

An Empirical Study on the Design Effort of Web Applications*

Luciano Baresi*, Sandro Morasca⁺, and Paolo Paolini*

(*) Dipartimento di Elettronica e Informazione - Politecnico di Milano
piazza Leonardo da Vinci, 32 - 20133 Milano (Italy)

baresi@elet.polimi.it, paolini@elet.polimi.it

(⁺) Dipartimento di Scienze Chimiche, Fisiche e Matematiche - Università degli Studi dell'Insubria
via Valleggio, 11 - 22100 Como (Italy)

sandro.morasca@uninsubria.it

Abstract

We study the effort needed for designing Web applications from an empirical point of view. The design phase takes an important part of the overall effort needed to develop a Web application, since the use of tools can help automate the implementation phase. We carried out an empirical study with the students of an advanced university class that used W2000 as a Web application design technique. Our first goal was to compare the relative importance of each design activity. Second, we tried to assess the accuracy of a priori design effort predictions and the influence of some factors on the effort needed for each design activity. Third, we also studied the quality of the designs obtained.

Keywords: Web design, Web metrics, W2000, Empirical Software Engineering

1. Introduction

In the last few years Web applications have been evolving from simple sites to fully distributed applications [2]. Sites were mainly aids to advertise products, institutions, and events. Nowadays applications let the user play an active role and are the key component to run the business of several companies.

In addition to their evolution, which would be enough to motivate a different attitude towards developing these applications, the time-to-market is imposing even faster development processes. The "code&fix" approach used by Web gurus in the early days is not good enough any longer. A more structured development process should be the base to

develop quality applications on an industrial scale. Among the different aspects that the process should cover, we must take into account the capability of predicting the complexity of the to-be-developed applications. This would help both shorten the development process and better allocate required resources.

In this paper, we illustrate an experimental study on the effort for designing Web applications, mainly its underlying site. We involved the students of an advanced university class on modeling Web applications to help us evaluate the design effort and the relationships with the used notation, tools, and designers' expertise. The subjects received the specifications of a pilot application and were asked to model it with W2000 [1], the modeling notation specific to Web applications developed at Politecnico di Milano. We also asked them to fill out a questionnaire before starting the design, to set their expertise and make them estimate the effort required during the various activities of the design phase. The subjects also filled out another questionnaire while completing the model to measure directly the actual effort.

Our experimental study is an exploratory one, and has a number of goals. Our first goal was to compare the relative importance of each design activity. The idea is that the knowledge of the ranking in effort among the design activities may help (1) identify possible improvement points in the design process and (2) better plan the design phase. Second, we wanted to assess the usefulness of the subjects' estimates in predicting the actual effort for the design phases. We also investigated factors that may influence the estimated and actual effort for each design activity. Third, we also studied the quality of the designs obtained and checked if there was any correlation between the subjects' self-assessment and the grades given by the instructor.

*This work has been partially supported by the ESERNET (IST-2000-28754) and UWA (IST-2000-25131) EC-IST Projects.

Being an exploratory study, its main objective is to identify hypotheses that deserve further attention, and not to confirm a well-established theory.

Despite the growing importance of Web applications, the state of the art of empirical studies in effort estimation of Web applications is still very preliminary, due to their relative novelty. For instance, papers [5, 6] illustrate two case studies on the evaluation of the development of 76 and 37 applications by using length and complexity metrics in Ordinary Least Square prediction models.

The rest of this paper is organized as follows. Section 2 briefly introduces W2000 to let readers understand the modeling features supplied to students. Section 3 describes the experimental setting and the hypotheses we investigated. Section 4 presents and discusses the results we obtained. Finally, Section 5 concludes the paper and summarizes our future work.

2. W2000

W2000 [1] is the latest evolution of HDM (Hypermedia Design Model, [4]). It extends the original proposal according to two main directions. Since Web applications are not read-only information repository anymore, roughly speaking, W2000 lets the user specify operations to add, delete, and modify stored data. Moreover, the whole extended notation comes with a UML-like ([3]) concrete syntax to better emphasize the similarities with object-oriented languages and supply users with a standard way to design their models.

In this paper, we do not use its newer modeling features, but we use W2000 as specification tool for conventional Web sites. According to this use, W2000 comprises three main models¹:

- The **Information model** specifies the contents (data) available to the user (*hyperbase*) and how the user can access it (*access structures*).
- The **Navigation model** rearranges the contents into chunks suitable for letting the user navigate through them and supplies the links among these elements.
- The **Presentation model** organizes the outcome from the navigation design into pages and links that are visible to the user.

Entities, the key element of the hyperbase, render data of interest as conceptual aggregations. They can be *single entities*, that is, ad-hoc singleton information elements, or *entity types*, which define templates in the same way classes do for objects. For example, if we think of an e-library, a

¹In this paper we comply with the OMG jargon, where a model defines a view on the whole specification.

typical single entity is the one that defines the information to be displayed in the first page to present the site; *book* and *author* would be other obvious entity types (Figure 1(a)).

Entities, the key element of the hyperbase, render data of interest as conceptual aggregations. They can be single entities, that is, ad-hoc singleton information elements, or entity types, which define templates in the same way classes do for objects. For example, if we think of an e-library, a typical single entity is the one that defines the information to be displayed in the first page to present the site; book and author would be other obvious entity types (Figure 1(a)).

Entities can be organized in generalization hierarchies and are specified through sets of *components*, which are pure organizational devices for grouping the contents of an entity into meaningful chunks. Components can further be decomposed in sub-components, but the actual contents can be associated with leaf nodes only. The result is a part-of hierarchy with an entity as root. In the example in Figure 1(a), a *book* could be defined through two components: one for the editorial information (title, author, publisher, price), and one for the summary. The contents of (leaf) components is specified in terms of *slots*, i.e., the attributes that define the primitive information elements. Again, the slots of Editorial Info components could be: title, author name, publisher name, number of pages, publication year, price, but also the thumbnail picture of the book and maybe a small picture of the author. Slots are not shown when we think of *in-the-large* specifications, since we define here only the overall structures of our models, but must be specified in *in-the-small* specifications². Slots are grouped in *segments* to foresee how the contents will be "consumed" by the user. For example a visual segment could comprise all those slots that must be displayed each time a book is presented in a single page. Notice that different slots could be used in different circumstances.

The relationships among entities are specified through *semantic associations*. They connect two entities to create the "infrastructure" for a possible navigation path, which has to be further specified in the navigation model. For example, a semantic association could be defined between *author* and *book* to trace who wrote what (Figure 1(a)). This means also that it will be possible to navigate from the page of a given author to the list of all his or her books.

Semantic associations have also proper, local, information, called *association centers*, which contain data to specify how to represent both single target elements, in a concise way, and the whole group of target elements that relate to the same source. For example, the center of the previous semantic association would contain all slots (preview segment) to render a single book in the page that lists all the

²A similar distinction between in-the-large and in-the-small specifications applies to all W2000 models. Roughly we can say that the former identify the structure while the latter add low-level details.

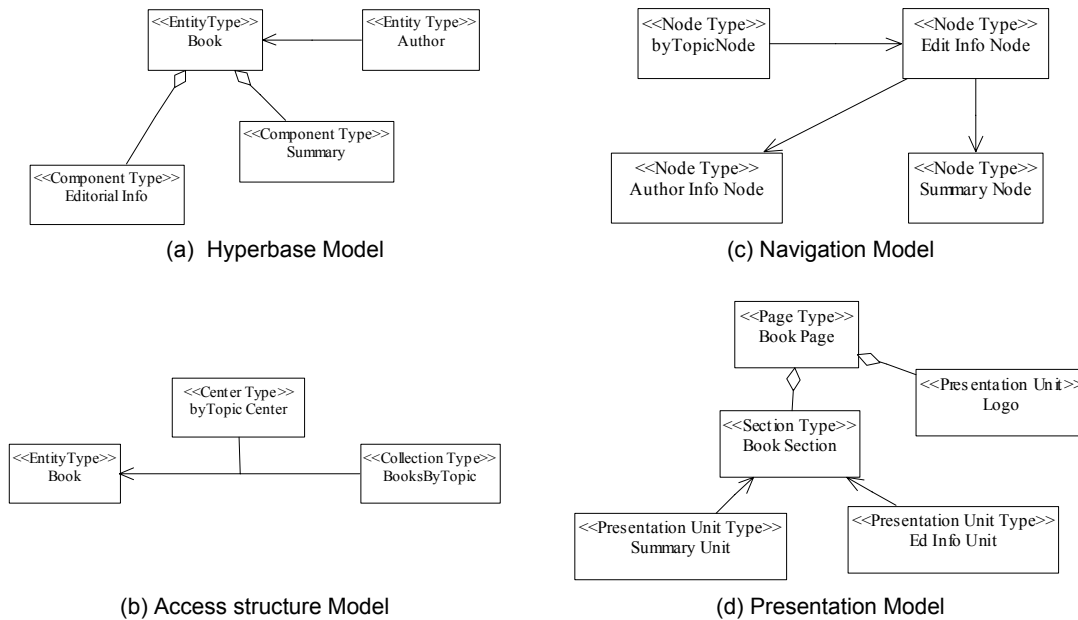


Figure 1. Excerpts of a simple W2000 model

books written by an author, but also information to characterize the page itself as a whole, say a brief introduction to the author's work.

Access structures are defined by organizing entities in *collections*, i.e., sets of information objects. A collection provides the user with a way to explore the application's contents. For example, besides accessing books by navigating from the page dedicated to the author, they could be organized by topic (the collection *booksByTopic* of Figure 1(b)) to browse through all available books ordered by topic. Otherwise, they could be organized by publisher, leading to the collection *booksByPublisher*, or there could be a collection *booksByAuthor* to list all books according to the alphabetical order of their authors. Each collection has a *center*, which, similarly to association centers, defines how to concisely represent the elements in the collection (once more, we could use the preview segment associated with the collection's contents).

The *Navigation model* reshapes the elements in the previous model to specify the *actual* information chunks. The information is organized in atomic units, called *nodes*: They do not define new contents, but come from entity components, semantic association, and collection centers. They contain the slots associated with the information elements they render. The simplest approach is that components, association centers and collection centers become nodes, but more sophisticated navigation models could require a finer granularity; thus information elements could be rendered through several nodes.

Two nodes are linked through a directed *accessibility relationship* to specify that the user can navigate from the source to the target node. In our e-library (Figure 1(c)), we could define a *byTopicNode* to render the collection; it would become the starting point to move to the books by means of *EditInfoNodes*. From these nodes, we could browse to the author info (*AuthorInfoNode*) and the summary (*SummaryNode*).

Nodes exist in the context of a *navigation cluster* that groups nodes and accessibility relationships to foster and facilitate the navigation among data (nodes). For example, different clusters originate from different collections of the same elements. The next element, and thus the instance of the accessibility relationship that should be traversed, could be different if books were ordered by author or by publisher. Clusters can be nested and can be further characterized according to the kind of information they render:

- *Structural clusters* consist of all the nodes derived from entity components;
- *Semantic clusters* comprise all nodes that come from sources, targets and centers of semantic associations;
- *Collection clusters* comprise all nodes that come from members and centers of collections;

In the simplest case, we could skip the presentation model and use the navigation model for this purpose.

The underlying assumption is that the application has been conceived for a single and specific channel and the

identity is the mapping function between navigation and presentation.

If we wanted a more sophisticated presentation model, *presentation units* are the smallest granules at this level. They can either come from nodes or add new contents that is defined at presentation level only for aesthetic/communication purposes. For example, the logo of the e-library does not need to be defined at the information level; it could simply be defined here. A *section* is a set of presentation units derived from nodes that belong to the same cluster. A page groups sections, even if they are not semantically related, from which it inherits links and navigation features.

Presentation units, sections, and pages can either be singleton ad-hoc elements or define types that can be instantiated as many times as necessary. These elements can also be sources or targets of *presentation links*, that is, a connection between two presentation elements to enable the navigation between them. According to the aforementioned concepts, we can further classify the links in a page as:

- *Focus links* to remain in the same page, but moving the page focus from a unit to another;
- *Intra-page links* to navigate between instances of the same page type;
- *Page links* to navigate between instances of different page types.

Figure 1(d) gives an excerpt of the pages for the e-library: Books are rendered to users through pages (*Book-Page*) that comprise the library logo – added in this model – and a section with all information about books. The section itself contains the editorial information (*EdInfoUnit*) and the summary (*SummaryUnit*).

3 Experimental Setting

The experiment documented in this paper was carried out at the Politecnico di Milano with students attending a class on how to design and implement advanced Web applications. They were taught to design their applications with W2000 and then implement them with the many available technologies. Since they were completely new to W2000, the experiment was not only aimed at measuring the design effort, but also – implicitly – the learning curve for W2000.

The experiment started with assigning the projects to the students. Roughly, all students were asked to work on the same project, that is, a hypothetical e-commerce application; what changed from project to project was the application domain: books, CDs, groceries, etc.

After reading the requirements, which were written in an informal style, we asked the students to fill out a questionnaire to:

- Acquire information on their general proficiency in computer-science-related college courses,
- Acquire information on their expertise on Web technologies and design methods, and
- Make them estimate the overall design effort, trying to split it according to the main models required by W2000.

We asked them also to fill a second questionnaire while completing their homework to report the actual effort spent in the different phases/models, list the tools they used³, and self-evaluate the quality of their work.

4 Results

Our experimental study is an exploratory one, rather than a confirmatory one. We investigated a number of experimental 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. Our main goals were related to discovering

- Which of these hypotheses were supported by empirical evidence in the context of our study and deserve further investigations;
- Which were not, and should therefore be either revisited (there might have been some reason why these hypotheses were not true in our experimental setting), or not investigated any longer (their likelihood is so small that they are probably not true in general).

We believe that it would be much too early to carry out a confirmatory study. As explained in the Introduction, the body of knowledge available in the literature about the effort related to developing Web applications, and especially the design phase, is very limited. Therefore, we report on both the hypotheses that our study confirmed and the hypotheses that were not supported by our study, so that our study can be taken as a starting point for future experimental activities.

As a first step in each of our data analyses, we checked the data points and we removed those that appeared to be corrupted, or clearly incorrect, or for which there was missing information, e.g., either the value of the independent variable or the value of the dependent variable was missing. As a second step in each data analysis, we carried out a very careful outlier analysis, i.e., we removed those few data points that were much too "far" from the others. This is a standard data analysis activity: it is carried out to

³It must be noticed that currently W2000 is not supported by any special-purpose modeling tool. Users are free to use what they prefer, but this could highly impact the quality and spent effort.

remove those few points that may unduly bias the results. The removal of outliers is absolutely necessary especially in exploratory studies like the one documented in this paper because of the current stage of quantitative knowledge on Web applications.

We now report on the experimental hypotheses we checked in our study and the results we obtained. The hypotheses we describe below are the so-called "alternative hypotheses" in the test of hypotheses. For brevity's sake, we do not report the so-called "null hypotheses," which can be obtained as the logical negation of the alternative hypotheses. The presentation of our experimental results is organized as follows:

- Rationale and illustration of our hypotheses,
- Experimental results,
- Discussion.

Our data analyses can be classified in two categories: (1) analyses related to the comparisons of distributions and (2) analyses related to correlations between random variables.

1. When comparing two distributions, we use the following statistics to illustrate the results:

- N , the number of data points of the distribution,
- M , the median of the distribution,
- n , the mean value of the distribution,
- σ , the standard deviation of the distribution,
- p , the statistical significance (p-value) of the hypothesis according to which the median of one of the distributions is greater than the median of the other distribution.

2. We used Ordinary Least Squares (OLS) to find out possible correlations among independent and dependent variables. The following statistics are used to illustrate the experimental results on correlations:

- N , the number of data points used to build the model; N changes from model to model because a different number of outliers are excluded from different models; this number provides an idea of the statistical basis on which our results are based;
- R^2 , which measures the goodness-of-fit of the model as the percentage of variance that is explained by the model;
- E , the estimates of the regression coefficients, one for each independent variable X in the model plus one for the intercept;

- σ , the standard deviations of the regression coefficients;
- p , which is the statistical significance of each independent variable X in the OLS model: given an independent variable X , its p-value provides an idea of the probability that X has an impact on the dependent variable Y by chance, so the smaller the value of the p-value, the more likely that X really has an impact on Y ; it is usual to consider statistically significant the impact on Y of those independent variables X for which $p < 0.05$, i.e., there is less than 5% probability that they have an impact on Y by chance.

We now report on the hypotheses we studied. However, in our study, we had to exclude the effort related to the presentation part, since few respondents provided data for the actual presentation effort. At any rate, as explained in Section 2, a presentation model is not always needed, unless a sophisticated kind of presentation is required. Thus, this exclusion may not have exceedingly biased our results.

Actual Information Effort vs. Actual Navigation Effort: distributions

Rationale We wanted to identify the model that takes the largest amount of time to be designed in general and for each respondent. Thus, we checked the following hypothesis.

Hypothesis 1 The median of the actual effort related to the information model (*ActInfoEff*) is higher than the median of the actual effort related to the navigation model (*ActNavEff*).

In this data analysis and the following ones on the comparison of distributions, we used the median and not the mean because we wanted to use statistics that were less sensitive to possible data approximations in the data provided by the students. As a consequence, we stayed on the safe side, and we used statistical tests (known as "non-parametric" tests) that do not depend on any specific hypotheses on the data distributions. These tests are less powerful than those (known as "parametric" tests) that we could have used with the means of the distributions, i.e., it is less likely to show that a hypothesis is statistically significant with a non-parametric test than with a parametric test. At any rate, we would like to note that all of the following results would have been obtained by stating the hypotheses in terms of the means and by using parametric tests.

Results The following table summarizes the statistics of our results. We used 49 data points, since we had 49 re-

spondents for both *ActInfoEff* and *ActNavEff*. All the actual and estimated effort data represent work hours.

Variables	<i>N</i>	<i>M</i>	<i>m</i>	σ	<i>p</i>
<i>ActInfoEff</i>	49	15	18.4	10.2	< 0.0001
<i>ActNavEff</i>		8	10.4	6.7	

Descriptive statistics for *ActInfoEff* and *ActNavEff*

The data show that median of *ActInfoEff* is much greater than the median of *ActNavEff*. To make sure that this circumstance is statistically significant, we used the Mann-Whitney test for the medians. Since this probability is less than 0.0001, as shown in the *p* column of the previous table, we can conclude that our hypothesis is certainly acceptable at the 0.05 level. This is also supported by the fact that 45 out of 49 subjects reported a value of *ActInfoEff* greater than *ActNavEff*. By using Wilcoxon’s matched pairs signed rank test we obtained even better significance results.

Discussion The information model is definitely more time consuming than the navigation model. This confirmed our initial hypothesis, which was based on the fact that students appeared to have greater problems in designing the underlying structure and data of the application than the navigation model, since they are certainly more used to navigating on the Web than building systems and therefore they have a better intuition about what needs to be done. The building of the information model precedes the building of the navigation model, so the subjects were probably more concerned with starting right than sketch a solution whose problems would show up later in the other design activities.

Estimated Effort vs. Actual Effort: distributions

Rationale The hypothesis we studied was whether the subjects were overly optimistic, i.e., they tended to underestimate the effort needed to design an application.

Hypothesis 2 For each effort category, the median of the actual effort is higher than the median of the estimated effort.

Results We first study the distributions of the 44 subjects that provided valid data for computing both the total estimated effort (*EstEff*) and total actual effort (*ActEff*). The distributions are shown in Figure 2.

The statistics for the comparison of *ActEff* and *EstEff* are shown in the following table:

Variables	<i>N</i>	<i>M</i>	<i>m</i>	σ	<i>p</i>
<i>EstEff</i>	44	15.5	18.1	9.4	0.0006
<i>ActEff</i>		26.75	29.6	15.5	

Statistics for *EstEff* and *ActEff*

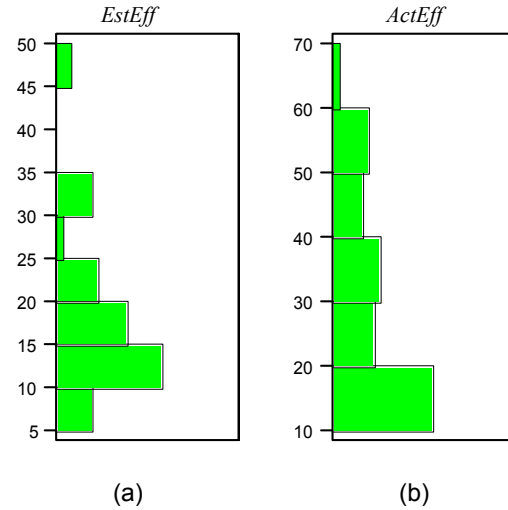


Figure 2. Distributions of *EstEff* and *ActEff*

The median value of *ActEff* is much greater than that of *EstEff*. Since 33 subjects provided a value of *ActEff* greater the median of the distribution of *EstEff*, the statistical test described for the previous hypothesis shows that the *p*-value of Hypothesis 2 is 0.0006, so we can conclude at the 0.05 level that the data show that the median of *ActEff* is greater than the median of *EstEff*, i.e., the estimates are overly optimistic. This is also confirmed by the fact that a large majority of subjects (36 out of 44) reported a greater value for *ActEff* than *EstEff*.

For the single categories, the descriptive statistics are presented in the following tables:

Variables	<i>N</i>	<i>M</i>	<i>m</i>	σ	<i>p</i>
<i>EstDataEff</i>	44	5	6.4	3.6	< 0.001
<i>ActDataEff</i>		9	10.3	5.6	

Statistics for *EstDataEff* and *ActDataEff*

Variables	<i>N</i>	<i>M</i>	<i>m</i>	σ	<i>p</i>
<i>EstStructEff</i>	44	5	5.4	3.3	0.003
<i>ActStructEff</i>		7	8.7	5.3	

Statistics for *EstStructEff* and *ActStructEff*

Variables	<i>N</i>	<i>M</i>	<i>m</i>	σ	<i>p</i>
<i>EstInfoEff</i>	44	9.5	11.7	6.5	< 0.001
<i>ActInfoEff</i>		15	19	10.3	

Statistics for *EstInfoEff* and *ActInfoEff*

Variables	<i>N</i>	<i>M</i>	<i>m</i>	σ	<i>p</i>
<i>ActInfoEff</i>	44	6	6.4	3.3	0.003
<i>ActNavEff</i>		8	10.6	6.9	

Statistics for *EstNavEff* and *ActNavEff*

Discussion Summarizing, the estimates for all the models were overly optimistic, and this circumstance was statistically significant at the 0.05 level. One of the causes of these results may probably be the subjects' inexperience. However, underestimation is a common phenomenon in software projects at all levels (from the micro-level of individual developers up to the macro-level of complete projects) even when experienced developers and managers are involved. The relevant factors of underestimation should be identified and quantified.

Estimated Effort vs. Actual Effort: correlation

Rationale We wanted to check whether effort estimates were actually useful to predict actual effort.

Hypothesis 3 For each effort category, there is a positive correlation between the estimated effort and the actual effort.

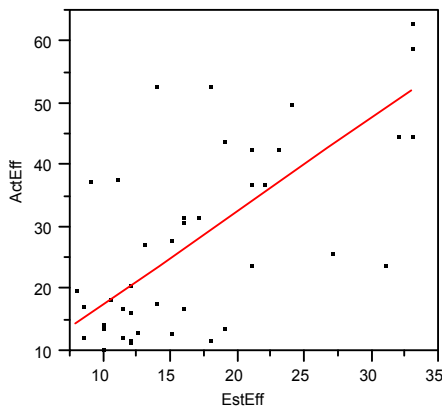


Figure 3. EstEff vs. ActEff.

Results We started by studying the overall efforts *EstEff* and *ActEff*. Figure 3 graphically shows the regression line whose equation is:

$$ActEff = 2.31 + 1.51 EstEff$$

and the following table summarizes the model statistics. Five outliers were identified during the analysis, so our results and model are based on $N = 39$ data points.

Coefficient	E	σ	p	R^2	N
<i>Intercept</i>	2.31	4.15	0.58	0.54	39
<i>EstEff</i>	1.51	0.23	< 0.0001		

Model statistics for $X = EstEff, Y = ActEff$

In this table, the insufficient statistical significance of the *Intercept* shows that we cannot rule out the (null) hypothesis that $Intercept = 0$, while the statistical significance of the coefficient of *EstEff* is very good, i.e., we can safely conclude that *EstEff* can be used to predict *ActEff*. The value of the estimate of the coefficient of *EstEff*, which is greater than 1, confirms that the subjects tended to underestimate the actual effort.

For brevity's sake, we do not report on the other correlations for each category of effort. All of them turned out to be statistically significant.

Discussion Our results show that it is possible to use the estimated values as predictors for the actual ones. We need to caution the reader that this does not mean that an accurate predictive model can be built for each of these kinds of effort, but only that the estimated effort in each category is correlated to the actual effort in that category, even though it does not explain all of the variance. As low time-to-market and effort are even more vital in Web applications than in other businesses, due to the intense competition among Web developers, ways for accurately estimating the effort would be very valuable for project planning and marketing activities. Our results are a starting point, i.e., the estimated effort for each category may be a component of a model that predicts the effort for that category. Other variables need to be identified, so that we can also explain the underestimation problems studied by Hypothesis 2. Next, we study some of these possible factors.

Impact of the subjects' technical knowledge: correlation

Rationale The following hypotheses investigate whether there is an influence of the subjects' technical knowledge on the estimated and actual effort values, and on the difference between estimated and actual effort. Also, we studied whether there was an influence of the students' technical background on estimated learning effort, actual learning effort and the difference between the two. Learning effort, whether estimated or actual, is the sum of the learning efforts related to the single models. Our idea was that greater skills are associated with smaller efforts and smaller differences.

Hypothesis 4 There is a negative correlation between the subjects' technical knowledge 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.

Results We quantified the subjects' technical knowledge (*TechKnow*) as the number of Web-related languages

(HTML, XML, etc.) and techniques (JSP, ASP, etc.) that they knew. However, none of the correlations turned out to be statistically significant except for the correlation between the technical knowledge and the difference between actual and estimated effort (*DeltaEff*), whose statistics are in the following table:

Coefficient	<i>E</i>	σ	<i>p</i>	R^2	<i>N</i>
<i>Intercept</i>	18.2	4.2	0.0001	0.14	38
<i>TechKnow</i>	-3.63	1.52	0.02		

Model statistics for $X = \textit{TechKnow}$, $Y = \textit{DeltaEff}$

Discussion It is somewhat surprising that some of the correlations that we believed to exist did not turn out to be statistically significant. For instance, the existence of a relationship between technical knowledge and actual effort is not proved by our results. It is interesting to note that neither a negative correlation nor a positive correlation existed between technical knowledge and actual effort. (One might have explained a positive correlation as the result of stronger commitment to the project.) At any rate, the results in the previous show that there is a negative correlation between technical knowledge and the difference between actual and estimated effort, as shown by the sign of the estimate of the coefficient (-3.63). Further studies are certainly required even for this correlation, because the goodness-of-fit of the relationship is too small to base any predictions on an univariate model that tries to predict *ActEff* based only on *TechKnow*. In our study, we tried to build a multivariate model that included *TechKnow* and *EstEff* as independent variables and *ActEff* as dependent variable. However, even though all of the coefficients of the model turned out to be statistically significant, the value of the goodness-of-fit as measured by R^2 did not improve significantly. Thus, *TechKnow* does not really add much to the prediction accuracy of already obtained via *EstEff*.

Impact of the subjects' proficiency: correlation

Rationale The following hypotheses investigate whether there is an influence of the subjects' school proficiency on the estimated and actual values, and on the difference between estimated and actual effort, and also on the learning efforts. Our idea was that a better proficiency is associated with smaller efforts and smaller differences.

Hypothesis 5 There is a negative correlation between the subjects' school proficiency and the estimated learning effort, actual learning effort, difference between actual and estimated learning effort, estimated effort, actual effort, difference between actual and estimated effort.

Results We quantified the subjects' school proficiency by either their average grade (*Avg*) or the number of computer science exams (*CSE*) they had passed. Four correlations turned out to be statistically significant, as shown by the following tables:

Coefficient	<i>E</i>	σ	<i>p</i>	R^2	<i>N</i>
<i>Intercept</i>	16.1	5.1	0.003	0.08	50
<i>Avg</i>	-0.43	0.21	0.047		

Model statistics for $X = \textit{Avg}$, $Y = \textit{EstLearnEff}$

Coefficient	<i>E</i>	σ	<i>p</i>	R^2	<i>N</i>
<i>Intercept</i>	-77.1	45.9	0.1	0.13	36
<i>Avg</i>	4.4	1.9	0.028		

Model statistics for $X = \textit{Avg}$, $Y = \textit{ActEff}$

Coefficient	<i>E</i>	σ	<i>p</i>	R^2	<i>N</i>
<i>Intercept</i>	-6.85	17.1	0.7	0.11	40
<i>CSE</i>	1.77	0.82	0.038		

Model statistics for $X = \textit{CSE}$, $Y = \textit{ActEff}$

Coefficient	<i>E</i>	σ	<i>p</i>	R^2	<i>N</i>
<i>Intercept</i>	-12.86	11.41	0.27	0.1	39
<i>CSE</i>	1.13	0.56	0.048		

Model statistics for $X = \textit{CSE}$, $Y = \textit{DeltaEff}$

Discussion There is a negative correlation between *Avg* and *EstLearnEff*, which means 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 *Avg* and *ActLearnEff*. Another interesting result is that better proficiency (whether measured by *Avg* or *CSE*) seems to imply higher levels of actual effort. We expected a negative correlation, but actually found a positive one. This is not totally surprising in an exploratory study like ours. One possible explanation is that higher proficiency may be associated with higher levels of commitment in carrying out the homework, and this would explain the positive correlation. At any rate, further studies are required, also because the values of goodness-of-fit (as measured by R^2) are quite low, so these independent variables may only be used as parts of multivariate models, since they do not explain a significant part of the variance alone. As a closing note, we tried to use the subjects' proficiency along with their technical knowledge to build a predictive model for estimated or actual effort, but we did not obtain any statistically significant results.

Estimated vs. Actual Learning Effort: correlation

Rationale We wanted to check if *EstLearnEff* could be used as a predictor for the actual learning effort.

Hypothesis 6 There is a positive correlation between *EstLearnEff* and *ActLearnEff*.

Results The table summarizes the results:

Coefficient	E	σ	p	R^2	N
<i>Intercept</i>	2.8	1.2	0.027	0.14	36
<i>EstLearnEff</i>	0.51	0.21	0.022		

Model statistics for $X = EstLearnEff, Y = ActLearnEff$

Discussion The results show that the hypothesis is sufficiently supported by the data. However, R^2 is too low to predict *ActLearnEff* based only on *EstLearnEff*. Other predictors will have to be studied and used along with *EstLearnEff*. Even though we do not show the corresponding hypothesis here, we also checked if there was a tendency to underestimate learning effort, like with all other effort categories. However, the difference between the two medians was not statistically significant, so there does not seem to be any underestimation, in this case.

Estimated Learning vs Estimated Effort Categories

Rationale We wanted to check whether there was a correlation between the estimated learning effort and the estimated modeling effort per each category. For clarity, it must be noted that the estimated learning effort is not a part of the estimated modeling effort.

Hypothesis 7 For each effort category, there is a positive correlation between the learning effort and the corresponding modeling effort.

Results The results are summarized in the following tables:

Coefficient	E	σ	p	R^2	N
<i>Intercept</i>	2.35	0.31	< 0.0001	0.13	52
<i>EstDLearnEff</i>	0.36	0.13	0.007		

Model statistics for $X = EstDLearnEff, Y = EstDataEff$

Coefficient	E	σ	p	R^2	N
<i>Intercept</i>	1.345	0.35	0.0004	0.24	47
<i>EstSLearnEff</i>	0.83	0.22	0.0005		

Model statistics for $X = EstSLearnEff, Y = EstStructEff$

Coefficient	E	σ	p	R^2	N
<i>Intercept</i>	1.94	0.43	< 0.0001	0.20	50
<i>EstNLearnEff</i>	0.68	0.19	0.0009		

Model statistics for $X = EstNLearnEff, Y = EstNavEff$

Discussion All the results seem to confirm Hypothesis 7, even though the values of R^2 are too low to make it possible to predict the dependent efforts based only on the corresponding independent efforts. At any rate, the estimated learning efforts have an effect on the estimated modeling efforts, so one way of improving the estimated modeling efforts would be to improve the learning estimates. This may be done by providing the students with better conceptual and practical tools for estimating their learning effort.

4.1 Self-Assessment vs. Grading

We asked the students to provide an assessment of the work they had done, to check if there was a correlation between their self-assessment and the grading done by the instructor. No correlation was found between the two scales, i.e., the subjects were not able to provide a reliable assessment of the quality of their work, even though many of them had already taken several computer science classes.

4.2 Validity of the experiment

Like in any empirical study, we need 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, and classes

These factors must be examined as for their influence on the internal and external validity of the empirical study.

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

1. *Subjects.* The subjects were not selected before-hand, i.e., they were a veritable cross-section of the graduate students that attend Web-related classes.
2. *Applications.* The subjects could choose the application they preferred. However, all of the applications only differed in their specific details, but they were all related to e-commerce.
3. *Availability of tools.* The students did not use any specific tools for W2000. They standard usual text and graphics editors.
4. *Notation.* All the subjects used the same notation.

5. *Classes*. There was little difference in the percentage of classes attended by the subjects, i.e., they attended almost all the classes.

We can conclude that our results were not biased by the choice of subjects or applications (in the context of e-commerce). The results were obviously influenced by the lack of tools, the notation, and the percentage of class attendance, i.e., we could have obtained different results 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.

External validity The question may arise as to how representative our empirical study is in the population of empirical studies on Web application effort estimation. Even though the results might have been biased from an internal validity point of view by the three factors availability of tools, notation, and classes, we need to discuss the possible factors that may influence the outcome of an empirical study such as ours.

1. *Subjects*. The subjects were representative of the population of Web designers and developers, since Web designers and developers are usually taken from young college undergraduates or graduates.
2. *Applications*. Many Web applications are actually in the same application context as the ones we studied, i.e., e-commerce ones.
3. *Availability of tools*. Professional tools may be used for designing Web applications, so this factor could have been a threat to the external validity of our study. However, little could be done, because of the cost of commercial tools.
4. *Notation*. We used a specific notation, W2000. However, this notation has many aspects in common with UML, which is becoming a software development standard notation.
5. *Classes*. The specific education 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. In addition, it must be said that Web designers and developers are trained personnel, who need to continuously update their knowledge.

5 Conclusions and Future Work

In this paper, we have illustrated an empirical study carried out on the effort required to design web applications. We have shown that the information model appears to take

the largest effort in the design phase. In addition, our results show that (at least in our environment) the a priori effort estimates can be used to predict the actual effort. A clear tendency to underestimation of effort has been clearly shown.

A number of predictors have been identified, so this study can be used as a starting point. Further research is clearly required to identify further predictors that may explain a larger percentage of effort variance than is now possible. To this end, our future research plans include:

- Using measures for internal design attributes; based on the students' projects, we will investigate whether internal design attributes (e.g., size, complexity, cohesion, coupling) are related to effort;
- Including an automated measurement tool in the tool that supports the W2000/HDM notation;
- Investigating prediction for other external attributes;
- Investigating the following of web development phases, e.g., implementation and verification;
- Replicating and refining this study, based on the experience we have acquired.

References

- [1] L. Baresi, F. Garzotto, and P. Paolini. "Extending UML for Modeling Web Applications". In Proceedings of 34th Annual Hawaii International Conference on System Sciences (HICSS-34). IEEE Computer Society, 2001.
- [2] G. Booch. "The Architecture of Web Applications", 2001 www.developer.ibm.com/library/articles/booch_Web.html.
- [3] G. Booch, I. Jacobson, and J. Rumbaugh. "The Unified Modeling Language User Guide", The Addison-Wesley Object Technology Series, 1998.
- [4] F. Garzotto, P. Paolini, D. Schwabe. "HDM - A Model-Based Approach to Hypertext Application Design", TOIS 11(1) (1993), pp.1-26.
- [5] E. Mendes, S. Counsell, and N. Mosley. "Measurement and Effort Prediction of Web Applications", Proc. Second ICSE Workshop on Web Engineering, 4 and 5 June 2000; Limerick, Ireland, 2000. January-March 2001.
- [6] E. Mendes, N. Mosley, S. and Counsell. "Web Metrics - Estimating Design and Authoring Effort". IEEE Multimedia, Special Issue on Web Engineering, January-March, 50-57, 2001.