Evaluating the Productivity and Reproducibility of a Measurement Procedure[1]

Nelly Condori-Fernández[1], Oscar Pastor[1]

[1] Department of Information Systems and Computation
Valencia University of Technology
Camino de Vera, s/n, Valencia, Spain
{nelly,opastor}@dsic.upv.es

Abstract. This paper reports an empirical study that used computer science major students as experimental subjects to evaluate the productivity and the reproducibility of RmFFP. This is a functional size measurement procedure designed according to the COSMIC-FFP method for object-oriented systems that are specified using the OO-Method approach. The results show that the productivity of RmFFP is acceptable when compared to other procedures found in the literature. Furthermore, RmFFP produces reproducible functional size assessments.

Keywords: Measurement procedure, COSMIC-FFP, Functional Size, experiment.

1 Introduction

Nowadays, Functional Size Measurement (FSM) plays an important role in software project management due to its extensive use in industry to monitor progress and performance, determine overall productivity, better manage software portfolios, assist in planning, etc. However, despite the fact that measurement method evaluation is crucial in ensuring high-quality size measures, very few of the proposed functional size measures have been evaluated in an experimental way [1], [2].

A new FSM method can be evaluated using the ISO/IEC 14143-3 standard [3], where performance properties (e.g. reproducibility, repeatability, accuracy, convertibility) are given.

In the last few years a number of proposals that measure object-oriented systems based on the COSMIC-FFP standard method have been proposed, such as Bévo et al. [4], Jenner [5], Poels [6], Diab et al.[7], Nagano et al. [8], Azzouz et al. [9], and Habela et al. [10]. However we have found very few empirical studies on the performance properties of COSMIC-FFP. With respect to the productivity of COSMIC-FFP, Nagano carried out an initial analysis to evaluate whether COSMIC-FFP is ease to use [1]. Diab evaluated the accuracy and repeatability of COSMIC-FFP

[1] Supported by the Ministry of Science and Technology under the DESTINO project, ref. TIN2004-03534.

in the measurement of real-time systems by means of the automation of the proposed procedure [7].

This paper focuses on the evaluation of a functional size measurement procedure for object-oriented systems called RmFFP [11] that was designed by mapping the concepts of the COSMIC-FFP method [12] onto the primitives of the OO-Method approach [13]. OO-Method is an automatic software production method that follows the model-driven architecture paradigm. The evaluation of RmFFP was carried out by means of an empirical study that investigates the productivity and reproducibility of RmFFP in estimating functional size from requirement specifications. Productivity is the quantity of size units that can be measured per unit of time. Reproducibility is the closeness of the agreement between results of measurements of the same measurand carried out under changed conditions of measurement [3].

This paper is organized as follows: Section 2 presents an overview of the RmFFP measurement procedure. Section 3 describes the empirical study used to evaluate RmFFP in terms of productivity and reproducibility. Section 4 discusses the analysis and interpretation of the results. Finally, Section 5 sets out our conclusions and indicates further work to be carried out.

2 A Size Measurement Procedure

The RmFFP procedure [11] was proposed in order to estimate the functional size of object-oriented systems from functional requirements specifications obtained using the Requirements Model [14]. This model includes *the Functions Refinement Tree* (FRT), which is a hierarchical decomposition of the business functions of the system. The leaves of this tree represent the functions of the desired system and are the entry point to be considered as *Primary Use Cases*. *Secondary Use Cases* are also possible, which are important for organizing and managing complexity through relationships among use cases stereotyped as EXTEND and INCLUDE. The *Sequence Diagrams* are built semi-automatically from use cases. A Sequence Diagram is represented by means of a set of messages between the required classes to perform the system behaviour. These messages are labelled with different stereotypes (signal, service, query, connect), which allows subsequent identification of the different elements of the OO-Method Conceptual Schema.

RmFFP starts with the definition of the measurement context, which includes purpose, scope, and measurement viewpoint. The scope of RmFFP comprises the functionality to be included in a particular measurement. The measurement viewpoint corresponds to the ‘end-user’ viewpoint, which will focus on an OO-Method requirements specification.

At this point, RmFFP starts a mapping phase to identify the significant primitives of the Requirements Model that contribute to the system’s functional size according to the concepts of the COSMIC-FFP metamodel. For this purpose we defined sixteen mapping rules whose principal purpose is to reduce misinterpretation of COSMIC-FFP generic concepts and facilitate automating of the RmFFP procedure. For instance, each use case is identified as a functional process, each message of the sequence diagram is identified as a data movement type, etc.

As the data movement is the fundamental component of the COSMIC-FFP method, we also defined four additional rules for eliminating duplicated data movements, which are explained with more detail in [15].

Once the data movements have been correctly identified, we proceed with the measurement phase, whose purpose is to produce a quantitative value that represents the functional size of software of a requirements specification. To do this we apply the measurement function, which consists of assigning a numerical value of 1 Cfsu (Cosmic Functional Size Unit) to each data movement. We defined four rules to add these quantified data movements, considering the relationships type between use cases in order to calculate the size of a (use case) functional process and the size of the entire system.

3 Evaluation of the application of RmFFP

In this section, we describe an empirical study carried out to evaluate the application of RmFFP in terms of productivity and reproducibility. In designing the experiment we used the experimental process provided by Wohlin et al. [16].

3.1 Experiment Planning

In order to define the goal of our empirical study we used the Goal/Question/Metric (GQM) template [17], which is described as follows:

"To analyze RmFFP for the purpose of evaluating its productivity and reproducibility from the viewpoint of the researcher in the context of Computer Science students measuring OO-Method requirements specifications".

Two research questions were addressed by this empirical study, the first being whether RmFFP is efficient, and the second being whether RmFFP is reproducible.

Selection of subjects. The subjects were computer science students at the Valencia University of Technology with similar backgrounds in the use of the OO-Method Requirements Model. These subjects were students enrolled in the "Software Development Environments" course from February to June of 2005. The experiment was organized as a mandatory part of this course. Two groups of students were formed because some students could not regularly attend class due to work commitments in companies. The first group was made up of 18 students who had no links to companies (e.g. work experience) and the second group was made up of 17 students who had some connection with companies.

Selection of variables. The independent variable is the variable for which the effects should be evaluated. In our study, this variable corresponds to the FSM procedure and as single treatment: RmFFP.

The dependent variables selected to evaluate RmFFP are as follows:

- *Measurement productivity:* this is obtained by calculating the number of size units, Cfsu, that can be measured per unit of time (e.g. per hour). The time recorded was the time required to apply the mapping rules and measurement rules of RmFFP.
- *Reproducibility* this is obtained by calculating the extent of variability existing in the measures obtained by different subjects and the same measurand.

Hypotheses formulation. The following hypotheses regarding the research questions were defined:

- Hypothesis 1: RmFFP is efficient when compared to reports found in the literature.
- Hypothesis 2: The functional size measures are reproducible applying RmFFP under changed conditions of measurement.

Experimental Tasks. Two groups of experimental tasks were carried out during the training task and measurement task.

The purpose of the *training* was for the subjects to develop the expertise required to measure using RmFFP. To do this, we carried out a training procedure with both groups. The training method used was demonstration/practice [18].

For the demonstration part, we considered the following tasks: (a) Presentation of the OO-Method Requirements Model, (b) use of the RETO tool (that supports the OO-Method Requirements Model), (c) presentation of the RmFFP measurement procedure, and (d) illustration of the use of RmFFP with an illustrative example of a case study.

For the practice part, we considered the following tasks: (a) Application of the theory presented in the case study and guided by the instructor, (b) guided application of RmFFP to a case study (the students could clarify their doubts), and (c) verification of knowledge learned by the student by working out an assigned case study. The time used for all the tasks included in this first session was eight hours distributed over four days.

With respect to the *measurement task*, each subject used the RmFFP guide to measure an OO-Method requirements specification (rent a car). This task was used to collect functional size and measurement time.

Before the subjects took the test, the experiment was conducted with another small group of people in order to improve it and ensure that the documentation was well designed. No changes were necessary as a result of this pre-test.

Instrumentation. The instrumentation used in this experiment included the experimental object and training materials. The experimental object was an OO-Method requirements specification of the Car Rental application. The functional specifications of this case study were developed by the students; however, for this experiment, we used the specification proposed by the course teacher. The training materials[2] were the following: a set of instructional slides on the OO-Method Requirements Model and the RmFFP procedure; a case-study that describes an example of the application of RmFFP, a measurement guide, and another case study to verify the training carried out.

[2] http://www.dsic.upv.es/~nelly/materials

3.2 Experimentation Operation

Execution. The experiment took place in a classroom. The interaction among subjects was controlled to avoid plagiarism. There was no time limit set for measuring the OO-Method requirements specification. We also allowed the use of the material used in the training sections.

Data Recording and Validation. The data recorded by the students was the functional size of the specification given and the time used to carried out this measurement. Once the data were collected, we verified whether the results were valid. We noted that two students made serious mistakes during the application of RmFFP (outliers). For instance, they confused the concepts of INCLUDE and EXTEND (relationships between use cases) for the application of the aggregation functions. Therefore, we did not take into account these two tests in the analysis of the results.

4 Analysis and Interpretation

The analysis and interpretation of the results are divided in four phases according to the research questions previously stated.

4.1 Analysis of the actual productivity

To calculate the measurement productivity of a subject, we divided the subject assessment by the measurement time. The obtained productivity rate for each subject measuring the experimental object is shown in Figure 1. These values oscillate between 90-190Cfsu/hour; the average measurement productivity was 131.48 Cfsu/hour.

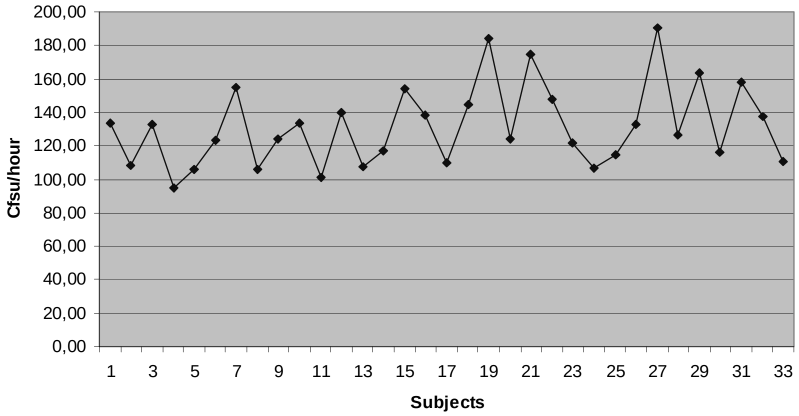


Fig. 1. Rate measurement productivity for each subject using RmFFP procedure

To evaluate the productivity of RmFFP, we considered the analysis carried out by Nagano about measurement productivity[3] applying COSMIC-FFP to 13 requirements specifications [1]. The average measurement productivity obtained was 45 Cfsu/hour. This value was about three times less compared to the Productivity of RmFFP. However, several factors could affect this result, such as:

- The functional specification provided to the subjects in order to size the system requirements. We used the OO-Method Requirements Model, which is based on UML notation with some stereotypes. Nagano used the natural language for the functional specification of the switching systems.
- The mapping rules defined specifically for the OO-Method context that allowed the reduction of the generality of COSMIC-FFP. Nagano applied directly the generic rules of COSMIC-FFP.
- The subjects were well-versed in the OO-Method Requirements Model and the RmFFP measurement procedure.

We also considered the different counting levels for Function Point Analysis (FPA) published by the company Total Metrics [19]. According to these levels, the productivity of an estimator can vary between 200-750 FP/day. As a day is assumed to have 8 working hours, the obtained productivity rate is approximately between 25-93.75 FP/hour, which will depend on the experience of the measurer.

Given that the correlation between function points and Cfsu is linear, and convertibility factor between FPA [20] and COSMIC-FFP is close to one [12], the obtained productivity using RmFFP (131.48 Cfsu/hour) was also higher than the rate reported by Total Metrics. However, we cannot yet draw definite conclusions about the measurement productivity obtained with RmFFP due to external conditions such as the complexity of the experimental object (functional specification), level of experience, etc.

4.2 Analysis of the actual reproducibility

To measure the degree of variation between assessments produced by different subjects using RmFFP, we used a practical statistical equation similar to the one proposed by Kemerer [21]. This equation is calculated by taking the difference in absolute value between the size value produced by a subject i and the average value produced by the other n- 1 subjects in the sample, divided by this average value. The scores (REP_i) closest to zero indicate least variability in the measurement or most reproducibility. These scores were thus obtained for each observation by applying the following equation:

$$REP_i = \left| \frac{\sum_{k=1,k\neq i}^{n} \frac{Values_k}{n-1} - Value_i}{\sum_{k=1,k\neq i}^{n} \frac{Values_k}{n-1}} \right| \qquad \textbf{(1)}$$

[3] Nagano used the term speed for productivity

These REPi values obtained for each subject are presented in the Appendix. The mean of variability (REP_i) for thirty three observations was 5.1 % (See Table 1). This value was compared to the result reported by Kemerer (26.53%) in a study about IFPUG FPA reliability [21], where the reproducibility obtained using RmFFP was higher.

Table 1. Descriptive statistics for RmFFP reproducibility.

Statistics	Reproducibility
Mean	0.051
Standard deviation	0.04164
Min	0.000
Max	0.15

To evaluate the actual reproducibility of RmFFP, we checked the normality of the score obtained (REP_i). To do this, we used the Shapiro-Wilk test. This test was applied to our data and we found that they were normal.

Next, we tested the hypothesis (H2). To do this, we used the t-student statistic since we were working with normal data; we considered the value cero as reference value to compare with the scores assigned by the subjects.

The results of the t-test, shown in Table 2, allow for the rejection of the null hypothesis because the value t obtained was outside the interval. Furthermore, the level of significance obtained was very high ($p < 0.001$). Therefore, we can claim with 95% confidence that the data obtained would satisfy the hypothesis that RmFFP is reproducible.

Table 2. One Sample t-test for the reproducibility.

Statistic	Reproducibility
Mean Difference	0.05091
95% Conf. Interval for the diff.	0.0361 (lower) 0.0657 (upper)
t	7.024
1-tailed p-value	.000

4.3 Validity of results

It is important to ensure that the experimental results are valid for the target population. In this section, we discuss the more important threats [16] related to conclusion validity, internal validity, and external validity of our empirical study.

Conclusion validity. Threats to conclusion validity are concerned with issues that affect the ability to draw the correct conclusion about relations between the treatment and the outcome of the experiment. The following threats were considered:

- Reliability of the application of treatments to subjects. There is a risk that the application is not similar for different persons applying the treatment on different occasions. In our experiment, RmFFP (treatment) was applied following a prescribed procedure for the two defined groups. Hence, the risk of obtaining dissimilar applications for different subjects and occasions was low.

- Random heterogeneity of subjects. All the subjects in the experiment had approximately the same level of experience working with the OO-Method Requirements Model. However, this homogeneity reduces the external validity of the experiment.

Internal validity. Threats to internal validity are influences that can affect the independent variable with respect to causality, without the researcher's knowledge. The following threats were considered:

- Selection. Depending on how the subjects are selected from a larger group, selection effects can vary. In the experiment, the subjects were selected for convenience, i.e., they were students enrolled in the "Software Development Environments" course. This course was selected because it was a specialized teaching unit. Furthermore, the students had the necessary preparation and training, and the experimental task itself fitted well into the scope of this course.
- Instrumentation. This is the effect caused by the artefacts used for experimentation execution. The instruments used for the experiment were verified. First, the specifications of the case studies (objects) were reviewed by an expert of the OO-Method Requirements Model. Second, the measurement guide (guideline) and the survey were verified in advance with a small group of people in order to improve its understandability.

External validity. It is concerned with generalization of the results to industrial practice. Here the following threats were considered:

- Interaction of selection and treatment. This is the effect of not having a representative population in the experiment with which to generalize. In our case, we accept that more experiments with a larger number of subjects (e.g., professionals) will be necessary.
- Interaction of setting and treatment. This is thee effect of not having representative material. In the experiment, we tried to use a representative OO-Method requirement specification of a real case in the MIS functional domain. This specification differs to some extent from "classical" specifications because RmFFP uses sequence diagrams with stereotyped messages as artefacts. However, we believe that RmFFP could also be used in UML sequence diagrams on a manual basis.

5 Conclusions and Future Works

This paper describes an empirical study that evaluates the productivity and reproducibility of RmFFP for estimating the functional size of object-oriented systems from software requirements specifications.

With respect to the efficiency analysis, the data collected indicate that the productivity of the subjects using RmFFP is several times higher than the productivity rate obtained by Nagano [1]. This way, we confirm that a measurement method due to its generic character is least efficient than a measurement procedure, such as RmFFP. We also found that the RmFFP productivity is higher than the industry rates found with IFPUG FPA, which were reported by the company Total Metrics [19].

With respect to the reproducibility analysis, we have corroborated that users of RmFFP produce reproducible assessments. This result can be explained by the training carried out with the subjects. Moreover, the complementary rules defined to control the duplicity of data movements allowed to reduce some interpretation problems of the RmFFP guidelines that could appear during the measurement procedure application.

Finally, we are aware that it is necessary carry out more empirical studies with industry professionals in order to confirm our initial results. In addition, our measurement procedure is currently being automated to be incorporated in the RETO tool, which captures user requirements and generates elements of an OO-Method conceptual schema.

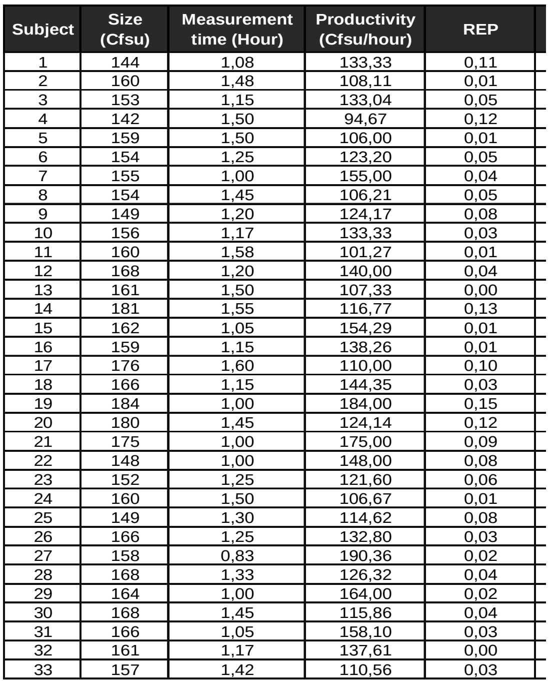
References

1. Nagano,S.; Mase, K., Watanabe Y., Watahiki T., Nishiyama S., Validation of Application Results of COSMIC-FFP, in Australian Software Conference on Measurement (ASCOM), Australia, 2001.
2. Abrahao S., Poels G., Pastor O. A Functional Size Measurement Method for Object-Oriented Conceptual Schemas: Design and Evaluation Issues. Software & System Modelling, Volume 5, Issue 1, Springer Verlag, 2005.
3. ISO, "ISO/IEC 14143-3: Information technology - Functional size measurement - Part 3: Verification of functional size measurement methods", 2003.
4. Bévo V., Lévesque G., and Abran A. UML Notation for Functional Size Measurement Method. In Proc. 9th International Workshop on Software Measurement, Canada, September 8-10, 1999, pp. 230-242.
5. Jenner M.S., "Automation of Counting of Functional Size Using COSMIC-FFP in UML," Proc. 12th Int'l Workshop Soft. Measurement, pp. 43-51, 2002.
6. Poels G., 2003. Functional Size Measurement of Multi-Layer Object-Oriented Conceptual Models, Working Paper, Gent University.
7. Diab H., Koukane F., Frappier M., St-Denis R., 2004, µcROSE: Automated Measurement of COSMIC-FFP for Rational Rose Real Time, Information and Software Technology, Volume 47, Issue 3, 1 March 2005, Pages 151-166.
8. Nagano S., Ajisaka T., Functional metrics using COSMIC-FFP for object-oriented real-time systems, in 13th International Workshop on Software Measurement (IWSM), Montreal, Canada, 2003.
9. Azzouz S., Abran A., "A Proposed Measurement Role in the Rational Unified Process and its Implementation with ISO 19761: COSMIC-FFP" in Software Measurement European Forum, Rome, Italy, 2004.
10. Habela P., Glowacki E., Serafinski T., Adapting Use Case Model for COSMIC-FFP based Measurement, in the 15th International Workshop on Software Measurement, Montreal, Canada, Shaker-Verlag, 2005.
11. Condori-Fernández N., Abrahao S., Pastor O., Towards a Functional Size Measure for Object-Oriented Systems from Requirements Specifications Abstract Functional Size Measurement. IEEE Quality Software International Conference, Germany, 2004.

12. Measurement Manual: The COSMIC Implementation Guide for ISO/IEC 19761: 2003, version 2.2.
13. Pastor O., Gomez J., Insfran E., Pelechano V., 2001. “The OO-Method approach for information systems modelling: from object-oriented conceptual modelling to automated programming”, Information Systems 26, pp. 507-534.
14. Insfrán E., Pastor O. and Wieringa R., 2002. Requirements Engineering-Based Conceptual Modelling. Journal Requirements Engineering (RE), Springer-Verlag, pp. 61-72.
15. Condori-Fernández N., Abrahao S., Pastor O., The Problem of Data Movements Duplicity in a Measurement Procedure, In Proc. of 9th Iberoamerican Workshop on Requirements Engineering and Software Environments, La Plata-Argentina, 2006 (in Spanish).
16. Wohlin C., Runeson P., Höst M., M. C. Ohlsson, B. Regnell, and A. Wesslén, Experimentation in Software Engineering: An Introduction, 2000.
17. Basili V. R. and Rombach H. D., 1988. The TAME Project: Towards Improvement-Oriented Software Environments, IEEE Transactions on Software Engineering, pp. 758-773.
18. DOE Handbook, Alternative Systematic Approaches to Training, 1074-95, January 1995.
19. Total Metrics, "Levels of Counting", Australia, August 2001.
20. IFPUG, 1999. Function Point Counting Practices Manual, Release 4.1, International Function Points Users Group, Mequon, Wisconsin, USA.
21. Kemerer C. F., "Reliability of Function Points Measurement", Communications of the ACM, vol. 36, pp. 85-97, 1993.

Appendix

Table A1. Data set used in the experiment



Subject	Size (Cfsu)	Measurement time (Hour)	Productivity (Cfsu/hour)	REP
1	144	1,08	133,33	0,11
2	160	1,48	108,11	0,01
3	153	1,15	133,04	0,05
4	142	1,50	94,67	0,12
5	159	1,50	106,00	0,01
6	154	1,25	123,20	0,05
7	155	1,00	155,00	0,04
8	154	1,45	106,21	0,05
9	149	1,20	124,17	0,08
10	156	1,17	133,33	0,03
11	160	1,58	101,27	0,01
12	168	1,20	140,00	0,04
13	161	1,50	107,33	0,00
14	181	1,55	116,77	0,13
15	162	1,05	154,29	0,01
16	159	1,15	138,26	0,01
17	176	1,60	110,00	0,10
18	166	1,15	144,35	0,03
19	184	1,00	184,00	0,15
20	180	1,45	124,14	0,12
21	175	1,00	175,00	0,09
22	148	1,00	148,00	0,08
23	152	1,25	121,60	0,06
24	160	1,50	106,67	0,01
25	149	1,30	114,62	0,08
26	166	1,25	132,80	0,03
27	158	0,83	190,36	0,02
28	168	1,33	126,32	0,04
29	164	1,00	164,00	0,02
30	168	1,45	115,86	0,04
31	166	1,05	158,10	0,03
32	161	1,17	137,61	0,00
33	157	1,42	110,56	0,03