Experimental evaluation of an object-oriented function point measurement procedure

Silvia Abrahão a,*, Geert Poels b

a Department of Computer Science and Computation, Valencia University of Technology, Camino de Vera, s/n, 46022, Valencia, Spain
b Faculty of Economics and Business Administration, Ghent University – UGent, Hoveniersberg 24, 9000 Ghent, Belgium

Received 26 May 2005; received in revised form 2 June 2006; accepted 9 June 2006
Available online 26 July 2006

Abstract

This paper presents an empirical study that evaluates OO-Method Function Points (OOmFP), a functional size measurement procedure for object-oriented systems that are specified using the OO-Method approach. A laboratory experiment with students was conducted to compare OOmFP with the IFPUG – Function Point Analysis (FPA) procedure on a range of variables, including efficiency, reproducibility, accuracy, perceived ease of use, perceived usefulness and intention to use. The results show that OOmFP is more time-consuming than FPA but the measurement results are more reproducible and accurate. The results also indicate that OOmFP is perceived to be more useful and more likely to be adopted in practice than FPA in the context of OO-Method systems development. We also report lessons learned and suggest improvements to the experimental procedure employed and replications of this study using samples of industry practitioners.

Keywords: Functional size measurement; Function point analysis; OO-method function points; Object-orientation; Empirical software engineering

1. Introduction

Functional Size Measurement (FSM) methods measure the size of software systems by quantifying their functional user requirements. These are the user practices and procedures that the software system must perform to fulfill the user needs [1]. The most commonly used FSM method is Function Point Analysis (FPA), introduced in the late seventies in IBM by Albrecht [2], but now universally supported by the International Function Point Users Group (IFPUG) [3].

FPA assumes a model of functional user requirements that is expressed in terms of logically related groups of data (e.g. product data) and transaction functions that use or act upon data groups (e.g. query product data, add new product, change product data, assign to product line, etc.). Such a model is relatively easy to construct when functional user requirements are specified using structured analysis and functional decomposition techniques (e.g. Data Flow modeling, Entity-Relationship modeling). It is less straightforward to apply when object-oriented (OO) techniques are employed. The elementary unit of functional user requirements in OO analysis is the object type, which describes the properties of a class of similar objects (e.g. products). As an object type encapsulates both data and behavior, the dichotomy in the functional user requirements assumed by FPA (i.e. data groups versus transaction functions), is not readily observable in an OO model of the requirements. This makes it difficult to uniquely interpret and consistently apply the FPA measurement rules, which are based on the interactions between transactions and data groups.

Two approaches to FSM in OO systems development can be distinguished. In the first approach, new FSM methods have been developed for the specific purpose of measuring OO software systems. These methods capture and
quantify functional size by means of constructs and models that are expressed in terms of the OO paradigm. Some of the methods in the first approach define measures that deviate significantly from function points, for instance by sizing systems in terms of frequencies of OO construct occurrences such as object classes and messages (e.g. [4–10]). Other methods are analogous to FPA in that they share common aspects with the FPA structure, for instance a weighting schema to express the relative functional size contributions of different types of OO constructs (e.g. [11–14]).

The second approach consists of methods that reuse the FPA constructs and model, but map them onto the language used by the OO analysis method for which they are developed (e.g. [15–17]). These FSM methods do not define new measures, but offer a measurement procedure\(^1\) that can be used in combination with FPA. Their main objective is to prescribe the supposedly correct interpretation of the FPA measurement rules for a target application context (i.e., an OO analysis method), and hence facilitate and improve the performance of FSM in this context. The main advantage of these methods over the first approach is that they result in a function point value that can be compared to other function point values, obtained using different measurement procedures (such as the ‘regular’ procedure proposed by IFPUG [3]) and across application contexts. The advantage is significant as over the years thousands of projects have been measured with FPA and a substantial amount of these historical data are publicly available for analysis and benchmarking purposes (e.g. in the ISBSG repository [19]).

In [20] we have presented an FPA-to-OO mapping that we developed as a function point measurement procedure for software systems specified using OO-Method [21]. OO-Method is an automated software production method that takes a model-driven approach to systems development based on conceptual modeling, model transformation, and code generation. Our proposal, called OO-Method Function Points (OOmFP), consists of a set of measurement rules and a procedure prescribing how to apply these rules on an OO-Method conceptual representation of the system specifications.

According to March and Smith’s framework for research in IT [22], OOmFP is a design science artifact resulting from a build research activity. As a method it is a set of steps used to perform a task (i.e. FSM), based on a set of underlying constructs (i.e. the FPA concepts) and a model of the solution space (i.e. FPA concepts mapped onto OO-Method conceptual modeling primitives). March and Smith argue that information technologies (as design science research outputs in IT) “once built, must be evaluated scientifically” (p. 258). Specifically for methods, evaluation considers “operationality (the ability to perform the intended task or the ability of humans to effectively use the method if it is not algorithmic), efficiency, generality, and ease of use” (p. 261). Hence, after the build activity, the evaluate research activity must answer the basic question how well the new information technology works.

This paper presents an evaluation of OOmFP as a design science artifact. The framework for the study is the Method Evaluation Model (MEM) of Moody [23], which is a model for evaluating information system (IS) design methods. Although this model was developed for evaluating IS design methods, its constructs can be described in terms of other criteria to evaluate other types of methods. Hence, we used the MEM to verify if and how much OOmFP improves the performance of function point measurement of systems specified using OO-Method. A laboratory experiment was conducted in which we compared the performance of students employing OOmFP against that of students employing the ‘regular’ IFPUG procedure for function point measurement. Following the MEM, the OOmFP and IFPUG function point measurement procedures were compared on objective as well as perception-based variables of task performance, including efficiency (i.e. effort required to apply FSM), reproducibility of the FSM results, accuracy of the results, perceived ease of use, and perceived usefulness. In addition, the intention to use a method was measured in an attempt to estimate the likely adoption in practice of OOmFP.

The paper is organized as follows. In Section 2, we briefly introduce the reader to OO-Method, OOmFP, and the MEM. In Section 3, we present the design of the laboratory experiment used to evaluate OOmFP. The data collected in the experiment are analyzed in Section 4. In Section 5, we describe the lessons learned from this first empirical evaluation study, assess the validity and generalisability of the study results, and suggest improvements to the experimental procedure that can be implemented in future replication studies. Finally, in Section 6 we present conclusions and outline future work.

2. Background

This section starts with a brief introduction to OO-Method and OOmFP. Next, we discuss the problem of how to evaluate FSM methods. The section ends with our solution approach in the form of an evaluation model based on the MEM [23].

2.1. OO-Method and OOmFP

OO-Method is a model-driven requirements and software engineering approach developed at the Valencia University of Technology [21]. In OO-Method, a clear distinction is made between two models: the Conceptual Model (centered on the problem space) and the Execution Model (centered on the solution space). The OO-Method development process is automated by implementing the provided set of mappings between conceptual primitives.

---

\(^1\) A set of operations, described specifically, used in the performance of particular measurements according to a given method of measurement”, as described in the International Vocabulary of Basic and General Terms in Metrology [18].
(problem space) and their software representations (solution space).

The Conceptual Model is used to capture the relevant properties of an object-oriented system in a graphical way. It is based on the specification of four complementary model views using an UML-compliant notation:

- **Object Model**: captures the structure of the system in terms of classes identified in the problem domain and their relationships.
- **Dynamic Model**: captures the behavior of the system in terms of valid object lives and interaction between objects.
- **Functional Model**: captures the semantics associated to the state changes of the objects.
- **Presentation Model**: captures the user interaction patterns for the system interface.

A system specification is described using these four model views, each one representing a particular perspective, but all together constituting the whole conceptual schema of the required system. This conceptual schema is a formal representation of the concepts that characterize the system as required by its users in a manner that is independent of the system design and implementation.

As a conceptual representation of the user requirements, the four model views of an OO-Method conceptual schema contain all the elements that may possibly contribute to the functional size of the system. Hence, the conceptual schema is a sufficient input for FSM and thus FPA can be applied to an OO-Method conceptual schema to measure the system's functional size.

Experience reports on FSM of OO systems have recognized that FPA can be used, even if the functional user requirements are captured in a different way (i.e. using OO constructs and models) than usually assumed [24, 25]. However, this does not mean that it is straightforward to apply FPA in such contexts. Therefore, OO-Method Function Points [20] was developed to facilitate the function point measurement of an OO-Method conceptual schema. OOmFP provides a modeling of the OO-Method Conceptual Model concepts onto the FPA concepts. This mapping is a formalized in a set of measurement rules. A procedure has been devised to apply these measurement rules (Fig. 1). This step consists in identifying the elements in the OO-Method conceptual schema that contribute to the functional size of the system and to disregard those that do not contribute to functional size. The result of this step is a collection of uniquely identified and categorized data groups and transaction functions that can be quantified in the next step.²

Hence, in the OOmFP measurement step, a function point value is obtained through the application of two other subsets of measurement rules. First, measurement rules are used to identify the elements that help determining the complexity of the identified data groups and transaction functions. These elements along with the data groups and transaction functions to which they apply, form a system representation according to the OOmFP Measurement Abstract Model of user requirements. Next, a last subset of measurement rules, which are standard IFPUG counting rules [3], are used to rate the complexity of the data groups and transaction functions, to transform complexity ratings into function point values, and to aggregate the assigned values into an overall functional size value for the system.

We remark that the FPA adjustment phase is not included in OOmFP. The view on functional size measurement articulated in the ISO/IEC 14143 series of standards on FSM [1] prohibits the use of an adaptation factor that does not assess functional user requirements (but assesses, for instance, quality or technical requirements as is the case with FPA's adjustment factor). Therefore, OOmFP expresses functional size in unadjusted function point values.

### 2.2. Evaluating FSM methods

Despite the wide usage of established FSM methods such as FPA [3] and Mark II FPA [27], and the proliferation of newer ones such as COSMIC-FFP [28], there currently exists no methodology for their systematic evaluation. The best approximation to an FSM method evaluation framework is the 14143 series of standards for FSM issued by the ISO/IEC.

Part 1 [1] of these standards defines the concepts of FSM and establishes a basis against which existing and new FSM methods can be compared. Part 2 [29] provides a process for checking whether a candidate FSM method conforms to the concepts described in part 1. Part 3 [30] provides guidance for verifying the stated performance of an FSM method with respect to its repeatability, reproducibility, accuracy, convertibility, discrimination threshold, and applicability to functional domains. The standard specifies, for instance, requirements for verification methods to

---

² FPA categorizes data groups into internal logical files (ILF) and external interface files (EIF). Transaction functions are external input functions (EI), external output functions (EO) or external inquiry functions (EQ) [3].

³ These elements are data element types (DET) and record element types (RET) for data groups, and DETs and file types referenced (FTR) for transaction functions [3].

⁴ For more information on the different subsets of OOmFP measurement rules, and their relationship to the IFPUG rules for function point measurement, we refer to [20].
ensure that verification results are objective, impartial, consistent and repeatable.

Part 4 \cite{31} defines a reference model to be used when verifying an FSM method. Although this part provides some examples of reference objects against which an FSM method can be applied and compared, the formulation and execution of evaluation tests as well as the interpretation of their results is outside the scope of this document. Finally, Part 5 \cite{32} addresses the definition of functional domains for which FSM can be used. A functional domain is a class of software characterized by specific software requirements (e.g. enterprise systems software, embedded software, etc).

The editors for part 3 of the standards, Jacquet and Abran, have also presented in \cite{33} (not part of the standards) a process model for evaluating FSM methods. This model describes all processes and activities involved in the definition and validation of FSM methods. However, the model of Jacquet and Abran is very general and needs to be specialized and detailed depending on the specific purpose for which the FSM method was developed. For OOmFP this specific purpose is to facilitate the function point measurement of OO-Method conceptual schemas, thereby improving upon the IFPUG measurement procedure for FPA. The evaluation must therefore focus on the performance of OOmFP in achieving its objective, meaning provide a ‘better’ measurement procedure than IFPUG does within the OO-Method application context.

Some of the performance properties in part 3 of the standards offer relevant evaluation criteria. The OOmFP and IFPUG measurement procedures can be directly compared with respect to repeatability, reproducibility, and accuracy. The other properties (i.e. convertibility, discrimination threshold, and applicability to functional domains) apply to the functional size measure used by an FSM method, hence cannot be used to differentiate between OOmFP and IFPUG (as both use Function Points (FPs) as a measure).

The retained properties relate to the effectiveness of FSM. According to March and Smith \cite{22}, efficiency, generality, and ease of use must also be considered when evaluating methods. Although part 3 of the ISO/IEC standards for FSM and the process model of Jacquet and Abran provide a starting point for evaluating OOmFP, what is missing is a structured view of the quality of a FSM measurement procedure. A simple list of isolated quality criteria hides relationships between quality properties. A quality model for FSM measurement procedures should explicitly describe how quality properties of different dimensions (e.g. effectiveness versus efficiency) are related and what their impact is on overall quality. Furthermore, a systematic evaluation based on such a model should not only seek objective evidence of the performance of a method in achieving its objectives. It should also test the user’s response to a new method and allow predicting its acceptance in practice. The importance of measuring perception-based performance variables (e.g. ease of use) for assessing software developers’ acceptance of new methods has been stressed in recent software engineering literature (see e.g. \cite{34}).

2.3. An evaluation model based on the MEM

Moody’s Method Evaluation Model (MEM) \cite{23} provides a suitable basis for a multidimensional quality model of FSM measurement procedures (Fig. 2). The MEM was originally proposed as an evaluation model for IS design methods. The main contribution of the MEM, compared
to alternative models, is that it incorporates two different aspects of method “success”: actual efficacy and actual usage. As argued in the previous subsection, both aspects must be considered when evaluating FSM methods or procedures.

In the MEM, efficacy is defined as a separate construct, different from efficiency and effectiveness. The efficacy construct is derived from Rescher’s notion of pragmatic success [35], which is defined as the efficiency and effectiveness to which a method achieves its objectives. Thus, the evaluation of the efficacy of a method requires measurement of both effort required (efficiency) and the quality of the results (effectiveness).

The core of the MEM, called the Method Adoption Model (MAM), is based on the Technology Acceptance Model (TAM) [36], a well-known and thoroughly validated model for evaluating information technologies 5. The constructs of the MEM are:

- **Actual Efficacy**, which consists of two parts:
  - Actual Efficiency: the effort required to apply a method. This represents an input variable to the MAM.
  - Actual Effectiveness: the degree to which a method achieves its objectives. This also represents an input variable to the MAM.

- **Perceived Efficacy**, which consists of two perception-based MAM variables:
  - 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 achieve its intended objectives.

- **Intention to Use**: an intention-based MAM variable for predicting adoption in practice, defined as the extent to which a person intends to use a particular method.

- **Actual Usage**: a behavior-based variable, defined as the extent to which a method is used in practice. This represents an output variable from the MAM.

To evaluate FSM measurement procedures such as OOmFP, the constructs of the MEM must be operationalized for use with this kind of method. Efficiency can be defined as the time it takes to apply the measurement procedure to a system of a given size. Effectiveness can be defined in terms of the retained performance properties from part 3 [30] of the ISO/IEC standards for FSM, i.e. repeatability, reproducibility, and accuracy. The perception/intention-based variables of the MAM can be directly used. Only the behavior-based variable Actual Usage is not used in the evaluation of OOmFP, as it can of course not be measured for newly proposed methods (only for established ones).

### 3. Method

In terms of the GQM template for goal definition [37], our empirical study aimed at analyzing function point measurement for the purpose of evaluating the OOmFP measurement procedure (hereafter referred to as ‘OOmFP’) against the IFPUG measurement procedure (hereafter referred to as ‘FPA’) with respect to their actual and perceived efficacy and likely adoption from the point of view of the researcher. The context of the study is an advanced software engineering course at the Valencia University of Technology, where the OO-Method approach to systems modeling is taught.

Our study was conducted as a laboratory experiment. The design of this experiment was based on the framework for experimentation in software engineering research suggested by Wohlin et al. [38]. The broad research questions addressed are:

- RQ1: Is the actual efficacy of OOmFP higher than the actual efficacy of FPA?
- RQ 2: Is the intention to use and the perceived efficacy of OOmFP higher than the intention to use and the perceived efficacy of FPA?

#### 3.1. Variables and operational hypotheses

The independent variable is the measurement procedure used: OOmFP or FPA. Hence, the experiment employed two treatments: the function point measurement of a system with OOmFP and the function point measurement of the same system with FPA. The experimental data collected allows comparing the effects of both treatments.
There are two types of dependent variables for comparing the treatments: performance-based and perception/intention-based variables. Performance-based variables assess how well the experiment participants perform the experimental task. They are used to evaluate the actual efficacy of the measurement procedures. Perception/intention-based variables assess the participants’ perceptions of their performance and their subsequent intention to use OOmFP or FPA. These variables are used to evaluate the perceived efficacy of the measurement procedures, as well as their likely adoption in practice.

Three performance-based variables were selected. The first variable is measurement time which operationalizes the actual efficiency construct of the MEM. It is defined as the time taken by a subject to complete the measurement task. For a fair comparison of the measurement time, it was important to ensure the equivalence of the measurement task for both treatments. Although OOmFP is applied on an OO-Method conceptual schema (see Fig. 1), we decided not to include the time spent on the conceptual modeling of the user requirements. Our motivation for this decision is that OOmFP was specifically developed to be used in an OO-Method application context, and thus an OO-Method conceptual schema will be developed anyway (independent of the need for FSM). On the other hand, we also realized that forcing participants to apply FPA to the OO-Method conceptual schema would disadvantage this measurement procedure, given the aforementioned mapping problem between FPA and OO concepts.

Besides, in practice, when documented user requirements and the conceptual schema are available, users of FPA can chose which format to use as an input for FSM. Therefore, in the experiment, FPA was directly applied to the user requirements specifications and not to the OO-Method conceptual schema.

The second performance-based variable is reproducibility, defined as the agreement between the function point measurement results of different participants applying the same measurement procedure on the same system. This variable, taken from part 3 of the ISO standards for FSM [30], operationalizes the actual effectiveness construct of the MEM. We postulate that the more similar the function point values obtained by different raters, the more effective the measurement procedure is in terms of consistency.

However, even when the obtained values lie close to each other, they can be far from the ‘true value’ for functional size. Consequently, we selected accuracy from part 3 of the FSM standards [30] as a complementary effectiveness variable. Accuracy is defined as the agreement between the function point measurement results of different participants obtained by applying a same measurement procedure on a same system and the true function point value for that system. To obtain this ‘true value’, the requirements specification documents were measured using the IFPUG function point measurement procedure by four Certified Function Point Specialists (CFPS), who were blind to the research questions tested in the experiment.

We wish to remark that repeatability, which is an alternative way to express consistency of measurement, is not considered in this experiment. This is because repeatability is measured through successive applications of the same measurement procedure under the same conditions of measurement (i.e. the same subject measuring the same system at different times). If measurement is performed by humans, we may assume that they remember (at least to some extent) having measured the same system. The learning effects that may result from such recall are difficult to control in an experiment of this kind.

As perception/intention-based variables, we used the Perceived Ease of Use (PEOU), Perceived Usefulness (PU) and Intention to Use (ITU) variables of the MAM.

The hypotheses tested in the experiment are summarized below. The first three hypotheses relate to the first research question while the last three hypotheses relate to the second research question. Because the goal of the experiment was to evaluate OOmFP, which was proposed to improve FPA as a function point measurement procedure when using OO-Method, the direction of the effect has been stated in the hypotheses.

\[ H_{\text{Time}}: \text{OOmFP takes less time to apply than FPA.} \]
\[ H_{\text{Rep}}: \text{OOmFP results in a more reproducible function point measurement than FPA.} \]
\[ H_{\text{Acc}}: \text{OOmFP results in a more accurate function point measurement than FPA.} \]
\[ H_{\text{PU}}: \text{OOmFP is perceived to be more useful than FPA.} \]
\[ H_{\text{ITU}}: \text{OOmFP is more likely to be used than FPA.} \]

3.2. Participants

The participants in the experiment were twenty-two students in the final year of Engineering in Computer Science, majoring in Information Systems. The students aged between 22 and 24 years and they had similar backgrounds (they previously attended the same courses about software development and management). The subjects were chosen for convenience, i.e., they were students enrolled in the

---

6 http://www.ifpug.org/certification/.
7 Whether the function points value produced by an CFPS is really the ‘true value’ of the system’s functional size is a question concerning the validity of function points as a functional size measure, and given our focus on measurement procedure evaluation outside the scope of the paper. The certification is an assurance that the value produced by an CFPS is ‘true’ in the sense that function point measurement was applied correctly. Because variability (or error rates) by CFPS up to 10% is accepted, we contacted a number of CFPS from the USA, Australia and Brazil to provide us with a ‘true value’. From the total of four responses received, the mean function points value was used as the ‘true value’.
Software Development Environments course during the period of February until June of 2003.

The Software Development Environments course was selected because it was an advanced unit (where students learn advanced techniques about software development). The necessary preparation for the experiment and the experimental task itself fitted well into the scope of the course, as one of the course objectives is to learn students about FSM in the OO-Method development context. To reduce the risk of experimenter bias (i.e., students wanting to please us by indicating a preference for OOmFP), students were informed beforehand that their answers would not affect their grade for the course and that the experimental tasks were to be considered as exercises on FSM (also to motivate the students to perform well and take the tasks seriously).

3.3. Design and instrumentation

In order to balance differences in ability between the participants as well as to get more data, this experiment used a within-subject design. The students were randomly assigned into two groups, one group using FPA first and OOmFP second, and another group using OOmFP first and FPA second. This counterbalancing procedure allowed us to cancel out possible order or learning effects due to similarities in the treatments, in particular the requirement to measure the same system in both treatments.

The instrumentation of the experiment consisted of a user requirements specification and the corresponding OO-Method conceptual schema describing a project management system for a fictitious company. The requirements documents used in the experiment followed the IEEE 830 standard [39] and also included an Entity-Relationship diagram and screen prototypes. Prior to the experiment, the requirements specification and conceptual schema were verified in a small pilot study and the obtained feedback was used to improve the understandability of the experimental materials. We also stress that it was carefully verified that the conceptual schema obtained with OO-Method conformed to the user requirements specification. Therefore, the results obtained with both procedures can be compared as they receive the same user requirements as input, albeit in a different format.

During the experiment, participants could also refer to previously used FSM training materials in the form of instructional slides describing FPA and OOmFP, a complete example for both procedures, and a measurement guideline summarizing the measurement rules of both procedures.

Finally, there was a post-task survey with 14 items (see Appendix A) used to measure the constructs of the TAM. These items were based on the measurement instrument proposed by Moody [23], which was itself based on the original TAM instrument for PEOU, PU and ITU [36]. The items of the survey instrument were formulated using a 5-point Likert scale, using the opposing statements format. The order of the items was randomized and half the item statements negated to avoid monotonous responses. PEOU is measured using five items on the survey (items 1, 3, 4, 6, and 9). PU is measured using six items on the survey (items 2, 5, 8, 10, 11, and 13). Finally, ITU is measured using three items on the survey (items 7, 12, 14).

3.4. Operation and data collection procedures

All students participating in the experiment received training sessions on FPA and OOmFP. For each procedure, two sessions of 2 h were needed. In the first session, the measurement rules of the procedure were explained and demonstrated using toy examples. In the second session, the application of the measurement rules on a complete case study was demonstrated. We are aware that a total training of 4 h per measurement procedure is not sufficient to turn the students into experienced function point counters. The group of students available for the experiment therefore approaches a sample of novice users of FSM methods, rather than a sample of function point analysis experts. On the other hand, the main objective of the training was to enable a fair comparison between both function point measurement procedures. Hence, from the point of view of the study’s internal validity, more important than the intensity and duration of the training was that all participants were equally trained in OOmFP and FPA, so that this study variable was effectively controlled.

During the first run of the experiment, each participant performed the measurement task using either OOmFP or FPA (depending on the group to which the participant was allocated) and next completed the post-task survey. During the second run, the participants that used OOmFP in the first run, now used FPA and the participants that used FPA in the first run, now used OOmFP. Next, the post-task survey was completed for the second time. Since the same system was measured in both runs, participants did not receive feedback on their task performance before completing the post-task survey. Giving the students feedback after their first-run measurement task could have impacted their behavior and subsequent performance in the second run of the experiment. To maintain the comparability of the data collected during both runs, no feedback was given after the second-run measurement task either.

The consequence of giving students no feedback on their performance before completing the post-task surveys is that the perceptions of ease of use, usefulness and intention...
to use are based on perceptions of task performance (subjective), rather than actual task performance in terms of reproducibility and accuracy of the measurement results (objective). As this is true for both treatments, the lack of feedback before answering the questionnaire does not threaten internal validity. Furthermore, when completing the post-task survey for the second time, participants can base their answers on an implicit comparison of their perceptions of task performance for the two measurement procedures used (e.g., did I find the second procedure easier to use or more useful than the first procedure?). Answering the questionnaire in a relative, comparative sense does not threaten construct validity because we do not interpret the participant ratings in an absolute sense (i.e., as absolute value judgments).

To avoid a possible ceiling effect, there was no time limit for the experimental tasks. We also controlled that no interaction whatsoever between participants occurred.

The performance-based variables were collected using data collection forms. These forms were used to record the outputs of the measurement procedures and the time spent (controlled by us). Once the data were collected, we verified whether the forms were complete. As two students did not record their FPA results, we took into account only the responses of the twenty remaining students.

4. Data analysis and interpretation of results

4.1. First research question

Descriptive statistics for measurement time (in hours) and measurement results (function point values) are presented in Table 1. Both mean measurement time and function point value are higher for OOmFP than for FPA. The full data set is shown in Table B-1 of Appendix B.

We first tested hypothesis $H_{\text{Time}}$ related to the efficiency of the measurement procedures in terms of measurement time. The Kolmogorov–Smirnov test for normality was applied to the differences of the paired observations. As this distribution was normal, we used the Paired t-test to check for a difference in mean measurement time between OOmFP and FPA.

The result of the test\(^9\) does not allow us to reject the null hypothesis, meaning that we cannot empirically corroborate that OOmFP takes less time to apply than FPA. In fact, the test shows exactly the opposite. The mean measurement time for FPA is significantly lower than that for OOmFP (see Table 2).

Next, we tested hypothesis $H_{\text{Rep}}$ related to the effectiveness of the measurement procedures in terms of reproducibility of the measurement results. To assess reproducibility, we used a practical statistic similar to that proposed by Kemerer [40]. This statistic is calculated as the difference in absolute value between the function point value produced by a participant and the average function point value produced by the other $n-1$ participants in the sample (using the same measurement procedure), relative to this average value. Reproducibility scores (REP) were thus obtained for each observation by applying the following equation:

$$REP_i = \left| \frac{\sum_{k \neq i} FPValue_k}{n-1} - FPValue_i \right|,$$

(1)

The lower the $REP_i$ score for a participant, the closer the function point value obtained by that participant is to the values obtained by the other participants. The obtained $REP_i$ scores for FPA and OOmFP are presented in Table B-2 of Appendix B.

The Kolmogorov–Smirnov test indicated a non-normal distribution of the differences in paired reproducibility scores. Therefore, we used the non-parametric Wilcoxon signed rank test for the difference in median reproducibility scores. The result of the one-tailed test (see Table 3) allows empirically corroborating that OOmFP (median value 0.04) results in more reproducible function point measurements than FPA (median value 0.09).

Next, we tested hypothesis $H_{\text{Acc}}$ related to the effectiveness of the measurement procedures in terms of accuracy. The values obtained by the four CFPSs were 178, 142, 156 and 153 FPs. Therefore, we used the mean CFPS value (157 FPs) to measure the subjects accuracy. To compare the function point value produced by a participant against the mean value obtained by the CFPS, the Magnitude of Relative Error (MRE) was used [41–44]. Thus, accuracy scores were obtained by applying the following equation for each observation:

$$MRE_i = \left| \frac{\text{CFPS FPValue} - \text{FPValue}_i}{\text{CFPS FPValue}} \right|.$$

(2)

The lower the MREi score is for a participant, the closer the function point value produced by the participant is to the mean function point value obtained by the CFPSs, which is considered as the ‘true value’ of functional size. The obtained $MRE_i$ for FPA and OOmFP are presented in Table B-2 of Appendix B.

\(^9\) All the collected data were analyzed according to the following levels of significance: not significant ($z > 0.1$), low significance ($z < 0.1$), medium significance ($z < 0.05$), high significance ($z < 0.01$) and very high significance ($z < 0.001$).
Table 2
Paired samples t-test rank for differences in mean measurement times (FPA versus OOmFP; z = 0.05)

<table>
<thead>
<tr>
<th>Dependent variable</th>
<th>Mean difference</th>
<th>Standard deviation</th>
<th>Standard error mean</th>
<th>95% conf. interval of the difference</th>
<th>t</th>
<th>One-tailed p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Measurement</td>
<td>–0.596</td>
<td>0.7716</td>
<td>0.17254</td>
<td>–0.957 (lower)</td>
<td>–3.457</td>
<td>0.0015</td>
</tr>
<tr>
<td>Time</td>
<td>–0.350</td>
<td>0.235</td>
<td>0.054</td>
<td>–0.685 (lower)</td>
<td>–2.035</td>
<td>0.042</td>
</tr>
</tbody>
</table>

Table 3
One-tailed Wilcoxon signed rank test for differences in median reproducibility and accuracy scores (FPA versus OOmFP; z = 0.05)

<table>
<thead>
<tr>
<th>Dependent variable</th>
<th>Mean rank</th>
<th>Sum of ranks</th>
<th>z</th>
<th>One-tailed p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reproducibility</td>
<td>5.25</td>
<td>10.50</td>
<td>–3.409</td>
<td>0.001</td>
</tr>
<tr>
<td>Accuracy</td>
<td>3.50</td>
<td>7.00</td>
<td>–3.662</td>
<td>0.001</td>
</tr>
</tbody>
</table>

Because the Kolmogorov–Smirnov test indicated a non-normal distribution of the paired differences in MREi between OOmFP and FPA, we again used the Wilcoxon signed rank test for the difference in median accuracy measurements. The result of the test (see Table 3) allows us to empirically corroborate that OOmFP results in more accurate measurements than FPA. The median magnitude of relative error was 18% for FPA, whereas it was only 4% for OOmFP.

4.2. Second research question

The second research question concerned the participants’ perception of OOmFP’s efficacy and intention to use OOmFP, as compared to FPA. The perception/intention-based dependent variables were measured according to subject ratings, using a measurement instrument based on the literature. The construct validity of this instrument was evaluated using an inter-item correlation analysis [45].

The evaluation is based on two criteria: convergent validity and discriminant validity. Convergent validity (CV) refers to the convergence among the different items that are meant to measure a same construct. The correlation between such items should be high. The CV value of an item is calculated as the average correlation between the scores for that item and the scores for the other items that are intended to measure the same construct. Discriminant validity (DV) refers to the divergence of the items used to measure different constructs. The correlation between items used to measure different constructs should be low. The DV value of an item is calculated as the average correlation between the scores for that item and the scores for the items that are intended to measure another construct. Low DV values indicate high discriminant validity.

According to [45] an item’s CV value must be higher than its DV value; otherwise the scores on the item should not be used in the data analysis.

An inter-item correlation analysis (see Table B-3 of Appendix B) shows that for all PEOU and ITU items the CV value was higher than the DV value (with a minimum difference of 0.20). The average convergent validity of all items was 0.57 for PEOU and 0.62 for ITU. For PU, five out of six items had a CV value higher than the DV value. The remaining item, Q11, has a same value for CV and DV. For this reason, the Q11 item was excluded from the data analysis. Removing Q11 increased the average convergent validity of the PU items from 0.40 to 0.47.

The use of multiple items to measure a same construct also requires the examination of the reliability or internal consistency among the measurements [46]. We evaluated reliability using Cronbach’s $\alpha$. For this analysis, item Q11 was excluded. All constructs have an alpha value equal to or greater than 0.7 (0.82 for PEOU, 0.70 for PU, and 0.70 for ITU), a common reliability threshold for exploratory research [47]. The overall Cronbach’s $\alpha$ obtained for the instrument (excluding item Q11) was 0.74.

A participant’s score for a perception/intention-based variable was calculated as the mean of the scores assigned to the different items of that variable. Descriptive statistics for the three variables of interest, perceived ease of use (PEOU), perceived usefulness (PU), and intention to use (ITU), are presented in Table 4. The full data set is shown in Table B-4 of Appendix B. Higher scores mean more positive perceptions and intentions towards a measurement procedure. Mean scores for all three variables are higher for OOmFP than for FPA.

Before testing H PEOU, H PU and H ITU, the Kolmogorov-Smirnov test for normality was applied to the differences of the paired observations. As all distributions were normal, we used the parametric Paired t-test to evaluate the statistical significance of the observed differences in mean PEOU, PU and ITU.

The results of the one-tailed t-tests (see Table 5) do not allow us to empirically corroborate that OOmFP is perceived to be easier to use than FPA (H PEOU) given that the test result was not significant ($z > 0.1$). On the other hand, the test results allow us to corroborate that OOmFP is perceived as being more useful than FPA and the participants’ intention to use it is higher than with FPA. The statistical significance was found to be very high ($z < 0.001$) for H PU and H ITU.

Table 4
Descriptive statistics for perception-based variables

<table>
<thead>
<tr>
<th>Descriptive statistics</th>
<th>Perceived ease of use</th>
<th>Perceived usefulness</th>
<th>Intention to use</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>FPA</td>
<td>OOmFP</td>
<td>FPA</td>
</tr>
<tr>
<td>Mean</td>
<td>2.7100</td>
<td>2.90</td>
<td>2.4800</td>
</tr>
<tr>
<td>Standard deviation</td>
<td>0.4517</td>
<td>0.7033</td>
<td>0.4606</td>
</tr>
<tr>
<td>Minimum</td>
<td>1.80</td>
<td>1.60</td>
<td>1.60</td>
</tr>
<tr>
<td>Maximum</td>
<td>3.60</td>
<td>4.40</td>
<td>3.60</td>
</tr>
</tbody>
</table>
4.3. Summary and interpretation of the results

The results of the data analysis can be summarized as follows:

- The function point measurements obtained using OOmFP were more reproducible and accurate than those obtained using the IFPUG measurement procedure;
- The IFPUG function point measurement procedure took less time to apply than OOmFP;
- OOmFP was perceived to be more useful and more likely to be adopted in practice than the IFPUG measurement procedure;
- There was no difference in perceived ease of use between the OOmFP and IFPUG function point measurement procedures.

For the performance-based variables, the results suggest that OOmFP accomplishes its objectives, but takes more effort to apply than the IFPUG procedure. A plausible reason why the experiment participants spent more time applying OOmFP is that its measurement rules are very detailed and were defined taking into account four dimensions of an object-oriented system (i.e. data, behavior, process and presentation). To apply each of the measurement rules, participants needed to verify the corresponding conceptual model specification. For example, to apply the rule “counting 1 FTR for each new class referenced in the formula of a precondition”, they had to check all the preconditions defined in the Dynamic Model specification of the system being measured. It is thus likely that the higher effectiveness/lower efficiency of OOmFP versus the IFPUG procedure is a result of the high degree of detail of the OOmFP measurement procedure, resulting in less interpretation freedom on behalf of the measurer, but also requiring a meticulous investigation of the conceptual schema.

From a practical point of view, the relative inefficiency of OOmFP vis-à-vis the IFPUG procedure is not really a problem, given that the use of OOmFP can be supported by a tool called OlivaNova Function Point Counter [49]. By the same token, the observed advantages of using OOmFP with respect to effectiveness are less relevant when function point measurement is tool-supported and thus less reliant on human judgment.

With respect to the perception/intention-based variables perceived usefulness and intention to use, the results suggest that participants perceived OOmFP to be an improvement over the ‘regular’ IFPUG measurement procedure to size OO systems specified using OO-Method. The lack of empirical support for a higher perceived ease of use is not surprising given that OOmFP took more time to apply than the IFPUG procedure. But even then, the perceived ease of use of OOmFP was not lower than that of the IFPUG procedure.

Before we draw conclusions based on the experiment results, the limitations of the study and their impact on the study validity need to be discussed. The next section outlines the lessons learned from conducting the experimental comparison of the OOmFP and IFPUG function point measurement procedures, and suggests some improvements to the experimental procedure applied.

5. Lessons learned

The main limitation of the study is the use of student participants, in particular students with no prior experience in function point analysis and having received only limited training in function point counting (i.e. 4 h for each measurement procedure employed in the experiment). The use of student participants in our study does not present a threat to the study’s internal validity. The students participating in our experiment were final year students, which can be considered as being close to the real population under study (i.e. software developers, project managers).10 Furthermore, the level of training received was the same for both treatments, allowing us to control this variable.

The use of students does reduce the external validity and thus generalizability of the study results. The group of students participating in the experiment are only representative of novice users of FSM. It is possible that novice FSM users prefer measurement procedures in the form of a ‘script’ to follow (as OOmFP presents) more than experienced FSM users do. In addition, we do not know if the same performance differences as in the experiment would be observed at higher levels of training. Therefore, the current study needs to be replicated using, preferably, practitioners experienced in function point analysis. If not available, replication studies can be performed with students, but at higher levels of function point counting training in order to confirm or disconfirm the currently obtained results and to increase external validity.

A related problem and possible threat to the study’s internal validity, though only with respect to the perception-based variables, is experimenter bias. The student participants, well versed in OO-Method, might perceive OOmFP as more useful than the IFPUG measurement

---

10 The use of student participants in empirical software engineering studies is relatively well accepted as they are the next generation of professionals, see e.g. [48].
procedure because they associate it with the familiar OO-Method approach and related tools (all developed at their university), and not because of their perceived performance on the experimental task. They might also indicate a higher intention to use OOmFP (instead of the standard IFPUG function point measurement procedure) just to please the experimenters.

Although the students were informed that their participation in the experiment would have no effect on their course grade, we acknowledge that to really eliminate experimenter bias, a replication using an 'independent' sample of study participants would be ideal. An improvement to the experimental procedure we employed, could be the use of the double-blind procedure, where the identity of both participants and experimenters is hidden. A difficulty would be that such a group of participants would need to be trained in OO-Method as well, as OO-Method practitioners would probably link OOmFP to the research group that developed OO-Method (making the double-blind procedure ineffective). An alternative change to our experimental procedure is to also measure user preferences towards particular FSM measurement procedures independent of the performance of experimental tasks. With such additional measurements, it can be investigated whether OOmFP's higher perceived usefulness and intention to use is due to (biased) subjective preferences or to perceived differences in task performance with the IFPUG procedure (or to both reasons).

Another lesson we have learned from this study is that the FSM results reported by the participants need to be sufficiently detailed to allow for an in-depth analysis of task performance. This issue is related to the construct validity of the accuracy measure we used. In our interpretation of the results, the higher accuracy with OOmFP was assumed to be a function of the measurement procedure offered, which, like a 'script', directs the user to the functionality to include in the function point count. So, it might be that novice counters could find functionality with OOmFP that they would miss using the IFPUG procedure. An alternative explanation could be that the OOmFP users found functionality that did not really exist (i.e., which should not be counted according to the FPA view on functional size) and that compensated for real but missed functionality. Although unlikely given the rigorous verification of the OOmFP rules performed before (see [20]), to rule out this alternative explanation requires an analysis of the details of the function point counts (looking for 'false positives' and 'false negatives' instead of just using the end results of the counting process). For our study, such a post-hoc analysis of the detailed FSM results was not possible because of insufficient data recording.

A final limitation confining the generalizability of any conclusions drawn from our experiment results is related to the experimental materials used. We tried to use a system that is representative of future uses of OOmFP. Considering the function point values obtained by the CFPSs, the system is rather small (purposely though, to manage the feasibility of an experiment with students). It is at this moment an open question whether our results scale up to larger applications.

6. Conclusions and future work

A study was presented that evaluated OOmFP, a new function point measurement procedure for systems development using the OO-Method method. The goal was to verify if and how much OOmFP improves the performance of function point measurement of object-oriented systems specified using OO-Method. The results indicate that, for novice users of functional size measurement methods, OOmFP is more time-consuming than the 'standard' IFPUG procedure for function point counting, but the measurement results are more reproducible and accurate. The results also indicate that OOmFP is perceived to be more useful and is more likely to be adopted in practice than IFPUG-FPA in the context of OO-Method systems development.

From a practical perspective, we are aware that this study provides only first evaluation results. We have outlined some critical factors and lessons learned from this study that must be handled in future experiments. An important limitation to the generalizability of the results was the use of students as participants. These students can at most be seen as being representative of novice users of functional size measurement methods. Nevertheless, this experiment has value as a pilot study for fixing the experimental procedure before trying it with industry practitioners. More experiments with a larger number of participants (e.g., students that are better trained in FPA and/or experienced FPA users) must be carried out to confirm the preliminary results obtained.

From a research perspective, our study adds new insights into the problem of how to evaluate alternative functional size measurement methods or measurement procedures. Our evaluation approach is original in the sense that it combines objective measures of user task performance with measures of user perceptions and intentions. The basis of this approach is provided by the Method Evaluation Model which was adapted to the functional size measurement context to assess the different factors that explain the user intention to use a particular method.

Apart from conducting replication studies, in future work we also plan to investigate the generality of OOmFP as a function point measurement procedure for object-oriented systems. Generality was a property suggested by March and Smith [22] that was not evaluated in this study as the application context was restricted to OO-Method. We wish to find out which of the OOmFP measurement rules are specific to OO-Method and which are not.

Acknowledgement

We thank Claudia Hazan, David Cleary, Richard Johnston, and John Ogilvie for their valuable contribution on the estimation of the requirements specification document.
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

1. I found the procedure for applying the method complex and difficult to follow
   O O O O O I found the procedure for applying the method simple and easy to follow

2. I believe that this method would reduce the time required to measure object-oriented systems.
   O O O O O I believe that this method would increase the time required to measure object-oriented systems.

3. Overall, I found the FSM method difficult to use
   O O O O O I found the measurement rules of the method confusing and difficult to understand

4. Overall, I found the FSM method to be useful
   O O O O O Overall, I did NOT find the FSM method to be useful

5. I found the FSM method difficult to learn
   O O O O O I found the FSM method easy to learn

6. I will use this method if I have to measure object-oriented systems in the future
   O O O O O I will NOT use this method if I have to measure object-oriented systems in the future

7. I think that this method would NOT improve the accuracy of estimates of object-oriented systems
   O O O O O I think that this method would improve the accuracy of estimates of object-oriented systems

8. I found it difficult to apply the FSM method to the case study
   O O O O O I found it easy to apply the FSM method to the case study

9. Overall, I think this method does NOT provide an effective way of measuring the functional size of OO systems during the requirements phase.
   O O O O O Overall, I think this method provides an effective way of measuring the functional size of OO systems during the requirements phase.

10. Using this method would improve my performance in measuring OO systems
    O O O O O Using this method would NOT improve my performance in measuring OO systems

11. I would be easy for me to become skilful in using this FSM method
    O O O O O I would be difficult for me to become skilful in using this FSM method

12. Overall, I think this method is an improvement to the FPA method
    O O O O O Overall, I think this method is NOT an improvement to the FPA method

13. I intend to use this FSM method in the future
    O O O O O I do NOT intend to use this FSM method in the future

Appendix B. Measurement results

See Appendix Tables B-1–B-4.

Table B-1
Measurement time and function point values, per treatment (FPA or OOmFP)

<table>
<thead>
<tr>
<th>Subject</th>
<th>First run procedure</th>
<th>Second run procedure</th>
<th>Measurement time in FPA (h)</th>
<th>Measurement time in OOmFP (h)</th>
<th>Size in FPA (FPS)</th>
<th>Size in OOmFP (FPs)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>FPA</td>
<td>OOmFP</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>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>2.25</td>
<td>2.95</td>
<td>115</td>
<td>172</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>2.50</td>
<td>2.00</td>
<td>124</td>
<td>155</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>2.75</td>
<td>2.85</td>
<td>143</td>
<td>151</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>2.37</td>
<td>4.15</td>
<td>128</td>
<td>160</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>2.33</td>
<td>4.23</td>
<td>114</td>
<td>158</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>2.37</td>
<td>2.50</td>
<td>118</td>
<td>148</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>2.40</td>
<td>2.25</td>
<td>145</td>
<td>160</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>2.50</td>
<td>2.00</td>
<td>159</td>
<td>165</td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>OOmFP</td>
<td>FPA</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>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>2.50</td>
<td>2.95</td>
<td>165</td>
<td>163</td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>2.54</td>
<td>3.40</td>
<td>174</td>
<td>171</td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>2.85</td>
<td>2.75</td>
<td>132</td>
<td>163</td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>2.25</td>
<td>2.25</td>
<td>114</td>
<td>158</td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>1.85</td>
<td>3.50</td>
<td>147</td>
<td>173</td>
<td></td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>2.45</td>
<td>3.15</td>
<td>121</td>
<td>148</td>
<td></td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>2.15</td>
<td>4.00</td>
<td>125</td>
<td>163</td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>2.00</td>
<td>3.33</td>
<td>127</td>
<td>154</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table B-2
Reproducibility (REPi) and accuracy (MREi) scores, per treatment (FPA or OOmFP)

<table>
<thead>
<tr>
<th>Subject</th>
<th>REPi in FPA</th>
<th>REPi in OOmFP</th>
<th>MREi in FPA</th>
<th>MREi in OOmFP</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.03</td>
<td>0.06</td>
<td>0.18</td>
<td>0.04</td>
</tr>
<tr>
<td>2</td>
<td>0.06</td>
<td>0.01</td>
<td>0.20</td>
<td>0.01</td>
</tr>
<tr>
<td>3</td>
<td>0.14</td>
<td>0.08</td>
<td>0.27</td>
<td>0.10</td>
</tr>
<tr>
<td>4</td>
<td>0.07</td>
<td>0.03</td>
<td>0.21</td>
<td>0.01</td>
</tr>
<tr>
<td>5</td>
<td>0.08</td>
<td>0.06</td>
<td>0.09</td>
<td>0.04</td>
</tr>
<tr>
<td>6</td>
<td>0.03</td>
<td>0.030</td>
<td>0.18</td>
<td>0.02</td>
</tr>
<tr>
<td>7</td>
<td>0.15</td>
<td>0.01</td>
<td>0.27</td>
<td>0.01</td>
</tr>
<tr>
<td>8</td>
<td>0.11</td>
<td>0.08</td>
<td>0.25</td>
<td>0.06</td>
</tr>
<tr>
<td>9</td>
<td>0.10</td>
<td>0.100</td>
<td>0.08</td>
<td>0.02</td>
</tr>
<tr>
<td>10</td>
<td>0.21</td>
<td>0.04</td>
<td>0.01</td>
<td>0.05</td>
</tr>
<tr>
<td>11</td>
<td>0.09</td>
<td>0.06</td>
<td>0.23</td>
<td>0.04</td>
</tr>
<tr>
<td>12</td>
<td>0.08</td>
<td>0.07</td>
<td>0.22</td>
<td>0.08</td>
</tr>
<tr>
<td>13</td>
<td>0.26</td>
<td>0.02</td>
<td>0.05</td>
<td>0.04</td>
</tr>
<tr>
<td>14</td>
<td>0.34</td>
<td>0.08</td>
<td>0.11</td>
<td>0.09</td>
</tr>
<tr>
<td>15</td>
<td>0.00</td>
<td>0.02</td>
<td>0.16</td>
<td>0.04</td>
</tr>
<tr>
<td>16</td>
<td>0.15</td>
<td>0.01</td>
<td>0.27</td>
<td>0.01</td>
</tr>
<tr>
<td>17</td>
<td>0.12</td>
<td>0.09</td>
<td>0.06</td>
<td>0.10</td>
</tr>
<tr>
<td>18</td>
<td>0.09</td>
<td>0.08</td>
<td>0.23</td>
<td>0.06</td>
</tr>
<tr>
<td>19</td>
<td>0.06</td>
<td>0.02</td>
<td>0.20</td>
<td>0.04</td>
</tr>
<tr>
<td>20</td>
<td>0.04</td>
<td>0.04</td>
<td>0.19</td>
<td>0.02</td>
</tr>
</tbody>
</table>

Table B-3
Correlation between survey items (construct validity analysis)

<table>
<thead>
<tr>
<th>Perceived ease of use</th>
<th>Perceived usefulness</th>
<th>Intention to Use</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q1</td>
<td>Q2</td>
<td>Q3</td>
<td>Q4</td>
</tr>
<tr>
<td>Q5</td>
<td>Q6</td>
<td>Q7</td>
<td>Q8</td>
</tr>
<tr>
<td>Q9</td>
<td>Q10</td>
<td>Q11</td>
<td>Q12</td>
</tr>
<tr>
<td>Q13</td>
<td>Q14</td>
<td>CV</td>
<td>DY</td>
</tr>
<tr>
<td>Valid?</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table B-4
PEOU, PU and ITU scores, per treatment (FPA or OOmFP)

<table>
<thead>
<tr>
<th>Subject</th>
<th>PEOU in FPA</th>
<th>PEOU in OOmFP</th>
<th>PUin FPA</th>
<th>PUin OOmFP</th>
<th>ITU in FPA</th>
<th>ITU in OOmFP</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3.60</td>
<td>3.20</td>
<td>2.60</td>
<td>4.00</td>
<td>3.67</td>
<td>4.33</td>
</tr>
<tr>
<td>2</td>
<td>3.20</td>
<td>2.60</td>
<td>2.20</td>
<td>3.20</td>
<td>1.67</td>
<td>3.67</td>
</tr>
<tr>
<td>3</td>
<td>3.20</td>
<td>2.60</td>
<td>1.80</td>
<td>3.20</td>
<td>1.67</td>
<td>3.33</td>
</tr>
<tr>
<td>4</td>
<td>3.00</td>
<td>3.40</td>
<td>2.00</td>
<td>4.60</td>
<td>1.33</td>
<td>4.67</td>
</tr>
<tr>
<td>5</td>
<td>2.80</td>
<td>2.60</td>
<td>2.80</td>
<td>3.40</td>
<td>3.00</td>
<td>4.33</td>
</tr>
<tr>
<td>6</td>
<td>2.40</td>
<td>2.00</td>
<td>2.20</td>
<td>3.00</td>
<td>2.33</td>
<td>4.00</td>
</tr>
<tr>
<td>7</td>
<td>3.40</td>
<td>1.60</td>
<td>3.00</td>
<td>3.40</td>
<td>3.67</td>
<td>3.67</td>
</tr>
<tr>
<td>8</td>
<td>2.60</td>
<td>3.40</td>
<td>2.20</td>
<td>3.20</td>
<td>2.33</td>
<td>4.00</td>
</tr>
<tr>
<td>9</td>
<td>2.40</td>
<td>3.80</td>
<td>2.60</td>
<td>3.60</td>
<td>2.67</td>
<td>3.33</td>
</tr>
<tr>
<td>10</td>
<td>2.80</td>
<td>4.40</td>
<td>3.60</td>
<td>4.40</td>
<td>3.67</td>
<td>3.67</td>
</tr>
<tr>
<td>11</td>
<td>1.80</td>
<td>2.80</td>
<td>2.60</td>
<td>3.40</td>
<td>2.33</td>
<td>3.33</td>
</tr>
<tr>
<td>12</td>
<td>2.20</td>
<td>2.60</td>
<td>2.80</td>
<td>3.20</td>
<td>3.33</td>
<td>4.33</td>
</tr>
<tr>
<td>13</td>
<td>2.80</td>
<td>2.60</td>
<td>2.80</td>
<td>2.80</td>
<td>2.33</td>
<td>3.33</td>
</tr>
<tr>
<td>14</td>
<td>2.60</td>
<td>2.40</td>
<td>2.00</td>
<td>3.00</td>
<td>2.33</td>
<td>3.33</td>
</tr>
<tr>
<td>15</td>
<td>2.20</td>
<td>2.60</td>
<td>2.80</td>
<td>3.60</td>
<td>2.00</td>
<td>4.00</td>
</tr>
<tr>
<td>16</td>
<td>2.40</td>
<td>3.80</td>
<td>2.40</td>
<td>3.60</td>
<td>2.67</td>
<td>3.33</td>
</tr>
<tr>
<td>17</td>
<td>3.20</td>
<td>3.40</td>
<td>3.40</td>
<td>3.80</td>
<td>3.00</td>
<td>4.33</td>
</tr>
<tr>
<td>18</td>
<td>2.60</td>
<td>2.60</td>
<td>2.40</td>
<td>3.80</td>
<td>3.00</td>
<td>3.67</td>
</tr>
<tr>
<td>19</td>
<td>2.60</td>
<td>2.0</td>
<td>2.80</td>
<td>2.80</td>
<td>2.67</td>
<td>3.00</td>
</tr>
<tr>
<td>20</td>
<td>2.40</td>
<td>3.60</td>
<td>1.60</td>
<td>4.00</td>
<td>267</td>
<td>4.67</td>
</tr>
</tbody>
</table>
References


[36] F.D. Davis, Perceived usefulness, perceived ease of use and user acceptance of information technology, MIS Quarterly 13 (1989).


