# Three Empirical Studies on Estimating the Design Effort of Web Applications

LUCIANO BARESI
Politecnico di Milano
and
SANDRO MORASCA
Università degli Studi dell'Insubria

15

Our research focuses on the effort needed for designing modern Web applications. The design effort is an important part of the total development effort, since the implementation can be partially automated by tools.

We carried out three empirical studies with students of advanced university classes enrolled in engineering and communication sciences curricula. The empirical studies are based on the use of W2000, a special-purpose design notation for the design of Web applications, but the hypotheses and results may apply to a wider class of modeling notations (e.g., OOHDM, WebML, or UWE). We started by investigating the relative importance of each design activity. We then assessed the accuracy of *a priori* design effort predictions and the influence of a few process-related factors on the effort needed for each design activity. We also analyzed the impact of attributes like the size and complexity of W2000 design artifacts on the total effort needed to design the user experience of web applications. In addition, we carried out a finer-grain analysis, by studying which of these attributes impact the effort devoted to the steps of the design phase that are followed when using W2000.

Categories and Subject Descriptors: D.2.8 [**Software Engineering**]: Metrics—*Product metrics*; G.3 [**Probability and Statistics**]—*Robust regression*; H.5.4 [**Information Interfaces and Presentation**]: Hypertext/Hypermedia

General Terms: Economics, Experimentation, Measurement

Additional Key Words and Phrases: Web application design, W2000, empirical study, effort estimation

This article was accepted by David Rosenblum.
The research presented in this article was partially funded by the IST project *QualiPSo*, sponsored by the EU in the 6th FP (IST-034763); the FIRB project *ARTDECO*, sponsored by the Italian Ministry of Education and University; and the project *La qualità nello sviluppo software*, sponsored by the Unversità degli Studi dell'Insubria.
Authors' addresses: L. Baresi, Dipartmento di Elettronica e Informazione, Politecnico di Milano, piazza Leonardo da Vinci, 32, I-20133 Milano (Italy); email: baresi@elet.polimi.it; S. Morasca, Dipartmento di Scienze della Cultura, Politiche e dell'Informazione, Università degli Studi dell'Insubria, via Valleggio, 11, I-22100 Como (Italy); email: sandro.morasca@uninsubria.it.

**ACM Reference Format:**
Baresi, L. and Morasca, S. 2007. Three empirical studies on estimating the design effort of web applications. ACM Trans. Softw. Eng. Methodol. 16, 4, Article 15 (September 2007), 40 pages. DOI = 10.1145/1276933.1276936 http://doi.acm.org/10.1145/1276933.1276936

## 1. INTRODUCTION

The Web and the applications that run on it have changed our everyday lives. Purchases, reservations, activity planning, etc. are now routinely carried out over the Web, and several companies view the Web as a key resource to run their businesses.

Modern Web applications are no longer mere hypertextual information repositories. They are complex distributed systems that use the Web as the interaction means and the Internet as the communication infrastructure. Pages do not only carry possibly dynamic information that the user can browse, but they model the *user experience*[1] for the services supplied by the application. Therefore, input parameters, user selections, and returned values must organized in a consistent manner in the Web pages browsed by the users.

Practical experience on Web development shows significant differences between Web applications and traditional software applications [Ginige and Murugesan 2001; Mendes and Mosley 2006; Kappel et al. 2006]. Some experts (e.g., Mendes and Mosley [2006]) view it as a new discipline, while others (e.g., Kappel et al.) describe it as a new application domain–or independent branch–of software engineering. They all agree on saying that Web applications are not just yet another example of conventional distributed systems, but their development has some distinctive characteristics. The dynamism that belongs to modern Web applications requires that suitable modeling–and implementation–techniques be conceived.

Due to the complexity of current Web applications, the design phase is a critical step of the development process. If the design phase of a Web application is carried out carefully, the subsequent phases may require a considerably lower amount of time and resources than the design phase itself, and can be automated by supporting tools, to some extent. This is the case of the many existing model-driven approaches (e.g., WebML [Ceri et al. 2000], UWE [Knapp et al. 2004], and OOHDM [Schwabe and Rossi 1998]), which require the writing of precise, unambiguous models to automatically generate many parts of the Web-based front-end of the application.

The user experience of Web applications plays a key role for their acceptability, and thus for their success. The back-end of these applications is often based on predefined frameworks, such as Sun J2EE, Microsoft .NET, and Web services-based infrastructures, while Web interfaces need to be designed

[1]The *user experience* describes the overall experience and satisfaction a user has when using a product or system, and commonly refers to a combination of software and business topics. The design of the user experience defines a sequence of screen presentations, user interactions, and system responses that meet user goals and tasks while satisfying business and functional requirements [Conallen 2002; Garrett 2002].

explicitly. However, designing the user experience goes well beyond designing the user interface of a Web application. It shapes the way the user navigates in a Web application and interacts with it, thus deeply affecting the working habits of the user. Therefore, the user experience must be carefully designed for a Web application to be successful, so the design phase takes an even higher proportion of the entire development process than in other types of applications.

Given the relevance of the design phase in the development of Web applications, the ability of accurately estimating the design effort may provide a Web development organization with a competitive advantage, as the design effort is tightly related to the design cost. More generally, there is a need for *Web estimation*, that is, the estimation of the costs of the various development phases of Web applications, as a few recent publications [Mendes et al. 2005a; 2002a; Reifer 2000; Murugesan and Deshpande 2002; Mendes and Mosley 2006] have stressed.

Thus, the final goal of the research documented in this article is to assess the cost of the design phase of Web applications in a quantitative way and build estimation models for it. To this end, we carried out three empirical studies[2] on the effort required for designing the user experience of modern Web applications.

These studies were carried out during advanced university classes on modeling Web applications with students as subjects. Two of these empirical studies were carried out at the school of computer science of the Politecnico di Milano, which is one of the most important engineering school in Italy, and one empirical study was carried out at the school of communication sciences of the Università della Svizzera Italiana, which is a university that has been recently founded in the Italian-speaking part of Switzerland. Having students from schools related to two different disciplines has also allowed us to understand and evaluate if, how, and to what extent the students' backgrounds impact their perception of the modeling task. In the empirical studies, the subjects used W2000 [Baresi et al. 2006], a special-purpose notation for the design of Web applications, but the hypotheses and results may apply to a wider class of modeling notations (e.g., OOHDM, WebML, or UWE), which are based on similar underlying concepts, and can also be viewed as special-purpose customizations (i.e., profiles) of UML.

This article extends and complements the work already presented in two other papers, which were based on the results from the first of the three empirical studies we describe here. In previous research, we studied the effort required to model Web applications with W2000 [Baresi et al. 2002], and the relationship between internal measures of Web applications and effort [Baresi et al. 2003].

[2]In this article, we use the general term "empirical study" instead of "experiment" because our studies, like the vast majority of quantitative studies in Software Engineering, cannot be technically classified as "experiments," which would require, for instance, the randomization of the treatments to the subjects, or being able to control for some specific factors that may influence the results. Our empirical studies can be classified as "case studies" or "correlational studies." Our choice also agrees with the use of "Empirical Software Engineering" as the term currently used to refer to the part of Software Engineering that addresses these topics.

Since Web estimation is a relatively new field, our empirical studies were exploratory, rather than confirmatory ones. At any rate, we have carried out an initial study and two replications of it, to provide more evidence for the hypotheses that were supported in the initial study. We investigated a number of hypotheses that seemed likely to be true based on our beliefs and knowledge about the W2000 notation, the subjects' skills, and the steps of the design process used. In this article, we report on

—hypotheses that were supported by empirical evidence in the context of our studies and deserve further investigations;
—hypotheses that were not supported by empirical evidence, and should therefore be either revised (there might have been some reasons why these hypotheses were not true in our experimental setting), or not investigated any longer (they are not likely to be supported by evidence in other contexts either).

We believe that both categories of hypotheses, whether supported or unsupported, may contribute to increasing the knowledge in a field where there is little empirical, quantitative evidence. We used parametric and nonparametric (distribution-free) statistical techniques to confirm or refute hypotheses. We carried out a careful outlier analysis prior to using parametric techniques and we used nonparametric techniques to provide further support to the results obtained with parametric techniques because the number of data points in each empirical study was somewhat limited and we wanted to avoid results due to a small number of over-influential data points that may not have been completely filtered out by our outlier analysis. In addition to nonparametric techniques that have already been used in Empirical Software Engineering, we have used a less conventional technique, Robust Regression [Rousseeuw and Leroy 1987], that allowed us to build linear models in a more robust way than those obtained via Ordinary Least Squares Regression.

Here, we summarize our main results, which we will illustrate in detail in this article.

—*Identification of the most time-consuming design activity.* The design of the Web applications' data and their inter-relationships appears to take the largest effort for the engineering students, while the design of the presentation takes the largest effort for the communication sciences students.
—*Evaluation of the ability of the subjects to predict the effort needed to develop the W2000 artifacts.* The studies show an underestimation problem, with different degrees of severity, that is, the students appear to be optimistic in general. At any rate, the effort estimated by the students turned out to be a predictor for the actual effort. In addition, the estimated effort for developing each W2000 artifact is a good predictor for the actual development effort for that artifact. The same applies to the estimated and the actual learning efforts.
—*Identification of internal characteristics of W2000 artifacts that may influence the actual effort used to produce them*. We identified a number of internal characteristics of W2000 artifacts that were correlated with the estimated

and actual effort, but not all the results were consistent across the studies, so their practical usefulness still needs to be confirmed in further studies.

The remainder of this article is organized as follows. To make the article self-contained, Section 2 concisely introduces W2000 and its modeling features. Section 3 describes the settings of our empirical studies. Section 4 surveys the data analysis techniques we used and provides the terminology to interpret the statistical results. Section 5 presents our hypotheses and the specific results we obtained. Section 6 discusses the validity of our empirical studies, Section 7 provides an organized view of the lessons learned, Section 8 surveys the related work, and Section 9 summarizes the article and outlines directions for future work.

## 2. W2000

W2000 is an evolution and an extension of HDM (Hypertext Design Model [Garzotto et al. 1993]), which is one of the first modeling notations for Web applications and, as such, is behind some other proposals in the field. In this section, we concisely introduce the main features of W2000 that are relevant for our studies, with the aid of a few models for a Web-based conference management system. Interested readers can refer to Baresi et al. [2006] for an in-depth presentation. W2000 applies the *model-view-controller* paradigm to the design of Web applications, and it is based on four models: the *information model*, the *navigation model*, the *presentation model*, and the *operation model*. However, the *operation model*, used to define the operations provided by the application, was not part of our empirical studies, so we do not present it here.

### 2.1 Information Model

Conceptually, the starting point of a W2000 design is the *information model*, whose goals are the identification and organization of all the data with which the application deals.

The information model identifies the *entities* that characterize the application, that is, the conceptual "objects" that are of interest for the user. *Components* are used to structure the contents of *entities* hierarchically into meaningful chunks. Leaf components contain *slots*, which are typed attributes that identify the primitive information elements.

*Segments* group sets of slots as "macros" that can be reused as single elements. Figure 1 shows that all Paper entities have three components: all papers must have an Abstract and a first Submission, and they might have a CameraReady.

The relationship between Paper and Abstract is an aggregation in UML terms.

Figure 2 shows the slots of Abstract components. Types are defined only informally, but we need to distinguish among: slots with primitive types, like Number, Title, MainTopic, and SubmissionCategory, and slots–like Author–that are compound and whose structures are represented by the subtrees drawn below the slot names. Title, Abstract, and MainTopic may be grouped into a segment, called PaperInfo, to define a concise and atomic way to characterize submitted papers. Slots play the role of attributes in UML class diagrams.

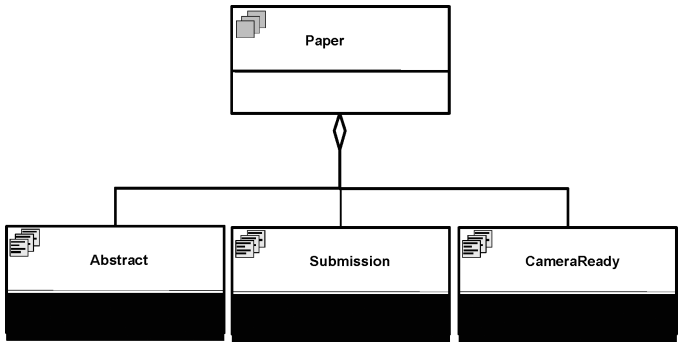


Fig. 1. Component tree for entity Paper.

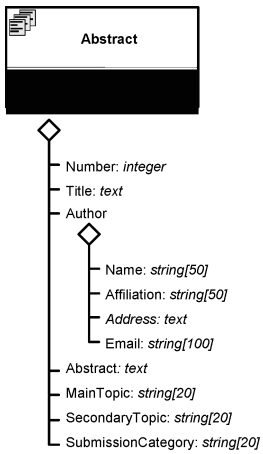


Fig. 2. Slots for Abstract components.

*Semantic associations* link pairs of *entities* and identify navigational paths between two related concepts. *Association centers* describe how to identify either the entire set of target instances as a whole or each individual element in the set. Figure 3 presents a semantic association between Author and Paper and two semantic association centers to allow the information flow between the two entities. Semantic associations are similar to standard UML associations, while their centers correspond to association classes.

*Collections* organize the information defined so far and group entities to ease navigation: the user can enter the whole collection and then move among its elements. Also, *collections* can have *centers*. For example, a collection like AllPapers groups all the papers submitted to the conference and does not depend on the user (e.g., a reviewer). Instead, a collection like PapersToReview is a "template" that defines a way to filter submitted papers according to the reviewers they are assigned to. Collections and entities are implicitly linked by means of UML aggregations. For instance, Figure 4 shows collection PapersByAuthor, which is a special-purpose way to access the papers of a particular

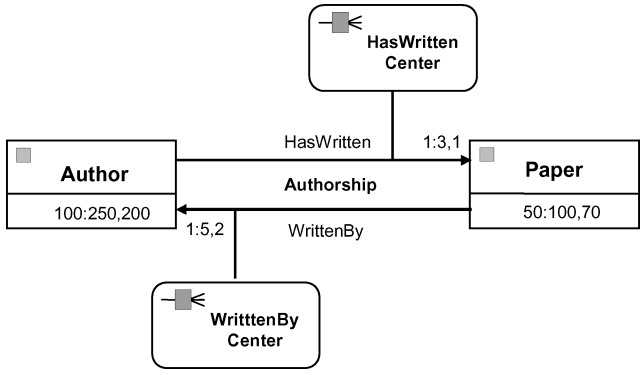


Fig. 3. Semantic association Authorship.

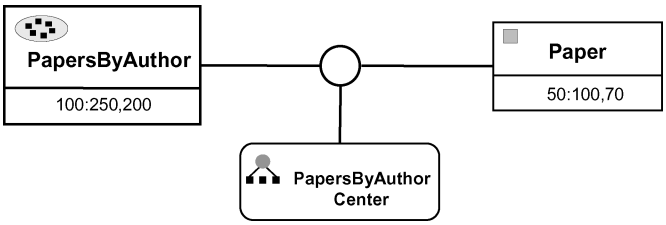


Fig. 4. Collection PapersByAuthor.

author as a whole. Again, Collection centers can be seen as UML association classes.

## 2.2 Navigation Model

The *navigation model* defines how the user can browse through the application. *Nodes* define atomic information units and reshape the elements in the information model to specify the *actual* information chunks. *Clusters* link together sets of *nodes* and define how the user can move around these elements. *Clusters* can be further organized in: *structural clusters* if all of their elements come from the same *entity*; *association clusters* if they render associations; *collection clusters* if they describe the topology of a collection; *transactional clusters* if they relate the set of nodes that the user should traverse to complete a business transaction.

Clusters and nodes are derived from the information model through a set of rules and design heuristics. For example, in the simplest case, we could imagine that each entity is mapped onto a structural cluster, whose nodes correspond to the leaf components of the entity (Figure 5(a)). We also assume that each node is reachable from all the other nodes in the cluster. However, this assumption does not hold if we need to reorganize the contents with a finer granularity. If the user moves to a PDA, the designer might prefer to split the information about papers into two nodes (Figure 5(b)): node Introduction contains information about the author and the main topics of the paper and node Abstract contains the abstract of the paper.

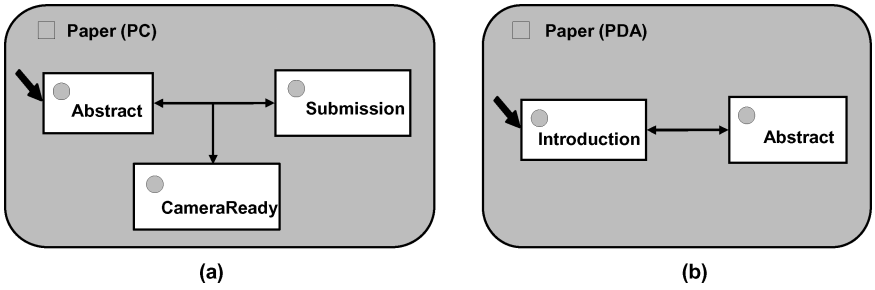


Fig. 5. Structural cluster Paper.

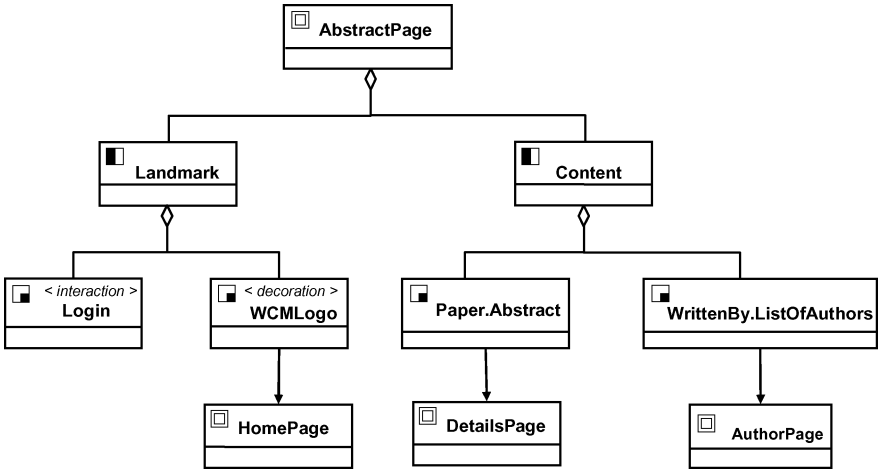


Fig. 6. Page Abstract.

## 2.3 Presentation Model

The *presentation model* organizes the application in a set of pages along with links among them. This model may be very similar to the *navigation model* if each *node* is rendered through a single *page*. However, pages are usually composed of *publishing units*, which usually render *nodes*, but can also be used to define forms, navigable elements, and labels. *Sections* group related *publishing units* to better structure a page and improve the degree of reuse of page fragments. *Pages* conceptually identify the screens as perceived by the user. *Links* connect pages and identify the actual navigation capabilities offered to users. *Links* can also "hide" the enabling of computations (i.e., the services provided by the back-end of the Web application).

Figure 6 shows the definition of page Abstract (a double square in the upper left corner denotes a page). Section Landmark (a half-filled square in the upper left corner denotes a section) contains the interaction unit Login (a quarter-filled square denotes a unit), which allows users to log in the application, and the decorator unit WCMLogo, which contains the conference logo and allows users to navigate to the HomePage. Section Content, on the other hand, embeds the content unit Paper.Abstract (derived from node Abstract of cluster Paper) and the

content WrittenBy.ListOfAuthors (derived from node ListOfAuthors of association center WrittenBy).

## 3. SETTINGS OF THE EMPIRICAL STUDIES AND DATA COLLECTION

We carried out two empirical studies at the Politecnico di Milano in Milan, Italy, with students attending an advanced class on Web application design during the fourth or fifth year of their curriculum: we will refer to these empirical studies as **PdM1** and **PdM2** in what follows. We also carried out one empirical study at the Università della Svizzera Italiana, in Lugano, Switzerland, with students attending an advanced class on Web application design during the second year of their curriculum: we will refer to this empirical study as **USI** in what follows.

The main difference between the students at the Politecnico di Milano and the ones at the Università della Svizzera Italiana is that the former are enrolled in a computer engineering curriculum, while the latter in a communication sciences curriculum. Web application designers may come from either background, and actually an increasing number of them now come from nontechnical schools [Conallen 2002].

Also, the students at the Politecnico di Milano were asked to develop their designs individually, while the students at the Università della Svizzera Italiana were grouped in teams composed of three people.

In all three empirical studies, the students were taught how to design Web applications with W2000 and implement them with the many technologies available. The empirical studies started by supplying the students with a concise and informal set of requirements for the application they were asked to model. All students were required to work on the same topic: a hypothetical e-commerce application, but each project had a different application domain: books, CDs, groceries, etc. So, all projects had very similar difficulties, and the differences across projects were only marginal. At the time the empirical studies were carried out, W2000 was not supported by any tool, so the students had to develop their designs by hand.

Along with the requirements, the students were given two questionnaires: an initial questionnaire that they were required to fill out before starting to design their applications and a final questionnaire that they had to fill out after completing their work. We made it clear to the students that the data they provided in the questionnaires would have no influence on the grade they would receive.

We designed the questionnaires to collect data on measures related to the empirical hypotheses that we wanted to investigate with our studies (see Section 5).

With the initial questionnaire, we wanted to

—make the subjects estimate the design effort required to model their applications, split according to the main W2000 models;

—measure the subject's general proficiency in computer science and Web technologies: as measures, we used the number of exams passed and the average grade obtained in those exams.

Table I. Effort Measures

| Actual | Estimated | Activities |
|---|---|---|
| *ActEff* | *EstEff* | All models |
| *ActEff_* | *EstEff_* | Information and presentation models |
| *ActInfoEff* | *EstInfoEff* | Information model |
| *ActNavEff* | *EstNavEff* | Navigation model |
| *ActPresEff* | *EstPresEff* | Presentation model |
| *ActLearnEff* | *EstLearnEff* | Learning for all kinds of models |
| *ActLearnEff_* | *EstLearnEff_* | Learning for information and presentation models |

With the second questionnaire, we wanted to

—make the subjects provide a self-evaluation of their designs: we used the Italian university grading system, in which exams are graded with a mark between 0 and 30 (a student passes an exam if his or her grade is at least 18)

—collect the actual effort used by the students to design their applications, split according to the main W2000 models.

It was recommended that the students collect effort data on a daily basis and then summarize the effort on the second questionnaire.

We collected data on the above measures in all three empirical studies. However, in **PdM1**, the estimated and actual effort data we managed to collect were only related to the W2000 Information and Navigation models. Too few subjects provided values for the effort related to the other W2000 models. The effort measures are described in Table I, where columns **Actual**, **Estimated**, and **Activities** describe the variables for the actual, estimated efforts, and the activities to which they are related, respectively.

We also collected measures (e.g., the number of entities, the number of navigation slots) to quantify the internal characteristics (e.g., size, structural complexity) of the W2000 design models developed by the students. These measures, derived from the empirical hypotheses we wanted to investigate (see Section 5), are listed in Table VI (see Section 5.5), organized by the W2000 model on which they are measured and the attribute they purport to measure. As a final remark on data collection, we must say that these measures on the internal characteristics W2000 models are available for **PdM1** and **PdM2**, but not for **USI**.

## 4. DATA ANALYSIS TECHNIQUES

Before describing the hypotheses and the results of our empirical studies (described in detail in Tables II, III, and IV), we now report on the specific statistics, data analysis techniques, and statistical tests we have used. Throughout this article, we have consistently used a threshold value 0.05, often used in the Empirical Software Engineering literature to decide whether the results of a statistical test of hypothesis are statistically significant. Thus, we reject the "null hypothesis" of a statistical test if $p \leq 0.05$, where $p$ denotes the statistical significance of the test, and we accept the "alternative hypothesis," that is, the hypothesis we wanted to check. Thus, a hypothesis is supported by the empirical evidence if we find a "small enough" value of $p$.

Table II. Median Effort Comparisons. The three sections of the table contain the results for PdM1, PdM2, and USI, respectively

| Actual vs. Estimated ($p-$value) | | Actual vs. Actual ($p-$value) | |
|---|---|---|---|
| *ActEff*$_-$ > *EstEff*$_-$ | <0.0001 | *ActInfoEff* > *ActNavEff* | <0.0001 |
| *ActInfoEff* > *EstInfoEff* | <0.0001 | *ActInfoEff* > *ActLearnEff* | <0.0001 |
| *ActNavEff* > *EstNavEff* | <0.0001 | *ActNavEff* > *ActLearnEff* | 0.0004 |
| *ActLearnEff* > *EstLearnEff* | 0.1402 | | |
| *ActEff*$_-$ > *EstEff*$_-$ | 0.0793 | *ActInfoEff* > *ActNavEff* | 0.0001 |
| *ActEff* > *EstEff* | 0.0444 | *ActInfoEff* > *ActPresEff* | 0.0005 |
| *ActInfoEff* > *EstInfoEff* | 0.1518 | *ActInfoEff* > *ActLearnEff* | 0.0005 |
| *ActNavEff* > *EstNavEff* | 0.0301 | *ActPresEff* > *ActNavEff* | 0.0998 |
| *ActPresEff* > *EstPresEff* | 0.0192 | *ActLearnEff* > *ActNavEff* | <0.2101 |
| *ActLearnEff* > *EstLearnEff* | 0.1335 | *ActLearnEff* > *ActPresEff* | <0.4098 |
| *ActEff*$_-$ > *EstEff*$_-$ | <0.0001 | *ActNavEff* > *ActInfoEff* | 0.1345 |
| *ActEff* > *EstEff* | <0.0001 | *ActPresEff* > *ActInfoEff* | 0.0070 |
| *ActInfoEff* > *EstInfoEff* | <0.0001 | *ActPresEff* > *ActNavEff* | 0.0159 |
| *ActNavEff* > *EstNavEff* | <0.0001 | *ActLearnEff* > *ActInfoEff* | <0.0001 |
| *ActPresEff* > *EstPresEff* | <0.0001 | *ActLearnEff* > *ActNavEff* | <0.0001 |
| *ActLearnEff* > *EstLearnEff* | 0.0016 | *ActLearnEff* > *ActPresEff* | <0.0001 |

Since the measures we collected were on the ratio level of measurement, we used parametric and nonparametric data analysis techniques, depending on the kind of statistical hypotheses we wanted to test. In addition to data analysis techniques that have been traditionally used in Empirical Software Engineering, we also used the Robust Regression (RR) technique [Rousseeuw and Leroy 1987], for which we provide a more detailed description in Section 4.2.1.

### 4.1 Descriptive Statistics and Related Tests

Descriptive statistics for sample distributions provide a concise idea of what the data look like and can be used for future comparisons. We use the example of the distribution of *ActEff*$_-$ of **PdM1** in Table V to illustrate the descriptive statistics we used. All effort data are expressed in person-hours. In this table, the learning effort is actually the sum of the learning efforts for each W2000 model.

—$N$, the number of data points of the sample. For instance, the number $N$ of data points of the sample of *ActEff*$_-$ is 49.

—central tendency statistics: $M$, $m$, and $LMS$, which denote the median, the mean, and the Least Median of Squares of the distribution, which is derived from the Robust Regression data analysis technique, which we describe in Section 4.2.1. For instance, the values of $M$, $m$, and $LMS$ of the distribution of *ActEff*$_-$ are 26.00, 28.79, and 18.50 person-hours, respectively.

—*Dispersion Statistics:* $\sigma$, the unbiased estimator of the standard deviation of the distribution (i.e., $\sigma = \sqrt{\sum_{i \in 1..N}(x_i - \overline{x})^2/(N-1)}$ for the distribution of a variable $X$ with $N$ data points $x_i$ and sample mean $\overline{x}$), and $\sigma_{LMS}$, a dispersion indicator of Robust Regression. For instance, the values of $\sigma$ and $\sigma_{LMS}$ of the distribution of *ActEff*$_-$ are 15.23 and 12.87, respectively.

Table III. Effort Data Analyses (OLS Results).
The three sections of the table contain the results for PdM1, PdM2, and USI, respectively

| Dependent Variable | Independent Variable | Ordinary Least Squares (OLS) | | | | | |
|---|---|---|---|---|---|---|---|
| | | $c_1$ | $c_0$ | $p_{c1}$ | $p_{c0}$ | $R^2$ | $N_{OLS}$ |
| *ActEff_* | *EstEff_* | 1.51 | 2.31 | <.0001 | 0.5813 | 0.54 | 39 |
| | *EstLearnEff_* | 2.62 | 10.83 | <.0001 | 0.0103 | 0.36 | 38 |
| | *Component* | 2.31 | 4.06 | 0.0011 | 0.56 | 0.31 | 31 |
| | *Segment* | −3.52 | 30.6 | 0.0014 | 0.0001 | 0.21 | 28 |
| | *NavSlot* | 0.31 | 18.3 | 0.0053 | 0.0001 | 0.22 | 33 |
| *ActInfoEff* | *EstInfoEff* | 1.10 | 6.24 | 0.0008 | 0.059 | 0.27 | 39 |
| | *Component* | 1.27 | 2.98 | 0.0007 | 0.4148 | 0.35 | 29 |
| | *Entity* | −1.42 | 24.02 | 0.0851 | <.0001 | 0.10 | 30 |
| | *Segment* | 2.08 | 18.13 | 0.0113 | <.0001 | 0.23 | 27 |
| | *SlotsPerSACenter* | 2.07 | 12.87 | 0.0668 | <.0001 | 0.10 | 33 |
| *ActNavEff* | *EstNavEff* | 1.03 | 3.22 | 0.0006 | 0.0877 | 0.27 | 40 |
| | *NLink* | 0.12 | 4.53 | 0.0367 | 0.0278 | 0.14 | 32 |
| | *SlotsPerNode* | 1.65 | 7.73 | 0.0741 | 0.0001 | 0.09 | 35 |
| *ActLearnEff* | *EstLearnEff* | 0.76 | 1.66 | <.0001 | 0.1082 | 0.41 | 39 |
| *ActEff_* | *EstEff_* | 0.86 | 8 | 0.0054 | 0.313 | 0.33 | 22 |
| | *EstLearnEff_* | −0.06 | 30.52 | 0.9522 | 0.0017 | 0.00 | 22 |
| | *EstLearnEff* | −0.59 | 36.86 | 0.365 | 0.0003 | 0.04 | 23 |
| | *Entity* | 2.65 | 11.38 | 0.05 | 0.24 | 0.17 | 23 |
| | *SACenter* | 2.8 | 18.09 | 0.0041 | 0.001 | 0.34 | 22 |
| | *SlotsPerSACenter* | 5.42 | 17.71 | 0.0315 | 0.0077 | 0.2 | 23 |
| | *SlotsPerNode* | 8.91 | 5.41 | 0.0111 | 0.6029 | 0.27 | 23 |
| *ActEff* | *EstEff* | 0.71 | 15.30 | 0.0137 | 0.1616 | 0.27 | 22 |
| | *EstLearnEff* | −0.90 | 53.91 | 0.3388 | 0.0003 | 0.04 | 23 |
| | *Entity* | 4.54 | 11.02 | 0.0195 | 0.4124 | 0.23 | 23 |
| | *SACenter* | 5.23 | 18.40 | 0.0004 | 0.0154 | 0.48 | 21 |
| | *SlotsPerSACenter* | 8.39 | 23.83 | 0.0216 | 0.0117 | 0.23 | 23 |
| | *SlotsPerNode* | 16.45 | −2.16 | 0.0018 | 0.8892 | 0.36 | 24 |
| | *CollCenter* | −3.88 | 66.27 | 0.0199 | 0 | 0.24 | 22 |
| | *Slot* | 0.93 | 4.15 | 0.0056 | 0.797 | 0.30 | 24 |
| | *NLink* | 0.35 | 24.91 | 0.0606 | 0.0175 | 0.17 | 21 |
| | *NCluster* | 1.93 | 9.92 | 0.0081 | 0.4329 | 0.30 | 22 |
| *ActInfoEff* | *EstInfoEff* | 0.72 | 6.64 | <.0001 | 0.0429 | 0.52 | 24 |
| | *Component* | 0.70 | 8.45 | 0.0231 | 0.0846 | 0.23 | 22 |
| | *Entity* | 1.51 | 8.54 | 0.0628 | 0.1464 | 0.16 | 22 |
| *ActNavEff* | *EstNavEff* | 1.37 | −2.12 | <.0001 | 0.313 | 0.79 | 26 |
| | *SlotsPerNode* | 3.77 | 1.89 | 0.0124 | 0.6864 | 0.26 | 23 |
| *ActPresEff* | *EstPresEff* | 1.03 | 0.40 | 0.0005 | 0.8851 | 0.47 | 22 |
| | *SACenter* | 3.23 | −0.37 | 0.0001 | 0.923 | 0.51 | 23 |
| | *SlotsPerNode* | 3.30 | 3.19 | 0.0577 | 0.5505 | 0.16 | 23 |
| | *SlotsPerSACenter* | 4.40 | 2.54 | 0.0029 | 0.4666 | 0.35 | 23 |
| *ActLearnEff* | *EstLearnEff* | 0.73 | 4.26 | 0.0002 | 0.0499 | 0.56 | 19 |
| *ActEff_* | *EstEff_* | 0.84 | 23.55 | <.0001 | 0.0147 | 0.39 | 34 |
| | *EstLearnEff_* | 0.05 | 56.98 | 0.9655 | <.0001 | 0 | 30 |
| | *EstLearnEff* | 0.66 | 48.51 | 0.345 | <.0001 | 0.03 | 29 |
| *ActEff* | *EstEff* | 0.85 | 38.49 | <.0001 | 0.0092 | 0.41 | 34 |
| | *EstLearnEff* | 1.34 | 77.74 | 0.3468 | 0.0003 | 0.03 | 28 |
| *ActInfoEff* | *EstInfoEff* | 0.82 | 11.14 | <.0001 | 0.0068 | 0.50 | 32 |
| *ActNavEff* | *EstNavEff* | 0.78 | 11.83 | <.0001 | 0.0041 | 0.48 | 32 |
| *ActPresEff* | *EstPresEff* | 0.83 | 16.25 | <.0001 | 0.002 | 0.43 | 33 |
| *ActLearnEff* | *EstLearnEff* | 0.96 | 1.06 | <.0001 | 0.2504 | 0.91 | 29 |

Table IV. Effort Data Analyses (RR and Association Analysis Results).
The three sections of the table contain the results for PdM1, PdM2, and USI, respectively

| Dependent Variable | Independent Variable | Robust Regression (RR) | | | | | Association | | |
|---|---|---|---|---|---|---|---|---|---|
| | | $c_1$ | $c_0$ | $\sigma$ | $R^2_{RR}$ | $R^2_{MS}$ | $\tau_b$ | $\rho$ | N |
| *ActEff_* | *EstEff_* | 2.45 | −12.58 | 9.47 | 0.82 | 0.58 | 0.41 | 0.55 | 44 |
| | *EstLearnEff_* | 2.86 | 3.07 | 9.17 | 0.82 | 0.58 | 0.32 | 0.41 | 44 |
| | *Component* | 0.43 | 12.43 | 8.62 | 0.72 | 0.42 | 0.13 | 0.19 | 36 |
| | *Segment* | 1.70 | 12.57 | 10.89 | 0.58 | 0.13 | −0.08 | −0.11 | 36 |
| | *NavSlot* | 0.03 | 16.29 | 10.57 | 0.62 | 0.22 | 0.14 | 0.19 | 36 |
| *ActInfoEff* | *EstInfoEff* | 0.5 | 6.5 | 7.22 | 0.75 | 0.39 | 0.42 | 0.58 | 44 |
| | *Component* | 0.24 | 8.18 | 4.63 | 0.66 | 0.19 | 0.11 | 0.16 | 36 |
| | *Entity* | −0.62 | 14.94 | 5.56 | 0.69 | 0.28 | −0.06 | −0.10 | 36 |
| | *Segment* | −0.14 | 10.34 | 6.57 | 0.61 | 0.07 | −0.07 | −0.11 | 36 |
| | *SlotsPerSACenter* | 0 | 11.25 | 7.25 | 0.57 | 0 | 0.06 | 0.09 | 36 |
| *ActNavEff* | *EstNavEff* | 0.92 | 2.25 | 3.67 | 0.76 | 0.55 | 0.30 | 0.44 | 44 |
| | *NLink* | −0.03 | 6.29 | 3.26 | 0.55 | 0.24 | 0.02 | 0.06 | 36 |
| | *SlotsPerNode* | −0.37 | 6.06 | 3.58 | 0.49 | 0.14 | 0.23 | 0.28 | 36 |
| *ActLearnEff* | *EstLearnEff* | 0.87 | 0.56 | 2.68 | 0.62 | 0.54 | 0.37 | 0.48 | 44 |
| *ActEff_* | *EstEff_* | 0.76 | 6.67 | 5.22 | 0.94 | 0.89 | 0.55 | 0.66 | 27 |
| | *EstLearnEff_* | 2 | 8.5 | 6.19 | 0.88 | 0.75 | 0.28 | 0.39 | 27 |
| | *EstLearnEff* | 1.58 | 3.79 | 9.28 | 0.87 | 0.71 | 0.23 | 0.32 | 26 |
| | *Entity* | 4.75 | −0.12 | 14.8 | 0.72 | 0.35 | 0.30 | 0.40 | 26 |
| | *SACenter* | 3 | 14.75 | 11.44 | 0.75 | 0.42 | 0.38 | 0.51 | 26 |
| | *SlotsPerSACenter* | 2 | 21.5 | 14.64 | 0.57 | 0 | 0.38 | 0.47 | 26 |
| | *SlotsPerNode* | 9 | −2 | 18.08 | 0.61 | 0.10 | 0.39 | 0.50 | 26 |
| *ActEff* | *EstEff* | 0.68 | 10.31 | 7.95 | 0.94 | 0.90 | 0.60 | 0.77 | 26 |
| | *EstLearnEff* | 1.90 | 14.74 | 19.30 | 0.63 | 0.41 | 0.21 | 0.29 | 26 |
| | *Entity* | 7 | 1 | 18.84 | 0.61 | 0.42 | 0.32 | 0.44 | 26 |
| | *SACenter* | 5.25 | 14.12 | 15.41 | 0.78 | 0.67 | 0.39 | 0.52 | 26 |
| | *SlotsPerSACenter* | 15 | 13 | 21.36 | 0.61 | 0.42 | 0.36 | 0.46 | 26 |
| | *SlotsPerNode* | 10 | 5 | 20.31 | 0.61 | 0.42 | 0.40 | 0.52 | 26 |
| | *CollCenter* | −0.9 | 37.05 | 23.54 | 0.66 | 0.20 | −0.18 | −0.26 | 25 |
| | *Slot* | 0.42 | 25.16 | 18.90 | 0.62 | 0.1 | 0.29 | 0.44 | 25 |
| | *NLink* | 0.5 | 18 | 14.72 | 0.88 | 0.71 | 0.02 | −0.02 | 25 |
| | *NCluster* | 1.79 | 10.81 | 16.99 | 0.76 | 0.42 | 0.07 | 0.06 | 25 |
| *ActInfoEff* | *EstInfoEff* | 1 | −2 | 3.29 | 0.94 | 0.84 | 0.63 | 0.75 | 26 |
| | *Component* | 1.58 | −6.67 | 8.88 | 0.80 | 0.37 | 0.27 | 0.40 | 25 |
| | *Entity* | 2.5 | 1.75 | 8.53 | 0.78 | 0.28 | 0.31 | 0.42 | 25 |
| *ActNavEff* | *EstNavEff* | 0.91 | 0.23 | 1.47 | 0.94 | 0.83 | 0.61 | 0.71 | 26 |
| | *SlotsPerNode* | 3 | −0.5 | 5.36 | 0.83 | 0.30 | 0.33 | 0.41 | 25 |
| *ActPresEff* | *EstPresEff* | 1 | 1.5 | 3.66 | 0.84 | 0.75 | 0.61 | 0.79 | 26 |
| | *SACenter* | 1.5 | 5 | 7.14 | 0.71 | 0.51 | 0.50 | 0.65 | 25 |
| | *SlotsPerNode* | 6 | −9 | 7.45 | 0.62 | 0.36 | 0.37 | 0.47 | 25 |
| | *SlotsPerSACenter* | 3.5 | 3.25 | 7.83 | 0.57 | 0.28 | 0.31 | 0.41 | 25 |
| *ActLearnEff* | *EstLearnEff* | 1 | −0.5 | 0.55 | 0.98 | 0.97 | 0.68 | 0.81 | 23 |
| *ActEff_* | *EstEff_* | 1.08 | −3.5 | 6.46 | 0.82 | 0.74 | 0.46 | 0.62 | 35 |
| | *EstLearnEff_* | 1.51 | 34.91 | 29.27 | 0.66 | 0.34 | 0.24 | 0.35 | 34 |
| | *EstLearnEff* | 1.14 | 30.64 | 24.31 | 0.69 | 0.35 | 0.29 | 0.41 | 33 |
| *ActEff* | *EstEff* | 0.94 | 12.51 | 13.09 | 0.92 | 0.86 | 0.43 | 0.56 | 35 |
| | *EstLearnEff* | 0.70 | 68.29 | 45.03 | 0.51 | 0.10 | 0.27 | 0.39 | 33 |
| *ActInfoEff* | *EstInfoEff* | 1 | 2.5 | 4.27 | 0.97 | 0.92 | 0.51 | 0.66 | 35 |
| *ActNavEff* | *EstNavEff* | 0.93 | 1.71 | 3.76 | 0.95 | 0.94 | 0.44 | 0.59 | 35 |
| *ActPresEff* | *EstPresEff* | 1 | 5 | 6.97 | 0.75 | 0.75 | 0.42 | 0.56 | 35 |
| *ActLearnEff* | *EstLearnEff* | 1 | 0 | 0 | 1 | 1 | 0.76 | 0.89 | 33 |

Table V. Effort Data (person-hours): Descriptive Statistics. The three sections of the table contain the descriptive statistics for PdM1, PdM2, and USI, respectively

| | $M$ | $m$ | $\sigma$ | $LMS$ | $\sigma_{LMS}$ | $N$ |
|---|---|---|---|---|---|---|
| *ActEff_* | 26.00 | 28.79 | 15.23 | 18.50 | 12.87 | 49 |
| *ActInfoEff* | 15.00 | 18.39 | 10.16 | 11.25 | 6.54 | 49 |
| *ActNavEff* | 8.00 | 10.40 | 6.69 | 5 | 5.61 | 49 |
| *ActLearnEff* | 6.00 | 7.42 | 6.12 | 5.5 | 1.14 | 49 |
| *EstEff_* | 15.00 | 17.55 | 9.28 | 11.5 | 5.18 | 56 |
| *EstInfoEff* | 9.00 | 11.17 | 6.33 | 6.5 | 3.51 | 56 |
| *EstNavEff* | 6.00 | 6.38 | 3.46 | 4.5 | 2.31 | 56 |
| *EstLearnEff* | 5.50 | 6.85 | 4.97 | 4 | 2.42 | 56 |
| *ActEff_* | 33.00 | 38.11 | 24.04 | 25 | 11.70 | 26 |
| *ActEff* | 43.75 | 54.02 | 33.08 | 33 | 16.68 | 26 |
| *ActInfoEff* | 19.00 | 22.92 | 14.02 | 15 | 5.76 | 26 |
| *ActNavEff* | 11.50 | 15.19 | 11.87 | 9 | 4.04 | 26 |
| *ActPresEff* | 13.50 | 15.90 | 10.26 | 11 | 5.16 | 26 |
| *ActLearnEff* | 15.00 | 16.93 | 11.28 | 13 | 3.91 | 23 |
| *EstEff_* | 29.00 | 34.59 | 22.11 | 26.25 | 4.76 | 27 |
| *EstEff* | 38.00 | 48.22 | 31.56 | 41 | 5.19 | 27 |
| *EstInfoEff* | 18.00 | 21.22 | 12.50 | 14.5 | 5.44 | 27 |
| *EstNavEff* | 11.00 | 13.37 | 10.61 | 10.5 | 1.16 | 27 |
| *EstPresEff* | 10.00 | 13.63 | 10.22 | 7.5 | 4.28 | 27 |
| *EstLearnEff* | 13.00 | 17.94 | 15.20 | 13.75 | 2 | 27 |
| *ActEff_* | 61.00 | 69.22 | 38.77 | 46 | 24.81 | 36 |
| *ActEff* | 97.00 | 109.83 | 57.60 | 70.5 | 44.77 | 36 |
| *ActInfoEff* | 29.50 | 34.06 | 19.51 | 23.5 | 9.88 | 36 |
| *ActNavEff* | 33.00 | 35.16 | 22.14 | 25.5 | 12.72 | 36 |
| *ActPresEff* | 35.50 | 40.61 | 22.49 | 35 | 0 | 35 |
| *ActLearnEff* | 40.00 | 47.91 | 26.85 | 32 | 15.22 | 34 |
| *EstEff_* | 43.50 | 46.25 | 22.43 | 34.5 | 14.65 | 36 |
| *EstEff* | 70.00 | 69.23 | 34.75 | 59.5 | 20.67 | 35 |
| *EstInfoEff* | 21.50 | 23.00 | 12.09 | 16.5 | 7.89 | 36 |
| *EstNavEff* | 22.50 | 23.25 | 11.23 | 17.5 | 8.78 | 36 |
| *EstPresEff* | 23.00 | 24.65 | 14.01 | 19 | 6.71 | 35 |
| *EstLearnEff* | 12.00 | 15.85 | 13.51 | 8.5 | 5.49 | 34 |

We used the Wilcoxon Matched Pairs test for the medians to compare distributions. For instance, we used this statistical test to compare whether the actual effort for building the information model was greater than the estimated effort for building the information model, in a statistically significant way. The Wilcoxon Matched Pairs test is a nonparametric one, so it can be safely applied to all variables that are measured at least on the ordinal level of measurement, regardless of any assumptions on their distributions. For instance, Table II shows that, according to the Wilcoxon Matched Pairs test, the median of *ActInfoEff* is greater than the median of *ActNavEff* with a p-value less than 0.0001 in **PdM1**.

### 4.2 Statistical Association/Correlation between Variables

The distinction we make in this article between statistical association and statistical correlation is that a statistical association can be defined for ordinal

variables, while a statistical correlation can be defined for variables that are at least on the interval level of measurement. By studying associations, we want to find out whether, for instance, when independent variable $X$ increases, the dependent variable $Y$ increases as well. Associations do not specify any kind of relationship between $X$ and $Y$, while correlation analysis (e.g., linear correlation) does.

Association analysis [Gibbons 1993] typically uses nonparametric data analysis techniques, while correlation analysis typically uses parametric techniques. Nonparametric techniques do not make any assumptions on the distribution of the data, for example, they do not assume that the data distribution is normal, like a number of parametric techniques do. Also, nonparametric techniques are less biased than parametric ones by the influence of few "outliers," that is, points that are "far" from the other points and whose value may be the result of some unlikely circumstances or data collection problems.

Parametric techniques, on the other hand, usually provide more information-bearing results: for instance, they also describe how steep the increase of the dependent variable is when the independent variable increases. Also, the statistical tests associated with parametric techniques usually require fewer data points to reject or accept the null hypothesis. Technically, statistical tests based on parametric techniques have a higher statistical power than those based on nonparametric ones. In other words, one is more likely to confirm an existing relationship between two variables with a test based on parametric techniques than with one that is not based on parametric techniques. However, one first needs to be reasonably sure that there are no overinfluential outliers and that the data come from a known underlying distribution (e.g., the normal distribution). In general, more powerful statistical tests are especially useful in an application field like ours because they require smaller data sets to confirm or refute a statistical hypothesis, and data sets usually have limited size in our field.

We have used several indicators and nonparametric and parametric techniques because, in an exploratory study like ours, we wanted to have additional checks and confirmations on the evidence that we may have obtained with just one technique. For instance, when we study the correlation between two variables with a parametric technique like Ordinary Least Square Regression, we also check whether the variables are associated with nonparametric techniques, to make sure that the results are not the effect of few outliers. In addition, we use Robust Regression to check how "reliable" the values for the coefficients of the Ordinary Least Square line are.

We used the following two nonparametric association statistics: (1) Kendall's $\tau_b$, the variant of Kendall's $\tau$ that can take into account ties in the data, and (2) Spearman's rank correlation[3] $\rho$. These statistics can provide an assessment of the degree of association between two variables $X$ and $Y$. Both statistics range

[3]Spearman's rank correlation $\rho$ can be seen as a *correlation* measure between the *ranks* of two variables $X$ and $Y$, but it is an *association* (in our sense) measure between the values of $X$ and $Y$. This explains why it is called a *rank correlation* measure, while it is actually used as an *association* measure.

between $-1$ and $+1$: a positive value of either statistic indicates that, when $X$ increases, $Y$ increases as well; a negative value that, when $X$ increases, $Y$ decreases; a null value, no association. In addition, statistical tests based on $\tau_b$ and $\rho$ can be used to support or refute the existence of positive or negative association, with the specified degree of statistical significance. For instance, it is possible to compare the two competing statistical hypotheses:

—$H_0$ ("*Null Hypothesis*"): $\tau_b \leq 0$ (i.e., there is no positive association)
—$H_1$ ("*Alternative Hypothesis*"): $\tau_b > 0$ (i.e., there is a positive association)

We used both indicators because, even though Spearman's rank correlation $\rho$ is better known, Kendall's $\tau_b$ has a more immediate interpretation as an association indicator and uses a test whose power is greater than the test used for $\rho$. This is not just a statistical detail, but it may be necessary in an application like ours with limited-sized data sets.

Table IV describes the nonparametric results we obtained via $\tau_b$, $\rho$, and $N$, the number of data points. For instance, the values of $\tau_b$ and $\rho$ for the association between $EstEff\_$ and $ActEff\_$ for **PdM1** are 0.41 and 0.55 respectively, based on $N = 44$ data points.

Even when we obtain a strong association value like $\tau_b = 0.9$, and a linear relationship exists between $X$ and $Y$, we still do not know whether the linear relationship between $X$ and $Y$ is like, say, $Y = 0.5X + 10$ or $Y = 500X + 10$. Therefore, we used linear correlation data analysis techniques to find whether it is possible to assess the steepness of the relationship between two variables, at least approximately. To this end, we used Ordinary Least Squares (OLS) and Robust Regression (RR).

The use of regression techniques such as OLS and RR is motivated by the idea that it is possible to find whether there is some sort of precise relationship between an independent variable $X$ and a dependent variable $Y$ that can be used for prediction purposes.

Univariate OLS is a very well-known technique, which is used to find whether there is a statistically significant linear correlation between two variables. Table III describes the OLS results we obtained via the following statistics: $c_1$, the estimate of the regression coefficient of the independent variable; $c_0$, the estimate of the intercept; $p_{c1}$ and $p_{c0}$, the statistical significance of the coefficient and the intercept, respectively: $p_{c1}$ also provides the statistical significance of the entire linear regression model; $R^2$, which measures the goodness-of-fit of the model as the percentage of variance that is explained by the model; $N$, the number of data points. For instance, in Table III, the values of $c_1$ and $c_0$ for the OLS between $EstEff\_$ and $ActEff\_$ for **PdM1** are 1.51 and 2.31, respectively, that is, the OLS regression straight line is $ActEff\_ = 1.51 \cdot EstEff\_ + 2.31$. The statistical significance $p_{c1}$ of $c_1$ and the entire OLS regression model is less than 0.0001 and the statistical significance $p_{c0}$ of $c_0$ is 0.5813. The goodness-of-fit is measured by $R^2 = 0.54$. The model is based on $N_{OLS} = 39$ data points.

4.2.1 *Outliers and Robust Regression.* As a first step in each of our OLS data analyses, we checked the data points and removed those that appeared to be corrupt or with incomplete information, that is, the value of either the

independent or the dependent variable was missing. As a second step in each data analysis, we carried out a careful outlier analysis, that is, we removed those few data points that were too "far" from the others and may unduly bias the results. This standard data analysis activity is necessary especially in exploratory studies like the one documented in this article because of the current stage of quantitative knowledge on Web applications. Outliers are a well-known problem in data analysis. It can be shown that even a single data point can bias the results to any degree, if that data point is "far" enough from the rest of the data [Rousseeuw and Leroy 1987]. Therefore, the presence of outliers may either blur and weaken the presence of an existing relationship, to the point that the relationship is no longer detectable, or, on the contrary, may cause the detection of a nonexistent relationship only because of the presence of few, over-influential data points. Even in the same data set, outliers, corrupt data points, or data points with missing information may be different according to the specific data analysis carried out. For instance, in Table III, the value of $N_{OLS}$ varies depending on the independent variable: we have $N_{OLS} = 39$ for the correlation between *EstEff*_ and *ActEff*_, and $N_{OLS} = 31$ for the correlation between *Component* and *ActEff*_. However, the number of data points in our application is sometimes limited, so removing too many data points as outliers may be a problem too. To alleviate this problem, we used the jackknife Mahalanobis distances of the data points to identify outliers via the JMP statistical analysis tool. The Mahalanobis distance of a point $P$ in a set of data points is a measure of how far $P$ is from the so-called "centroid" of the set of data points, which provides a concise idea of the location of the data points. The jackknife Mahalanobis distance of $P$ in a set of data points is a measure of how far $P$ is from the set of data point after $P$ has been removed from the set of data points. The idea is that $P$ attracts the centroid, so it should not be taken into account when assessing the distance of $P$ from all the other points. JMP uses a threshold value that is based on Fisher's F-distribution. Threshold values based on other distributions (e.g., chi-square) may tend to take too many data points as outliers, so JMP's threshold is more conservative in that it highlights a considerably smaller number of data points as possible outliers. All those points whose distances are above the threshold are candidate outliers. However, only one candidate outlier gets removed, and the jackknife Mahalanobis distances of all the points of this new data set are recomputed. The algorithm ends when no more candidate outliers are found, or a prespecified maximum percentage of points have been removed from the original data set. It is important to note that removing candidate outliers does not necessarily improve the statistical significance or the goodness-of-fit of the OLS model, which may actually worsen. Nevertheless, a uniform policy for dealing with outliers must be used for all models, since identifying an OLS model with a high goodness-of-fit that is due only to the presence of few outliers would be a problem.

In addition, we used Robust Regression (RR) [Rousseeuw and Leroy 1987], which is a data analysis technique that is robust to outliers. We used RR in combination with OLS, as a way to further corroborate the results we obtained with OLS.

Robust Regression is a linear regression technique based on the *median* of the squared residuals. The basic idea stems from the fact that OLS regression is based on the minimization of the *average* value of the squared residuals $(\sum_{i\in 1..N}(y_i - ax_i - b)^2)/N$, which is equivalent to minimizing the sum of squared residuals $\sum_{i\in 1..N}(y_i - ax_i - b)^2$, as is commonly said for OLS. The problem with the average as an indicator of central tendency is that it can be biased by a small numbers of outliers—even only one. Other, robust indicators of central tendency are used, such as the median. Thus, RR is based on the idea that one should find the straight line that minimizes the *median* of the squared residuals $med_{i\in 1..N}\{(y_i - ax_i - b)^2\}$.

In addition, it is possible to find an indicator of central tendency for a distribution based on this idea. Again, the average $m$ of a distribution is the value that minimizes the average of the squared unidimensional residuals $\sum_{i\in 1..N}(y_i - m)^2/N$. Likewise, a robust indicator $LMS$ for the central tendency of a distribution may be found by minimizing the median of the squared residuals $med_{i\in 1..N}\{(y_i - LMS)^2\}$. It is worth noting that $LMS$ and the median are built upon different formulas, so the value of $LMS$ does not usually coincide with the value of the median.

RR comes with goodness-of-fit indicators, $\sigma_{RR}$ and $R^2_{RR}$ [Rousseeuw and Leroy 1987]. $\sigma_{RR}$ is defined as follows:

$$\sigma_{RR} = \sqrt{\frac{\sum_{i\in 1..N} w_i r_i^2}{\sum_{i\in 1..N} w_i - v}}, \tag{1}$$

where $v$ is the total number of variables in the model, that is, the number of independent variables plus one, and $w_i = 1$ if $|r_i| \leq 2.5s^0$, where $s^0$ is defined as

$$s^0 = 1.4826\left(1 + \frac{5}{N - v}\right) med_{i\in 1..N}\{|r_i|\}. \tag{2}$$

The values 2.5 and 1.4826 that appear in the above formulas are defined in Rousseeuw and Leroy [1987] based on considerations related to the normal distribution. The value of $\sigma_{RR}$ ranges between 0 and $\infty$, and it represents an average deviation computed on a potentially smaller number of observations than $N$. Given a dependent variable $y$, $\sigma_{RR}$ is used in this article to compare the goodness-of-fit of univariate RR models each using a different independent variable.

The other goodness-of-fit indicator, $R^2_{RR}$ is defined as

$$R^2_{RR} = 1 - \left(\frac{med_{i\in 1..N}\{|r_i|\}}{mad(y)}\right)^2, \tag{3}$$

where $mad(y) = med_{i\in 1..N}\{|y_i - med_{j\in 1..N}\{|y_j|\}|\}$ is the median absolute deviation from the median. Thus, $R^2_{RR}$ shows the improvement that a RR model provides over predicting the value of $y$ for each data point with a constant

value: the median of $y$. $R^2_{RR}$ ranges between 0 (there is no improvement) to 1 (the RR model explains all uncertainty in the data set).

We introduced an additional indicator,

$$R^2_{MS} = 1 - \left( \frac{med_{i \in 1..N}\{|r_i|\}}{mad_{LMS}(y)} \right)^2, \quad (4)$$

where $mad_{LMS}(y) = med_{i \in 1..N}\{|y_i - LMS(y)\}$, that is, the median absolute deviation from the $LMS$ of $y$. The reason is that $R^2_{MS}$ shows the improvement that a RR model provides over predicting the value of $y$ for each data point simply as the $LMS$ of $y$, that is, a simpler linear RR model that has only the intercept but no independent variable. This is more similar to OLS analysis, in which $R^2_{OLS}$ actually quantifies the improvement that an OLS model provides in explaining the uncertainty (i.e., the variance) of the dependent variable over a simpler model that predicts that the value of the dependent variable in each data point is given by the *mean* of $y$. As our data analysis shows too, $R^2_{RR} \geq R^2_{MS}$, that is, $R^2_{RR}$ shows a greater improvement than $R^2_{MS}$, because $R^2_{RR}$ quantifies the improvement of using a RR linear model over using the median (which does not minimize the median of the residuals), while $R^2_{MS}$ quantifies the improvement of using a RR linear model over using the LMS (which does minimize the median of the residuals).

As RR does not come with statistical tests that may be used to assess the statistical significance of the regression coefficients, we used RR as an additional way to obtain a robust slope for the regression straight line derived via OLS, one that may be less influenced by outliers than OLS.

Table IV describes our RR results via the following statistics: $c_1$, the estimate of the regression coefficient of the independent variable; $c_0$, the estimate of the intercept; $\sigma_{RR}$, $R^2_{RR}$, and $R^2_{MS}$, the goodness-of-fit indicator; $N$, the number of data points. For instance, the RR analysis of *ActEff*$_-$ vs. *Component* for **PdM1** shows that the RR straight line is $ActEff_- = 0.43 \cdot Component + 12.43$, with $\sigma_{RR} = 8.62$, $R^2_{RR} = 0.72$, and $R^2_{MS} = 0.42$, based on $N = 36$ data points.

## 5. HYPOTHESES AND RESULTS

**PdM1** was the initial empirical study and **PdM2** and **USI** were subsequent studies. Our goal was to find out consistent patterns, which will need to be confirmed in future empirical studies. We started with a number of hypotheses that we believed to be true in our application contexts. After we carried out **PdM1**, we discarded all those hypotheses that turned out to be clearly unsupported by the data analysis, and we kept all those that were supported (at the 0.05 significance level) and those data with a p-value close to 0.05, to test them again.

In this section, we report on the hypotheses we set and the corresponding results. For each hypothesis, our explanation is organized as follows:

**Question**. We describe the research question that prompted us to investigate the hypothesis.

**Rationale**. This is the reason why we believe that the research question is relevant.

**Hypothesis**. We first describe the underlying belief whose truthfulness we wanted to check with our empirical analysis. We then describe the statistical hypothesis we tested, which is the "alternative" ($H_1$) hypothesis in the statistical test of hypotheses. The "null" ($H_0$) hypothesis can always be found as the logical negation of the "alternative" hypothesis. This shortened description of the statistical test of hypothesis turns out to be especially useful because, in several cases, we tested a family of hypotheses instead of a single hypothesis. Spelling out all the details of the single hypotheses would not add anything to the presentation. In the description of the hypotheses, we also mention the actual variables used to quantify the attributes appearing in the hypotheses.
**Results**. We concisely present the results we obtained for each of the empirical studies in which the hypothesis was tested, along with information about the collected data that may be useful for the interpretation of the results. Unless explicitly mentioned, all of the results we present are statistically significant at the 0.05 level.
**Discussion**. We discuss the positive and negative results we found, along with any consistencies or discrepancies across the empirical studies.

As a matter of fact, we have also set more detailed hypotheses than the ones we discuss here. For brevity's sake, we will only briefly report on them where appropriate.

### 5.1 Comparison among Efforts

**Question.** Does one of the W2000 models consistently take more effort to be created than the others?

**Rationale.** If we find out which W2000 model takes the most effort, we can allocate more resources to it. We also know on which design phase there may be the largest potential for effort reduction.

**Hypothesis.** We believed that the Information Model would be the most effort-consuming one, since it is the "heart" of the entire W2000 design. The statistical hypothesis we tested is: *the median of ActInfoEff is higher than the median of the actual effort related to any other model*.

**Results.** Table II contains the statistical significance of the comparisons between medians, according to the Wilcoxon Matched Pairs test. For instance, the statistical significance of the alternative hypothesis $ActEff\_ > EstEff\_$ for **PdM1** is less than 0.0001. In **PdM1** and **PdM2**, the median of *ActInfoEff* was consistently greater than the medians of the other actual effort components. In **USI**, the median of *ActInfoEff* was consistently *smaller* than the medians of the other actual effort components. Specifically, the median of *ActInfoEff* was smaller than the medians of *ActPresEff* and *ActLearnEff* in a statistically significant way. The median of *ActLearnEff* was the highest median, and this was also statistically significant in the comparisons against *ActInfoEff* and *ActNavEff*.

**Discussion.** There may be several causes for this difference. The main motivation behind the results is the cultural difference between the students in **PdM1** and **PdM2** and those in **USI**. Even though the design of a Web application

is a multi-faceted problem, which would require experts with different backgrounds, each culture tends to emphasize a particular facet. Engineers are always concerned with making applications work "correctly:" They know what is behind the user experience and they think about it in terms of its relation with the back-end. The result is that a good data model (i.e., the information model) becomes the starting point for designing a reliable application. The user experience, along with its usability, is only a consequence. In contrast, communication experts flip the problem and start conceiving their applications from the other end: They only have a vague idea of what is behind the scene and their designs start and end with the definition of the user-experience. The other layers (models) are not fundamental for their message and thus they can be underemphasized.

Even though all subjects were asked to design a Web application from scratch, different communities–and thus different expertise–associate different weights with W2000 models (but this consideration applies to many modeling techniques for Web applications). This means that they take different perspectives on the development process. In what could be called a *top-down* approach, one starts from the Web pages and then moves to rendered data; in a *bottom-up* approach, one starts modeling relevant information and arrives at the concrete pages by means of transformations and refinements. In contrast, with conventional software development, in this case, the choice does not depend on the starting point (new system vs. legacy components), but comes from the different background.

This different attitude is also supported by the difference in the learning effort estimates. The subjects in **USI** found the notation much more difficult than the students in **PdM1** and **PdM2**. In the first case, the subjects were asked to learn a modeling discipline: in a sense, they felt they had to adopt an engineering perspective to carry our their projects, thus they thought it was difficult. The students in **PdM1** and **PdM2** are more used to modeling and design notations and understood the problem as something they were used to (even though they did not consider presentation problems in detail, especially in **PdM1**).

The difference in learning effort may also be explained by the fact that the students at **USI** were second-year ones, while the students of **PdM1** and **PdM2** were fourth- and fifth-year students.

### 5.2 Subjects' Ability to Estimate and Predict Actual Efforts

In this section, we present the hypotheses and results we obtained about the ability of our subjects to estimate the actual effort:

—we first investigated how close the estimates of our groups of subjects as a whole were to the actual values

—we then evaluated whether the estimates could be used to build prediction models

—we checked if a few factors related to the subjects could influence the estimates, actual values, and differences between estimates and actual values.

5.2.1 *Effort Underestimation.*

**Question.** Are the subjects overly optimistic, that is, do they tend to underestimate the effort needed to design an application and the efforts related to each effort category?

**Rationale.** This is a common problem in software and Web engineering, in which even experienced technicians and managers tend to underestimate effort. The knowledge of the fact that there is a tendency to underestimate effort and its components may be used in two ways:

—be prepared for a high likelihood that the actual effort will be larger than the estimated one

—study the causes that make Web application designers underestimate the design effort and remove them, if possible, or alleviate the problem.

**Hypothesis.** We believed that there actually is a tendency to underestimate effort, so this is the statistical hypothesis we tested: *for each effort category, the median of the actual effort is higher than the median of the estimated effort*.

**Results.** In the three cases, the students were asked to estimate their effort after they had classes on using W2000. In **PdM1**, we managed to collect data only on the effort for the information and navigation models, so we use *EstEff_* and *ActEff_* instead of *EstEff* and *ActEff*. For the sake of comparison, we here report on the results on *EstEff_* and *ActEff_* also for **PdM2** and **USI**. The results are in Table II.

*PdM1.* The subjects tended to underestimate the design effort: The median of *ActEff_* was approximately 50% greater than the median of *EstEff_* and the difference was statistically significant. The same applies to the differences between the effort categories, except for the learning effort.

*PdM2.* The results are less clear than for **PdM1**. The median of *ActInfoEff* was not greater than the median of *EstInfoEff* in a statistically significant way. That had a relevant impact on the fact that the median of *ActEff_* was not greater than the median of *EstEff_* in a statistically significant way, despite the fact that the median of *ActNavEff* turned out to be greater than the median of *EstNavEff* in a statistically significant way. The median of *ActPresEff* was greater than the median of *EstPresEff* in a statistically significant way, so the median of *ActEff* turned out to be greater than the median of *EstEff* in a statistically significant way. In contrast, the median of *ActLearnEff* was not greater than the median of *EstLearnEff* in a statistically significant way.

*USI.* The subjects tended to underestimate the design effort, based on the analysis of *EstEff_* vs. *ActEff_*, and *EstEff* vs. *ActEff*. Specifically, the median of *ActEff* is more than 30% greater than the median of *EstEff*. The same applies to the differences between the effort categories.

**Discussion.** There appears to be an underestimation problem in all three of our studies, regardless of the fact that **PdM1** and **PdM2** involved estimates from individuals, while **USI** had estimates from teams, and the fact that **PdM1**

and **PdM2** involved engineering students, while **USI** involved communication sciences students. To be precise, the underestimation hypotheses are not supported for all development effort components of **PdM2**, but there is fairly good evidence that there is an underestimation pattern in **PdM2** too, since the underestimation hypothesis was not supported for the information model effort alone, in addition to the learning effort. At any rate, all of our subjects applied a naive approach and the result was that they underestimated the problem, even though engineering students were from later years and personnel with skills in communication is not usually asked to plan a technical project and estimate the effort needed to complete it. The students in **USI** provided estimates whose median value was comparatively larger than the median value of the estimates provided by the **PdM1** and **PdM2** students. Still, their estimates were larger because the students worked in teams of three people, so the estimated effort of each team was likely to be higher. By looking at the per capita effort, **USI** students actually have a lower estimated value than **PdM1** or **PdM2** students. The data in Table II also show that the medians of the estimated efforts for **PdM1** students were approximately half the medians for the corresponding efforts for **PdM2** students. One possible explanation is that the subjects had talked with the subjects from the previous year (i.e., the ones that participated in **PdM1**) and so they may have provided initial estimates that took the underestimation problem into account. Still, underestimation problems can be found in **PdM2**.

#### 5.2.2 *Usefulness of Effort Estimates.*

**Question.** Are effort estimates actually useful to predict actual effort?

**Rationale.** Effort estimates are often produced in a subjective way. Nevertheless, they can often be useful to predict actual effort, even though there may be several causes of deviation between the estimated and the actual values. This question is different from the previous one, which investigated if a set of subjects tend to underestimate the design effort. The question we are discussing here aims at checking whether it is possible to build prediction models for the actual effort based on the estimated effort, even though the estimated effort may be off target on average.

**Hypothesis.** Our belief was that effort estimates are useful to predict the actual effort. This is the hypothesis we tested: *for each effort category, there is a positive correlation between the estimated effort and the actual effort*.

**Results.** We describe only the most interesting results in some detail, for the sake of conciseness. The results are in Tables III and IV.

*PdM1* $EstEff\_$ is a good predictor of $ActEff\_$, even though it appears to consistently underestimate it: the model derived from OLS regression is $ActEff\_ = 1.51 \cdot EstEff\_ + 2.31$ (with goodness-of-fit $R^2_{OLS} = 0.54$) and the model derived from RR is $ActEff\_ = 2.45 \cdot EstEff\_ - 12.58$ (with goodness-of-fit $R^2_{RR} = 0.82$ and $R^2_{MS} = 0.58$). The single estimated effort categories are also good predictors for the corresponding actual

effort category. It may appear that the OLS and RR coefficients of the models (1) with dependent variable *ActInfoEff* and dependent variable *EstInfoEff* and (2) with dependent variable *ActNavEff* and dependent variable *EstNavEff* are much closer to 1 than the coefficients of *EstEff_* in the OLS and RR equations for *ActEff_*. However, that may be due to the fact that the intercepts in either model are considerably higher, which tilts the regression lines in such a way as to make the slopes less steep. At any rate, the values of $\tau_b$ and $\rho$ also show the existence of an association between the estimated and the corresponding actual effort categories and total effort as well.

*PdM2* *EstEff_* is a good predictor for *ActEff_*, and so is *EstEff* for *ActEff*. The OLS regression model $ActEff\_ = 0.86 \cdot EstEff\_ + 8$ (with goodness-of-fit $R^2_{OLS} = 0.33$) and the model derived from RR is $ActEff\_ = 0.76 \cdot EstEff\_ + 6.67$ (with goodness-of-fit $R^2_{RR} = 0.94$ and $R^2_{MS} = 0.89$), seem actually to indicate a (small) overestimation problem, since the coefficients of *EstEff_* are smaller than 1 and the intercept is relatively small. Similar considerations apply to the prediction of *ActEff* by *EstEff*, though the intercepts of the OLS and the RR models are certainly higher. This again shows that the overestimation problem in **PdM2** is not as clear as in **PdM1** and in **USI**. The values of $\tau_b$ and $\rho$ also show the existence of an association between the estimated and the corresponding actual effort categories and total effort as well. As for the various subcategories, the estimated effort is a good predictor of the actual effort, and the slopes obtained with OLS and RR are fairly close.

*USI*. All estimated efforts appear to be good predictors of the corresponding actual efforts. For instance, the OLS line $ActEff = 0.85 \cdot EstEff + 38.49$ with goodness-of-fit $R^2_{OLS} = 0.41$) and the RR line $ActEff = 0.94 \cdot EstEff + 12.51$ (with goodness-of-fit $R^2_{RR} = 0.92$ and $R^2_{MS} = 0.86$) show an underestimation problem because the intercepts (especially the intercept of the OLS line) appear to be relatively large, even though the coefficients of *EstEff* are not larger than 1. The nonparametric analysis confirms these results.

**Discussion.** Our analyses show that it is possible to use the estimated effort as a predictor for the actual effort in OLS models, even though we have an underestimation problem in two applications. As the values of $R^2$ show, the estimated values seem to account for a fairly significant part of the variance, even though other factors clearly need to be taken into account. The differences among the studies can be explained like in the comparison among the means of estimated and actual effort.

5.2.3 *Influencing Factors.*

**Question.** Which subjects' characteristics may influence the effort estimates, the actual effort, and the accuracy of the effort estimates?

**Rationale.** Identifying the factors that may influence the effort estimates, the actual effort, and the accuracy of the effort estimates can be used to act on

those factors to produce more accurate estimates or to identify "trustworthy" subjects. This is relevant because effort estimation is one of the most difficult and important tasks in Web (and software) engineering.

**Hypothesis.** Our belief was that effort estimates, the actual effort, and the accuracy of the effort estimates were influenced by a number of subjects' factors. We believed that subjects' characteristics would influence design effort and learning effort as well. Our idea was that greater skills and better school proficiency are associated with smaller efforts between actuals and estimates (since the subjects would be able to complete the assignment in less time) and smaller differences (since the subjects would be able to make better estimates). Thus, this is the hypothesis we tested: *there is a negative correlation between*

—*the subjects' technical knowledge (measured by the number of Web-related languages (HTML, XML, etc.) and techniques (JSP, ASP, etc.) that they knew) and school proficiency (measured by their average grade or the number of computer science exams they had passed)* on the one hand, and

—*the estimated learning effort, actual learning effort, difference between actual and estimated learning effort, estimated effort, actual effort, and difference between actual and estimated effort,* on the other hand.

**Results.** The results we obtained in **PdM1** were not really encouraging. The only correlations we found were a negative correlation between technical knowledge and the difference between actual and estimated effort, and between proficiency and estimated learning effort. Surprisingly, there seems to be a positive correlation between proficiency and actual effort, contrary to our hypothesis. In addition, the goodness-of-fit of these results was too small to base any predictions on univariate models. Few of the hypotheses turned out to be confirmed, and with a very low exploratory power, due to its low goodness-of-fit, and this is why we did not pursue checking this hypothesis in the other studies. For conciseness, we do not report the detailed results in this article.

**Discussion.** The negative correlation between technical knowledge and the difference between actual and estimated effort seems to indicate that subjects with greater technical knowledge are better at providing an estimate for the actual effort. The negative correlation between proficiency and *EstLearnEff* shows that subjects with higher proficiency tended to believe that they would have an advantage in the learning activities. However, this advantage did not materialize, since there is no correlation between the average of their previous grades and *ActLearnEff*. As for the positive correlation between proficiency and *ActEff_*, one possible explanation is that higher proficiency may be associated with higher levels of commitment in carrying out the homework. This may also help explain why we did not obtain results for the other hypotheses, since there was the superposition of two effects: on the one hand, better skills and proficiency made our subjects more productive; on the other hand, the subjects with better skills and proficiency were also more committed and devoted more effort.

### 5.3 Subjects' Ability to Estimate the Learning Effort

Since our studies had students as subjects, we also investigated their ability to estimate their learning effort.

**Question.** How good are the subjects at estimating the learning effort for Web design?

**Rationale.** Our subjects were students, so they were used to estimating the effort it would take them to learn a specific topic, at least approximately.

**Hypothesis.** To this end, we adopted the same underestimation approach that we used for the other. So, our first hypothesis was that the median of *ActLearnEff* is greater than the median of $EstLearn$ effort. If we obtain a statistically significant result, we have support to reject the claim that the students can estimate their learning effort well, otherwise, we cannot reject this claim. The second hypothesis is that there is a positive correlation between *EstLearnEff* and *ActLearnEff*, to make sure that we can use the estimation results for prediction purposes.

**Results.** The first hypothesis is not supported in **PdM1** and **PdM2** so we cannot reject (and we have some evidence for) the claim that students know how to estimate the learning effort, but the results for **USI** actually show that $ActLearnEff > EstLearnEff$. In **PdM1**, **PdM2**, and **USI**, the positive correlation hypothesis is supported by the OLS, the RR, and the nonparametric analyses, with satisfactory goodness-of-fit ($R^2$) values, especially for **USI**.

**Discussion.** The results supported our hypotheses as to **PdM1** and **PdM2**, but not for **USI**. **USI** subjects were younger and not used to software modeling at all. They had to predict the effort for something completely new to them without the experience gained after years at university. As for **PdM1** and **PdM2**, we can say that students in their last college years are usually good at estimating the effort they need to learn a new topic, or better to deal with a given course. Moreover, even if they were new to Web applications, they were already familiar with the problem of modeling software applications with UML-like notations.

### 5.4 Self-Grading

**Question.** Are the students good at evaluating their own design models?

**Rationale.** Artifact quality evaluation is an important issue in software and Web engineering and is usually carried out on a largely subjective basis. Thus, we wanted to check how good the subjects/designers in our empirical studies were in evaluating their own artifacts.

**Hypothesis.** We believed that our students were expert enough to provide an assessment of their own artifacts that would agree with the grading provided by their professor. Thus, our hypothesis was that *there is a positive correlation between the grade provided by the students and the grade provided by the professor*.

**Results.** In **PdM1**, the results did not show any correlation or even an association, so we did not replicate this part of the study in the following empirical studies.

**Discussion.** It is possible that the students did not really want to provide us with a reliable self-evaluation of their artifacts, because they did not want to appear either too "bold" or too "shy," for fear that their self-evaluation would influence the professor's evaluation one way or the other. The evaluation of the products is often used to evaluate the persons who have contributed to that product, so this problem is similar to the problem in industrial organization, where practitioners may sometimes be reluctant to provide data about their own work and products for fear these data will be used against them.

### 5.5 Characteristics of W2000 Models and Actual Effort

Here, we investigate whether there are any relationships between the measures for the internal attributes of W2000 models and the actual effort. We study the entire actual effort first and then we investigate its components, that is, the actual information, navigation, and presentation effort. Data and analyses are available for **PdM1** and **PdM2**, but not for **USI**. The internal attributes and corresponding measures for W2000 artifacts are in Table VI. For space reasons, we do not report here on all the hypotheses we tested, so the table contains a superset of the measures that appear in the hypotheses of this article. The interested reader may refer to Baresi et al. [2003], which is about **PdM1**, for a more complete illustration of the hypotheses we have set and tested.

**Question.** Does the size of information, navigation, presentation models, their complexity, or the granularity adopted to decompose the information model influence the actual development effort? Does reuse influence the total effort to develop W2000 (Web application) models?

**Rationale.** These factors are commonly believed to be highly influential in a number of application fields of empirical software engineering [Boehm 1981; Boehm et al. 2000; Fenton 1991; Fenton and Pfleeger 1996]. In addition to considering the size of produced artifacts, we must distinguish among information, navigation, and presentation. Complexity can be added at any level: One can start with a simple information model, and add complexity with a very detailed navigation or presentation model. Complexity may also come from the granularity adopted to select the components of the information model (i.e., the way entities are partitioned), and the nodes of the navigation model. In contrast, slots defined in the information model can be reused in the other models by means of properly defined segments, which are the W2000 construct conceived to support reuse across the different models.

**Hypothesis.** Our hypothesis can be expressed as follows.

(1) *The actual effort increases with: the size of the information, the navigation, or the presentation model, the complexity of the information or the navigation model, and the granularity decomposition of the information model (measured by the attributes shown in Table VI).*

Table VI. Measures for Internal Attributes of W2000 Models. The three sections of the table contain the results for Information, Navigation, and Presentation models, respectively

| Attribute | Measure | Definition |
|---|---|---|
| Size | **entities** | # of entities in the model |
| | **components** | # of components in the model |
| | **infoSlots** | # of slots in the model |
| Average size | **slotsSACenter** | avg. # of slots per sem. assoc. center |
| | **slotsCollCenter** | avg. # of slots per collection center |
| Average granularity of decomposition | **componentsEntity** | avg. # of components per entity |
| | **slotsComponent** | avg. # of slots per component |
| Complexity | **collections** | # of collections in the model |
| | **Sassociations** | # of sem. associations in the model |
| Data cohesion | **SACenters** | # of sem. assoc. centers in the model |
| Reuse | **segments** | # of segments in the model |
| Size | **nodes** | # of nodes in the model |
| | **navSlots** | # of slots in the model |
| Average size | **nodesCluster** | avg. # of nodes per cluster |
| | **slotsNode** | avg. # of slots per node |
| Complexity | **navLinks** | # of links in the model |
| Structuredness | **clusters** | # of clusters in the model |
| Size | **pages** | # of pages in the model |
| | **pUnits** | # of publishing units in the model |
| Complexity | **prLinks** | # of links in the model |
| Structuredness | **sections** | # of sections in the model |

(2) *The actual effort decreases with: the reuse in the models (measured by segments as shown in Table VI).*

**Results.** In **PdM1**, variables *components* and *navSlots* are positively correlated with the actual effort, and *segments* is negatively correlated with the actual effort. All the measures for the other attributes were not correlated with the actual effort. In **PdM2**, *Entity*, *SACenter*, *SlotperNode*, and *SlotperSACenter* are associated and correlated with *ActEff* and $ActEff_{-}$.

If we analyze the effort with respect to the main W2000 models, in **PdM1**, variables *components* and *segments* are linearly correlated with *ActInfoEff*, and *entities*, and *slotsCollCenter* are "weakly" linearly correlated with it (their p-values are close to but not below 0.05). However, there is no statistically significant association between any of these variables and *ActInfoEff*. In addition, there is a remarkable difference between the coefficients of the regression lines obtained with OLS and RR, so the linear correlations we found may be the effect of some undetected outliers. Variables *navLinks* and *slotsNode* are linearly correlated with *ActNavEff*, and *slotsNode* is also associated with it. The coefficients and intercepts of the regression lines obtained with OLS and RR with *slotsNode* are somewhat close, so there may not be a large influence due to outliers.

In **PdM2**, the association and correlation of two measures, *components* and *entities*, and *ActInfoEff* are statistically significant. However, the OLS coefficients are quite far from those we obtained with RR, maybe due to the effect of outliers. Variable *slotsNode* is both associated and correlated with *ActNavEff*.

The difference in the coefficients between the regression lines obtained with OLS and RR may be still due to the presence of undetected outliers. No association or correlation exists between *navLinks* and *ActNavEff*. *SACenter* and *SlotPerSACenter* are both associated and correlated with *ActNavEff*, while, *slotsNode* is associated but weakly correlated with it.

**Discussion.** We have obtained different correlation and association results in **PdM1** and **PdM2** for the total efforts, while some similarities may be found in the predictors for the effort categories. These results might partially be caused by the fact that our subjects were students: their skills are very different, and even more importantly not all students want to allocate the same effort to a given course. Also, this highlights the need for further empirical studies in an exploratory study like ours. Results that seem to be supported in an initial study need to undergo additional empirical studies. At any rate, it is important that even results that initially are not supported by both a correlation and association analysis should not be discarded and should undergo further analyses. Previous studies (e.g., Mendes et al. [2002b] and Costagliola et al. [2004]) did not seem to find significant correlations between internal measures and effort data either.

## 6. VALIDITY OF THE EMPIRICAL STUDIES

Like in any empirical study, we need to examine the possible factors that may have biased our results. We believe that the following factors may influence an empirical study like ours, from both an internal and an external point of view: subjects, applications, availability of tools, notation, classes, data collection, and constructs. We focus on how these factors may have influenced the internal and external validity of our study. Internal validity is related to the study in itself, for example, to the fact that the right conclusions have been drawn within the study, the data analysis has been carried out correctly, no self-selection or "mortality" effects occur in the sample, etc. External validity is related to the generalizability of the results of the study to some target population, that is, the population of professional web designers in our case. Threats to both internal and external validity need to be examined and discussed in every empirical study, to assess the extent of their impact.

### 6.1 Internal Validity

Here, we need to examine whether the factors could pose a threat to the internal validity of the empirical studies.

**Subjects.** No initial selection of the subjects was carried out, so no initial bias was introduced. The degree of self-selection among the subjects was also limited. In our presentation, we have chosen to show the results of the three empirical studies separately. If we had lumped together all the data points in one sample, we could have introduced a bias in the selection of the students, since they came from two engineering courses and one communication sciences course. The students in each empirical study had been exposed to the same techniques in previous classes, so the background of the subjects within each

empirical study can be considered sufficiently homogeneous, though there are differences between engineering and communication students. In addition, the students were enrolled in different years at college, so our results may have been influenced by that factor too, and that is an additional reason why we have kept the results separate.

**Applications.** Even though we let the subjects choose the application they preferred, all applications only differed in their specific details, but they were all related to e-commerce.

**Availability of Tools.** At the time of the empirical studies, no automated supporting tools existed for W2000. So, the students used standard text and graphics editors.

**Notation.** All the subjects used the same notation.

**Classes.** There was little difference in the percentage of classes attended by the subjects, that is, they attended almost all the classes.

**Data Collection.** The data we collected seem to be trustworthy enough. For instance, the results show that the students did not just provide a copy of their effort estimates as their actual efforts. We did not disclose any of our hypotheses to the students, and we made it clear that the data we collected would not have any influence on the grade the students would obtain. The accuracy of the reported data was clearly not perfect, but it was an economically viable compromise between the needs of the subjects and the needs of the researchers [Carver et al. 2003]. A stricter data reporting mechanism might have been too imposing on the students and would have got in the way of their work.

**Constructs.** The risk is that the measures for the internal attributes of W2000 artifacts (see Table VI) do not adequately quantify the attributes they purport to measure. The measures we used are fairly simple, to capture the most important elements and attributes of the W2000 models in a straightforward way. We cannot totally exclude that other measures for the same attributes (e.g., other size measures) would be correlated with the dependent variables.

Thus, we believe that our studies were not biased by the choice of subjects, applications (in the context of e-commerce), and data collection mechanism. The results were clearly influenced by the lack of tools, the notation, and the percentage of class attendance, that is, we could have obtained different results in different empirical studies with students if tools had been available, a set of different notations had been used, and our sample of students had attended from 0% to 100% of the classes instead of attending basically all the classes.

### 6.2 External Validity

We need to check how representative our empirical studies are in the population of empirical studies on Web application effort estimation. In other words, we need to examine the factors that may make it difficult to extend the results of these studies to web development organizations.

**Subjects.** Since no preselection was carried out, our subjects can be considered representative of the entire population of students attending advanced web design classes. Furthermore, the subjects may be considered also somewhat representative of the population of Web designers and developers since many Web designers are still young professionals who got their degrees recently.

**Applications.** E-commerce applications, like the ones in our studies, are typical Web applications. Those proposed to our subjects present all the typical features of traditional e-commerce applications, but with a controlled size since they were projects proposed to students.

**Availability of Tools.** The design of Web applications may be carried out with the aid of automated tools in industrial environments. However, professional tools may be used for designing Web applications, so this factor could have been a threat to the external validity of our studies.

**Notation.** W2000 has many aspects in common with UML, which is a standard software development notation. UML and W2000 share the same underlying concepts. W2000 can be seen as a proper UML profile, that is, a special-purpose customization of the notation. It is more complex than UML because of its layered nature. UML does not force the use of the MVC pattern to design complex software systems, even if it would be good common practice, while W2000 forces designers to define the different models to address the main aspects of a complex Web application (i.e., information, navigation, and presentation). Notice that the same applies to other notations for modeling Web applications: for example, both UWE and WebML can be explained by means of proper UML meta-models, that is, UML profiles. Again, this means that these notations too can be seen as proper customizations of UML and thus have many aspects in common with UML.

**Classes.** The specific educational content provided to the students attending the classes in which the empirical studies were run may not be entirely representative. However, little could be done about this, since we could certainly not prevent a part of the students from attending the classes, and the instructors' first responsibility is to provide students with education on techniques and methods that the instructors deem best for their students. In addition, it must be said that Web designers and developers are trained personnel, who need to continuously update their knowledge.

**Data Collection.** The data may not necessarily be of worse quality than the data collected in industrial organizations, where people are required to fill out timesheets that may sometimes be biased for a number of reasons. Actually, the students may be thought of having less at stake than practitioners during data collections (even though that may not be the students' perspective).

**Constructs.** Other measures may be used depending on the notation chosen. At any rate, the attributes quantified by these measures may not necessarily change.

Therefore, our studies may be fairly representative as for choice of subjects, applications (in the context of e-commerce), and data collection mechanism,

but less so as for tools, notation, percentage of class attendance (i.e., college education), and constructs.

## 7. LESSONS LEARNED

The studies support the idea that people with different background adopt particular mental attitudes while designing Web applications. In our case, computer scientists (**PdM1** and **PdM2**) concentrate on the information model, that is, the foundations of the application, while experts in communication (**USI**) privilege the user interfaces. This may support the idea that a team with mixed expertise produces better Web application. The different background, along with the related mental attitude, is also supported by the fact that the subjects in **PdM1** and **PdM2**, who are used to software modeling and design, appear to be better at estimating the effort to become proficient in W2000 than **USI** subjects. In contrast, we have no correlation between the actual background of the subjects and their capability to accurately predict the design efforts.

Our idea, borrowed from conventional software development, that there is a tendency to underestimate the effort required to fully model a realistic Web application, is fully supported in **PdM1** and **USI**, and at least partially supported in **PdM2**. **PdM2** still shows the need for further empirical studies (possibly with professional developers as subjects). These results demonstrate that the "economics" behind software projects should be taught better to students and studies like ours might provide valuable insights to this end.

However, our empirical studies consistently highlight the good fit of estimated efforts as indicators for the actual ones. The exploratory approach and the need for further experiments do not allow us to propose general models to predict the actual effort given the estimated one, but we are pretty confident that these first three empirical studies give good evidence of the feasibility of such a model.

The statistical analyses used to correlate the effort with internal attributes of application designs did not give the results we expected. The different studies correlate different size measures with the actual effort. There are no general hypotheses supported by the analyses we conducted, but we still think that the designer's background impact the perception of complexity. We would have expected that computer scientists, who tend to embed the complexity of their applications in the information model and then adopt a "simplified" navigation and presentation, had privileged the number of entities, components, and collections as candidates for possible correlations with the actual effort. Similarly, we would have expected that experts in communication, who usually adopt a simple information model, and then concentrate on its navigation and presentation, had considered pages, sections, and units as "their" preferred indicators of complexity. Unfortunately, this is not fully confirmed by the statistical analyses presented in the previous sections. Similarly, the correlation between segments (i.e., reuse) and effort is not supported by all the experiments.

Even though we studied W2000, we are confident that these lessons apply to other notations for Web applications as well. OOHDM, W2000, WebML and other notations come from HDM, which was one of the first notation proposed

for modeling complex Web applications. This means that even if concepts are not exactly the same, these notations are based on the same underpinnings. For example, all these notations foster a layered view of a Web application and propose similar modeling concepts. Moreover, the underestimation problem is not peculiar of Web development, but it applies to the wider domain of software development. Needless to say, our studies are exploratory in their essence and thus all results need further investigations.

## 8. RELATED WORK

In the last few years, a number of studies have appeared in the scientific literature on effort estimation and building prediction models for Web applications.

### 8.1 Frameworks and Taxonomies

Despite the novelty of the field, a few papers have appeared in the literature to provide a framework for Web measures, on which effort studies are based.

Dhyani et al. [2002] provide a survey of a number of Web-related measures. Specifically, measures are classified as related to the graph structure (of the entire Web or parts thereof, such as Web sites), Web page significance (on the relevance and usefulness of Web pages), usage characterization (on the dynamic usage of Web sites), Web page similarity (on the consistency across Web pages), Web page search and retrieval (on the performance of Web search and retrieval services), and information theoretic (related to information needs, production, and consumption). Among these categories, the graph structure measures are those that are most closely related to the measures defined in this article. However, the survey does not mention any empirical evidence for assessing the actual usefulness of these metrics.

Calero et al. [2004] propose WQM (Web Quality Model) as a means to organize the wide set of metrics proposed for quantifying the quality of Web applications. WQM is an important starting point to frame our contribution in this context. The model is organized around three main dimensions: web features, life cycle processes, and quality characteristics. Our work addresses the first dimension, in which the proposed framework further distinguishes among presentation, navigation, and content, which are exactly the three main layers around which W2000 models are organized.

Mendes et al. [2005a] propose a taxonomy for size measures of hypermedia and Web applications, which are mostly based on high-level characteristics of the final artifacts such as the number Web pages. In this article, we have defined measures for the *design* of Web application and we have adopted a finer-grained and layered approach.

### 8.2 Empirical Studies

A number of empirical studies have been carried out. Size has often been believed to be the main effort driver in software engineering applications, so a number of studies have focused on the influence of size on the effort for developing Web applications. The studies have used different kinds of size

measures, specifically either related to the structure of a Web application or its functionality.

The *Tukutuku* Benchmarking Project [2006] by Mendes et al. collects information about completed projects worldwide. On-line forms allow people to enter information about their projects. These data have originated a significant amount of work and a dedicated monograph on *theory and practice of metrics and measurements for Web development* [Mendes and Mosley 2006]. In this book, there is an interesting chapter on Web effort estimation where Mendes et al. [2006] introduce the concepts related to effort estimation, and present a case study on building and validating an effort estimation model based on the data collected in the Tukutuku database.

The proposed model is built by means of a manual stepwise procedure. The two selected variables are the total number of new Web pages and the total number of high-effort features/functions in the application. Together, they explain 76% of the variation in total effort, but the cross-validation revealed the scarce accuracy–*MMRE* too high and *Pred*(25) too low–of the proposed models. The final suggestion is to use the proposed models to obtain an estimated effort, and then tailor it by considering factors like previous experience with similar projects and the skills of developers. Even if our proposal concentrates on the artifacts produced while designing applications, and this work addresses the final artifacts directly, these results support the first hypotheses of our work, that is, the relative novelty of problem and the need for more empirical studies–further data–before being able to create significant models.

Mendes et al. [2005b] try to identify size measures and cost drivers for early Web cost estimation based on current practices of several Web companies worldwide. They analyzed 133 Web projects and obtained that the two most common size metrics used for Web cost estimation are "total number of Web pages" (70%) and "which features/functionality to be provided by the application" (66%). These results, after validation, were used to prepare Web project data entry forms to gather data on Web projects worldwide. After gathering data on 67 real Web projects worldwide, multivariate regression applied to the data confirmed that the number of Web pages and features/functionality provided by the application to be developed were the two most influential effort predictors.

Mendes et al. [2003b] propose a comparative study of cost estimation models for Web applications. The study compares the prediction accuracy of three CBR (case-based reasoning) techniques to estimate the effort to develop Web applications and to choose the one with the best estimates, starting from a data set that includes several size measures. It also compares the prediction accuracy of the best CBR technique against two prediction models built with stepwise regression and regression trees, and concludes that stepwise regression produces the best predictions.

The use of early size measures for the estimation of the development cost of Web application is investigated by Fewster and Mendes [2003]. Having early size measures that are related to cost would provide Web development companies with a competitive advantage in pricing their Web applications. A survey was carried out with data from 133 companies from several countries, to identify the main drivers that companies use to quote a price for their Web applications.

By far, the total number of Web pages and which features/functionalities turned out to be most used pieces of information for pricing Web applications. The same paper also addressed the comparison of the accuracy of company specific models vs. estimation models built with data from several different companies. Mendes and Kitchenham [2004] show that a company-specific model may provide a higher accuracy.

Kitchenham and Mendes [2004] study productivity measurement with multiple size measures, for different aspects of size. The idea is that productivity can be defined as the ratio between the estimated effort for a project to the actual effort. The estimated effort is computed via a regression-based effort estimation model in which all the size measures used as independent variables have a statistically significant impact on effort. The expected value of this indicator for a project is 1. Thus, though introducing a productivity notion that somewhat differs from the usual one, the approach comes with a built-in assessment criterion for the productivity of a project: if a project has a productivity index lower than 1, then its productivity is lower than expected; if the productivity index is greater than 1, its productivity is higher than expected. The approach is applied to the Web application projects of the Tukutuku data set.

Costagliola et al. [2004] adopt a different approach. They propose the COSMIC-FFP (Cosmic Full Function Points, [COSMIC 2003]–a variant of Function Points [Albrecht and Gaffney Jr. 1983; Garmus and Herron 2001]–to estimate the development effort of Web applications. Mendes et al. [2002b] provided a formal method to adopt COSMIC-FFP to measure the size of static hypermedia Web applications, while Costagliola et al. [2004] extend it to dynamic applications. The proposal starts from the UML profile proposed by Conallen [2002] for modeling Web applications and defines rules to measure the functional size of client-server applications. The empirical evaluation is based on 32 projects from students attending a Web engineering course, and is performed by using an Ordinary Least-Squares regression analysis. The linear regression analysis shows a high adjusted $R^2 = 0.763$, which indicates that 76.3% is the variance of the dependent variable *effort* that is explained by the model. Moreover, a high $F\ value$ and a low $p-value$ indicate that the prediction is possible with a high degree of confidence. This proposal and our approach share "adjusted" UML diagrams as starting point and students as subjects. The main difference is that we want to investigate the relations between significant elements of W2000 models and the global effort, while this proposal adapts a well-known sizing model to a specific modeling notation for Web applications.

Abrahão et al. [2003] define and validate measures for size and structural complexity of navigational models. After a theoretical validation of the measures, an empirical study was carried out with students. The dependent attribute chosen was the maintainability of navigational models, quantified in terms of both maintenance time and measures of the ISO9126 maintainability sub/characteristics. The measures for size and structural complexity turned out to be correlated or associated with the dependent measures.

Abrahão et al. [2004] validate OOmFPWeb as a functional size measurement method for Web applications *per se*, but it is not investigated whether it is also a predictor of Web development effort. An empirical investigation is carried

out to assess, for example, the estimated and the actual ease of use and the effectiveness of the size measure.

Umbers and Miles [2004] address the estimation of Web development effort and use COSMIC-FFP as the sizing measure. Starting from unadjusted size and productivity estimates, multiplicative coefficients are used to obtain adjusted estimates for size and productivity.

Mendes et al. [2002b] propose a case study in which they describe size metrics to characterize length, complexity, and functionality of Web applications and generate effort prediction models. Results suggest that in general all categories present a similar prediction accuracy, but none of them had the required statistical significance level for the absolute residuals. They counted the number of COSMIC-FFP functional size units using the final implementation and conclude that measuring the functionality of those applications was very much related to the number of links that the application has. But the low statical significance suggests that further investigation is needed to determine whether functionality is correlated to connectivity.

Reifer (WebMo, [Reifer 2000; 2002]) tailors the COCOMO II model to Web development by revising the cost factors. Notice that WebMo introduces *Web objects* as a size measure for these applications. *Web objects* are an extension and adaptation of Function Points to Web applications, by taking into account additional features that are typical of Web applications such as multimedia files, web building blocks, scripts, and links, which are added to the five function types of the original Function Point approach.

Ruhe et al. [2003a] propose an empirical validation of Web Objects for the estimation of the development effort for Web applications. The empirical study was carried out on twelve projects from a small web development company. By using leave-one-out cross-validation, twelve OLS models (each using eleven projects) were built with Web objects and twelve models were built with Function Points. The accuracy of the models obtained with Web Objects was compared to the accuracy of the models obtained with Function Points in terms of Mean Magnitude of Relative Error. The difference between the MMRE's turned out to be statistically significant and showed an advantage of Web Objects over Function Points. The accuracy of expert opinion was also compared against that of the OLS models. Though Web Object-based models appeared to be more accurate than expert opinion, the results were not statistically significant.

Mangia and Paiano [2003] propose a sizing model for Web applications deeply based on W2000. They concentrate on a limited number of W2000 features and suggest a hierarchy of sizing models to measure the functionality, navigational capabilities, presentation features, and multimedia contents of W2000-based Web applications. The authors do not document any empirical study: They only propose their sizing model and claim, in the Conclusions, that its application on real projects always had a Magnitude of Relative Error of less than 12%.

Via the replication of previous studies, Mendes et al. [2003a] show how adaptation rules and Feature Subset Selection can be used in analogy-based estimation of web cost to improve the accuracy of estimation for datasets characterized by the absence of outliers and small collinearity. The results were not as good on a dataset with different characteristics.

Mendes and Kitchenham [2004] investigate whether it is useful to use a cost model built on Web applications taken from several companies to estimate the cost of the projects of a company that were not used to build the cost model. The study shows that the cost model provided much worse results than a cost model built based on projects coming from the same company. Also, the cost model built based on the project from that company provided poor results when applied to the projects of the other companies. This provides evidence that cost models for Web applications cannot be reused "as-is" across companies. The investigation confirms that more studies are needed before a generally applicable cost model for Web applications can be found and used (if that is even a realistic goal). In addition, the paper says that "All company estimates were underestimates." which shows that the underestimation phenomenon is common to software professional and students.

Similarly, Ruhe et al. [2003b] propose an adaptation of the COBRA method (COst estimation, Benchmarking, and Risk Assessment [Briand et al. 1998]) to the Web domain and apply it to the data from 12 projects developed by a small Australian company.

## 9. CONCLUSIONS AND FUTURE WORK

In this article, we have illustrated three empirical studies carried out on the effort required to design web applications. Students are not perfect subjects, but the peculiarities of the problem domain allowed us to obtain interesting results. Moreover, the different background and curricula of our subjects helped us obtain a good mix of the professionals usually involved in the development of Web applications.

Because of the relative novelty of the field, our empirical studies were exploratory, rather than confirmatory ones. We investigated a number of hypotheses that seemed likely to be true based on our beliefs and knowledge about the W2000 notation, the subjects' skills, and the steps of the design process used. Nevertheless, the hypotheses and results may apply to a wider class of modeling notations for Web applications, which can be viewed as special-purpose customizations (i.e., profiles) of UML.

Though with some inconsistencies across the empirical studies, a number of predictors have been identified, so our studies can be used as a starting point for further research in the field. As already stated by other similar studies, we need further empirical studies to identify further predictors that may explain a larger percentage of effort variance than is now possible.

Therefore, future work will include further empirical studies with students that will help us confirm the hypotheses that were supported in our research by providing even more evidence. Empirical studies with software professionals will be carried out, based on the current and future results. This will allow us to build prediction models that are practically useful for software companies. We do not believe that the prediction models obtained for one companies will be readily reusable in other contexts. However, by carrying our empirical studies in a number of industrial contexts and studying their similarities and differences, we will identify not only individual predictors for effort, but build possibly

multivariate prediction models. The future studies will address other notations, in addition to W2000, because it will be necessary to use the specific notations used by the software companies involved in the empirical studies. It is expected that most of them can be viewed as UML profiles, so they will share a number of common features. During this process, it may be necessary to investigate and define new measures because the existing ones may not be fully applicable to all of these other notations and because these other notations will have additional constructs that will be useful to measure. In addition, new measures may help us better quantify internal software attributes than the measures we have used in this article.

## ACKNOWLEDGMENTS

We wish to thank the anonymous referees for their useful suggestions that have helped us improve the article.

## REFERENCES

ABRAHÃO, S. M., CONDORI-FERNÁNDEZ, N., OLSINA, L., AND PASTOR, O. 2003. Defining and validating metrics for navigational models. In *Proceedings of the 9th IEEE International Software Metrics Symposium (METRICS 2003)* (Sydney, Australia, Sept. 3–5). IEEE Computer Society Press, Los Alamitos, CA, 200–210.

ABRAHÃO, S. M., POELS, G., AND PASTOR, O. 2004. Evaluating a functional size measurement method for web applications: An empirical analysis. In *Proceedings of the 10th IEEE International Software Metrics Symposium (METRICS 2004)* (Chicago, IL, Sept. 11–17). IEEE Computer Society Press, Los Alamitos, CA, 358–369.

ALBRECHT, A. J. AND GAFFNEY, JR., J. E. 1983. Software function, source lines of code, and development effort prediction: A software science validation. *IEEE Trans. Softw. Eng. 9,* 6, 639–648.

BARESI, L., MORASCA, S., AND PAOLINI, P. 2002. An empirical study on the design effort of web applications. In *Proceedings of the 3rd International Conference on Web Information Systems Engineering (WISE 2002)* (Singapore, Dec. 12–14). IEEE Computer Society Press, Los Alamitos, CA, 345–354.

BARESI, L., MORASCA, S., AND PAOLINI, P. 2003. Estimating the design effort of web applications. In *Proceedings of the 9th IEEE International Software Metrics Symposium (METRICS 2003)* (Sydney, Australia, Sept. 3–5). IEEE Computer Society Press, Los Alamitos, CA, 62–72.

BARESI, L., COLAZZO, S., MAINETTI, L., AND MORASCA, S. 2006. Web engineering, W2000: A modeling notation for complex Web applications. In *Web Engineering*, Mendes, E. and Mosley, N. Eds., Springer-Verlag, 335–364.

BOEHM, B. M., HOROWITZ, E., MADACHY, R., REIFER, D., CLARK, B. K., STEECE, B., BROWN, A. W., CHULANI, S., AND ABTS, C. 2000. *Software Cost Estimation with Cocomo II*. Prentice-Hall, Upper Saddle River, NJ.

BOEHM, B. W. 1981. *Software Engineering Economics*. Prentice-Hall, Upper Saddle River, NJ.

BRIAND, L., EL EMAM, K., AND BOMARIUS, F. 1998. COBRA: A hybrid method for software cost estimation, benchmarking, and risk assessment. In *Proceedings of the 20th International Conference on Software Engineering* (Kyoto, Japan, Apr. 19–25). IEEE Computer Society Press, Los Alamitos, CA, 390–399.

CALERO, C., RUIZ, J., AND, PIATTINI, M. 2004. A web metrics survey using WQM. In *Web Engineering—4th International Conference, ICWE 2004* (Munich, Germany, July 26–30), *Proceedings*. Lecture Notes in Computer Science, vol. 3140. Springer-Verlag, Berlin, Germany, 147–160.

CARVER, J., JACCHERI, L., MORASCA, S., AND SHULL, F. 2003. Issues in using students in empirical studies in software engineering education. In *Proceedings of the 9th IEEE International Software Metrics Symposium (METRICS 2003)* (Sydney, Australia, Sept. 3–5). IEEE Computer Society Press, Los Alamitos, CA, 239–248.

CERI, S., FRATERNALI, P., AND BONGIO, A. 2000. Web modeling language (WebML): A modeling language for designing Web sites. *Comput. Netw. 33,* 1–6 (June), 137–157.

CONALLEN, J. 2002. *Building Web Applications with UML*, 2nd ed. Addison-Wesley, Boston, MA.

COSMIC. 2003. *COSMIC-FFP Measurement Manual, vers. 2.2*. Common Software Measurement International Consortium. http://www.cosmicon.com.

COSTAGLIOLA, G., FERRUCCI, F., GRAVINO, C., TORTORA, G., AND VITIELLO, G. 2004. A COSMIC-FFP based method to estimate web application development effort. In *Web Engineering—4th International Conference, ICWE 2004* (Munich, Germany, July 26–30). Lecture Notes in Computer Science, vol. 3140. Springer-Verlag, Berlin, Germany, 161–165.

DHYANI, D., NG, W. K., AND BHOWMICK, S. S. 2002. A survey of web metrics. *ACM Comput. Surv. 34,* 4, 469–503.

FENTON, N. 1991. *Software Metrics: A Rigorous Approach*. Chapman and Hall, London, UK.

FENTON, N. AND PFLEEGER, S. L. 1996. *Software Metrics: A Rigorous and Practical Approach*, 2nd ed. International Thomson Computer Press, London, UK.

FEWSTER, R. M. AND MENDES, E. 2003. Portfolio management method for deadline planning. In *Proceedings of the 9th IEEE International Software Metrics Symposium (METRICS 2003)* (Sydney, Australia, Sept. 3–5). IEEE Computer Society Press, Los Alamitos, CA, 325–334.

GARMUS, D. AND HERRON, D. 2001. *Function point analysis: Measurement Practices for Successful Software Projects*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA.

GARRETT, J. J. 2002. *The Elements of User Experience: User-Centered Design for the Web*. New Riders Press, Berkeley, CA.

GARZOTTO, F., SCHWABE, D., AND PAOLINI, P. 1993. HDM—A model based approach to hypermedia application design. *ACM Trans. Inform. Syst. 11,* 1 (Jan.), 1–26.

GIBBONS, J. D. 1993. *Nonparametric Measures of Association*. SAGE Publications, Thousand Oaks, CA.

GINIGE, A. AND MURUGESAN, S. 2001. Guest editors' introduction: Web engineering—an introduction. *IEEE MultiMedia 8,* 1, 14–18.

KAPPEL, G., PRÖLL, B., REICH, S., AND RETSCHITZEGGER, W. 2006. *Web Engineering: The Discipline of Systematic Development of Web Applications*. Wiley, New York, NY.

KITCHENHAM, B. A. AND MENDES, E. 2004. Software productivity measurement using multiple size measures. *IEEE Trans. Softw. Eng. 30,* 12, 1023–1035.

KNAPP, A., KOCH, N., ZHANG, G., AND HASSLER, H.-M. 2004. Modeling business processes in web applications with argouwe. In *Proceedings of UML 2004—The Unified Modelling Language: Modelling Languages and Applications. 7th International Conference* (Lisbon, Portugal, Oct. 11–15). Lecture Notes in Computer Science, vol. 3273. Springer-Verlag, Berlin, Germany, 69–83.

MANGIA, L. AND PAIANO, R. 2003. MMWA: A software sizing model for web applications. In *Proceedings of the 4th International Conference on Web Information Systems Engineering (WISE 2003)* (Rome, Italy, Dec. 10–12). IEEE Computer Society Press, Los Alamitos, CA, 53–61.

MENDES, E., COUNSELL, S., AND MOSLEY, N. 2005a. Towards a taxonomy of hypermedia and web application size metrics. In *Web Engineering, 5th International Conference, ICWE 2005* (Sydney, Australia, July 27–29). Lecture Notes in Computer Science, vol. 3579. Springer-Verlag, Berlin, Germany, 110–123.

MENDES, E. AND KITCHENHAM, B. A. 2004. Further comparison of cross-company and within-company effort estimation models for web applications. In *Proceedings of the 10th IEEE International Software Metrics Symposium (METRICS 2004)* (Chicago, IL, Sept. 11–17). IEEE Computer Society Press, Los Alamitos, CA, 348–357.

MENDES, E. AND MOSLEY, N. 2006. *Web Engineering*. Springer-Verlag, Berlin, Germany.

MENDES, E., MOSLEY, N., AND COUNSELL, S. 2002a. Comparison of web size measures for predicting web design and authoring effort. *IEE Proc.—Softw. 149,* 3, 86–92.

MENDES, E., MOSLEY, N., AND COUNSELL, S. 2003a. A replicated assessment of the use of adaptation rules to improve web cost estimation. In *Proceedings of the 2003 International Symposium on Empirical Software Engineering (ISESE 2003)* (Rome, Italy, Sept. 30–Oct. 1). IEEE Computer Society Press, Los Alamitos, CA, 100–109.

MENDES, E., MOSLEY, N., AND COUNSELL, S. 2005b. Investigating web size metrics for early web cost estimation. *J. Syst. Softw. 77,* 2, 157–172.

MENDES, E., MOSLEY, N., AND COUNSELL, S. 2006. *Web Engineering*. Springer-Verlag, Berlin, Germany (Chapter: Web Effort Estimation), 31–76.

MENDES, E., WATSON, I., TRIGGS, C., MOSLEY, N., AND COUNSELL, S. 2002b. A comparison of length, complexity and functionality as size measures for predicting web design and authoring effort. In *Proceedings of the IEEE Metrics Symposium, 2002*. IEEE Computer Society Press, Los Alamitos, CA.

MENDES, E., WATSON, I., TRIGGS, C., MOSLEY, N., AND COUNSELL, S. 2003b. A comparative study of cost estimation models for web hypermedia applications. *Empir. Softw. Eng. 8,* 2, 163–196.

MURUGESAN, S., AND DESHPANDE, Y. 2002. Meeting the challenges of web application development: The web engineering approach. In *Proceedings of the 22rd International Conference on Software Engineering, ICSE 2002* (Orlando, FL, May 19–25). ACM, New York, 687–688.

REIFER, D. 2000. Web-development: Estimating quick-time-to-market software. *IEEE Softw. 17,* 8 (Nov./Dec.), 57–64.

REIFER, D. 2002. Ten deadly risks in internet and intranet software development. *IEEE Softw. 18,* 2 (Mar./Apr.), 12–14.

ROUSSEEUW, P. J. AND LEROY, A. M. 1987. *Robust Regression and Outlier Detection*. Wiley, New York.

RUHE, M., JEFFERY, D. R., AND WIECZOREK, I. 2003a. Using web objects for estimating software development effort for web applications. In *Proceedings of the 9th IEEE International Software Metrics Symposium (METRICS 2003)* (Sydney, Australia, Sept. 3–5). IEEE Computer Society Press, Los Alamitos, CA, 30–39.

RUHE, M., JEFFERY, R., AND WIECZOREK, I. 2003b. Cost estimation for web applications. In *Proceedings of the 25th International Conference on Software Engineering* (Portland, OR, May 3–10). IEEE Computer Society Press, Los Alamitos, CA, 285–294.

SCHWABE, D. AND ROSSI, G. 1998. An object oriented approach to web-based applications design. *Theory Pract. Obj. Syst. 4,* 4, 207–225.

TUKUTUKU BENCHMARKING PROJECT 2006. Home page, http://www.cs.auckland.ac.nz/tukutuku/.

UMBERS, P. AND MILES, G. 2004. Resource estimation for web applications. In *Proceedings of the 10th IEEE International Software Metrics Symposium (METRICS 2004)* (Chicago, IL, Sept. 11–17). IEEE Computer Society Press, Los Alamitos, CA, 370–381.

Received February 2006; revised August 2006; accepted December 2006