Estimating 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

Our study focuses on the effort needed for designing Web applications. The effort required for the design phase is an important part of the total development effort of a Web application, whose implementation can be (partially) automated by tools.

We carried out an empirical study with the students of an advanced university class that used W2000, as special-purpose object-oriented design notation for the design of Web applications. We investigated the impact of a few attributes (e.g., size, complexity) of the W2000 design artifacts built during the design phase on the total effort needed to design web applications and we identified a few attributes that may be related to the total design effort. In addition, we carried out a finer-grain analysis, by studying which of these attributes have an impact on the effort devoted to the steps of the design phase that are followed when using W2000.

Keywords:
Web Application Design, W2000, Empirical Study, Effort Estimation

Principal author:
Luciano Baresi
Dipartimento di Elettronica e Informazione
Politecnico di Milano
piazza Leonardo da Vinci, 32 - 20133 Milano (Italy)
Telephone: +39 02 2399 3638
Fax: +39 02 2399 3411
E-mail: baresi@elet.polimi.it

Estimating 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

Our study focuses on the effort needed for designing Web applications. The effort required for the design phase is an important part of the total development effort of a Web application, whose implementation can be (partially) automated by tools.

We carried out an empirical study with the students of an advanced university class that used W2000, as special-purpose object-oriented design notation for the design of Web applications. We investigated the impact of a few attributes (e.g., size, complexity) of the W2000 design artifacts built during the design phase on the total effort needed to design web applications and we identified a few attributes that may be related to the total design effort. In addition, we carried out a finer-grain analysis, by studying which of these attributes have an impact on the effort devoted to the steps of the design phase that are followed when using W2000.

Keywords: Web Application Design, W2000, Empirical Study, Effort Estimation

1 Introduction

Modern Web applications are complex distributed systems that use the Web as the interaction means and the Internet as the communication infrastructure. They offer both *pages*, through which the user can navigate, and *services*, with which the user can interact: We can buy goods, make reservations for hotel rooms or tickets, and trade on stock markets through these applications. Several companies already view Web applications as a key component to deliver quality services to their customers.

The intertwining between hypertextual information repositories and business models is transforming Web application design from the definition of a set of hypertexts to a complex software engineering task. Practical experience on web development shows that there are significant differences between traditional software applications and web applications. Web applications continuously change their requirements, contents, and functionality, so we cannot apply traditional software engineering approaches unchanged. Web applications call for special-purpose disciplined approaches. In the context of these approaches, a few recent publications [10, 9, 6] stress the need for *Web estimation*. The ability of accurately estimating the cost of the various development phases of Web applications may provide a web development organization with a competitive advantage.

The design phase of a Web application is often a critical one and uses a large percentage of the development resources. If the design phase of a Web application is carried out carefully, the subsequent phases may require considerably less resources than the design phase itself, and can be automated to some degree, especially when supporting tools are used.

The goal of the research documented in this paper is to build estimation models for the design phase of Web applications. In this paper, we describe an empirical study on the effort required for designing complex Web applications that we carried out during an advanced university class on modeling Web applications. We have been able to identify

---

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

a number of attributes that may have an impact on design effort. Based on these attributes, we have defined measures that have allowed us to build univariate and multivariate correlational models for the total design effort and the effort used in the steps of the design phase.

We worked with the students of an advanced university class on modeling Web applications. The empirical study had several goals, as we now concisely describe. (1) We wanted to evaluate the ability of the subjects to predict the effort needed to develop the W2000 artifacts and find out the factors that affect these predictions and their accuracy. (2) We wanted to evaluate the accuracy of the students' self grading of the artifacts. (3) We wanted to investigate the relationship between internal attributes of the artifacts produced during Web application design and the effort required.

We published the results about goals (1) and (2) in a previous paper [3]. The specific results are described in the Related Work section of the present paper (Section 6). Here, we focus on goal (3), i.e., we want to show that it is possible to build univariate and multivariate statistical models based on Web application design measures. We studied, and built models for, the effort necessary to carry out the entire design phase. In addition, we studied and built models for the effort required to design the so-called *information model* and *navigation model* required by the W2000 web design notation.

We would like to point out that, since the field of the estimation of the cost of developing Web applications is a relatively new one, our experimental study is an exploratory one, rather than a confirmatory one. Its main objective is to identify and provide empirical support for hypotheses that deserve further attention, and not to confirm a well-established theory, which can actually be built based on several empirical studies like ours.

The rest of this paper is organized as follows. To make the paper self-contained, Section 2 concisely introduces W2000 and its modeling features, Section 3 describes the experimental setting, Section 4 presents our hypotheses and the results we obtained, Section 5 discusses the validity of our experiments, Section 6 surveys the related work, and Section 7 concludes the paper and summarizes.

2 W2000

W2000 [2] is the UML-like modeling notation for Web applications developed at Politecnico di Milano. Due to space limitations, we cannot fully describe W2000 (readers can refer to [1] for an in-depth presentation), but we use the excerpts of W2000 models presented in Figure 1 to introduce the main concepts.

W2000 fosters *separation of concerns* by organizing a complete specification into four different models. Conceptually, the starting point is the *information model*, whose goals are the identification and organization of all data that the application should deal with. The former goal is up to the *hyperbase model*, while the latter belongs to the *access structures model*.

The *hyperbase model* identifies the *entities*[1] that characterize the application. They define conceptual "objects" that are of interest for the user. *Components* are then used to structure the contents of *entities* into meaningful chunks. They can further be decomposed into sub-components, but the actual contents can be associated with leaf nodes only. *Slots* identify primitive information elements and are the typed attributes that specify the contents of leaf components. *Segments* define "macros": They identify sets of slots that can be reused in different *Components* and models. Figure 1(a) defines the entity *type* Concert, which consists of three components: Description comprises the main information associated with the concert, Cast presents the singers, and Audio is an audio fragment of the concert. Notice that we decided to detail only the first component, but all elements must be fully described.

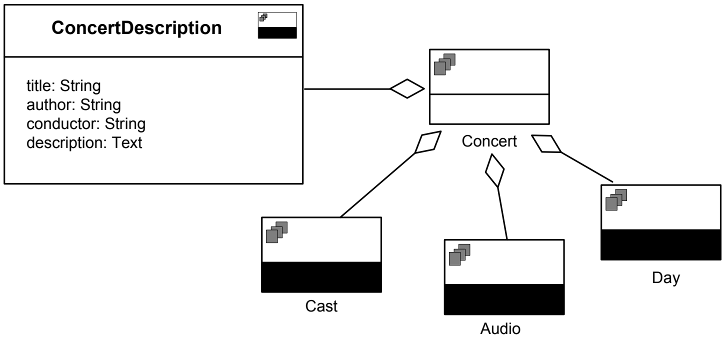
Semantic associations link pairs of *entities* and identify navigational paths between two related concepts. *Association centers* describe the set of "target" elements identified by a *semantic association*. In a 1 to n association, defines how to identify either the entire set of targets as a whole or each individual element in the set. Figure 1(b) presents a simple association that links the *single* entity Theater with a set of Concerts.

The *access structures model* is used to organize the information defined so far. It specifies the main access points to the application. It comprises only *collections* that define groups of elements that should be perceived as related by the user. *Collections* organize application data in such a way that the user can fully exploit them. Also *collections* can have *centers*. Figure 1(c) describes the collection of all offered concerts, that is, the whole season. It groups Concerts so that they can be perceived as a single element. The center specified how each concert is rendered in the collection (i.e., the slots that are used to describe/identify it).

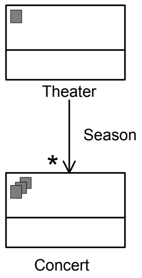
When we move to define how the user can browse through the application, we build the *navigation model*, which reshapes the elements in the previous model to specify the *actual* information chunks. The main modeling elements are the *nodes* that define atomic information units. Usually, they do not define new contents, but they render information already defined in the *information model*. *Clusters* link together sets of *nodes* and define how the user can move around these elements. They can be further organized

---

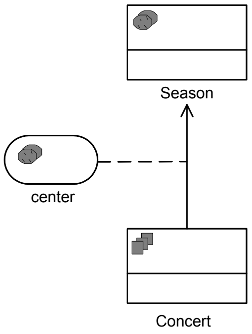
[1] All W2000 elements can be either *typed* or *single*. They belong to the first group when define *classes* of similar elements; they belong to the second group when define singletons, that is, elements that should be available as single instances only.



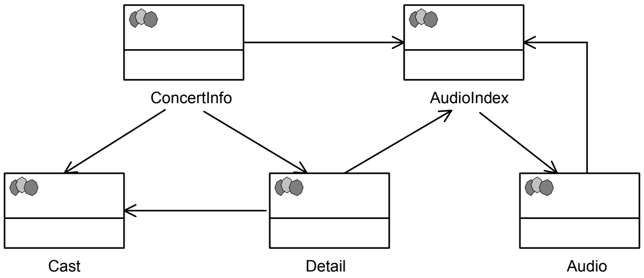
(a) Entity **Concert**



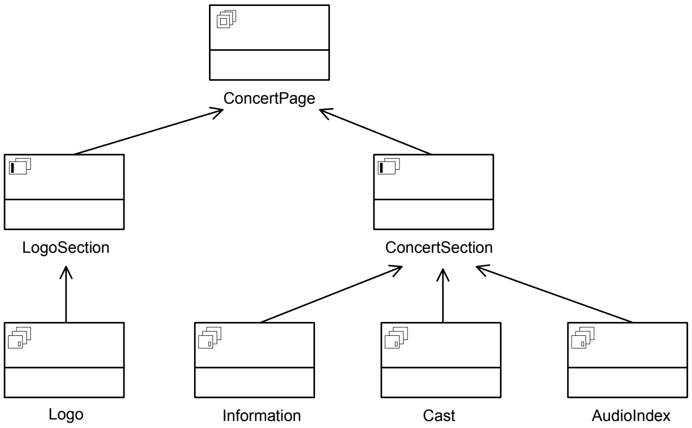
(b) Semantic Association **Concerts**



(c) Collection **Season**



(d) Cluster **Concert**



(e) Page **Concert**

Figure 1. Excerpts of W2000 example models

in: *structural clusters* if all their elements come from the same *entity*; *association clusters* if they render associations; *collection clusters* if they describe the topology of a collection, and *transactional clusters* if they relate the set of nodes that the user should traverse to complete a business transaction. *Clusters* are also useful to identify how the current context varies while the user interacts with the application. Figure 1(d) defines a simple cluster: From the ConcertInfo node, the user can move to the Cast node, to the Details node, or to AudioIndex node, which is an entry point for the set of Audio nodes that contain the real audio fragments. The cluster identifies only the main paths through nodes; others can be identified while designing the actual pages.

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 we assume that each *node* is rendered through a particular *page*.

The *presentation model* offers *publishing units* that are the smallest information elements that are visualized on pages. They usually render *nodes*, but they 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 *operations* defined in the *operations model*) on application data. Figure 1(e) describes how the page that renders a concert should be organized: it comprises two sections, the TheaterLogo and ConcertSection, respectively. The former section contains only a presentation unit, the Logo that is added in this model. The latter section comprises three presentation unit: Information materializes the ConcertInfo node, Cast renders the Cast node, and Audio corresponds to the AudioIndex.

The *operations model* describes all the *operations* that can be performed by the user on the application data and are associated with *links*, thus also with *pages*. Each *operation* should be clearly specified in terms of pre- and post-conditions. W2000 simply suggests rigor and completeness. For example, each *operation* should identify how it modifies data (not only in terms of *entities*, but also considering back-end data), navigation (e.g., new *nodes* and *links*) and presentation (new *pages*). But each operation should also identify – if needed – the new page to which the user is taken after executing the operation.

Operations are presented here to give a complete description of W2000, but they are not further discussed and used in this paper.

3 Setting of the empirical study

This paper documents an empirical study that was carried out at the Politecnico di Milano with students attending an advanced class on Web application design during the fourth or fifth year of their curriculum. The students were taught to design their Web applications with W2000 and implement them with the many technologies available.

The empirical study started with supplying the students with a brief and informal set of requirements for the application that they were asked to model. All students were required to work on the same topic: a hypothetical e-commerce application, but each project (student) had a different application domain: books, CDs, groceries, etc.

Along with the requirements, the students were given two questionnaires. 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. With the first questionnaire, which the students were required to fill before they started working on the project, we wanted to measure their general proficiency in computer science and Web technologies and make them estimate the design effort required to model their applications. We asked them to split the design effort according to the main W2000 models.

The second questionnaire had to be filled after completing their work and had to report on the effort actually spent to model the application. It must be considered that when we carried out the experiment, W2000 was not supported by any tool and this lack could have impacted their work. We also asked the students to self-evaluate their models.

When the students handed their models over to us, we collected a number of web design measures that we derived based on the empirical hypotheses that we wanted to investigate, as we explain in the next section. Here is the list of measures we collected:

- **Information model**
  - *Size measures*
    - **entities:** number of entities in the model
    - **components:** number of components in the model
    - **infoSlots:** number of slots in the model
  - *Average size measures*
    - **slotsSACenter:** average number of slots per semantic association center
    - **slotsCollCenter:** average number of slots per collection center in the model
  - *Average granularity of decomposition*
    - **componentsEntity:** average number of components per entity

    - ∗ **slotsComponent:** average number of slots per component
  - *Complexity measure*
    - ∗ **SAssociations:** number of semantic associations in the model
  - *Data cohesion*
    - ∗ **SACenters:** number of semantic association centers in the model
  - *Reuse measure*
    - ∗ **segments:** number of segments in the model

- **Navigation model**
  - *Size measures*
    - ∗ **nodes:** number of nodes in the model
    - ∗ **navSlots:** number of slots in the model
  - *Average size measures*
    - ∗ **nodesCluster:** average number of nodes per cluster
    - ∗ **slotsNode:** average number of slots per node
  - *Complexity measure*
    - ∗ **navLinks:** number of links in the model
  - *Structuredness*
    - ∗ **clusters:** number of clusters in the model

- **Presentation model**
  - *Size measures*
    - ∗ **pages:** number of pages in the model
    - ∗ **pUnits:** numer of publishing units in the model
  - *Complexity measure*
    - ∗ **prLnks:** number of links in the model
  - *Structuredness*
    - ∗ **sections:** number of sections in the model

4 Results

Due to the limited availability of knowledge about the effort related to developing Web applications, especially the effort related to the design phase, we believe that it would be too early to carry out a confirmatory study, so our empirical study is an exploratory one.

We investigated a few empirical hypotheses that we believed likely to be true based on our beliefs, knowledge about the W2000 notation, and the steps of the design process used. Therefore, we report on both the hypotheses confirmed by our study 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. Along the lines of [5], we first stated ”empirical hypotheses” that referenced software attributes and not actual measures, because we were more interested in studying the impact of factors on the effort needed for each design phase, rather than studying the existence of correlations between specific measures and effort.

We initially planned to study the effort related to each design step and the effort related to the entire design phase. Specifically, we studied the following *dependent attributes*:

- information effort, i.e., the effort necessary to design the information model

- navigation effort, i.e., the effort necessary to design the navigation model

- presentation effort, i.e., the effort necessary to design the presentation model

- total effort, i.e., the effort necessary for the entire design phase

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. In what follows, the total effort is composed of the information and the navigation 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.

We also studied effort for the single sub-activities (e.g., modeling of the hyperbase and its access structures) within each design step. We do not report on this here because we believe that this analysis may be too-fine grained at the present time, and it would be more useful to study and report on higher level activities, i.e., the building of the information model, the building of the navigation model, the building of the entire design.

For each dependent attribute, we stated a number of empirical hypotheses that linked dependent attributes to independent attributes. For instance, one empirical hypothesis stated that the information effort (dependent attribute) increases when the size (independent attribute) of the information model increases. The dependent and the independent attribute were then instantiated via measures. These measures were used to build statistical models. In subsections 4.1, 4.2, and 4.3, we describe the empirical hypotheses and the results we obtained for total, information, and navigation effort. In each subsection, 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:

- c_x, the estimates of the regression coefficients, one for each independent variable X in the model plus one for the intercept.

- p_{cx}, 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 $p < 0.05$ or $p < 0.10$, i.e., there is less than 5% or 10% probability that they have an impact on Y by chance.

  In an exploratory paper like this, we fix a 0.10 threshold for the statistical significance of the impact of the independent variables, to mitigate the risk of discarding too many independent variables that do have an impact on the dependent variables, but whose p-value is between 0.05 and 0.10 in our study. Further confirmatory studies may use stricter threshold values such as 0.05, to provide more evidence that these independent variables actually have an impact on the dependent variables.

- 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 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.

Prior to carrying out OLS analyses, we removed those data points that appeared to be corrupt, or clearly incorrect, or for which information was missing, i.e., the value of either the independent or the dependent variable was missing. The number of valid data points is 36. Then, before each OLS analysis, we carried out a very careful outlier analysis, to remove those few data points that were too “far” from the others. This is a standard data analysis activity: it is carried out to eliminate the in influence of those few points that may unduly bias the results. Removing outliers is absolutely necessary in exploratory studies because one needs to build sensible hypotheses that may be confirmed in future studies and rule out hypotheses that may appear to hold only thanks to too few data points.

The next three subsections will be structured as follows:

- Empirical hypotheses
- Statistically significant results
- Discussion

4.1 Total effort

Empirical hypotheses

Here are the empirical hypotheses we investigated:

HTE1: the higher the size of the information or the navigation model, the higher the total effort

HTE2: the higher the complexity of the information or the navigation model, the higher the total effort

HTE3: the higher the reuse in the models, the lower the total effort

HTE4: the higher the granularity decomposition of the information model, the higher the total effort

These empirical hypotheses and the empirical hypotheses shown in Sections 4.2 and 4.3 motivate the choice of the web application measures listed in Section 3, which quantify the independent attributes [3] of the empirical hypotheses.

Statistically significant results

Table 1 contains the results of our univariate OLS analyses. Each row contains the statistics for a different OLS model. For instance, in the row corresponding to component, c_0 is the estimate of the intercept, c_1 is the estimate of the coefficient of the independent variable component, p_{c0} is the statistical significance of the statistical hypothesis $Intercept \neq 0$, p_{c1} is the statistical significance of the statistical hypothesis $c_1 \neq 0$, R^2 is the OLS model goodness-of-fit, and N is the number of data points used to build the model whose independent variable is component.. N may differ across rows because of the different number of outliers removed to build the models. The models are reported in decreasing order of goodness-of-fit. We did not report the overall p-value for each model separately because it always coincides with p_{c1}. In other words, what is assumed to be the overall p-value for a model is related to the probability that the independent variable does have an impact on the dependent variable. Thus, the fact that $p_{c0} = 0.56$ for the model based on components only means that we cannot really conclude that $c_0 \neq 0$, but this does not affect the fact that we can realistically conclude that it is very unlikely that $c_1 \neq 0$, i.e., it is very likely that components is correlated to the total effort.

We have also identified a statistically significant multivariate model (see Table 2), which we provide for completeness. The number of data points used to build the model is 30, and the model’s $R^2 = 0.74$.

X	c_0	c_1	p_{c0}	p_{c1}	R^2	N
components	4.06	2.31	.56	.0011	.31	31
navSlots	18.3	.31	.0001	.0053	.22	33
segments	30.6	-3.52	.0001	.0014	.21	28

Table 1. Statistics for univariate OLS models for *total effort*

Term	Estimate	p
Intercept	59.2	0.0011
entities	-3.82	0.0115
components	2.53	0.0042
SACenters	3.411	0.0069
slotsSACenter	-7.06	0.0030
segments	-4.13	0.0003
clusters	-1.75	0.0008
nodesCluster	-10.4	0.0434
navSlots	0.469	0.0002

Table 2. Statistics for multivariate OLS models for *total effort*

Discussion

Univariate OLS models The positive signs of the c1 coefficients in the rows corresponding to components and navSlots in Table 1 show that there is an influence of size on the total effort (hypothesis HTE1). This influence has not been found to be statistically significant in the context of OLS for the other size measures. For instance, the impact of the number of entities in the information model or nodes in the navigation model does not appear to be statistically significant in the context of OLS. This shows that building theories is actually an iterative process where one starts with general ideas (e.g., size has an impact on effort), which then needs to be refined based on the evidence obtained. Complexity, as measured via our complexity measures, doesn't seem to have a statistically significant influence on the total effort, so hypothesis HTE2 is not supported. On the contrary, the influence of reuse on the total effort seems to be the expected one according to hypothesis HTE3. One can save effort by having some reusable model parts. This is an important point, since it quantitatively shows that there is value added in reusing when carrying out web design. No correlation was found between the decomposition granularity of the information model and the total effort (hypothesis HTE4).

Multivariate OLS model The coefficients of the model are very significant from a statistical viewpoint. Due to possible interactions among the independent variables it is not entirely possible to interpret the results by looking at the coefficients separately. However, it is worth noting that the sign of the coefficient of the independent variable segments is negative. This seems to confirm the idea that reuse may help save effort.

4.2 Information effort

4.2.1 Empirical hypotheses

HIE1: the higher the size of the information model, the higher the information effort

HIE2: the higher the average size related to aggregations, the higher the information effort

HIE3: the higher the granularity of decomposition, the higher the information effort

HIE4: the higher the complexity of the information model, the higher the information effort

HIE5: the higher the data cohesion of the information model, the lower the information effort

HIE6: the higher the reuse in the models, the higher the information effort

While hypotheses HIE1, HIE4, and HIE5 state common beliefs that are supposed to be true for many models and software systems, hypotheses HIE2, HIE3, and HIE6 deserve some further comments. As for HIE2, the idea is that information models where the average information content of the aggregations is "heavier" require a higher information effort to design the structure properly. HIE3 is somewhat similar to HIE2, i.e., information models where more care has been devoted to decomposition should require a higher information effort.

The intuition behind HIE6 is that the information effort increases with reuse is actually due to the fact that the reusable elements are actually identified and put together when building the information model and this may take some extra effort.

Statistically significant results

The results of univariate analyses are in Table 3. No statistically significant multivariate models have been identified.

Discussion

The influence of size (HIE1) on the information effort seems to be mixed. The positive coefficient of component demonstrates the positive influence of the number of components on the information effort. On the other hand, the negative coefficient of entities shows that the information effort decreases when the number of entities increase. We

X	c_0	c_1	p_{c0}	p_{c1}	R^2	N
components	2.98	1.29	.41	.0007	.35	29
segments	18.1	2.08	.0001	.011	.23	27
entities	24.0	-1.42	.0001	.085	.10	30
slotsCollCenter	12.9	2.07	.0001	.067	.10	33

Table 3. Statistics for univariate OLS models for *information effort*

can provide a possible explanation of this phenomenon, as follows. The same problem can be "solved" in different ways: more entities, maybe single-component ones, identify a trivial design, and thus less effort; less entities require more effort to manage complexity and clearly design their structure in terms of components. This clearly shows the need for refining empirical hypothesis HIE1 to identify more precise hypotheses that address single aspects of size.

HIE2 seems to be partially true, since the average of slots per collection center is statistically significant and its coefficient is positive, though it is the independent variable that explains the least amount of variation of the dependent variable. HIE3, HIE4, and HIE5 are not supported by our empirical validation, so granularity of decomposition, complexity, and data cohesion (as quantified by the measures we collected) do not seem to have any statistically significant effects on information effort. As for HIE6, the number of segments, i.e., reusable parts of the model, has the expected impact on the information effort. Even though the information effort increases, having reusable components helps curb down the total effort. Thus, it is worth investing in identifying reusable parts in the information model.

4.3 Navigation effort

Empirical hypotheses

HNE1: the higher the size of the navigation model, the higher the navigation effort

HNE2: the higher the average size of aggregations, the higher the navigation effort

HNE3: the higher the complexity of the navigation model, the higher the navigation effort

HNE4: the higher the degree of structuredness in the navigation model, the higher the navigation effort

HNE5: the higher the reuse in the navigation model, the lower the navigation effort

Some of the above hypotheses deserve further comments. Hypothesis HNE2 is analogous to hypotheses HIE2 and HIE3 that we stated for information effort. Hypothesis HNE4 is based on the fact that carefully structuring a navigation model will require more effort.

X	c_0	c_1	p_{c0}	p_{c1}	R^2	N
navLinks	4.53	.124	.0.28	.037	.14	32
slotsNode	7.73	1.65	.0001	.074	.09	35

Table 4. Statistics for univariate OLS models for *navigation effort*

Term	**Estimate**	p
Intercept	6.58	0.027
slotsSACenter	-2.84	0.012
slotsCollCenter	-2.77	0.012
segments	-1.35	0.0007
infoSlots	0.253	0.0003
nodes	-0.230	0.01
navSlots	0.253	0.0001

Table 5. Statistics for multivariate OLS models for *navigation effort*

4.3.1 Statistically significant results

Table 4 contains the results of our univariate OLS analyses.

We have also been able to identify a statistically significant multivariate model, whose statistics are in Table 5. The number of data points used to build the model is 30, and the model's $R^2 = 0.73$.

Discussion

Only hypotheses HNE2 and HNE3 are supported by our data. This confirms that a more complex navigation model, i.e., one in which there is a greater amount of links between the nodes, will require a higher amount of effort (HNE3). Also, navigation models with a higher size of aggregations will require a higher effort. Size (HNE1), structuredness (HNE4), and reuse (HNE5) in isolation do not seem to have any statistically significant effect on navigation effort. The multivariate model seems to show that there reuse might actually have the expected impact, since its coefficient is negative. However, like for the multivariate model for total effort, we need to be cautious in the interpretation of this variable alone in the context of the multivariate model. At any rate, it is a fact that all the evidence we have about reuse is in the expected direction: reuse is likely to increase the information effort, but it is also likely to decrease the navigation effort, and more importantly, total effort.

5 Validity of the empirical study

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, classes, and construct validity.

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

5.1 Internal validity

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

Subjects No initial selection of the subject was carried out, so no bias was apparently introduced.

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 study, 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, i.e., they attended almost all the classes.

Construct validity The risk is that the measures used 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 (for instance, other size measures) would be correlated with the independent variables.

Thus, 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 clearly 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.

5.2 External validity

We need to check how representative our empirical study is 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 this study to web development organizations.

Subjects Since no pre-selection 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 representative of the population of Web designers and developers, since Web designers and developers are usually taken from young college undergraduates or graduates.

Applications E-commerce applications, like the ones in our study, are typical of web development.

Availability of tools The design of web application 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 study. However, little could be done, because of the cost of commercial tools.

Notation W2000 has many aspects in common with UML, which is becoming a software development standard notation.

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.

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

6 Related work

The relative novelty of Web applications and the difficulty of collecting significant data in a such a frenetic world justify the poorness of the state of the art of *Web estimation*.

To the best of our knowledge, few proposals exist on empirical studies in effort estimation of Web applications. Mendes et al. [7, 8] 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. Their experiments consider only the resulting systems (pages and links), while we focus our analysis on design artifacts.

In a previous paper [3], we studied the effort required to model Web applications with W2000. We show that the information model appears to take the largest effort in the design phase and that a priori estimates can be used to predict the actual effort. We highlight also a clear tendency to underestimate the actual effort.

Mendes et al. [9] compare the prediction accuracy of three case-based reasoning (CBR) techniques to estimate the effort to develop Web hypermedia applications. They also compare the best CBR technique, according to their results, against three common prediction models: multiple linear regression, stepwise regression and regression trees.

If we move a step further towards effort prediction models, Reifer (WebMo, [10, 11]) tailors the COCOMO II model to Web development by revising the cost factors. Notice that WebMo introduces *Web objects* as size measure for these applications.

Similarly, Ruhe et al. [12] propose an adaptation of the COBRA method (COst estimation, Benchmarking, and Risk Assessment, [4]) to the Web domain and apply it to the data from 12 projects developed by a small Australian company.

7 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 correlational models can be built that link measures of the attributes of web designs and various categories of design effort.

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:

- Including an automated measurement tool in the environment that supports the W2000 notation. Currently, W2000 is supported by a special-purpose add-in for Rational Rose. We are also implementing a new environment based on open-source products;
- Investigating the impact of development tools (previous bullet);
- Investigating prediction for other external attributes;
- Investigating subsequent web development phase, e.g., implementation and verification;
- Replicating and refining this study, based on the experience we have acquired.

References

[1] L. Baresi, E. Bianchi, L. Mainetti, M. Maritati, and A. Paro. UWA Hypermedia - User Manual. Technical Report UWA-15, Politecnico di Milano, 2001.

[2] L. Baresi, F. Garzotto, and P. Paolini. From Web Sites to Web Applications: New Issues for Conceptual Modeling. In *Proceedings of the International Workshop on The World Wide Web and Conceptual Modeling, co-located with the 19th International Conference on Conceptual Modeling*, 2000.

[3] L. Baresi, S. Morasca, and P. Paolini. An Empirical Study on the Design Effort of Web Applications. In *Proceedings of the* 3^{rd} *International Conference on Web Information System Engineering*, pages 345–354. IEEE-CS Press, 2002.

[4] L. Briand, K. El Emam, and F. Bomarius. COBRA: A Hybrid Method for Software Cost Estimation, Benchmarking, and Risk Assessment. In *Proceedings of 20th International Conference on Software Engineering*, pages 390–399. ACM Press, 1998.

[5] L. Briand, S. Morasca, and V. Basili. An Operational Process for Goal-Driven Definition of Measures. *IEEE Transactions on Software Engineering*, 28(12):1106 – 1125, December 2002.

[6] Y. Deshpande. Consolidating Web Engineering as a Discipline. *Software Engineering Australia*, pages 31–34, 2002.

[7] E. Mendes, S. Counsell, and N. Mosley. Measurement and Effort Prediction of Web Applications. In *Proceedings of the* 2^{nd} *ICSE Workshop on Web Engineering*, 2001.

[8] E. Mendes, N. Mosley, and S. Counsell. Web Metrics - Estimating Design and Authoring Effort. *IEEE Multimedia, Special Issue on Web Engineering*, pages 50–57, January/March 2001.

[9] E. Mendes, I. Watson, C. Triggs, N. Mosley, and S. Counsell. A Comparison of Length, Complexity & Functionality as Size Measures for Predicting Web Design & Authoring Effort. In *Proceedings of the IEEE Metrics Symposium, 2002*. IEEE-CS Press, 2002.

[10] D. Reifer. Web-Development: Estimating Quick-Time-to-Market Software. *IEEE software*, 17(8):57–64, November/December 2000.

[11] D. Reifer. Ten Deadly Risks in Internet and Intranet Software Development. *IEEE software*, 18(2):12–14, March/April 2002.

[12] M. Ruhe, R. Jeffery, and I. Wieczorek. Cost Estimation for Web Applications. In *Proceedings of the* 25^{th} *International Conference of Software Engineering*. ACM Press, 2003. To appear.