Comparative Evaluation of Functional Size Measurement Methods: An Experimental Analysis

Silvia Abrahão¹, Geert Poels² and Oscar Pastor¹

¹Department of Computer Science and Computation
Valencia University of Technology
Camino de Vera, s/n, 46022, Valencia, Spain
Email: sabrahao, opastor@dsic.upv.es

²Faculty of Economics and Business Administration
Ghent University
Hoveniersberg 24, 9000 Ghent, Belgium
Email: geert.poels@ugent.be

Abstract – A number of Functional Size Measurement (FSM) methods have been proposed in the literature, but so far there has been no systematic evaluation of these methods. A major criticism is that little attention has been paid to the empirical validation of FSM methods. By empirical validation we refer to the evaluation of the efficacy of the method and its likely adoption in practice using experimental techniques and statistical data analysis. This paper describes a laboratory experiment which compares Function Points Analysis, a standard FSM method supported by the International Functional Point Users Group (IFPUG FPA) and OO-Method Function Points (OOmFP), a recently proposed FSM method for sizing object-oriented (OO) software systems that are developed using the OO-Method approach. The goal is to investigate whether OOmFP results in better size assessments and is more likely to be adopted in practice, within the context of an OO-Method development process. As OOmFP and IFPUG FPA are FSM methods, only the functional size of a software system is quantified, meaning that only the functional system requirements as seen from the user’s perspective are considered as contributing to system size.

The methods are compared using a range of performance-based and perception-based variables, including efficiency (effort required to apply the methods), reproducibility, accuracy, perceived ease of use, perceived usefulness and intention to use. An important contribution is the development and empirical testing of a theoretical model for evaluating FSM methods in general.

The results show that OOmFP is more time-consuming than IFPUG FPA but the measurement results are more consistent and accurate. Also, OOmFP is perceived to be a useful FSM method in the context of OO-Method systems development. Moreover, the theoretical model proposed might help to bridge the gap between research and practice in Empirical Software Engineering research, as it addresses the issue of method adoption in practice, which has been ignored by ESE researchers.


1. INTRODUCTION

Functional Size Measurement (FSM) methods are intended to measure the size of software by quantifying the functional user requirements that the software delivers. A FSM method measures the logical external view of the software from the user’s perspective by evaluating the amount of functionality to be delivered. The capability to accurately quantify the size of software in an early phase of the development project and to control the functionality delivered during the software lifecycle, is critical to software project managers for evaluating potential risks, developing accurate project estimates and having early project indicators.

The most commonly used FSM method is IFPUG Function Point Analysis (IFPUG FPA) [23]. It is based on the method proposed by Alan Albrecht [3]. This technique was developed
specifically to measure the amount of data that each function accesses as an indicator of functional size. One of the most serious practical limitations of existing FSM methods is their inability to cope with object-oriented systems. A number of approaches have been proposed in the literature to address this issue [5], [6], [19], [20], [21], [24], [30], [33], [34], [38], [47], [48], [52], [54], [57], [59], [60], [63], but so far, none of these has been widely accepted in practice.

A new method, OO-Method Function Points (OOmFP) [1], [43] has been proposed to overcome these difficulties in the context of an automated software production method called OO-Method [44]. OOmFP was designed to conform to the IFPUG FPA counting rules [23] since it redefines the IFPUG counting rules in terms of the concepts used in OO-Method. As design science is not complete without evaluating the research outputs [36], this paper reports on an empirical study that investigated whether OOmFP is better than IFPUG FPA in sizing OO software systems that are developed according to the OO-Method.

Because of the lack of generally accepted and sufficiently rigorous validation processes for FSM methods it seems difficult to evaluate new proposals, both on a practical and on a theoretical level. In response, the ISO published a series of standards for FSM. Part 1 [25] of the standard provides the concepts of FSM and establishes a basis against which existing and new FSM methods can be compared. Part 2 [26] provides a process for checking whether a candidate FSM method conforms to the concepts described in part 1. Part 3 [27] provides a framework for verifying the statements of an FSM method and/or for conducting tests with the following performance criteria: repeatability and reproducibility, accuracy, convertibility, discrimination threshold, and applicability to functional domains. However, this standard does not offer any support to compare different FSM methods in order to decide which one is the best.

Hence, how can we determine whether a particular FSM method is better than another FSM method? Empirical studies can help determine the efficacy of proposed theories and methods [55] [62]. This is not only useful for researchers in order to prove their theories, but also for practitioners to select a new technology. We believe that future research directions for empirical software engineering should take into account not only the efficacy of a method in achieving its objectives, but also the likelihood of the method being adopted in practice.

Following these ideas, a theoretical model for assessing the efficacy and acceptance of different FSM methods is presented. This model is based on the Method Evaluation Model (MEM) [40], a model for evaluating IS design methods, which incorporates both aspects of method “success”: actual efficacy and likely adoption in practice. It combines Rescher’s theory of pragmatic justification [49], a theory for validating methodological knowledge, and the Technology Acceptance Model of Davis [17], a model for explaining and predicting user acceptance of information technology.

The constructs of the MEM were adapted to evaluate the efficacy and acceptance of FSM methods. Efficacy is defined as the efficiency and effectiveness of a FSM method in achieving its objectives. Evaluation of the efficacy of a FSM method requires measurement of effort required (inputs) and the quality of the measurement results (outputs). Thus, when we compare different FSM methods, the “best” one can be defined as the most efficient and/or the most effective and/or the most likely to be adopted in practice (though the latter variable depends on the former two according to the MEM).

The model for evaluating FSM methods provides a range of performance-based and perception-based variables, including efficiency (effort required to apply the methods), reproducibility, accuracy, perceived ease of use, perceived usefulness and intention to use. This model was applied in the design of a laboratory experiment that was carried out in the Valencia University of Technology.

This paper is organized as follows. In Section 2 an overview of FSM methods is given and efforts to empirically validate them are discussed. The focus is on variants of FPA for measuring OO systems. In Section 3 the main principles of IFPUG FPA and OOmFP are described. We then provide an overview of the MEM in Section 4. This is followed by a description of our experimental procedure to compare OOmFP and IFPUG FPA in Section 5. This procedure is
described in terms of a Theoretical Model that is proposed for evaluating FSM methods. The experiment results are analysed and interpreted in Section 6. Next, in Section 7, we discuss our findings and reflect upon our procedure and make some suggestions as to how it might be improved for future investigations. Lastly, the conclusions are presented in Section 8.

2. FUNCTIONAL SIZE MEASUREMENT METHODS

Function Point Analysis (FPA) proposed by Albrecht [3] can be considered as the first FSM method. FPA’s view on functional size is that a software system consists of logical files and functions that perform transactions using or altering the data in the logical files. The amount of data (types) ‘handled’ by a logical file or transaction function determines the amount of functionality that the software delivers, hence its functional size.

FPA has evolved from Albrecht’s method and is currently supported by the International Function Point User Group (IFPUG), which has proposed detailed rules for applying FPA [23]. Hereafter, when we refer to FPA, we always refer to the FPA method in its current form as supported by IFPUG.

A number of variants have been proposed taking alternative views on functional size, the most important of which are Mark II FPA [58], Full Function Points (FFP) [53] and recently COSMIC FFP [2].

Mark II FPA [58] aimed to improve on Albrecht’s size measurement, and to be compatible with ideas from structured analysis and design. The intention of FFP [53] is to extend IFPUG FPA to capture the functional size of real-time applications. COSMIC-FFP [2] was proposed to measure real-time or business application software that is organized in different layers and peer items.

Moreover, in order to cope with object-oriented systems measurement several approaches have been proposed. We group these approaches in three categories of FSM methods for OO systems. A first category is composed of methods that are compliant to IFPUG FPA. They reformulate the IFPUG rules in terms of OO concepts to facilitate the function points counting process. The final result of a size assessment is (or should be) the same as what would have been obtained by directly applying IFPUG FPA. A second category contains FSM methods that are not compliant to IFPUG FPA, but take a view on functional size that is related to the IFPUG FPA view. For these methods, the underlying model of the items that are considered to contribute to functional size is a data-oriented abstraction similar to, but not the same as that of IFPUG FPA. The count that is obtained would, however, not be considered a valid count according to the IFPUG rules for FPA. Apart from these two approaches, there is a third category of methods that take a fundamentally different view on functional size by no longer distinguishing between data and transactions, but considering the object (actually the class definition of an object) as the main item that contributes to functional size. These three categories are given below.

In the first category we find a proposal of IFPUG [24], and other proposals by Lehne [34], Fetcke [19], Uemura et al. [57], and Abrahão and Pastor [1].

IFPUG defined a case study [24], which illustrates the application of the IFPUG counting rules to an object-oriented analysis and design. This case study uses Object Models in which the methods of classes are identical to the specified in the requirements document. Under this assumption, methods can be directly counted as transactions. Lehne [34] presents an experience report in function points counting for object oriented analysis and design using a method called OOram. Using this method a system specification is obtained from three conceptual models: a system user model, a system requirements model (based on use cases), and a system design model. However, the mapping between FPA concepts and the primitives of the OOram conceptual models was not completely established.

Fetcke [19] demonstrates the applicability of FPA as a FSM method for OO specifications. He suggested a mapping to the OO-Jacobson method (based on use cases) [28] starting from an abstract Function Points Model. This mapping has been formulated as a small set of rules that
support the current IFPUG counting rules. Uemura et al. [57] propose FPA measurement rules for design specifications based on UML (Unified Modeling Language) and demonstrate a function point measurement tool, whose input products are design specifications developed using the Rational Rose CASE tool.


In the second category are the proposals of Whitmire [59], [60], ASMA [6], Kammelar [30], Antoniol and Calzolari [5], and Ram and Raju [48].

Whitmire [60] adapted the concepts of FPA to an OO approach. Whitmire based his approach on an extended Class Diagram including messages sent between classes. An object is a collection of data (attributes and properties) and functional logic (methods). The data define the state of the object and the methods define the behaviour of the object.

The Australian Software Metrics Association (ASMA) [6] takes an approach similar to that proposed by Whitmire. Methods implement services that are delivered by objects to the client. These methods are considered as transactions. The complexity of methods is weighted based on the number of accessed attributes and inter-object communications. Objects are treated as files, their attributes determining their complexity. In the proposal of Kammelar [30] the functional size is expressed in component object points (COP’s) obtained from elements which may be candidates for reusable software components. The measurement steps and rules are presented but no example is commented.

In 1996, Whitmire [59] proposes 3-D Function Points, an extension of FPA that measures three dimensions of a software project: data, control and functionality. This is accomplished by adding transformations (algorithms) and transitions (changes in application state) to the conventional FPA focus on data. While 3D Function Points addresses many of the problems encountered using traditional FPA and it measures different dimensions of functionality, size measurement is only accomplished at the class level.

Antoniol and Calzolari [5] propose Object-Oriented Function Points, a method analogous to FPA, but with measurement rules that are based on the elements of a static Object Model. A relevant aspect of this method is its flexibility. An organization may experiment with several measurement procedures and find one that is more suited to its environment. However, Object-Oriented Function Points does not distinguish the different types of transactions (input, output or inquiry) as in FPA. The operations are treated as generic method requests between objects. Ram and Raju [48] propose Object Oriented Design Function Points as a variation of IFPUG FPA to measure OO systems. As in [5], this method does not distinguish different types of transactions.

Finally, in the third category are the proposals of Laranjeira [33], Rains [47], Thomson et al. [54], Zhao and Stockman [63], Hastings [21], Sneed [52], Gupta and Gupta [20], and Minkiewicz [38].

Laranjeira [33] reviewed traditional FSM methods. His study showed difficulties in applying FPA to non-business information systems. Consequently, he proposed a FSM method based on layers of abstraction. Increasing levels of detail are revealed at each layer leading to more precise estimates. The only evaluation of Laranjeira’s method, however, is a demonstration proof (by Laranjeira himself) on two projects. The proposals of Rains [47] and Hastings [21] for measuring OO systems focus on implementation aspects (based on ADA package characteristics) and lack generalization.

In the proposal of Thomson et al. [54] each system requirement is expressed as a single abstract functional capability (expressed in use cases) that is measured using a set of metrics: number of use cases, events, classes, etc. However, no example is shown to explain the approach. Zhao and Stockman [63] proposes Object-Function Point Analysis as a extension of the Laranjeira approach with physical size factors and reuse size factors.

Object-Points proposed by Sneed [52] are measured by weighting the following elements: object types, object attributes, object relations, methods, messages, parameters in messages,
message source, message destinations and percentage of reuse. The obtained value is weighted using 10 influential factors. Gupta and Gupta define Object Points [20] as a new measure with a structure similar to FPA. But instead of user functionality, objects are the items that are measured. Objects are weighted based on their “complexity”, derived from so called effective attributes, instance and message connections.

Predictive Object Points (POPs) [38] incorporates three dimensions of OO systems: the amount of functionality the software delivers, communication between objects and reuse through inheritance. Unlike FPA-based methods, which are based on the data and transactions model, POPs are based on objects and their characteristics.

A main criterion for the success of a proposed method is its adoption in practice. Some methods for sizing OO systems that were discussed above, like 3D Function Points [59], and Predictive Object Points [38] have been developed in the context of companies (e.g. Boeing, Price Systems). No details about their actual usage in these companies are publicly available, however.

Some empirical studies that investigate the ability of a FSM method to produce a size value that can be used for effort prediction, have been published in the literature. For instance, Moser et al. [41] present an empirical study using 36 projects that demonstrate that the System Meter method which explicitly takes reuse into account predicts effort substantially better than a model based on function points. However, as far as we know, no study has been published that contains a rigorous empirical evaluation of the suitability of IFPUG FPA for measuring the functional size of OO systems.

For the other, academic proposals presenting FPA variants or extensions for OO systems, no systematic evaluation of their ‘validity’ could be found in the literature. The only validation of any kind are proofs of concept presented by the own method developers, mostly in the same paper that proposes the new FSM method. Typical for this situation is Laranjeira [33] who applied his method to two projects. Some methods never transcend the stage of original proposal, and the paper in which they are proposed does not even offer an illustrative example of how the rules should be applied [27], [34], [42], [45], [47], [48], [52], [54], [63].

The work presented here is different in the sense that we evaluate OOmFP using a controlled experiment based on a general evaluation model for FSM methods that is grounded in theory. As a first-category FSM method, OOmFP is compliant to IFPUG FPA. At the same time it is meant to improve upon IFPUG FPA for the function point counting of OO systems that are developed using the OO-Method approach. Therefore, IFPUG FPA is treated as the benchmark FSM method against which the new proposal, OOmFP is evaluated. To predict the success of OOmFP in terms of its actual usage, the actual and perceived efficacy of applying this method and the intention to use the method in practice are evaluated and compared to IFPUG FPA. It is this attention to the necessary evaluation component of a design science project that distinguishes our work from the other proposals discussed in this section.

3. OVERVIEW OF IFPUG FPA AND OOMFP

In this section we first present IFPUG FPA, a standard FSM method, and next OOmFP, a specific method for sizing OO systems developed using the OO-Method approach.

3.1 IFPUG Function Points Analysis

In IFPUG Function Point Analysis (IFPUG FPA), the functional size is calculated by counting the number of five types of components, namely: External Inputs (e.g., transaction types entering data into the system), External Outputs (e.g., report types producing information output), External Inquiries (e.g., types of inquiries supported), Internal Logical Files (data maintained by the system), and External Interface Files (e.g., data accessed by the system but not maintained by it). These
components are weighted (according to their complexity – low, average or high), and their weights are summed. This value is then adjusted using fourteen general system characteristics to produce a function point measurement.

For the purposes of this experiment we do not take into account the adjustment phase. As demonstrated by Jeffery and Stathis [29] and Lokan [35] the adjustment factor does not improve the FSM method and its relationship with effort. It is useful for understanding project cost drivers and for comparing similar projects, but the doubts about its construction are not balanced by any practical benefit. Moreover, the ISO view on FSM also excludes any adaptation factor that does not directly measure functional size.

Fetcke et al. propose [18] a generalized representation for FSM methods. According to this representation, FSM requires two steps of abstraction: first the elements that contribute to the functional size of the system are identified (identification step) and second these elements are mapped into numbers (measurement step). The aim of the first step is to select the elements in the requirements documentation that add to the functional size of the system and to abstract from those that do not contribute to functional size. The second step constructs an abstract model of the elements identified in the first step according to the meta-model of the FSM method that is used. This meta-model reflects the particular view on functional size taken by the FSM method. The elements in the abstract model are subsequently mapped into numbers representing the (relative) amount of functionality that is contributed to the functional size of the system. Finally the numbers are aggregated into an overall functional size value.

Figure 1 shows the IFPUG FPA Measurement Procedure according to the generalized representation for FSM methods.

![Figure 1. An Abstraction of IFPUG FPA Measurement Procedure (adapted from [18] and [9])](image-url)

It must be noted that the distinction between the two steps of abstraction in IFPUG FPA is not precisely defined due to a lack of definitions and abstractions in the method.

In the **identification step**, given a User Requirements Specification, data models and process models are specified, and windows, screens and reports are designed. This is an implicit step that is not described as part of measurement process but the IFPUG Count Practices Manual [23] suggests that measurement may be based on one or more of the following components: user requirements, data models (Entity-Relationship Diagrams), process models (Data Flow Diagrams) and designs of windows, screens, and reports. In some examples, the requirements specification
document stand alone as the basis for the measurement, but most of the examples presented in the manual include a data or process model, and designs of windows, screens, and reports. Using these documents, models and designs, significant elements for measurement are identified.

In the measurement step, the previously identified elements are captured in an abstract model that is specified according to the IFPUG FPA Measurement Abstract Model. This ‘meta-model’ describes all elements that contribute to the functional size of the system, according to the IFPUG view. The construction of the abstract model is an implicit process that is performed by applying the IFPUG measurement rules, in particular those that require to identify and classify transactional functions (External Inputs, External Outputs, and External Inquiries) and logical files (Internal Logical Files, External Interface Files), and to rate their complexity. The abstract model is finally used to assign a numerical value to the system representing its functional size (i.e. translating complexity-rated and classified functions into function point values, aggregating the values) [46].

3.2 OO-Method Function Points

Basically, in an OO-Method system specification we can distinguish two models: the Conceptual Model and the Execution Model. When facing the conceptual modeling step of a given information system, we have to determine the components of the object society without being worried about any implementation considerations.

In the problem space level a precise system definition (formal specification) is obtained through transformation of a Conceptual Model. The Conceptual Model is based on the specification of four conceptual model views that describe the system functionality within a well-defined OO framework. This formal specification acts as a high-level system repository. Furthermore, using an Execution Model, an OO software system which is functionally equivalent to the formal specification is generated in an automated way.

OO-Method Function Points (OOmFP) was designed to conform to the IFPUG FPA counting rules. It redefines the IFPUG counting rules in terms of the conceptual primitives defined in OO-Method. Figure 2 shows a representation of the OOmFP Measurement Procedure.

![Figure 2. An Abstraction of OOmFP Measurement Procedure](image)
As in IFPUG FPA there are two main steps: identification and measurement, but now clearly separated. In fact the identification step is strictly spoken not part of OOmFP (although required by it before measurement can be performed) as it refers to OO-Method conceptual modeling.

In the **identification step**, given a User Requirements Specification, an OO-Method conceptual schema is built, including the following perspectives of an object-oriented system:

- **Data**: the information that is maintained by the system. This information is defined in the Object Model.
- **Process**: the computations that a system performs are defined in the Object Model along with a precise definition of the semantics associated to object state changes in the Functional Model.
- **Behaviour**: the dynamic interactions of a system in terms of valid object lives and inter-object interaction, defined in the Dynamic Model.
- **Presentation**: the user interactions with the system, defined in the Presentation Model.

Using these models, significant elements for measurement are identified. In the **measurement step**, an abstract model of requirements (i.e. identifying, classifying and weighting transactional and data functions) is specified according to the OOmFP Measurement Abstract Model. This is done by applying the OOmFP measurement rules.

Finally, the functional size is quantified (i.e. translating complexity-rated and classified functions, and aggregating the values). As shown in the figure, the measurement step starts from the elements identified in an OO-Method conceptual schema. As a consequence, the functional size of the OO system is calculated in the problem space. For more information on the method usage and measurement rules refer to [1] and [43].

### 4. THE METHOD EVALUATION MODEL

The Method Evaluation Model [40] is a theoretical model, which incorporates both actual efficacy and likely adoption in practice. At the top level, the model consists of the same constructs as the Technology Acceptance Model (TAM) [17], but now adapted for evaluating methods. These constructs are:

- **Perceived Ease of Use**: the degree to which a person believes that using a particular method would be free of effort.
- **Perceived Usefulness**: the degree to which a person believes that a particular method will be effective in achieving its intended objectives.
- **Intention to Use**: the extent to which a person intends to use a particular method.

These central constructs are called the Method Adoption Model (shown in Figure 3). This model can be extended with additional constructs that provide inputs to and outputs from this model. These additional constructs are:

- **Actual Efficiency**: the effort required to apply a method. This represents an input variable to the Method Adoption Model.
- **Actual Effectiveness**: the degree to which a method achieves its objectives. This also represents an input variable to the Method Adoption Model.
- **Actual Usage**: the extent to which a method is used in practice. This represents an output variable from the Method Adoption Model.

Actual Usage of a method will be determined by Intention to Use. This follows from empirical studies of the TAM, which have reported a strong significant causal link between behavioural intention to use and actual behaviour [37].

According to Figure 3, actual efficacy only has an indirect effect on decisions to use a method, through perceptions of efficacy.
This theoretical model recognises that for a method to be successful, it not only has to improve task performance, but also that people need to be agreeable to use it. In this model, the “ultimate” dependent variable is Usage in Practice - this is when research has an impact on practice. This is determined by Intention to Use, which is in turn determined by perceptions of Ease of Use and Usefulness. Actual Efficiency and Effectiveness (pragmatic success as defined by Rescher) determine intentions to use a method only via perceptions of ease of use and usefulness.

5. EXPERIMENT DESIGN

In this section we describe the experiment we have carried out to empirically evaluate the proposed FSM method (OOmFP) in comparison with the IFPUG FPA. The experimental design was guided by the framework for experimental software engineering suggested by Wholin et al. [61].

The goal of the experiment was to investigate which method, OOmFP or IFPUG FPA, results in the best functional size assessment of object-oriented systems when measuring the user requirements. In other words, which method has the highest efficacy and/or likely adoption in practice.

In terms of the Goal/Question/Metric (GQM) template for goal-oriented software measurement [8], the goal pursued in this experiment is: analyze functional size measurements for the purpose of evaluating OOmFP and IFPUG FPA with respect to their efficacy and adoption from the point of view of the researcher. The context of the experiment is the OO-Method approach to the conceptual modeling and development of object oriented systems as performed by students in the Department of Computer Science at the Valencia University of Technology.

The broad research questions addressed by this experiment are:

- **Research Question 1:** Is the actual efficacy of OOmFP in measuring the functional size of OO systems based on their functional user requirements higher than that of IFPUG FPA?
Research Question 2: Is the perceived efficacy of OOmFP and the intention to use OOmFP higher than the perceived efficacy of IFPUG FPA and the intention to use IFPUG FPA?

Research Question 3: Is OOmFP likely to be adopted in practice?

Research Question 4: Is the Method Adoption Model a valid model and measurement instrument for evaluating FSM methods?

As previously mentioned the evaluation of the efficacy of a method requires measurement of the effort required (inputs) and the quality of the results (outputs). Thus, the research questions are broken down in the following questions:

- Research Question 1 (Efficiency): how efficient is OOmFP compared to IFPUG FPA for sizing object-oriented systems based on requirements specifications?
- Research Question 1 (Effectiveness): how effective is OOmFP compared to IFPUG FPA for sizing object-oriented systems based on requirements specifications? In order to measure effectiveness we select the following criteria described in Part 3 of the standard ISO/IEC for functional size measurement [27]:
  - Reproducibility: does OOmFP produce more consistent measurements of functional size when used by different people than IFPUG FPA? Reproducibility refers to the use of the method on the same product in the same environment by different subjects. The obtained results should be identical.
  - Accuracy: does OOmFP produce more accurate measurements of functional size from requirements specifications than IFPUG FPA?
- Research Question 2 (Perception of Efficacy and Intention to Use): is the perceived efficacy of OOmFP and the intention to use OOmFP higher than with IFPUG FPA?
- Research Question 3 (Acceptance): Is OOmFP likely to be adopted in practice?
- Research Question 4 (Validity): is the Method Adoption Model a valid theoretical model and measurement instrument for evaluating FSM methods?

5.1 Planning

5.1.1 Selection of subjects

The subjects that participated in this study were 22 students in the final year of Engineering in Computer Science with a major in Information Systems. Final year students were used as proxies for practitioners for the following reasons:

- Accessibility: the possibility of getting practitioners is too difficult given time and cost constraints. In [14] the costs and benefits of empirical studies with students are discussed.
- Similarity: final year students were the closest to practitioners [22]. Students are the next generation of professionals and are therefore close to the population under study [32].

The students aged between 22 and 24 years and they had similar backgrounds (they attended the same courses about project development and management). The subjects were chosen for convenience, i.e., they were students enrolled in the Software Development Environments course during the period of February until June of 2003. This course was selected because it was an advanced unit (where students learn advanced techniques about software development), also, the necessary preparation and task itself fitted well into the remit of this course.

Originally, the content of this course includes conceptual modeling and development approaches, but we have included a new topic about functional size measurement. Thus, during this course, these students were trained in the use of UML and OO-Method approaches for the purpose of conceptual modeling, and IFPUG FPA and OOmFP for the purpose of functional size measurement and project estimation.
5.1.2 Variables selection

The independent variable is a single variable, which is the method used by subjects to size an OO system. The independent variable has two levels, corresponding to the different sizing methods being compared: OOmFP [1] and IFPUG FPA [23].

We distinguish between two types of dependent variables: performance based (objective) and perception based (subjective) variables. The classification of dependent variables in this experiment is shown in Table 1. Performance based variables measure how well subjects perform the experimental task. These variables operationalize the actual efficacy constructs in the Method Evaluation Model, i.e. the external variables that are used as inputs to the Method Adoption Model. Three performance-based variables are distinguished:

- Sizing effort (D1): the time taken by a subject to complete the measurement task.
- Reproducibility (D2): the agreement between the measurement results of different subjects using the same method.
- Accuracy (D3): the agreement between the measurement results and the true value.

Perception based variables measure the perceived efficacy and subsequent intention to use OOmFP and IFPUG FPA. These variables operationalize the constructs of the Method Adoption Model. Three perception-based variables are distinguished:

- Perceived Ease of Use (D4): the degree to which a subject believes that using a particular method would be free of effort.
- Perceived Usefulness (D5): the degree to which a subject believes that a particular method will be effective in achieving its intended objectives.
- Intention to Use (D6): the degree to which an individual intends to use a particular method as a result of his perception of the method’s efficacy.

<table>
<thead>
<tr>
<th>Dependent Variables</th>
<th>Efficiency</th>
<th>Effectiveness</th>
<th>Adoption</th>
</tr>
</thead>
<tbody>
<tr>
<td>Performance Based</td>
<td>Sizing Effort</td>
<td>Reproducibility</td>
<td>Accuracy</td>
</tr>
<tr>
<td>Perception Based</td>
<td>Perceived Ease of Use</td>
<td>Perceived Usefulness</td>
<td>Intention to Use</td>
</tr>
</tbody>
</table>

5.1.3 Experimental Treatments

The treatments correspond to the two levels of the independent variable: use of OOmFP versus use of IFPUG FPA to size an OO system. We choose a within-subject design experiment to control for differences in human ability.

In a within-subject design each subject contributes an observation for each treatment, which is an additional advantage given the small sample size. In order to cancel out a possible learning effect due to similarities in the treatments (i.e. the relatedness of both FSM methods), we consider the sequence in which the tests are given to the subjects as a factor.

In addition, there is a learning effect due to the same requirements specification is used for all students. If one system is used, the order of applying the methods might introduce a confounding effect. However, this learning effect is also cancelled taking into account the sequence of the tests as a factor.

Using the counterbalancing procedure, the subjects were randomly assigned in two groups, with equal numbers in each group and tests presented in a different order:

- Experimental Group 1: IFPUG FPA first and OOmFP second (n=11)
- Experimental Group 2: OOmFP first and IFPUG FPA second (n=11)
5.1.4 Instrumentation

The instrumentation used in this experiment includes the experimental object, training materials and a post-task survey.

The experimental object was a requirement specification document for building a new Project Management System (PMS) for a fictitious company. This document describes the requirements for the system using the IEEE 830 standard [4]. Using this standard a software requirements specification is described in terms of functionality (what is the software supposed to do from the user's perspective), external interfaces (interaction of the software with people, hardware and other software), performance (such as availability, response time, etc.), quality attributes (such as portability, maintainability, security, etc.), and design constraints imposed by implementation (required standards, implementation language, policies for database integrity, operating environments, etc.).

The PMS should support the following transactions: project maintenance (create, change, delete, tasks assignment), types of task maintenance (create, change, delete), tasks maintenance (create, change, delete), user maintenance (create, change, delete, change password), and inquiries (users, projects and type of tasks).

An example of a functional requirement is: when the company starts a new project the responsible employee of the project must enter the data in the system. A project is created with the following data: identification number, description, name of responsible employee, start date, estimated duration, estimated final date, actual duration (sum of the duration of the tasks made until now), cost (sum of all tasks costs associated to the project), situation (0=developing, 1=little delay, <10% more than the estimated time, 2=very delay, >= 10% more than the estimated time, 3=finished), observations.

The following training materials were prepared for all subjects: a set of instructional slides, describing each FSM method and the procedure for applying it; a worked example showing how the methods could be applied to an example; and a measurement guideline summarizing the measurement rules of the methods.

The worked example for the IFPUG FPA training session included a requirements specification document of a Banking System as well as the specification of an ER model and screens of the system. The worked example for OOmFP included the requirement specification of a Library System and a complete specification of an OO-Method conceptual model.

We also defined a post-task survey with 14 closed questions, which were the items used to measure the constructs of the Method Adoption Model [39]. These items were formulated using a 5-point Likert scale, using the opposing statements question format. The order of the items was randomized and half the questions negated to avoid monotonous responses.

The perceived ease of use is measured using 5 items on the survey (Questions 1, 3, 4, 6, and 9). The perceived usefulness is measured using 6 items on the survey (Questions 2, 5, 8, 10, 11, and 13). Finally, the intention to use is measured using 3 items on the survey (Questions 7, 12, 14). The survey is shown in Appendix A.

One of the major advantages of using the MAM and the measurement instrument associated is that it is based on previous works were similar surveys were used and validated in the context of technology adoption [17], [37] and IS methods adoption [40]. Recently, a field study that uses the same kind of instrument to explain software developer acceptance of methodologies was published [50].

5.1.5 Experimental Tasks

The experiment included the following experimental tasks: conceptual modeling task, sizing task, and post-task survey.

In the conceptual modeling task, each experimental group was given the requirements specification and asked to specify a conceptual schema using the OO-Method approach. Subjects
were allowed to use the Oliva Nova™ Modeling Software to document their models. Moreover, they were allowed to refer to the training materials while performing this task.

In the sizing task, each experimental group used IFPUG FPA and OOmFP in a different order. Sizing with OOmFP includes a measurement task where the measurement rules of the method are applied. However, sizing with IFPUG FPA includes two activities: a modeling task, where an ER model and screen prototypes are developed and a measurement task, where the measurement rules are applied.

Finally, in the post-task survey, when the subjects had finished the sizing task for a method, they were asked to complete a survey for evaluating the FSM method they had used. Table 2 shows the sequence of the training sessions and experimental tasks performed.

Table 2. Summary of Training Sessions and Experimental Tasks

<table>
<thead>
<tr>
<th>Training Sessions and Experimental Tasks</th>
<th>Experimental Group 1</th>
<th>Experimental Group 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trial 1</td>
<td>Training session in conceptual modeling</td>
<td></td>
</tr>
<tr>
<td>Trial 2</td>
<td>Conceptual Modeling Task</td>
<td></td>
</tr>
<tr>
<td>3 Training session in IFPUG FPA</td>
<td>Training session in OOmFP</td>
<td></td>
</tr>
<tr>
<td>4 Sizing Task with IFPUG FPA</td>
<td>Sizing Task with OOmFP</td>
<td></td>
</tr>
<tr>
<td>- Modeling Task</td>
<td>- Measurement Task</td>
<td></td>
</tr>
<tr>
<td>- Measurement Task</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5 Post-task survey for IFPUG FPA</td>
<td>Post-task survey for OOmFP</td>
<td></td>
</tr>
<tr>
<td>6 Training session in OOmFP</td>
<td>Training session in IFPUG FPA</td>
<td></td>
</tr>
<tr>
<td>7 Sizing Task with OOmFP</td>
<td>Sizing Task with IFPUG FPA</td>
<td></td>
</tr>
<tr>
<td>- Measurement Task</td>
<td>- Modeling Task</td>
<td></td>
</tr>
<tr>
<td>- Measurement Task</td>
<td>- Measurement Task</td>
<td></td>
</tr>
<tr>
<td>8 Post-task survey for OOmFP</td>
<td>Post-task survey for IFPUG FPA</td>
<td></td>
</tr>
</tbody>
</table>

5.1.6 Hypothesis formulation

In order to evaluate whether the proposed method (OOmFP) is more efficient and/or effective than IFPUG FPA and (more) likely to be adopted in practice, we planned to test a number of hypotheses. Included are also hypotheses related to the validation of the MAM as an evaluation instrument for FSM methods. Figure 4 shows the theoretical model being tested in this experiment. It summarizes the hypotheses that will be tested and the relationships between the independent variable and the dependent variables and the causal relationships between the constructs of the MAM.

Research Question 1: Comparison Between Methods (Efficiency)

Hypothesis 1: This hypothesis tests the relationship between the independent variable (FSM method) and Sizing Effort (D1).

- Null hypothesis, $H_{1n}$: OOmFP will take more or as much time than IFPUG FPA.
- Alternative hypothesis, $H_{1a}$: OOmFP will take less time than IFPUG FPA.

To allow for a fair comparison of the efficiency of both methods, we have to take separate timings of different activities for both methods. Separating these times is necessary because the FSM methods compared in the experiment require different activities in the identification and measurement steps that lead from the user requirements to the functional size value.

For each subject we have collected the time (in work-hours) spent to complete the (conceptual) modeling and measurement tasks associated with each FSM method. The “modeling time” for the application of IFPUG FPA corresponds to the time the subjects spent to develop an Entity-Relationship diagram and screen prototypes. This modeling is required in the sizing task for
IFPUG FPA. In fact, the subjects made their ER models using the implicit knowledge gathered when they built in the previous phase their OO-Method conceptual models. For OOmFP, the “modeling time” is the time the subjects spent on building a complete OO-Method conceptual model specification, which includes all aspects that contribute to functional size. The requirements modeling considered here is part of the conceptual modeling task of the experiment. The “measurement time” in both methods is the time that the subjects spent to apply the method measurement rules (as part of the sizing tasks).

Given the research questions and context of the experiment, we do not include the OO-Method conceptual modeling time in the sizing effort required for the application of the FSM methods. As OOmFP is specifically intended to be used in conjunction with OO-Method, constructing an OO-Method conceptual model is done anyway, regardless whether this FSM method or another is used. We assume, however, that if IFPUG FPA is used, an ER data model and a set of screen prototypes are needed, before the elements that contribute to functional size can be identified.

---

Figure 4. Theoretical Model

---

1 As remarked before, this is an implicit requirement of the IFPUG rules for FPA. Due to a lack of examples in the IFPUG manual, counting function points directly from the original requirements document is obviously more difficult and hence there is a danger that requiring subjects to do so would bias the experimental results in favour of OOmFP (especially with respect to reproducibility and accuracy).
Research Question 1: Comparison Between Methods (Effectiveness)

**Hypothesis 2:** This hypothesis tests the relationship between the independent variable and Reproducibility (D2).
- Null hypothesis, \( H_{2n} \): OOmFP will produce less or equally consistent assessments than IFPUG FPA.
- Alternative hypothesis, \( H_{2a} \): OOmFP will produce more consistent assessments than IFPUG FPA.

The effectiveness of a FSM method depends amongst others on the inter-rater reliability of the measurements [31]. We postulate that the closer the measurements obtained by different raters, the more effective the FSM method is.

**Hypothesis 3:** This hypothesis tests the relationship between the independent variable and Accuracy (D3).
- Null hypothesis, \( H_{3n} \): OOmFP will produce less or equally accurate assessments than IFPUG FPA.
- Alternative hypothesis, \( H_{3a} \): OOmFP will produce more accurate assessments than IFPUG FPA.

Even when the obtained measurements are (nearly) identical, they can be far away from the true value for functional size. So another dimension of effectiveness is related to the accuracy of the method in producing the ‘right’ value. This third hypothesis requires an objective standard against which to evaluate accuracy. Thus, we can only compare OOmFP and IFPUG FPA if there is some third (and supposedly ‘right’) way of assessing functional size. In order to do this, the system is sized with IFPUG FPA by a Certified Function Point Specialist (CFPS), and we take this count as our standard.

The CFPS is a formal recognition of a level of expertise in the area of Function Point Analysis. A CFPS is acknowledged as having the skills necessary to perform consistent and accurate function point counts and comprehension of the most recent counting practices.

In order to apply FPA, the CFPS was given the original requirements documentation. The CFPS was not familiar with OO-Method, and hence no OO-Method conceptual model was provided. Once the function point count was obtained from the CFPS, we compared how close the measurements are that are produced by the subjects that use IFPUG FPA and by the subjects that use OOmFP.

Research Question 2: Comparison Between Methods (Perception of Efficacy and Intention to Use)

**Hypothesis 4:** This hypothesis tests the relationship between the independent variable (FSM method) and Perceived Ease of Use (D4). This hypothesis follows from H1 (i.e. differences in actual efficiency between the FSM methods) and the causal relationship between actual efficiency and Perceived Ease of Use (i.e. perception of efficiency) as shown in Figure 4 (see Hypothesis 10 below).
- Null hypothesis, \( H_{4n} \): Participants will perceive OOmFP to be more or equally difficult to use than IFPUG FPA.

---

2 Although there is of course no certainty that the CFPS produced the ‘true’ value of functional size, the certification is our best guarantee for the trust that we have in the functional size measurement that was obtained.
• Alternative hypothesis, $H_{4a}$: Participants will perceive OOmFP to be easier to use than IFPUG FPA.

As the post-task survey for a FSM method is applied directly after the sizing task associated with that method, it is hypothesized that the perception of efficiency will be influenced by the experiences with applying the FSM method (according to the Method Evaluation Model). Therefore the relationship between the independent variable and Perceived Ease of Use that is tested is an indirect one. This observation also holds for the next hypotheses.

**Hypothesis 5:** This hypothesis tests the relationship between the independent variable and Perceived Usefulness ($D_5$). This follows from $H2$ and $H3$, and the causal relationship between actual effectiveness and Perceived Usefulness defined in the MEM (see $H11$ and $H12$ below). In addition, this follows from the relationship between Perceived Ease of Use and Perceived Usefulness in the MEM (see $H13$ below).

- Null hypothesis, $H_{5n}$: Participants will perceive OOmFP to be less or equally useful than IFPUG FPA.
- Alternative hypothesis, $H_{5a}$: Participants will perceive OOmFP to be more useful than IFPUG FPA.

**Hypothesis 6:** This hypothesis tests the relationship between the independent variable and Intention to Use ($D_6$). This follows from $H4$, $H5$ and the relationships between Perceived Ease of Use, Perceived Usefulness and Intention to Use defined in the MEM (see $H14$ and $H15$ below).

- Null hypothesis, $H_{6n}$: Participants will be less or equally likely to use OOmFP than IFPUG FPA.
- Alternative hypothesis, $H_{6a}$: Participants will be more likely to use OOmFP than IFPUG FPA.

**Research Question 3: Acceptance of OOmFP**

The following hypotheses test whether the proposed method is likely to be adopted in practice. So the question of likely adoption in practice is here more an absolute one than the relative question of which method (OOmFP or IFPUG FPA) is more likely to be adopted in practice in research question 2. Even if the results of the experiment demonstrate that OOmFP is more likely to be adopted in practice than IFPUG FPA, it is not sure whether the subjects really intend to adopt it. Therefore the following hypotheses are introduced.

**Hypothesis 7:** This hypothesis tests how efficient subjects thought the OOmFP was in achieving its objectives.

- Null hypothesis, $H_{7n}$: Participants will perceive OOmFP to be difficult to use.
- Alternative hypothesis, $H_{7a}$: Participants will perceive OOmFP to be easy to use.

**Hypothesis 8:** This hypothesis tests how effective subjects thought the OOmFP was in achieving its objectives.

- Null hypothesis, $H_{8n}$: Participants will perceive OOmFP to be not useful.
- Alternative hypothesis, $H_{8a}$: Participants will perceive OOmFP to be useful.

**Hypothesis 9:** This hypothesis tests if the subjects intend to use OOmFP in the future.

- Null hypothesis, $H_{9n}$: Participants do not intend to use OOmFP.
- Alternative hypothesis, $H_{9a}$: Participants intend to use OOmFP.
Research Question 4: Validation of the Method Adoption Model

Finally, we test a number of hypothesized relationships between the dependent variables in our experiment, based on the *causal relationships* defined in the Method Evaluation Model. This represents a test of the predictive and explanatory power of the model. However, we remark that only the MAM and its inputs from actual efficacy are tested. Actual Usage of a method will be determined by Intention to Use. This follows from empirical studies of the Technology Acceptance Model (TAM) [17], which have reported a strong significant causal link between behavioural intention to use and actual behaviour [37].

**Hypothesis 10:** Measurement Time is a performance-based measure of efficiency, whereas Perceived Ease of Use is a perception-based measure of efficiency. According to the Method Evaluation Model, perceptions of efficiency are determined by actual efficiency.
- Null hypothesis, $H_{10n}$: Perceived Ease of Use is not determined by Measurement Time.
- Alternative hypothesis, $H_{10a}$: Perceived Ease of Use is determined by Measurement Time.

**Hypothesis 11:** Reproducibility is a performance-based measure of effectiveness, whereas Perceived Usefulness is a perception-based measure of effectiveness. According to the Method Evaluation Model, perceptions of effectiveness are determined by actual effectiveness.
- Null hypothesis, $H_{11n}$: Perceived Usefulness is not determined by Reproducibility.
- Alternative hypothesis, $H_{11a}$: Perceived Usefulness is determined by Reproducibility.

**Hypothesis 12:** Accuracy is a performance-based measure of effectiveness, whereas Perceived Usefulness is a perception-based measure of effectiveness. According to the Method Evaluation Model, perceptions of effectiveness are determined by actual effectiveness.
- Null hypothesis, $H_{12n}$: Perceived Usefulness is not determined by Accuracy.
- Alternative hypothesis, $H_{12a}$: Perceived Usefulness is determined by Accuracy.

**Hypothesis 13:** This is one of the causal relationships defined in the Method Evaluation Model and based on the Technology Acceptance Model [17].
- Null hypothesis, $H_{13n}$: Perceived Usefulness is not determined by Perceived Ease of Use.
- Alternative hypothesis, $H_{13a}$: Perceived Usefulness is determined by Perceived Ease of Use.

**Hypothesis 14:** This is one of the causal relationships defined in the Method Evaluation Model and based on the Technology Acceptance Model [17].
- Null hypothesis, $H_{14n}$: Intention to Use is not determined by Perceived Ease of Use.
- Alternative hypothesis, $H_{14a}$: Intention to Use is determined by Perceived Ease of Use.

**Hypothesis 15:** This is one of the causal relationships defined in the Method Evaluation Model and based on the Technology Acceptance Model [17].
- Null hypothesis, $H_{15n}$: Intention to Use is not determined by Perceived Usefulness.
- Alternative hypothesis, $H_{15a}$: Intention to Use is determined by Perceived Usefulness.

5.2 Experiment Operation

5.2.1 Preparation

At the time the experiment was performed, the subjects had taken a course in conceptual modeling using OO-Method. Also, we ran separate training sessions on IFPUG FPA and OOmFP for each
treatment group prior to getting them to do the sizing tasks. However, the subjects were not aware of what aspects we intended to study, nor were they informed of the stated hypothesis.

The course in conceptual modeling was run in three sessions of two hours each one. In the first session we explained the primitives of the OO-Method conceptual models using some practical examples. In the second session we applied the concepts presented in the previous session in the resolution of a complete case study involving a Library system. Then, in the third session we demonstrated how the conceptual model obtained in the previous session can be build and documented using the Oliva Nova™ Modeling Software.

The training sessions on IFPUG FPA and OOmFP were run before subjects applied a given method (see Table 2). For each method two sessions of two hours were needed. In the first session we explain the measurement rules of the method and demonstrate their application using some toy examples. In the second session we demonstrate the application of the measurement rules on a complete case study (i.e. the Library system).

5.2.2 Execution

The subjects received all the materials as described before. We also explained how to do the experimental tasks. The experimental tasks were run on-line as part of a course [61]. For the conceptual modeling task with OO-Method we gave them a maximum of six hours (two sessions of three hours each one).

For the sizing task we gave them a maximum of about five hours (two sessions of two hours and thirteen minutes). The sequence of the experimental tasks can be seen in Table 2. Each subject had to do each test alone. It was allowed the use of the training materials when performing the tasks but we control that no interaction whatsoever between subjects occurred.

5.2.3 Data Recording and Validation

The performance-based dependent variables were measured using two different data collection forms.

The first form was used to record the outputs of the IFPUG FPA functional size measurement and the time spent to size the requirement specifications. For the IFPUG FPA method, we called “modeling time” the time that the subjects spent to develop the data oriented abstraction assumed by FPA. We called “measurement time” the time that the subjects spent to measure it.

The second form was used to record the outputs of the OOmFP functional size measurement and the time spent to size the requirements specifications. For the OOmFP method, we called “modeling time” the time that the subjects spent to develop the OO-Method models, and “measurement time” the time that the subjects spent to apply the OOmFP rules.

The following data was collected as part of the experiment:

- Subjects were timed as to how long it took them to complete the conceptual model task (for OOmFP) and sizing task (for IFPUG FPA and OOmFP). In the sizing task for IFPUG FPA we have collected separate times for modeling and measurement tasks.
- Results of the sizing task were analyzed for completeness.
- Responses to the post-task survey were coded in numerical form.

Once the data were collected, we verified whether the tests were complete. As two students used only OOmFP, we took into account the responses of twenty subjects.

6. ANALYSIS AND INTERPRETATION

In general, the results are described in terms of their statistical significance that is measured by alpha (α), which represents the probability that the result could have occurred by chance due to a Type I error. We define the following levels of significance:
• Not significant: $\alpha > 0.1$
• Low significance: $\alpha < 0.1$
• Medium Significance: $\alpha < 0.05$
• High significance: $\alpha < 0.01$
• Very high: $\alpha < 0.001$

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

6.1 Comparative Analysis of the Efficacy of the FSM methods

We first test hypothesis $H_{1a}$ related to the efficiency of the FSM methods. Given the differences in the steps of these FSM methods previously mentioned, we decide to use only the time related to the measurement time (presented in Appendix B). Descriptive statistics for the IFPUG FPA and OOmFP methods measurement time are presented in Table 3.

<table>
<thead>
<tr>
<th>Measurement Time</th>
<th>IFPUG FPA</th>
<th>OOmFP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of observations</td>
<td>20</td>
<td>20</td>
</tr>
<tr>
<td>Mean</td>
<td>2.4105</td>
<td>3.0070</td>
</tr>
<tr>
<td>Standard deviation</td>
<td>0.3334</td>
<td>0.6689</td>
</tr>
<tr>
<td>Minimum</td>
<td>1.85</td>
<td>2.00</td>
</tr>
<tr>
<td>Maximum</td>
<td>3.33</td>
<td>4.23</td>
</tr>
<tr>
<td>Percentiles (25th, 50th and 75th)</td>
<td>2.1750; 2.3850; 2.5300</td>
<td>2.3500; 2.9750; 3.3825</td>
</tr>
</tbody>
</table>

The Kolmogorov-Smirnov test for normality was applied to the differences of the paired observations. As this distribution was normal and measurement time is a continuous variable, we decided to use the Paired t-test to check for a difference in mean measurement time between OOmFP and IFPUG FPA. In order to evaluate the significance of the observed difference, we applied the test with a significance level of 5 %, i.e. $\alpha = 0.05$. The result of the test does not allow rejecting the null hypothesis ($H_{1n}$), meaning that we cannot empirically corroborate that OOmFP will take less time than IFPUG FPA.

In fact, the test shows that for the data collected the mean measurement time for IFPUG FPA is significantly lower than that for OOmFP (Table 4). The reason could be that the object used takes into account all perspectives of an OO system (data, process, and behaviour). Consequently, the subjects spent more time to apply all the OOmFP measurement rules. However, this problem can be solved by providing tool support for measurement rules application.

<table>
<thead>
<tr>
<th>Dependent variable</th>
<th>Mean</th>
<th>Std. Deviation</th>
<th>Std. Error Mean</th>
<th>95% Confidence Interval of the difference</th>
<th>t</th>
<th>1-tailed p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Measurement Time</td>
<td>-0.596</td>
<td>0.7716</td>
<td>0.17254</td>
<td>-0.957 (lower) -0.235 (upper)</td>
<td>-3.457</td>
<td>0.0015</td>
</tr>
</tbody>
</table>

Then, we test hypothesis $H_{2a}$ related to the effectiveness of the FSM methods, in terms of reproducbility. Table 5 shows descriptive statistics for the functional size produced (presented in Appendix B) for the IFPUG FPA and OOmFP methods. We remark that the values of the column “Size in IFPUG-FPA” are unadjusted function point values.
Table 5. Descriptive statistics functional size

<table>
<thead>
<tr>
<th>Reproducibility</th>
<th>IFPUG FPA</th>
<th>OOmFP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of observations</td>
<td>20</td>
<td>20</td>
</tr>
<tr>
<td>Mean</td>
<td>132.40</td>
<td>159.55</td>
</tr>
<tr>
<td>Standard deviation</td>
<td>17.44</td>
<td>7.964</td>
</tr>
<tr>
<td>Minimum</td>
<td>114</td>
<td>148</td>
</tr>
<tr>
<td>Maximum</td>
<td>174</td>
<td>173</td>
</tr>
<tr>
<td>Percentiles (25th, 50th and 75th)</td>
<td>121.00; 126.00; 144.50</td>
<td>151.75; 159.00; 164.50</td>
</tr>
</tbody>
</table>

In order to measure the degree of variation between assessments produced by different subjects using the same method we use a practical statistic similar to that proposed by Kemerer [31]. This statistic is calculated as the difference in absolute value between the count produced by a subject and the average count (for the same FSM method) produced by the other subjects in the sample, relative to this average count.

Reproducibility measurements (REP) were thus obtained for each observation by applying the following equation:

\[
REP_i = \frac{\text{Average Other assessments} - \text{Subject Assessment}}{\text{Average Other Assessments}}
\]

The obtained REP, for both methods are presented in the Appendix B. Then, the differences in reproducibility assessments obtained using both methods were described using the Kolmogrov-Smirnov test to ascertain if the distribution was normal. The skewness of the distribution is close to one (1.230), meaning more observations in the left tail than normal. Because of this indication of a non-normal distribution, we used the Wilcoxon signed rank test for the difference in median reproducibility assessments, which is a non-parametric alternative to the paired samples t-test. The result of the one-tailed test (see table 6) allows rejecting the null hypothesis (H2n), meaning that we empirically corroborate that OOmFP produce more consistent assessments than IFPUG FPA.

Table 6. 1-tailed Wilcoxon’s signed rank test for differences in median reproducibility assessments (IFPUG FPA versus OOmFP; \(\alpha = 0.05\))

<table>
<thead>
<tr>
<th>Dependent variable</th>
<th>Mean Rank</th>
<th>Sum of Ranks</th>
<th>z</th>
<th>1-tailed p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reproducibility</td>
<td>5.25</td>
<td>10.50</td>
<td>-3.409</td>
<td>0.000</td>
</tr>
</tbody>
</table>

Next, we test hypothesis H3a related to the effectiveness of the FSM methods, in terms of accuracy. The value obtained by the CFPS is 153 unadjusted function points. In order to compare the assessments produced by both methods and a Certified Function Point Specialist (CFPS) the Magnitude of Relative Error (MRE) was used [15], [10], [11], [51]. Accuracy measurements were obtained by applying the following equation for each observation:

\[
MRE_i = \frac{\text{CFPS assessment} - \text{Subject assessment}}{\text{CFPS assessment}}
\]

The obtained MRE, for both methods are presented in the Appendix B. Then, the differences in accuracy assessments obtained were described using the Kolmogrov-Smirnov test to ascertain if the distribution was normal. As the distribution was non-normal, we used the Wilcoxon signed rank test for the difference in median accuracy assessments. In order to evaluate the significance of the observed difference, we applied a statistical test with a significance level of 5 %. The result
of test (see Table 7) allows rejecting the null hypothesis ($H_{3n}$), meaning that we empirically corroborate that OOmFP produce more accurate assessments than IFPUG FPA.

### Table 7. 1-tailed Wilcoxon’s signed rank test for differences in median accuracy assessments (IFPUG FPA versus OOmFP; $\alpha = 0.05$)

<table>
<thead>
<tr>
<th>Dependent variable</th>
<th>Mean Rank</th>
<th>Sum of Ranks</th>
<th>z</th>
<th>1-tailed p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>4.75</td>
<td>9.50</td>
<td>-3.442</td>
<td>0.000</td>
</tr>
</tbody>
</table>

**6.2 Comparative Analysis of the Likelihood of Adoption in Practice of the FSM methods**

After that, we test hypotheses $H_{4a}, H_{5a},$ and $H_{6a}$ related to the perceptions of the FSM methods, in terms of perceived ease of use, perceived usefulness and intention to use. Descriptive statistics for the IFPUG FPA and OOmFP methods perceived ease of use are presented in Table 8.

### Table 8. Descriptive statistics perceived ease of use

<table>
<thead>
<tr>
<th>Perceived Ease of Use</th>
<th>IFPUG FPA</th>
<th>OOmFP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of observations</td>
<td>20</td>
<td>20</td>
</tr>
<tr>
<td>Mean</td>
<td>2.7100</td>
<td>2.90</td>
</tr>
<tr>
<td>Standard deviation</td>
<td>0.4517</td>
<td>0.7033</td>
</tr>
<tr>
<td>Minimum</td>
<td>1.80</td>
<td>1.60</td>
</tr>
<tr>
<td>Maximum</td>
<td>3.60</td>
<td>4.40</td>
</tr>
<tr>
<td>Percentiles ($25^{th}, 50^{th}$ and $75^{th}$)</td>
<td>2.40; 2.60; 3.1500</td>
<td>2.60; 2.60; 3.40</td>
</tr>
</tbody>
</table>

The mean values obtained show that OOmFP has a higher mean score than IFPUG FPA, meaning that is perceived to be easier to use than IFPUG FPA. In order to evaluate the significance of the observed difference, we applied a statistical test with a significance level of 5%, i.e. $\alpha = 0.05$. As the data were normal we decided to use Paired t-test to evaluate the statistical significance of the observed difference in mean perceived ease of use.

The result of the one-tailed test (see table 9) allows rejecting the null hypothesis ($H_{4n}$), meaning that we empirically corroborate that the participants perceived OOmFP as easier to use than IFPUG FPA.

### Table 9. 1-tailed Paired Samples t-test rank for differences in mean perceived ease of use (IFPUG FPA versus OOmFP; $\alpha = 0.05$)

<table>
<thead>
<tr>
<th>Dependent variable</th>
<th>Mean</th>
<th>Std. Deviation</th>
<th>Std. Error Mean</th>
<th>95% Confidence Interval of the difference</th>
<th>t</th>
<th>1-tailed p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Perceived Ease of Use</td>
<td>0.190</td>
<td>0.8620</td>
<td>0.1927</td>
<td>-0.213 (lower) 0.593 (upper)</td>
<td>0.986</td>
<td>0.01685</td>
</tr>
</tbody>
</table>

Table 10 show descriptive statistics for IFPUG FPA and OOmFP methods perceived usefulness. In the same way the Kolmogorov-Smirnov test for normality was applied to the differences of the paired observations. As this distribution was normal, the Paired t-test was selected to check for a difference in mean perceived usefulness between OOmFP and IFPUG FPA.
In order to evaluate the significance of the observed difference, we applied a statistical test with a significance level of 5 %, i.e. \( \alpha = 0.05 \). The result of the one-tailed test (see table 11) allows rejecting the null hypothesis (H_0), meaning that we empirically corroborate that participants perceive OOmFP to be more useful than IFPUG FPA.

Table 12 show descriptive statistics for the IFPUG FPA and OOmFP methods intention to use. Kolmogorov-Smirnov test for normality was applied to the differences of the paired observations. As this distribution was normal, the Paired t-test was selected to check for a difference in mean intention to use between both FSM methods. In order to evaluate the significance of the observed difference, we applied a statistical test with a significance level of 5 %, i.e. \( \alpha = 0.05 \).

The result of the one-tailed test (see table 13) allows rejecting the null hypothesis (H_0), meaning that we empirically corroborate that participants will be more likely to use OOmFP than IFPUG FPA.

### Table 10. Descriptive statistics perceived usefulness

<table>
<thead>
<tr>
<th>Perceived Usefulness</th>
<th>IFPUG FPA</th>
<th>OOmFP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of observations</td>
<td>20</td>
<td>20</td>
</tr>
<tr>
<td>Mean</td>
<td>2,4800</td>
<td>3,5250</td>
</tr>
<tr>
<td>Standard deviation</td>
<td>0,4606</td>
<td>0,4920</td>
</tr>
<tr>
<td>Minimum</td>
<td>1,60</td>
<td>2,80</td>
</tr>
<tr>
<td>Maximum</td>
<td>3,60</td>
<td>4,60</td>
</tr>
<tr>
<td>Percentiles (25th, 50th and 75th)</td>
<td>2,2000; 2,5000; 2,8000</td>
<td>3,2000; 3,4000; 3,8000</td>
</tr>
</tbody>
</table>

### Table 11. 1-tailed Paired Samples t-test rank for differences in mean perceived usefulness (IFPUG FPA versus OOmFP; \( \alpha = 0.05 \))

<table>
<thead>
<tr>
<th>Dependent variable</th>
<th>Mean</th>
<th>Std. Deviation</th>
<th>Std. Error Mean</th>
<th>95% Confidence Interval of the difference</th>
<th>t</th>
<th>1-tailed p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Perceived Usefulness</td>
<td>1,020</td>
<td>0,6582</td>
<td>0,1471</td>
<td>0,7119 (lower) 1,3281 (upper)</td>
<td>6,930</td>
<td>0,000</td>
</tr>
</tbody>
</table>

### Table 12. Descriptive statistics intention to use

<table>
<thead>
<tr>
<th>Intention to Use</th>
<th>IFPUG FPA</th>
<th>OOmFP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of observations</td>
<td>20</td>
<td>20</td>
</tr>
<tr>
<td>Mean</td>
<td>2,6333</td>
<td>3,6667</td>
</tr>
<tr>
<td>Standard deviation</td>
<td>0,6657</td>
<td>0,6839</td>
</tr>
<tr>
<td>Minimum</td>
<td>1,33</td>
<td>2,33</td>
</tr>
<tr>
<td>Maximum</td>
<td>3,67</td>
<td>4,67</td>
</tr>
<tr>
<td>Percentiles (25th, 50th and 75th)</td>
<td>2,3333; 2,6667; 3,0000</td>
<td>3,3333; 3,6667; 4,3333</td>
</tr>
</tbody>
</table>

The result of the one-tailed test (see table 13) allows rejecting the null hypothesis (H_0), meaning that we empirically corroborate that participants will be more likely to use OOmFP than IFPUG FPA.

### Table 13. 1-tailed Paired Samples t-test rank for differences in mean perceived usefulness (IFPUG FPA versus OOmFP; \( \alpha = 0.05 \))

<table>
<thead>
<tr>
<th>Dependent variable</th>
<th>Mean</th>
<th>Std. Deviation</th>
<th>Std. Error Mean</th>
<th>95% Confidence Interval of the difference</th>
<th>t</th>
<th>1-tailed p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intention to Use</td>
<td>1,0333</td>
<td>0,8976</td>
<td>0,2007</td>
<td>0,6132 (lower) 1,4535 (upper)</td>
<td>5,148</td>
<td>0,000</td>
</tr>
</tbody>
</table>
Finally, Figure 5 shows the mean responses to each question on the post-task survey for each FSM method. The obtained results show that the OOmFP method scored highest on all items except Q1 and Q2.

![Figure 5. Mean Responses to Post-task Survey](image)

Overall, the responses to the post-task surveys suggest that OOmFP is more useful and is more likely to be used in the future. However, scores on individual items are not meant to be interpreted in isolation as different items measure the same construct. Thus, we need to evaluate the construct validity and reliability of the items used to measure perception-based variables. This validates the measurement instrument associated with the Method Adoption Model.

**Construct Validity**

*Factor Analysis* is the most used technique for evaluating the construct validity for this type of measurement instrument. However, a sample size of 200 is recommended to achieve stability of the results [45]. As the sample size of this experiment is too small, *inter-item correlation analysis* was carried out. We assume that all items associated with a particular construct have equal weights.

Campbell and Fiske [12] proposed the concepts of convergent and discriminant validity in the context of construct validity:

- **Convergent validity (CV):** refers to the convergence among different indicators used to measure a particular construct. The correlation between indicators used to measure the *same* construct should be as high as possible. CV of an indicator is measured by the average correlation between the indicator and the other indicators that are used to measure the same construct.

- **Discriminant validity (DV):** refers to the divergence of indicators used to measure different constructs. The correlation between indicators used to measure different constructs should be as low as possible. DV of an indicator is measured by the average correlation between the indicator and the indicators that are used to measure a different construct.

Table 14 shows the inter-item correlations for all fourteen items that were used in the survey. The last three columns of the matrix show the results of the construct validity analysis. An item is valid if convergent validity is higher than discriminant validity (as suggested by Campbell and Fiske [12]).

The results of the validity analysis for each construct are:

- **Perceived Ease of Use (PEOU):** The obtained results show that for all items convergent validity was greater than discriminant validity. The average convergent validity of all
items was 0.57. Overall, convergent validity was more than twice the size of discriminant validity.

- **Perceived Usefulness (PU):** The obtained results show that the convergent validity was greater than discriminant validity for five of the six items. Q11 item have a same value for convergent and discriminant validity. The average convergent validity across all indicators was 0.40. Overall, convergent validity was greater than the size of discriminant validity.

- **Intention to Use (ITU):** The obtained results show that the convergent validity was high. The average convergent validity across all indicators was 0.62. Overall, convergent validity was more than twice the size of discriminant validity.

### Table 14. Correlation Between Survey Items (Construct Validity Analysis)

<table>
<thead>
<tr>
<th></th>
<th>PEOU</th>
<th>PU</th>
<th>ITU</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q1</td>
<td>1.00</td>
<td>0.61</td>
<td>0.42</td>
</tr>
<tr>
<td>Q3</td>
<td>0.61</td>
<td>1.00</td>
<td>0.46</td>
</tr>
<tr>
<td>Q4</td>
<td>0.42</td>
<td>0.46</td>
<td>1.00</td>
</tr>
<tr>
<td>Q6</td>
<td>0.37</td>
<td>0.40</td>
<td>0.48</td>
</tr>
<tr>
<td>Q9</td>
<td>0.47</td>
<td>0.51</td>
<td>0.56</td>
</tr>
<tr>
<td>Q10</td>
<td>0.51</td>
<td>0.47</td>
<td>0.56</td>
</tr>
<tr>
<td>Q11</td>
<td>0.54</td>
<td>0.51</td>
<td>0.56</td>
</tr>
<tr>
<td>Q13</td>
<td>0.31</td>
<td>0.15</td>
<td>0.24</td>
</tr>
<tr>
<td>Q12</td>
<td>0.39</td>
<td>0.06</td>
<td>0.07</td>
</tr>
<tr>
<td>Q14</td>
<td>0.40</td>
<td>0.06</td>
<td>0.07</td>
</tr>
</tbody>
</table>

The problem that emerged from the analysis was the low level of convergent validity for the item Q11. This item have a value for CV that is equal than its value for DV. The average convergent validity for Perceived Usefulness is 0.40. For this reason, Q11 was excluded from the analysis. With the exclusion of this item the results of the validity analysis are improved as shown in Table 15. This can be noted since the average convergent validity for Perceived Usefulness is now 0.47.

### Table 15. Correlation Between Survey Items (after Q11 removed)

<table>
<thead>
<tr>
<th></th>
<th>PEOU</th>
<th>PU</th>
<th>ITU</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q1</td>
<td>1.00</td>
<td>0.57</td>
<td>0.37</td>
</tr>
<tr>
<td>Q3</td>
<td>0.57</td>
<td>1.00</td>
<td>0.40</td>
</tr>
<tr>
<td>Q4</td>
<td>0.37</td>
<td>0.40</td>
<td>1.00</td>
</tr>
<tr>
<td>Q6</td>
<td>0.44</td>
<td>0.48</td>
<td>0.57</td>
</tr>
<tr>
<td>Q9</td>
<td>0.44</td>
<td>0.48</td>
<td>0.57</td>
</tr>
<tr>
<td>Q10</td>
<td>0.51</td>
<td>0.47</td>
<td>0.56</td>
</tr>
<tr>
<td>Q11</td>
<td>0.39</td>
<td>0.06</td>
<td>0.07</td>
</tr>
<tr>
<td>Q12</td>
<td>0.39</td>
<td>0.06</td>
<td>0.07</td>
</tr>
<tr>
<td>Q14</td>
<td>0.40</td>
<td>0.06</td>
<td>0.07</td>
</tr>
</tbody>
</table>

### Reliability

We also conducted a reliability analysis on the items used to measure the PEOU, PU and ITU variables. The reliability of a test is the extent to which can be relied upon to produce ‘true’ scores as a measure of its consistency/variability across occasions or with different sets of equivalent
items. For this analysis item Q11 was excluded. Table 16 shows the obtained results for each construct using Cronbach’s alpha, which is the most common measure of scale reliability. These values are all 0.7 or above as required for constructs to be deemed reliable as suggested by Nunally [42]. Also, Caplan et al. [13] argument that alphas as low as 0.5 are considered acceptable in some cases.

Table 16. Item reliabilities for Constructs

<table>
<thead>
<tr>
<th>Construct</th>
<th>CRONBACH’S α</th>
</tr>
</thead>
<tbody>
<tr>
<td>Perceived Ease of Use</td>
<td>0.82</td>
</tr>
<tr>
<td>Perceived Usefulness</td>
<td>0.70</td>
</tr>
<tr>
<td>Intention to use</td>
<td>0.70</td>
</tr>
</tbody>
</table>

In addition, the general Cronbach’s alpha obtained for the instrument was 0.74. As a result of this analysis, we conclude that the items on the survey (except Q11) are reliable and valid measures of the underlying constructs of the proposed theoretical model. Also, as a result from this analysis we may observe that if item 12 is deleted the reliability for Intention to Use will be 0.91, indicating that this item is not related with this constructor.

Therefore, these results also allow partially validating the Method Adoption Model (Research Question 4) in terms of the validity and reliability of the measurement instrument (the measurement part of the model).

6.3 Analysis of the Acceptance of OO-Method Function Points

The objective of this section is to evaluate the likely acceptance in practice of OOmFP (Research Question 3). This requires comparing the mean values of the constructs of the Method Adoption Model for OOmFP (shown in tables 8, 10 and 12) to verify if these values were significantly greater than 3.

We first test hypothesis H7a related to the perceived ease of use of the OOmFP method. The Kolmogorov-Smirnov test for normality was applied to the OOmFP perceived ease of use observations. As this distribution was normal, we decided to use the One sample t-test to check for a difference in mean perceived ease of use OOmFP and the value 3. In order to evaluate the significance of the observed difference, we applied a statistical test with a significance level of 5 %, i.e. $\alpha = 0.05$. The result for perceived ease of use (Table 17) does not allow rejecting the null hypothesis (H7n), meaning that we cannot empirically corroborate that participants will perceive OOmFP to be easy to use.

Table 17. 1-tailed One Sample t-test rank for differences in mean perceived ease of use

<table>
<thead>
<tr>
<th>Dependent variable</th>
<th>Mean</th>
<th>Std. Deviation</th>
<th>Std. Error Mean</th>
<th>95% Confidence Interval of the difference</th>
<th>t</th>
<th>1-tailed p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Perceived Ease of Use</td>
<td>2.9000</td>
<td>0.70338</td>
<td>0.1572</td>
<td>-0.4292 (lower) 0.2292 (upper)</td>
<td>-0.636</td>
<td>0.266</td>
</tr>
</tbody>
</table>

Then, we test hypothesis H8a related to the perceived usefulness. The Kolmogorov-Smirnov test for normality was applied to the OOmFP perceived usefulness observations. As this distribution was normal, we use the One sample t-test to check for a difference in mean perceived usefulness OOmFP and the value 3. In order to evaluate the significance of the observed difference, we applied a statistical test with a significance level of 5 %, i.e. $\alpha = 0.05$. The result for perceived
usefulness (Table 18) allows rejecting the null hypothesis \(H_8\), meaning that we can empirically corroborate that participants will perceive OOmFP to be useful.

**Table 18.** 1-tailed One Sample t-test rank for differences in mean perceived usefulness

<table>
<thead>
<tr>
<th>Dependent variable</th>
<th>Mean</th>
<th>Std. Deviation</th>
<th>Std. Error Mean</th>
<th>95% Confidence Interval of the difference</th>
<th>t</th>
<th>1-tailed p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Perceived Usefulness</td>
<td>3.5000</td>
<td>0.4920</td>
<td>0.1100</td>
<td>0.2697 (lower) 0.7303 (upper)</td>
<td>4.544</td>
<td>0.000</td>
</tr>
</tbody>
</table>

Finally, we test hypothesis \(H_9\) related to the intention to use. The Kolmogorov-Smirnov test for normality was applied to the OOmFP intention to use observations. As this distribution was normal, we use the One sample t-test to check for a difference in mean intention to use OOmFP and the value 3.

**Table 19.** 1-tailed One Sample t-test rank for differences in mean intention to use

<table>
<thead>
<tr>
<th>Dependent variable</th>
<th>Mean</th>
<th>Std. Deviation</th>
<th>Std. Error Mean</th>
<th>95% Confidence Interval of the difference</th>
<th>t</th>
<th>1-tailed p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intention to Use</td>
<td>3.6667</td>
<td>0.6839</td>
<td>0.1529</td>
<td>0.3466 (lower) 0.9868 (upper)</td>
<td>4.359</td>
<td>0.000</td>
</tr>
</tbody>
</table>

In order to evaluate the significance of the observed difference, we applied a statistical test with a significance level of 5\%, i.e. \(\alpha = 0.05\). The result for intention to use shown in Table 19 allows rejecting the null hypothesis \(H_{9n}\), meaning that we can empirically corroborate that participants will intend to use OOmFP.

### 6.4 Validation of Method Adoption Model

According to the Method Evaluation Model (MEM) there exist a number of hypothesized causal relationships between the dependent variables in our study. Collectively, the relationships between the constructs of the Method Adoption Model (MAM) and between these constructs and the external variables of the MEM explain the external behaviour towards a method (see Figure 6). If these relationships are validated, then the MAM can be used to predict the actual usage of a method.

The intention of this section is to validate the structural part of the MAM in terms of the causal relationships between its constructs and with its external variables, with the exception of Actual Usage. According to Davis [17] and Moody [40] it is not possible to validate the hypothesized relationship between Intention to Use and Actual Usage in an experimental context such as the one used here\(^3\).

We have chosen regression analysis to evaluate the MAN since the hypotheses to test are causal relationships between continuous variables. This evaluation was conducted in three phases using four separate regression analyses taking into account each dependent variable:

- Phase 1 (simple regression): the effect of Actual Efficiency (independent variable) on Perceived Ease of Use (dependent variable); this is a test of hypothesis H7.

---

\(^3\) Note that the third research question we wish to answer relates to the validation of the Method Adoption Model, not the Method Evaluation Model which would also require validating the causal relationship between Intention to Use and Actual Usage.
Phase 2 (multiple regression): the effect of Actual Effectiveness (Reproducibility and Accuracy independent variables) on Perceived Usefulness (dependent variable); this is a test of hypothesis H8 and H9.

Phase 3 (simple regression): the effect of Perceived Ease of Use (independent variable) on Perceived Usefulness (dependent variable); this is a test of H10.

Phase 4 (multiple regression): the effect of Perceived Ease of Use and Perceived Usefulness (independent variables) on Intention to Use (dependent variable); this is a test of H11 and H12.

Figure 6. Relationships between Dependent Constructs (adapted from Moody, 2001)

Phase 1: A Regression Model for Actual Efficiency vs Perceived Ease of Use

In this phase, we test hypothesis H10 to verify if perceptions of efficiency are determined by actual efficiency. In order to do this, Actual Efficiency (in terms of measurement time) was used as the independent (predictor) variable and Perceived Ease of Use as the dependent (predicted) variable. The regression equation resulting from the analysis is:

\[ \text{Perceived Ease of Use} = 5.40 - 0.83 \times \text{Measurement Time} \]

The details of the regression analysis are:

- \( R^2 = 0.63 \)
- F-statistic = 30.33
- Significance level (p) = 0.000

The result of the regression (see table 20) allows rejecting the null hypothesis (H10), meaning that we empirically corroborate that perceived of use is determined by measurement time.

<table>
<thead>
<tr>
<th>Model</th>
<th>Unstd. Coefficients (b)</th>
<th>Std. Error</th>
<th>Std. Coefficients (Beta)</th>
<th>t</th>
<th>Sig.</th>
<th>95% Confidence Interval for B</th>
<th>Correlations (r)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>5,405</td>
<td>0,465</td>
<td></td>
<td>11.614</td>
<td>0,000</td>
<td>4,427 (lower) 6,382 (upper)</td>
<td></td>
</tr>
<tr>
<td>Measurement time</td>
<td>-0,833</td>
<td>0,151</td>
<td>-0,792</td>
<td>-5,507</td>
<td>0,000</td>
<td>-1,151 (lower) -0,515 (upper)</td>
<td>-0,792</td>
</tr>
</tbody>
</table>
Regarding statistical significance, the regression equation was found to be very high significant with $\alpha < 0.001$. This means that H7 was strongly confirmed. The significance of the regression coefficient for measurement time indicates that perceptions of efficiency are (partially) determined by actual efficiency.

With respect to the predictive power of the model, Actual Efficiency explains 63% of the variance in Perceived Ease of Use. This is indicated by $R$ squared ($r^2$) that measures the proportion of the variance of the dependent variable predicted by the independent variable. The regression coefficient ($b$) defines the effect of the independent variable on the dependent variable. In this case, an increase in Measurement Time in one minute results in a decrease of 0.83 in Perceived Ease of Use. This means that the variables are inversely related, as expected.

The Pearson correlation coefficient ($r$) is the degree of linear relationship between two variables measured from the same individual. This defines the effect of the independent variable on the dependent variable. In this case, the correlation coefficient ($r$) obtained was -0.79, which represents a strong correlation between variables. Thus, all results show that the causal relationship between Actual Efficiency and Perceived Ease of Use is statistically significant. Figure 7 shows the scatterplot for measurement Time against Perceived Ease of Use scores with the regression line.

### Phase 2: A Regression Model for Actual Effectiveness vs Perceived Usefulness

In this phase, we test hypotheses H11a and H12a to verify if Perceived Usefulness (PU) is determined by Reproducibility and Accuracy. To do this, reproducibility and accuracy were used as independent (predictor) variables and PU as the dependent (predicted) variable. The regression equation resulting from the analysis is:

$$\text{Perceived Usefulness} = 3.61 + 3.23 \times \text{Reproducibility} - 4.57 \times \text{Accuracy}$$

The details of the regression analysis are:
- $R$ squared ($r^2$) = 0.13
- F-statistic = 1.3
- Significance level (p) = 0.1575

The result of the regression summarized in Table 21 does not allow rejecting the null hypotheses (H11a and H12a), meaning that we cannot empirically corroborate that perceived usefulness is determined by actual effectiveness.
Table 21. Multiple Regression between PU, Reproducibility and Accuracy  
(OOmFP; $\alpha = 0.05$)

<table>
<thead>
<tr>
<th>Model</th>
<th>Unstd. Coefficients (b)</th>
<th>Std. Error</th>
<th>Std. Coefficients (Beta)</th>
<th>t</th>
<th>Sig.</th>
<th>95% Confidence Interval for B</th>
<th>Correlations (r)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>3,608</td>
<td>0,215</td>
<td>16,8</td>
<td>0,000</td>
<td></td>
<td>3,154 (lower) 4,061 (upper)</td>
<td></td>
</tr>
<tr>
<td>Reproducibility</td>
<td>3,228</td>
<td>3,8</td>
<td>0,202</td>
<td>0,834</td>
<td>0,416</td>
<td>-4,940 (lower) 11,396 (upper)</td>
<td>0,069</td>
</tr>
<tr>
<td>Accuracy</td>
<td>-4,566</td>
<td>2,9</td>
<td>-374</td>
<td>-1,544</td>
<td>0,141</td>
<td>-10,805 (lower) 1,674 (upper)</td>
<td>-0,302</td>
</tr>
</tbody>
</table>

The regression coefficients for Reproducibility and Accuracy were not significant ($\alpha > 0.1$). This means that $H_{11}$ and $H_{12}$ (see Figure 6) were not confirmed. This also suggests that perceptions in usefulness are not determined by perceptions in actual effectiveness. A possible explanation is that the participants do not know the results of their measurement. Consequently, they have not the perception of usefulness of the method they applied. For future experiments a solution could be to present the results prior to getting them to do the surveys.

With respect to the predictive power of the model, Reproducibility and Accuracy explain only 13% of the variance in Perceived Usefulness as indicated by $r^2$. Figure 9 shows the scatterplot of the REP and MRE values against the Perceived Usefulness scores with the regression line.

![Figure 9. Scatterplot: Perceived Usefulness vs. Reproducibility and Accuracy](image)

Phase 3: A Regression Model for Perceived Ease of Use vs Perceived Usefulness

In this phase, we test $H_{13a}$ to verify if perceived usefulness is determined by perceived ease of use. Then, Perceived Ease of Use was used as the independent (predictor) variable and Perceived Usefulness as the dependent (predicted) variable. The regression equation resulting from the analysis is:

\[
\text{Perceived Usefulness} = 2,13 + 0,47 \times \text{Perceived Ease of Use}
\]

The details of the regression analysis are:
• R squared ($r^2$) = 0.46
• F-statistic = 15.09
• Significance level (p) = 0.0005

The result of the regression (see table 22) allows rejecting the null hypothesis ($H_{13n}$), meaning that we empirically corroborate that perceived usefulness is determined by perceived ease of use.

**Table 22.** Simple Regression between PEOU and PU  
(OMIP; $\alpha = 0.05$)

<table>
<thead>
<tr>
<th>Model</th>
<th>Unstd. Coefficients (b)</th>
<th>Std. Error</th>
<th>Std. Coefficients (Beta)</th>
<th>t</th>
<th>Sig.</th>
<th>95% Confidence Interval for B</th>
<th>Correlations (r)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>2.130</td>
<td>0.362</td>
<td>5.878</td>
<td>0.000</td>
<td></td>
<td>1.369 (lower) 2.892 (upper)</td>
<td></td>
</tr>
<tr>
<td>Perceived Ease of Use</td>
<td>0.472</td>
<td>0.122</td>
<td>0.675</td>
<td>3.884</td>
<td>0.001</td>
<td>0.217 (lower) 0.728 (upper)</td>
<td>0.675</td>
</tr>
</tbody>
</table>

Regarding statistical significance, the regression equation was found to be very high significant with $\alpha < 0.001$. This means that $H_{13}$ was strongly confirmed. The significance of the regression coefficient for Perceived Ease of Use indicate that perceptions in usefulness ($H_5$) are partially determined by perceptions in ease of use ($H_4$). With respect to the predictive power of the model, Perceived Ease of Use explains 46% of the variance in Perceived Usefulness as indicated by ($r^2$). Figure 8 shows the scatterplot of Perceived Ease of Use scores against Perceived Usefulness scores with the regression line.

**Figure 8.** Scatterplot: Perceived Ease of Use vs. Perceived Usefulness.

**Phase 4: A Regression Model for Perceived Ease of Use and Perceived Usefulness Vs. Intention to Use**

In this phase, we test hypotheses $H_{14a}$ and $H_{15a}$ to verify if Intention to Use (ITU) is determined by Perceived Ease of Use (PEOU) and Perceived Usefulness (PU). To do this, PEOU and PU were used as independent (predictor) variables and ITU as the dependent (predicted) variable. The regression equation resulting from the analysis is:

$$\text{ITU} = 0.71 + 1.13 \times \text{PU} - 0.34 \times \text{PEOU}$$
The details of the regression analysis are:

- $R^2 = 0.40$
- $F$-statistic = 5.6
- Significance level ($p$) = 0.007

The result of the regression summarized in Table 23 allows us empirically corroborate that intention to use is determined by perceived ease of use and perceived usefulness.

### Table 23. Multiple Regression between PEOU, PU and ITU

(OOMFP; $\alpha = 0.05$)

<table>
<thead>
<tr>
<th>Model</th>
<th>Unstd. Coefficients (b)</th>
<th>Std. Error</th>
<th>Std. Coefficients (Beta)</th>
<th>t</th>
<th>Sig.</th>
<th>95% Confidence Interval for B</th>
<th>Correlations (r)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>0.71</td>
<td>0.934</td>
<td></td>
<td>0.765</td>
<td>0.455</td>
<td>-1.256 (lower) 2.684 (upper)</td>
<td></td>
</tr>
<tr>
<td>Perceived Ease of Use</td>
<td>-0.34</td>
<td>0.249</td>
<td>-0.350</td>
<td>-1.368</td>
<td>0.189</td>
<td>-0.865 (lower) 0.185 (upper)</td>
<td>-0.315</td>
</tr>
<tr>
<td>Perceived Usefulness</td>
<td>1.13</td>
<td>0.355</td>
<td>0.810</td>
<td>3.166</td>
<td>0.006</td>
<td>0.573 (lower) 0.609 (upper)</td>
<td>0.609</td>
</tr>
</tbody>
</table>

Regarding statistical significance, the regression coefficient for Perceived Usefulness was found to be highly significant with $\alpha < 0.01$. This means that $H_{15}$ was confirmed. This indicates that perceptions in intention to use are partially determined by perceptions in usefulness. The results do not allow rejecting $H_{14b}$ as the regression coefficient for Perceived Ease of Use is not significant. Hence we found no support for the hypothesized causal relationship between Perceived Ease of Use and Intention to Use.

With respect to the predictive power of the model, Perceived Ease of Use and Perceived Usefulness together explain 40% of the variance in Intention to Use as indicated by $(r^2)$. Figure 10 shows the 3D scatterplot of Intention to Use scores against Perceived Ease of Use and Perceived Usefulness scores.

![Figure 10. Scatterplot: Intention to Use vs. Perceived Ease of Use and Perceived Usefulness](image-url)
In fact, we agree that there are other factors that affect the decision of people in using a given technology (i.e., FSM method). Bailey and Pearson identified 39 factors that can influence user satisfaction [7]. In our case, other factors that could influence the decision of people in use a FSM method are technology infrastructure adoption, government decisions, standardization, etc. However, as point out by Cheney et al [16] there are factors that are uncontrollable (i.e., organizational time frame), others are partially controllable (i.e., systems development backlog) and others fully controllable (i.e., end-user computing).

Our objective here was select variables that can be controlled like the behaviour of people using the method. We consider perceived ease of use and perceived usefulness due to the fact that these are the most important factors in explaining system use. The objective was to provide a base for tracing the impact of external variables (measurement time, reproducibility, and accuracy) on internal beliefs, attitudes, and intentions.

However, we are aware that it is necessary to replicate this experiment and to carry out new ones that consider other factors to draw better conclusions. For instance, in [50] factors such as voluntariness, compatibility and perceived behavioral are used to explain software developer acceptance of methodologies.

7. DISCUSSION

With respect to the comparison between methods (Research Questions 1 and 2), Figure 11 shows all the significant differences found between OOmFP and IFPUG FPA. The labels on the arrows show the variables on which significant differences were found, while the arrowheads show the direction in which the difference was found (from superior to inferior).

As observed in Figure 11, only H1 was not confirmed. As mentioned before the subjects spent more time to apply the OOmFP measurement rules. However, this problem can be solved by providing tool support for measurement rules application. Actually, OOmFP has now been automated in a tool called Oliva Nova Function Points Counter [56] that obtains automatically the functional size of OO systems from OO-Method conceptual schemas.

It should be noted that even if OOmFP is perceived to be more acceptable than IFPUG-FPA this does not indicate that it is likely to be adopted. Hence, the objective of Research Question 3 was to evaluate the likely acceptance in practice of OOmFP. The results summarised in Table 24 show that only H7 was not confirmed. The statistical significance was found to be very high ($\alpha < 0.001$) for H8 and H9.

Moreover, we made an analysis of the construct validity and reliability of the instrument used to measure the perception-based variables. In the construct validity analysis we found an invalid question (Q11) that was excluded from the analysis. In the reliability analysis all questions (except Q11) were found to be reliable and valid (cronbach’s $\alpha \geq 7$) measures of the underlying constructs in the proposed theoretical model.
However, these analyses also revealed that the majority of questions that measure a given construct have high correlations with questions that measure other constructs (i.e., divergent validity is high). For future experiments we need to be careful in the definition of the questions to measure these constructs. Also, the reliability analysis reveals that Q12 is more related to PEOU than ITU.

Regarding the validation of the MAM (Research Question 4), Table 25 summarises the results of the regression analysis in terms of statistical significance (p). With the exception of Actual Effectiveness → Perceived Usefulness relationship and Perceived Ease of Use → Intention to Use, all relationships were confirmed. The statistical significance was found to be very high ($\alpha < 0.001$) for H10 and H13, and medium ($\alpha < 0.05$) for H14 and H15.

Table 25. Regression Analysis Results

<table>
<thead>
<tr>
<th>Relationship</th>
<th>Significance (p)</th>
<th>Confirmed?</th>
</tr>
</thead>
<tbody>
<tr>
<td>H10: Actual Efficiency → Perceived Ease of Use</td>
<td>0.000</td>
<td>YES</td>
</tr>
<tr>
<td>H11 and H12: Actual Effectiveness → Perceived Usefulness</td>
<td>0.315</td>
<td>NO</td>
</tr>
<tr>
<td>H13: Perceived Ease of Use → Perceived Usefulness</td>
<td>0.001</td>
<td>YES</td>
</tr>
<tr>
<td>H14: Perceived Ease of Use → Intention to Use</td>
<td>0.094</td>
<td>NO</td>
</tr>
<tr>
<td>H15: Perceived Ease of Use → Intention to Use</td>
<td>0.003</td>
<td>YES</td>
</tr>
</tbody>
</table>

Figure 12 shows the results of the regression analysis in diagrammatical form. The relationship between Actual Effectiveness and Perceived Usefulness was found to be non-significant. The Method Adoption Model was therefore partially validated by the regression analysis.

Figure 12. Validation of Theoretical Model
7.1 Summary of Findings

The results for each research question are:

- **Research Question 1**: Out of three hypotheses, two were supported. OOmFP was found to be superior to IFPUG FPA in actual effectiveness as measured in terms of the reproducibility and accuracy of the functional size measurements.

- **Research Question 2**: All three hypotheses were supported. The results show that OOmFP is perceived to be easier to use, more useful and more likely to be adopted in practice than IFPUG FPA. The results for perceived usefulness and intention to use suggest that participants perceived OOmFP to be a significant improvement on IFPUG FPA for measuring OO systems developed using the OO-Method approach.

- **Research Question 3**: Out of three hypotheses, two were supported. The results show that OOmFP is perceived to be useful and is likely to be adopted in practice. However, the result for perceived ease of use was not statistically significant.

- **Research Question 4**: Three hypotheses were supported and two were not supported. The results show that the Method Adoption Model is a partially reliable and valid instrument for evaluating FSM methods. The relationships between Actual Effectiveness and Perceived Usefulness and between Perceived Ease of Use and Intention to Use were not found to be significant.

7.2 Model Revision

It is common practice to drop any variable or relationship from the model that is found to be non-significant [45]. This means that the relationships between Actual Effectiveness and Perceived Usefulness, and between Perceived Ease of Use and Intention to Use should be removed from the model. The model then becomes a three stage simple regression model. A regression analysis was carried out using Perceived Usefulness as the independent variable and Intention to Use as the dependent variable (i.e. removing Perceived Ease of Use from Phase 4 of the analysis). The regression equation which results is:

\[
\text{Intention to Use} = 0,88 \times \text{Perceived Usefulness} + 0,80
\]

The scatterplot of Perceived Usefulness against Intention to Use is shown below, with regression line, 95% confidence intervals:

![Figure 13. Scatterplot: Perceived Usefulness vs Intention to Use](image-url)
The details of the regression are:

- $R^2 = 0.33$
- $F$ statistic = 8.8
- Significance level ($p$) = 0.004

The result of the regression (see table 26) allows us empirically corroborate that intention to use is determined by perceived usefulness.

**Table 26. Simple Regression between PU and ITU**

<table>
<thead>
<tr>
<th>Model</th>
<th>Unstd. Coefficients ($b$)</th>
<th>Std. Error</th>
<th>Std. Coefficients ($Beta$)</th>
<th>$t$</th>
<th>Sig.</th>
<th>95% Confidence Interval for $B$</th>
<th>Correlation ($r$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>0.877</td>
<td>0.948</td>
<td></td>
<td>0.925</td>
<td>0.367</td>
<td>-1.115 (lower) 2.869 (upper)</td>
<td></td>
</tr>
<tr>
<td>Perceived Usefulness</td>
<td>0.797</td>
<td>0.268</td>
<td>0.573</td>
<td>2.969</td>
<td>0.008</td>
<td>0.233 (lower) 1.361 (upper)</td>
<td>0.573</td>
</tr>
</tbody>
</table>

The major advantage of removing Perceived Ease of Use is that it simplifies the model, and all relationships are now highly significant ($\alpha < 0.01$). The revised model is shown in Figure 14.

![Figure 14. Validation of Theoretical Model](image)

However, we argue against removing the link between Perceived Ease of Use and Intention to Use for two reasons. First, as discussed earlier, there is strong evidence that the relationship between Perceived Ease of Use and Intention to Use is a contingent relationship. Such relationships cannot be easily analysed using linear regression techniques. Second, the participants in this experiment (students) are different to the target population (practitioners). Practitioners may use different decision making processes to novices for making decisions about whether to use a method or not (this is investigated in [40]).

### 7.3 Validity Evaluation

In this section we discuss several issues that can affect the validity of the empirical study and how we attempted to alleviate them.

#### 7.3.1 Threats to conclusion validity

In order to control the risk that the variation due to individual differences is larger than due to the treatment we selected a homogeneous group of subjects. Another risk in the experiment implementation is that the subjects were trained one day and the experiment was run the next day.
Thus, they might inform each other about the other method. However, we believe that the risk of doing it is low.

### 7.3.2 Threats to Construct Validity

The performance-based dependent variables we used are criteria proposed in the ISO/IEC 14143-3 [27]. The perception-based dependent variables were measured according to subject ratings, using a measurement instrument based on the literature. Moreover, we evaluated the construct validity and the reliability of this measurement instrument.

### 7.3.3 Threats to Internal Validity

We informed the students that their grade on the course was not affected by the performance in the experiment, only by their attendance. Also, the following issues have been considered:

- Differences among subjects. Using a within-subjects design, error variance due to differences among subjects is reduced. Also, we use the counterbalancing procedure where subjects were randomly assigned in two groups. This procedure cancels out a possible learning effect due to similarities in the treatments and a confounding effect that might introduce the order of learning/applying the methods.
- Knowledge of the universe of discourse. We use a same requirement specification document for all subjects. It specifies the requirements of a Project Management System for a fictitious company. This is a known universe of discourse.
- Fatigue effects. On average, each subject took three hour per session to solve the experimental tests. We run separated sections of each treatment group in the same day. The section was run in parallel. So, fatigue was not very relevant.
- Persistence effects. In order to avoid persistence effects, the experiment was carried out by subjects who had never done a similar experiment.
- Subject motivation. We motivated students in the last year in Computer Science area to participate in the experiments, by explaining to them that functional size measurement is a topic used in practice by companies for developing project estimates and having early project indicators.

### 7.3.4 Threats to External Validity

The greater the external validity, the more the results of an empirical study can be generalised. Three threats have been identified which limit the ability to apply any such generalisation:

- Materials and tasks used. We tried to use a representative requirement specification of a real case in the MIS functional domain. However, more empirical studies are needed, using others user requirement specifications within this functional domain.
- Subjects. We are aware that experiments with practitioners must be carried out in order to be able to generalise these results.
- OO-Method approach. We used a functional size measurement method for object oriented systems that follows the OO-Method concepts. So, in order to generalize the results obtained in the context of OO-Method to others OO systems we need to include in future experiments the evaluation of convertibility criteria defined in [27].

---

4 Defined in ISO/IEC 14143-1 as a class of software based on the characteristics of Functional User Requirements.
8. CONCLUSIONS AND FURTHER WORK

This paper described a controlled experiment which compares IFPUG FPA and OOmFP. The goal was to investigate which method has the highest efficacy and/or likely adoption in practice for sizing OO system within the context of an OO-Method development process. The important results from the experiment are:

- **Effectiveness:** we have corroborated that within its context, OOmFP produces more consistent and accurate assessments than IFPUG FPA.
- **Acceptance:** The results show that the likely acceptance of OOmFP is higher than that of IFPUG FPA. However, perceived ease of use needs to be revised. The results allow partially validating the Method Adoption Model in terms of the validity and reliability of the measurement instrument.
- **Validation of the theoretical model:** as discussed in section 7 part of the MAM was validated.

Despite of the initial experiment validity we are aware that more experimentation is needed in order to reconfirm these results. Several threats to the validity of this study have been identified in the paper. In particular the choice of experimental objects (limited to a unique requirements specification). The greatest threat to the generalisability of the findings of this study was the use of students as experimental subjects. We are planning a new experiment using practitioners in the context of the Spanish Association of Software Metrics.

The findings are positive in terms of future revisions of the proposed theoretical model. This is the first study to empirically test theoretical models (i.e., MAM and TAM) in the context of FSM methods. This research contributes to the creation of a body of knowledge needed in Software Engineering by establishing the foundation of theories.

Further work includes the replication of the experiment using practitioners, the measurement of other criteria such as the convertibility of FSM methods and finally, the improvement of the proposed theoretical model (i.e., increasing exploratory power).

ACKNOWLEDGMENTS

The authors would like to thank Nelly Condori and Ann Maes for their help in the execution of the experiment. This work was partly funded by Faculty of Economics and Business Administration of Ghent University and by the Spanish National Agency, MCYT Project under grant TIC2001-3530-C02-01.

REFERENCES


APPENDIX A: FSM METHOD SURVEY INSTRUMENT

For each of the following paired statements, please mark a cross over the circle which most closely matches your opinion. There are no “right” answers to these questions—just give your honest opinion, based on your experience using the sizing method.

**PLEASE READ EACH QUESTION CAREFULLY BEFORE GIVING YOUR RESPONSE**

<table>
<thead>
<tr>
<th></th>
<th>I found the procedure for applying the method complex and difficult to follow</th>
<th>I found the procedure for applying the method simple and easy to follow</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>O O O O O</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>I believe that this method would reduce the time required to measure object-oriented systems.</th>
<th>I believe that this method would increase the time required to measure object-oriented systems.</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>O O O O O</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Overall, I found the FSM method difficult to use</th>
<th>Overall, I found the FSM method easy to use</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>O O O O O</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>I found the measurement rules of the method clear and easy to understand</th>
<th>I found the measurement rules of the method confusing and difficult to understand</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>O O O O O</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Overall, I found the FSM method to be useful</th>
<th>Overall, I did NOT find the FSM method to be useful</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>O O O O O</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>I found the FSM method difficult to learn</th>
<th>I found the FSM method easy to learn</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>O O O O O</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>I will use this method if I have to measure object-oriented systems in the future</th>
<th>I will NOT use this method if I have to measure object-oriented systems in the future</th>
</tr>
</thead>
<tbody>
<tr>
<td>7</td>
<td>O O O O O</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>I think that this method would NOT improve the accuracy of estimates of object-oriented systems</th>
<th>I think that this method would improve the accuracy of estimates of object-oriented systems</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td>O O O O O</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>I found it difficult to apply the FSM method to the case study</th>
<th>I found it easy to apply the FSM method to the case study</th>
</tr>
</thead>
<tbody>
<tr>
<td>9</td>
<td>O O O O O</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Overall, I think this method does NOT provide an effective way of measuring the functional size of OO systems during the requirements phase.</th>
<th>Overall, I think this method provides an effective way of measuring the functional size of OO systems during the requirements phase.</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>O O O O O</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>---</td>
<td>----------------------------------------------------------------</td>
<td>---</td>
</tr>
<tr>
<td>11.</td>
<td>Using this method would improve my performance in measuring OO systems</td>
<td></td>
</tr>
<tr>
<td>12.</td>
<td>I would be easy for me to become skilful in using this FSM method</td>
<td></td>
</tr>
<tr>
<td>13.</td>
<td>Overall, I think this method is an improvement to the FPA method</td>
<td></td>
</tr>
<tr>
<td>14.</td>
<td>I intend to use this FSM method in the future</td>
<td></td>
</tr>
</tbody>
</table>

Please provide free text answers to each of the following questions:

15. Do you have any suggestions as to how to make the FSM method easier to use?

16. Do you have any suggestions as to how to make the FSM method more useful for measuring the functional size of object-oriented systems?

17. What are the reasons why or why not you intend to use this method in a future?

Please write any other comments you would like to make about the sizing method or the experiment in the space below

__________________________________________________________________
__________________________________________________________________
__________________________________________________________________
__________________________________________________________________
__________________________________________________________________
APPENDIX B: FSM METHOD SURVEY INSTRUMENT

In order to improve the understandability we provide the detailed data used in this experiment in tables 27 and 28.

### Table 27. Measurement Time and Functional size assessments

<table>
<thead>
<tr>
<th>Subject</th>
<th>Measurement Time in IFPUG FPA</th>
<th>Measurement Time in OOmFP</th>
<th>Size in IFPUG-FPA</th>
<th>Size in OOmFP</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2.00</td>
<td>2.30</td>
<td>129</td>
<td>150</td>
</tr>
<tr>
<td>2</td>
<td>2.15</td>
<td>3.00</td>
<td>125</td>
<td>158</td>
</tr>
<tr>
<td>3</td>
<td>2.25</td>
<td>2.95</td>
<td>115</td>
<td>172</td>
</tr>
<tr>
<td>4</td>
<td>2.50</td>
<td>2.00</td>
<td>124</td>
<td>155</td>
</tr>
<tr>
<td>5</td>
<td>2.75</td>
<td>2.85</td>
<td>143</td>
<td>151</td>
</tr>
<tr>
<td>6</td>
<td>2.45</td>
<td>3.15</td>
<td>121</td>
<td>148</td>
</tr>
<tr>
<td>7</td>
<td>2.33</td>
<td>4.23</td>
<td>114</td>
<td>158</td>
</tr>
<tr>
<td>8</td>
<td>2.37</td>
<td>2.50</td>
<td>118</td>
<td>148</td>
</tr>
<tr>
<td>9</td>
<td>2.40</td>
<td>2.25</td>
<td>145</td>
<td>160</td>
</tr>
<tr>
<td>10</td>
<td>2.50</td>
<td>2.00</td>
<td>159</td>
<td>165</td>
</tr>
<tr>
<td>11</td>
<td>3.33</td>
<td>3.25</td>
<td>121</td>
<td>151</td>
</tr>
<tr>
<td>12</td>
<td>2.67</td>
<td>3.33</td>
<td>122</td>
<td>170</td>
</tr>
<tr>
<td>13</td>
<td>2.50</td>
<td>2.95</td>
<td>165</td>
<td>163</td>
</tr>
<tr>
<td>14</td>
<td>2.54</td>
<td>3.40</td>
<td>174</td>
<td>171</td>
</tr>
<tr>
<td>15</td>
<td>2.85</td>
<td>2.75</td>
<td>132</td>
<td>163</td>
</tr>
<tr>
<td>16</td>
<td>2.25</td>
<td>2.25</td>
<td>114</td>
<td>158</td>
</tr>
<tr>
<td>17</td>
<td>1.85</td>
<td>3.50</td>
<td>147</td>
<td>173</td>
</tr>
<tr>
<td>18</td>
<td>2.37</td>
<td>4.15</td>
<td>128</td>
<td>160</td>
</tr>
<tr>
<td>19</td>
<td>2.15</td>
<td>4.00</td>
<td>125</td>
<td>163</td>
</tr>
<tr>
<td>20</td>
<td>2.00</td>
<td>3.33</td>
<td>127</td>
<td>154</td>
</tr>
</tbody>
</table>

### Table 28. Functional size assessments, and reproducibility and accuracy calculation

<table>
<thead>
<tr>
<th>Subject</th>
<th>REP, in OOmFP</th>
<th>REP, in IFPUG-FPA</th>
<th>MRE, in OOmFP</th>
<th>MRE, in IFPUG-FPA</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.06</td>
<td>.03</td>
<td>.02</td>
<td>.16</td>
</tr>
<tr>
<td>2</td>
<td>.01</td>
<td>.06</td>
<td>.03</td>
<td>.18</td>
</tr>
<tr>
<td>3</td>
<td>.08</td>
<td>.14</td>
<td>.12</td>
<td>.25</td>
</tr>
<tr>
<td>4</td>
<td>.03</td>
<td>.07</td>
<td>.01</td>
<td>.19</td>
</tr>
<tr>
<td>5</td>
<td>.06</td>
<td>.08</td>
<td>.01</td>
<td>.07</td>
</tr>
<tr>
<td>6</td>
<td>.08</td>
<td>.09</td>
<td>.03</td>
<td>.21</td>
</tr>
<tr>
<td>7</td>
<td>.01</td>
<td>.15</td>
<td>.03</td>
<td>.25</td>
</tr>
<tr>
<td>8</td>
<td>.08</td>
<td>.11</td>
<td>.03</td>
<td>.23</td>
</tr>
<tr>
<td>9</td>
<td>.00</td>
<td>.10</td>
<td>.05</td>
<td>.05</td>
</tr>
<tr>
<td>10</td>
<td>.04</td>
<td>.21</td>
<td>.08</td>
<td>.04</td>
</tr>
<tr>
<td>11</td>
<td>.06</td>
<td>.09</td>
<td>.01</td>
<td>.21</td>
</tr>
<tr>
<td>12</td>
<td>.07</td>
<td>.08</td>
<td>.11</td>
<td>.20</td>
</tr>
<tr>
<td>13</td>
<td>.02</td>
<td>.26</td>
<td>.07</td>
<td>.08</td>
</tr>
<tr>
<td>14</td>
<td>.08</td>
<td>.34</td>
<td>.12</td>
<td>.14</td>
</tr>
<tr>
<td>15</td>
<td>.02</td>
<td>.00</td>
<td>.07</td>
<td>.14</td>
</tr>
<tr>
<td>16</td>
<td>.01</td>
<td>.15</td>
<td>.03</td>
<td>.25</td>
</tr>
<tr>
<td>17</td>
<td>.09</td>
<td>.12</td>
<td>.13</td>
<td>.04</td>
</tr>
<tr>
<td>18</td>
<td>.00</td>
<td>.03</td>
<td>.05</td>
<td>.16</td>
</tr>
<tr>
<td>19</td>
<td>.02</td>
<td>.06</td>
<td>.07</td>
<td>.18</td>
</tr>
<tr>
<td>20</td>
<td>.04</td>
<td>.04</td>
<td>.01</td>
<td>.17</td>
</tr>
</tbody>
</table>