properties. In the post-measurement survey the subjects were asked to complete a questionnaire to evaluate perception properties. The instruments used to conduct this experiment include:

- Experimental object: includes a requirements specification[1] document and source code for a web application of a fictional University. It includes functionality such as student admission, course creation, and instructor assignments. The specification

[1] https://drop.citic.cr/public.php?service=files&t=2dc019254fce636370a98e57e4d5d952 Password: SET

document describes the requirements for the system using the standard IEEE Recommended Practice for Software Requirements Specifications [31]. The requirements were described in terms of functionality. The application includes functionality such as student admission, course creation, and instructor assignments. The application support the following functions and transactions: Course maintenance (create, edit, delete, report, search, department assignment), instructor maintenance (create, edit, delete, report, course assignment, course assignment report), department maintenance (create, edit, delete, report, administrator assignment), and student maintenance (create, edit, delete, report, search).

- Training materials: includes a set of instructional slides that describe the FSM method and the measurement procedure, and a measurement example used in the training sessions. Besides, a measurement guideline for each FSM method was provided. Finally, a technical manual and user manual of the application was provided as well.
- Survey instrument[2]: includes 14 closed questions based on the survey presented in [24]. The items were formulated as a five-point Likert scale using an opposing statements question format. Perceived easy to use was measured using five items (Questions 1, 3, 4, 6, and 9), perceived usefulness using five items (Questions 2, 5, 8, 10, 11), and the intention to use using three items (Questions 7, 12, 13). We include one more item (14) to ask for the perception about how easy could be to automate the method according to the specifications. The order of the items was randomized and some of the questions were negated. Perceived adoption properties were calculated as the average of the questions that constitutes each construct.

4.1.7 Threats to validity

This section analyses the threats to the validity for this study and the actions undertaken to mitigate them. Internal validity is primarily focused on the validity of the actual study. External validity is concerned with subjects, measurement object, and measurement methods. Construct validity is about generalizing the result to the theory behind the experiment.

Internal validity: Differences among subjects were reduced selecting subjects with the same level of experience in FSM methods. The same requirement specification and source code was used for all subjects, both of them for the same application. Measurement time was self-reported by subjects in work effort (hours). Although the small number of subjects is a threat, the fact that the subjects are industry practitioners is an advantage for the study. The students were expected to deliver many data as part of their work with the course. Thus, there is a risk that the data is faked or simply not correct due to mistakes. Only one expert was used to count the functional size used as a “true value”. This is a threat to validity because expert counting could present variations.

[2] https://drop.citic.cr/public.php?service=files&t=6d703347f0318e51af35ea64e86345ca Password: SET

External validity: Although the application is a small example, the requirements document and the source code of the application were a very similar example of the practices in a real case development in the industry. The subjects were mainly developers and testers. They are not the population that normally use FSM measurement methods; however, they are very familiar with software engineering practices. All tasks were in the context of FPA and AFP methods in the MIS functional domain.

Construction validity: the dependent variables used in this study to evaluate the effectiveness of FSM methods are based on ISO 14143-3[22], and the technology acceptance model (TAM) [25] adapted in [23].

4.2 Operation

4.2.1 Preparation

The subjects were not aware of the aspects under study. They were informed that the researchers wanted to study the measurement process of the FPA and AFP methods but they did not have knowledge of the study´s hypotheses and from their point of view, they were solving a course exercise. All students were guaranteed anonymity. The survey material was prepared in advance. We ran training sessions during the course prior to getting them to perform the experimental tasks, in which the measurement rules were introduced and demonstrated using several practical examples.

4.2.2 Execution

The experiment was performed over a 4-week period, during which four training sessions (3 hours each one) on measurement were conducted. At the end of the training sessions, the measurement task material and the experimental object were presented and experimental tasks were explained. After that, the practitioners received all the materials. The data was primarily collected through results sheets previously prepared. They filled up the measurement results sheet during the experiment and the survey at the end. The experiment was run as part of a graduate-level metrics course and the students were graded on the exercise.

4.2.3 Data validation

Data was collected for 14 students (practitioners). The dependent variables were measured using different data collection forms. Two forms were used to collect measurement results of the FPA and AFP methods. The survey instrument where used to collect the data for the perception properties. We also collected data from an expert. All forms were correctly filled up and we took into account all responses. No data had to be removed because it was deemed valid or at least not questionable. Some of the subjects did not report effort so we treated those data as missing values.

5 Analysis and interpretation

5.1 Descriptive statistics

Table 1 shows descriptive statistics for the functional size in unadjusted function points (UFP), accuracy, and reproducibility. Finally, productivity (UFP/h) results are presented. For each column, FPA and AFP methods are compared. Data shows that the accuracy of both methods were similar, FPA MMRE was 6% and AFP MMRE was 8%. The same is presented with reproducibility; FPA result (8%) and AFP (9%) are very close. According to [30], the lowest productivity for first time counters is 200-300 FP/day (8 hours working). That is 23-37.5 FP/hour. The mean measurement productivities of 43.4 FPA/hour and 37.8 AFP/hour are about the same reported in industry. FPA and AFP method process produce similar results in terms of functional size, accuracy and reproducibility. These results are based on data presented on Appendix A.

Table 2 shows descriptive statistics for perceived properties comparing the FPA and AFP method. We presented the number of answers and its percentage [n (%)]. In general, the data indicate that there is no difference for perceived as easy to use (PEOU), perceived usefulness (PU), and perceived intention to use (ITU) between the methods. Besides, the subjects' opinion were divided or neutral within methods. However, Q3 and Q4 (PEOU), Q8 (PU), and Q7 (ITU) presents some interesting insights about methods. Q4 (I found the measurement rules of the FSM method clear and easy to understand) shows that, for this subject's sample, practitioners found the measurement rules confusing and difficult to understand for both methods. In general, they claim for more examples in practice about the process of application of the FSM methods. Additionally, Q3 (Overall, I found the FSM method easy to use) shows that practitioners found AFP method process difficult to use. Q8 (I think that this FSM method would improve the accuracy of estimates of applications) shows that practitioners perceive FPA could improve accuracy in software size estimations. Finally, Q7 (I will use this FSM method if I have to size applications in the future) shows that practitioners would be open to use FPA in the future.

Table 1 Descriptive statistics for functional size and performance

		UFP		Accuracy		Reproducibility		Productivity	
		FPA	AFP	FPA	AFP	FPA	AFP	FPA	AFP
n		7	7	7	7	7	7	7	7
Mean		96.43	105.43	0.06	0.08	0.08	0.09	43.45	37.84
Median		100.00	102.00	0.03	0.10	0.08	0.08	25.43	40.00
Std. Dev.		7.89	9.83	0.07	0.07	0.05	0.05	41.21	19.79
Min		85.00	91.00	0.00	0.01	0.01	0.40	10.99	17.24
Max		104.00	118.00	0.16	0.17	0.14	0.16	114.34	63.40
Percentile	25	86.00	98.00	0.01	17.91	0.04	0.40	17.91	17.74
	50	100.00	102.00	0.03	25.43	0.08	0.08	25.43	40.00
	75	103.00	115.00	0.15	90.53	0.12	0.14	90.53	62.05

Table 2 Descriptive statistics for perception properties (n=7)

Property	Question	FPA					AFP				
		(+)Positive		Neutral	Negative(-)		(+)Positive		Neutral	Negative(-)	
		1	2	3	4	5	1	2	3	4	5
PEOU	Q1	1 (0.07)	1 (0.07)	3 (0.21)	2 (0.14)	0 (0.00)	0 (0.00)	2 (0.14)	3 (0.21)	1 (0.07)	1 (0.07)
	Q3	1 (0.07)	0 (0.00)	3 (0.21)	3 (0.21)	0 (0.00)	0 (0.00)	0 (0.00)	1 (0.07)	**6 (0.42)**	0 (0.00)
	Q4	0 (0.00)	1 (0.07)	2 (0.14)	**4 (0.28)**	0 (0.00)	0 (0.00)	0 (0.00)	1 (0.07)	**3 (0.21)**	**3 (0.21)**
	Q6	0 (0.00)	2 (0.14)	2 (0.14)	2 (0.14)	1 (0.07)	0 (0.00)	2 (0.14)	1 (0.07)	2 (0.14)	2 (0.14)
	Q9	1 (0.07)	2 (0.14)	2 (0.14)	1 (0.07)	1 (0.07)	0 (0.00)	2 (0.14)	3 (0.21)	1 (0.07)	1 (0.07)
PU	Q2	1 (0.07)	1 (0.07)	2 (0.14)	3 (0.21)	0 (0.00)	0 (0.00)	2 (0.14)	3 (0.21)	2 (0.14)	0 (0.00)
	Q5	1 (0.07)	2 (0.14)	3 (0.21)	1 (0.07)	0 (0.00)	0 (0.00)	2 (0.14)	3 (0.21)	2 (0.14)	0 (0.00)
	Q8	**2 (0.14)**	**4 (0.28)**	1 (0.07)	0 (0.00)	0 (0.00)	0 (0.00)	1 (0.07)	4 (0.28)	2 (0.14)	0 (0.00)
	Q10	2 (0.14)	2 (0.14)	3 (0.21)	0 (0.00)	0 (0.00)	0 (0.00)	1 (0.07)	5 (0.35)	1 (0.07)	0 (0.00)
	Q11	1 (0.07)	4 (0.28)	2 (0.14)	0 (0.00)	0 (0.00)	1 (0.07)	2 (0.14)	4 (0.28)	0 (0.00)	0 (0.00)
ITU	Q7	1 (0.07)	3 (0.21)	3 (0.21)	**0 (0.00)**	**0 (0.00)**	0 (0.00)	2 (0.14)	2 (0.14)	1 (0.07)	**2 (0.14)**
	Q12	0 (0.00)	2 (0.14)	1 (0.07)	2 (0.14)	2 (0.14)	1 (0.07)	1 (0.07)	1 (0.07)	3 (0.21)	1 (0.07)
	Q13	1 (0.07)	1 (0.07)	2 (0.14)	2 (0.14)	1 (0.07)	0 (0.00)	2 (0.14)	1 (0.07)	2 (0.14)	2 (0.14)

5.2 Hypothesis testing

The normality test indicates that the functional size (.617), Reproducibility (.759), Accuracy (.022), perceived easy to use (.509), intention to use (.757), and productivity (.010) data belonged to normal distribution (Shapiro-Wilk test). The Levene test confirmed equality of variances.

First, the variance between the means was tested (Hypothesis 0). The results from the one-way ANOVA indicate that there is not enough evidence to reject the null hypothesis (p=0.083). There is no significant difference between the functional size results of the two methods, which supports the claim that AFP produces similar measurement results than FPA. The results from the test are shown in Table 3. Second, in order to evaluate the degree of variation in reproducibility, the statistic proposed in [13, 22, 31] was applied. The differences in means reproducibility measurements were tested (Hypothesis 1). The results from the one-way ANOVA indicate that there is not enough evidence to reject the null hypothesis (p=0.572). There is no significant difference between the reproducibility results of the two methods, which support the claim that AFP produces the same consistent measurement results as FPA. The results from the test are shown in Table 4. Third, MRE (Magnitude of Relative Error) was used to evaluate accuracy results. The functional size calculated by an expert was used as a "true value". The differences in means accuracy measurements were tested (Hypothesis 2). The results from the one-way ANOVA indicate that there is not enough evidence to reject the null hypothesis (p=0.554). There is no significant difference between the accuracy results of the two methods, which supports the claim that AFP produces the same accurate measurement results as FPA. The results from the test are shown in Table 5.

Table 3 Functional Size (UFP) ANOVA Test

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	283.500	1	283.500	3.568	0.083
Within Groups	953.429	12	79.452		
Total	1236.929	13			

Table 4 Reproducibility ANOVA Test

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	0.001	1	0.001	0.337	0.572
Within Groups	0.025	12	0.002		
Total	0.026	13			

Table 5 Accuracy ANOVA Test

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	0.002	1	0.002	0.370	0.554
Within Groups	0.052	12	0.004		
Total	0.054	13			

Hypothesis 3, 4, and 5 were tested by verifying when the scores assigned to the perception properties were better than the middle score (score = 3 in a 5-point Likert scale) [24, 30]. The scores of a subject were averaged over the items that are relevant for a property (perceived as easy to use, perceived as useful, and intention to use). For this analysis, the scores of a subject were averaged over the items that are relevant for a construct resulting in three scores for each subject (see Appendix A). These scores were then compared against the value 3 [24]. The results from the one-way ANOVA indicate that there is no significant difference for perceived as easy to use (p=0.388), and intention to use (p=0.491) between the methods ($\alpha = 0.05$). In order to check differences between perceived properties and the neutral value, one sample t-test was used with a significance level $\alpha = 0.05$. The results for the test shows that there was no evidence to conclude that the means differ for the neutral value (score =3).

6 Summary

In this study, the Function Point Analysis (FPA) and the Automated Function Points (AFP) measurement processes were evaluated and compared. Results applying each method were similar (MMRE 6-8%) and productivity rates were about the same reported in industry (43.4 FPA/h, 37.8 AFP/h). Our study did not find any significant differences between the FPA and AFP methods for functional size, reproducibility, and accuracy. The results on perceived adoption properties indicate that there is no significant difference for perceived easy to use, perceived usefulness, and intention to use between the two methods. The perceived properties versus a neutral value show that there was no evidence to conclude that the means differ for the neutral value. Our subjects believe there is a need for a more detailed guidance on how to apply the AFP

method. They claim that an automated tool for the AFP method could encourage organizations to start to collect functional size of their applications. In addition, results show that, for this subject's sample, practitioners found the measurement rules confusing and difficult to understand for both methods, and AFP method process difficult to use. However, they perceived that FPA could improve accuracy in software size estimations, and they would be open to use FPA in the future.

7 Conclusions

This paper described a controlled experiment to compare the FPA and AFP functional size measurement methods. The goal was to evaluate and compare the measurement process of the two methods on several performance and adoption properties. The results support the claim that AFP method process produces similar measurement results as FPA method process. The results corroborated the potential for developing automation tools for function point counting that could produce more consistent measurement results in conformance with the FPA counting guidelines. An automated and quick FPA counting tool will increase the adoption of the metric in industry. FSM methods are hardly automatable and the setup of a measurement procedure for each input to the measurement process is needed. Although encouraging results were obtained, further research is needed to corroborate performance results and to draw more conclusions on the perceived adoption properties. Replications should be conducted using more complex applications, using a bigger sample of subjects, and more than one counting expert in order to consider the variation interval for the functional size if the application.

8 Acknowledgments

This research was supported by the Costa Rican Ministry of Science, Technology and Telecommunications (MICITT).

9 References

1. Peixoto, C. E. L., Audy, J. L. N., & Prikladnicki, R. (2010, May). The importance of the use of an estimation process. In *Proceedings of the 2010 ICSE Workshop on Software Development Governance* (pp. 13-17). ACM.
2. Molokken, K., & Jorgensen, M. (2003, October). A review of software surveys on software effort estimation. In *Empirical Software Engineering*, 2003. ISESE 2003. Proceedings. 2003 International Symposium on (pp. 223-230). IEEE.
3. Boehm, B. W. (1981). *Software engineering economics*.
4. Low, G. C., & Jeffery, D. R. (1990). Function points in the estimation and evaluation of the software process. *Software Engineering, IEEE Transactions on*, *16*(1), 64-71.
5. Garmus, D., & Herron, D. (2001). *Function point analysis: measurement practices for successful software projects*. Addison-Wesley Longman Publishing Co., Inc.
6. Kitchenham, B. (1993) Using Function Points for Software Cost Estimation – Some Empirical Results. *10th Annual Conference of Software Metrics and Quality Assurance in Industry, Amsterdam*, Netherlands.

7. ISO. (2007). *ISO/IEC 14143-1- Information Technology - Software measurement - Functional Size Measurement. Part 1: Definition of Concepts*.
8. Albrecht, A. J. (1979, October). Measuring application development productivity. In *Proceedings of the Joint SHARE/GUIDE/IBM Application Development Symposium* (Vol. 10, pp. 83-92). Monterey, CA: SHARE Inc. and GUIDE International Corp.
9. Albrecht, A. J., & Gaffney, J. E. (1983). Software function, source lines of code, and development effort prediction: a software science validation. *Software Engineering, IEEE Transactions on*, (6), 639-648.
10. Jeng, B., Yeh, D., Wang, D., Chu, S. L., & Chen, C. M. (2011). A Specific Effort Estimation Method Using Function Point. *Journal of Information Science and Engineering*, *27*(4), 1363-1376.
11. OMG. (2014). Automated Function Points. Version 1.0.
12. Ellafi, R., & Meli, R. A Source Code Analysis-based Function Point Estimation Method integrated with a Logic Driven Estimation Method.
13. Abrahao, S., Poels, G., & Pastor, O. (2004, August). Assessing the reproducibility and accuracy of functional size measurement methods through experimentation. In Empirical Software Engineering, 2004. ISESE'04. Proceedings. 2004 International Symposium on (pp. 189-198). IEEE.
14. Wohlin, C., Runeson, P., Hst, M., Ohlsson, M. C., Regnell, B., & Wessln, A. (2012). Experimentation in software engineering. Springer Publishing Company, Incorporated.
15. Jones, C. (2013). Function points as a universal software metric. *ACM SIGSOFT Software Engineering Notes*, *38*(4), 1-27.
16. ISO. (2009). *ISO/IEC 20926, Software and systems engineering - Software measurement – IFPUG functional size measurement method.*
17. Jeffery, R., & Stathis, J. (1996). Function point sizing: structure, validity and applicability. *Empirical Software Engineering*, *1*(1), 11-30.
18. Lavazza, L., Morasca, S., & Robiolo, G. (2013). Towards a simplified definition of Function Points. *Information and Software Technology*, *55*(10), 1796-1809.
19. ISO. (2003). *ISO/IEC TR 14143-3:2003 Information technology -- Software measurement -- Functional size measurement -- Part 3: Verification of functional size measurement methods.*
20. A Abran, A., & Jacquet, J. P. (1999). A structured analysis of the new ISO standard on functional size measurement-definition of concepts. In Software Engineering Standards, 1999. *Proceedings. Fourth IEEE International Symposium and Forum on* (pp. 230-241). IEEE.
21. Jacquet, J. P., & Abran, A. (1997, June). From software metrics to software measurement methods: a process model. In *Software Engineering Standards Symposium and Forum, 1997. Emerging International Standards. ISESS 97, Third IEEE International* (pp. 128-135). IEEE.
22. Abrahao, S. M., & Director-Lopez, O. P. (2004). On the functional size measurement of object-oriented conceptual schemas: *design and evaluation issues*. Universidad Politecnica de Valencia (Spain).
23. Abrahao, S., Poels, G., & Pastor, O. (2004, September). Evaluating a functional size measurement method for Web applications: an empirical analysis. In Software Metrics, 2004. Proceedings. 10th International Symposium on (pp. 358-369). IEEE.
24. Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. MIS quarterly, 319-340.
25. Marín, B., Condori-Fernández, N., & Pastor, O. (2008, August). Towards a method for evaluating the precision of software measures. In *Eighth International Conference on Quality Software (QSIC), IEEE Computer Society Press* (pp. 305-310).
26. Pastor, O., Abrahão, S. M., Molina, J. C., & Torres, I. (2001). A FPA-like measure for object oriented systems from conceptual models. Current Trends in Software Measument, Ed. Shaker Verlag, 51-69.
27. Abrahão, S., Poels, G., & Insfran, E. (2008, July). A replicated study on the evaluation of a size measurement procedure for web applications. In *Web Engineering, 2008. ICWE'08. Eighth International Conference on* (pp. 217-223). IEEE.
28. Abrahao, S., & Poels, G. (2006, October). Further analysis on the evaluation of a size measure for Web applications. In *Web Congress, 2006. LA-Web'06. Fourth Latin American* (pp. 230-240). IEEE.
29. Basili, V. R., & Rombach, H. D. (1988). The TAME project: Towards improvement-oriented software environments. *Software Engineering, IEEE Transactions on,* 14(6), 758-773.
30. Kemerer, C. F. (1993). Reliability of function points measurement: a field experiment. *Communications of the ACM*, *36*(2), 85-97.
31. IEEE Computer Society. Software Engineering Standards Committee, & IEEE-SA Standards Board. (1998). IEEE Recommended Practice for Software Requirements Specifications. Institute of Electrical and Electronics Engineers.

Appendix A. Dataset used in the experiment															
IFPUG FPA								OMG AFP							
Sub-ject	**UFP**	**Rep**	**MRE**	**(FP/h)**	**PEOU**	**PU**	**ITU**	**Sub-ject**	**UFP**	**Rep**	**MRE**	**(FP/h)**	**PEOU**	**PU**	**ITU**
1	96	0.01	0.05	26.92	3.00	3.00	2.33	8	102	0.04	0.01	40.00	3.60	3.00	2.33
2	101	0.06	0.00	114.34	4.20	5.00	4.33	9	115	0.11	0.14	21.90	2.00	2.40	1.33
3	103	0.08	0.02	25.43	2.60	3.40	3.00	10	98	0.08	0.03	42.61	3.20	2.80	3.33
4	104	0.09	0.03	18.03	3.20	3.80	2.33	11	112	0.07	0.11	63.40	3.20	2.80	2.67
5	100	0.04	0.01	17.91	2.40	2.80	2.33	12	102	0.04	0.01	17.24	2.80	3.00	3.33
6	85	0.14	0.16	10.99	1.80	3.20	2.00	13	118	0.14	0.17	17.74	1.40	3.20	1.67
7	86	0.12	0.15	90.53	3.20	4.20	4.00	14	91	0.16	0.10	62.05	1.40	3.20	3.33
Unadjusted Function Points (UFP), Reproducibility (Rep), Perceived as easy to use (PEOU), Perceived as useful (PU), Intention to use (ITU)															