



# An Initial Experimental Assessment of the Dynamic Modelling in UML

MARI CARMEN OTERO

jipotvim@vc.ehu.es

JOSÉ JAVIER DOLADO

dolado@si.ehu.es

*Department of Computer Languages and Systems, University of the Basque Country, Spain*

**Abstract.** The goal of this empirical study is to compare the semantic comprehension of three different notations for representing the dynamic behaviour in unified modelling language (UML): (a) sequence diagrams, (b) collaboration diagrams, and (c) state diagrams. Eighteen students of Informatics analysed the three types of diagrams within three different application domains. We performed a  $3 \times 3$  factorial experimental design with repeated measures. The metrics collected were total time and total score. The main conclusion of this study is that the comprehension of the dynamic modelling in object-oriented designs depends on the diagram type and on the complexity of the document. The software project design written in the UML notation is more comprehensible, when the dynamic behaviour is modelled in a sequence diagram. While if it is implemented using a collaboration diagram, the design turns out to be less comprehensible as the application domain, and consequently, the document is more complex.

**Keywords:** Dynamic modelling, unified modelling language, semantic comprehension,  $3 \times 3$  Latin square.

## 1. Introduction

Nowadays, unified modelling language (UML) is becoming a standard adopted by many software development companies. But, although its use continues expanding, there is little empirical evidence that demonstrates the qualities attributed to it by its creators. Some opinions expressed at the ESSDE Workshop emphasised the need of empirical evaluation of the object-oriented (OO) artifacts and methods (Briand et al., 1999).

Given the complexity and the demanding nature of today's information systems, the OO analysis and design process requires techniques for dealing with this complexity and techniques that allow the early detection of errors in their models. An effective form of controlling complexity is to make use of the principle of separation of views that UML supports. Regarding the other requirement (detection of errors, in Travassos et al., 1999) a reading technique was studied, based on the characteristic of the traceability of the OO designs (documented in UML). Error detection was oriented towards perceiving the defects existing between the static structure diagrams and those of dynamic modelling.

The communicability, in terms first of reading and then of understanding, that transmits a design document written in a more or less formal notation, constitutes a

decisive factor for adopting a method. Also, one of the reasons for using UML is the construction of visually expressive static and dynamic models, which are easy to interpret by their users. Therefore, our investigation does not try to explore new reading techniques, but rather it is focused on experimenting “in vitro” with the semantic comprehension of the different dynamic diagrams related to modelling in UML.

## 2. Experimental Description

### 2.1. Goal of the Experiment

We have used the well-known Goal-Question-Metric (GQM) paradigm for organising the experiment (Basili and Rombach, 1988; Solingen and Berghout, 1999). There are five parameters in a GQM template used to specify an experiment: (1) object of study; (2) purpose of the study; (3) focus: the characteristics of the object of study that are of interest; (4) point of view and (5) context: the environment in which the study is performed.

Rephrasing these five characteristics in our research, we can say that our objective is: “To analyse the *dynamic models of OO designs written in the UML notation*, in order to *evaluate their semantic comprehension* from the viewpoint of the *software developer* in the context of an *advanced university classroom*”.

### 2.2. Experimental Variables

The experiment manipulates two independent variables:

- Diagram type used in design documents to represent the dynamic behaviour in UML (DIAGRAM): sequence, collaboration and state diagrams are evaluated.
- Application domain of the design documents (APPLICATION): a Simple Cellular Telephone (Martin, 1998), a Library System (Eriksson and Penker, 1998) and a Digital Dictaphone (Porres and Lilius, 1999) are studied. These “deliverable” designs make reference to part of these software projects.

The dependent variables that we will examine are:

- Total Time (TSEC): It is the sum of the time elapsed in answering each question (each time is measured in hundredth seconds and then the total time is rounded up in seconds).
- Total Score (NRESP): It is the number of correct answers (a 1 is counted for each correct answer and a 0 for each incorrect one).

The dependent variables are measured on a ratio scale and the independent variables on a nominal scale.

### 2.3. Hypotheses

It is clear that in this empirical study the semantic comprehensibility is measured by the two dependent variables mentioned before. Therefore we formulate in Table 1 the following null hypotheses:

*Table 1.* The intersection of a line with a column is a null hypothesis of this experiment.

	TSEC	NRESP
DIAGRAM $\times$ APPLICATION interaction effect There is no difference between $3 \times 3 = 9$ experimental conditions with respect to ...	$H_{01-TSEC}$	$H_{01-NRESP}$
DIAGRAM main effect There is no difference between the subjects using the three diagram types with respect to ...	$H_{02-TSEC}$	$H_{02-NRESP}$
APPLICATION main effect There is no difference between the subjects reading OO design documents with respect to ...	$H_{03-TSEC}$	$H_{03-NRESP}$

### 2.4. Experimental Material

The instruments needed to make this experiment are described as follows:

1. Notes and bibliography about the UML notation.
2. Three design documents written in the UML modelling language.

The three documents belong to three different applications: Simple Cellular TELEphone (Martin, 1998), LIBrary System (Eriksson and Penker, 1998) and Digital DICtaphone (Porres and Lilius, 1999). These designs, named TEL, LIB and DIC respectively, specify only one part of the dynamic behaviour of these software projects.

As a minimum, each one of them contains a use cases diagram, a packages diagram and a classes diagram. However, in the case of dynamic modelling, the behaviour of a use case has to be the same in the three diagram types.

This fact implies some additional work, that is, the transformation of one behaviour notation into another. For example, it is necessary to convert a dynamic model expressed by sequence diagram into collaboration one (see Figures 3 and 4 in Appendix). Also, we have to consider the construction of a sequence diagram in a state one. For this conversion, it may be necessary to include one or several state diagrams. Several classes are instantiated in the specification of the behaviour of a specific use case. Some of them change state and, consequently, the state diagrams corresponding to these classes must be included.

3. Three tests of multiple choice questions and answers. Each test is composed of five questions and each question is written on a different page together with the four possible answers.
4. Digital chronometers. Each student makes use of a Casio HS-3 digital chronometer to measure the time elapsed to answer each question.

### **2.5. Subjects**

Given the difficulty of experimenting with computing professionals, this empirical study was carried out with 18 last-year students of Informatics, in May 1999. Previous to the experimental phase, a formation period had to be implemented. In this stage they had to attend classes about UML regularly and, also, to participate in a training exercise related to the correct use of the material. Therefore, an OO design document and a questionnaire (different to those of the experiment) were used for this phase.

The training phase was carried out for 1 day with an OO design related to the operation of a Coffee Machine. An OO design document, a test of three questions and a digital chronometer were provided to all students. They were also given instructions about the order of the activities:

1. Set the chronometer at 00:00:00.00.
2. Start the chronometer.
3. Read the question thoroughly.
4. Evaluate the four solutions according to UML models that were provided; afterwards, write down the selected option in the box indicated.
5. Stop the chrono and write down the time.

In order to avoid false time recordings, the students were informed that the results obtained would not have any influence on the final grade. However, each round was performed under exam conditions.

### **3. Experimental Design**

As the two factors to manipulate have three levels, and the number of individuals is small, we perform a  $3 \times 3$  factorial design with repeated measures. It is clear that the total number of treatment combinations is 9, but it is not feasible to measure the subjects in all them (see Table 2). This is due to the restrictions that we impose on our experiment in order to avoid learning effects. Once a student has seen a particular design document it would be incorrect to allow him/her to use the same document in the next round of the experiment. If this happened there would be a practice effect

associated with the experimental material that would represent a severe threat to the validity of the experimental results. Under these restrictions and in order to use a complete factorial, our experimental design is partitioned into three blocks of size 3, as it is done in (Dean and Voss, 1999, pp. 462–464; Winer et al., 1991, pp. 590–595).

Using the blocking technique in a  $3 \times 3$  factorial we reduce the number of treatment combinations assigned to the blocks. However, this reduction implies the confounding of the effect of these blocks with the effect of the two factors (or even with the effect of their interaction), which results in the loss of some information. But, what approach should we take into account in building these blocks? A priori, the principle of confounding recommends that blocks should match higher-order interactions, as it happens in (Basili and Rombach, 1996; Roper et al., 1997). Both those experiments use the complete confounding with the  $A \times B$  interaction, assuming that the last one is negligible. The  $2 \times 2$  factorial analysis in (Basili and Rombach, 1996), that does not have complete confounding, confirms this idea, while that the  $3 \times 3$  interaction in (Roper et al., 1997) is statistically significant and should not be ignored.

Whenever possible, no interaction should be totally confounded (Cochran and Cox, 1980, p. 246). For this reason, we prefer the  $3 \times 3$  factorial design with  $A \times B$  (DIAGRAM  $\times$  APPLICATION) partially confounded. And since the  $A \times B$  interaction has two independent components:  $AB$  and  $AB^2$  (for more details see (Winer et al., 1991, p. 591)), we choose to confound the construction of the blocks with the  $AB^2$  component. The procedure for allocating the treatments to one of the three blocks is done according to the equation that defines this component:  $a + 2b = r(\text{mod}3)$ , while the subjects are assigned at random to these blocks.

The subjects numbered 1, 2, 3, 10, 11 and 12 in Table 2 were assigned to block 1 ( $r = 0$ ), which is formed by the following experimental conditions: Sequence ( $a = 0$ ) and TEL ( $b = 0$ ), Collaboration ( $a = 1$ ) and LIB ( $b = 1$ ), State ( $a = 2$ ) and DIC ( $b = 2$ ). The subjects 7, 8, 9, 16, 17 and 18 to block 2 ( $r = 1$ ); and the subjects 4, 5, 6, 13, 14 and 15 to block 3 ( $r = 2$ ). Since there are three treatments in each block (block size = 3), each student was measured three times and participated in three different rounds, which implied the use of three different design documents of the following applications (TEL: Cellular TELEphone, LIB: LIBrary System, DIC: Digital DIC-taphone). Therefore, in each session a subject answered the questions corresponding to the different designs related to the specific application domain and to the particular dynamic modelling diagram. The subjects recorded the time spent in answering these questions.

Table 2.  $3 \times 3$  factorial experimental design with repeated measures.

	Round I			Round II			Round II		
	TEL	LIB	DIC	TEL	LIB	DIC	TEL	LIB	DIC
Sequence	1,10	4,13	7,16	3,12	6,15	9,18	2,11	5,14	8,17
Collaboration	8,17	2,11	5,14	7,16	1,10	4,13	9,18	3,12	6,15
State	6,15	9,18	3,12	5,14	8,17	2,11	4,13	7,16	1,10

#### 4. Threats to the Experiment

A crucial step in the experimental design consists in minimising the impact of the threats to the internal validity, that is, of those factors that can affect the dependent variables without the researcher's knowledge. Next, we explain the way in which some factors are controlled in our experiment:

- Instrumentation effects caused by the differences in the experimental material. This factor will be measured by the (ANOVA).
- Selection effects due to the natural variation in the human performance.
- In each round, each student is assigned to one of the nine experimental conditions (see Table 2). In this way the differences in subject skills are equally spread across all the treatments of the experiment.
- Maturation effects due to the learning and practice effect as the experiment proceeds.

Manipulating the order of the diagram types and design documents the effects are minimised, because the subjects cannot improve their performance in the series of sessions. Also, in order to avoid the effect that fatigue and practice may have on the results, the experiment was carried out on three consecutive days.

This threat has not only been controlled through the experimental design, but it can also be measured through the variable "round" (maturation). This variable is independent from the variable "application" (instrumentation), since every day, that is, in each round three different design documents were used. Consequently, "round" may be considered as another independent variable.

The ANOVA shows that maturation effect is not statistically significant with respect to the dependent variables: TSEC ( $F = 0.188, p = 0.830 > \alpha$ ) and NRESP ( $F = 0.860, p = 0.429 > \alpha$ ).

- Presentation effects are due to the order in which the experimental material is distributed.

In this design no fixed order of presentation was chosen for neither of the two factors (DIAGRAM and APPLICATION). As indicated in Section 3, each day a different diagram type and design document were assigned to each subject. In this way, this threat is minimised, with the consequent risk of students exchanging information about the experimental material from one session to the other (plagiarism effect).

#### 5. Analysis of the Data

The main objective of the study is to uncover the extent up to which the diagram type and application domain factors have an influence on two dependent variables that explain the semantic comprehensibility of the OO designs (written in UML notation). Given the restrictions of our experiment, stated at beginning of Section 3, the

univariate setup is used for data configuration, as in (SPSS Inc., 1990, pp. 806–807). Applying the ANOVA technique to the data, we will contrast (test) the null hypotheses expressed in Table 1, using a significance level  $\alpha = 0.1$ . In fact, the analysis will be carried out with two perspectives. In the first one, the block is considered another additional factor. The information obtained under this first approach justifies the need for a second analysis, where the construction of the three blocks is part of the random experimental error.

### 5.1. First ANOVA

Our factorial design is equivalent to the five experimental plan of  $3 \times 3$  Latin square of repeated measures (Winer et al., 1991, pp. 702–704). Considering that the interactions with the block factor will be negligible, the ANOVA model for the analysis of this experiment is:

$$X_{ijkm} = \mu + A_i + B_j + C_k + D(C)_{m(k)} + (AB)_{ij}' + \varepsilon_{ijkm}$$

where  $X_{ijkm}$  is the observed response (with respect to each of the two dependent variables),  $\mu$  is the overall mean response,  $A_i$  is the main effect of diagram type,  $B_j$  is the main effect of application domain,  $C_k$  is the effect associated with the block factor,  $D(C)_{m(k)}$  are the effects associated with the subjects within the blocks and  $\varepsilon_{ijkm}$  is the experimental error. The symbol  $(AB)_{ij}'$  indicates that only partial information is available of the  $A \times B$  interaction, which corresponds to the  $AB$  component and is free of confounding.

This model is mixed, because the  $D$  factor (the subjects) is random and the rest are fixed, and balanced, having the same number of observations in each cell. The reasons for choosing this experimental model are:

- If the interaction of the factors has a significant value, there is only partial confounding with the block main effect.
- Otherwise, the factors are independent and the test corresponding to the main effect of the DIAGRAM factor or the APPLICATION factor is more powerful (sensitive), that is, it has a higher probability to reject  $H_0$  correctly.

#### 5.1.1. Dependent Variable: TSEC

Considering the results of Table 3, the analysis reveals that the  $AB$  component (not confounded) of the interaction between the diagram type and the application domain is significant (Sig. of  $F = 0.065$ ), in contrast to the  $AB^2$  component, which is not.

### 5.1.2. Dependent Variable: NRESP

The results of the ANOVA with respect to NRESP, summarised in Table 4, contradict those obtained with TSEC. The exploration of the data indicates that the  $AB^2$  component of the interaction between the two factors is statistically significant (Sig. of  $F = 0.040$ ). The  $AB^2$  effect is totally confounded with the block main effect. On the contrary, the  $AB$  component is not significant.

## 5.2. Second ANOVA

Given the considerable effect of one of the two components with respect to the total time (TSEC) and the total score (NRESP), it is necessary to obtain the combined information about the  $A \times B$  interaction. Therefore, the data should be analysed as a  $3 \times 3$  randomised block factorial (Cochran and Cox, 1980, p. 227). In this case, the assignment of the treatment combinations to the blocks is done at random, rather than obeying one equation.

### 5.2.1. Dependent Variable: TSEC

According to the ANOVA of Table 5, we should reject the null hypothesis  $H_{01-TSEC}$  with respect to the total time. The analysis reveals that the interaction between

Table 3. First ANOVA for the TSEC variable.

Source		Dependent variable					
		Total time in seconds (TSEC)					
		Type III SS	Dg	MS	$F$	Sig.	Power <sup>a</sup>
Intercept	Hypothesis	53465429.578	1	53465429.6	164.766	0.000	1.000
	Error	4867405.967	15	324493.731 <sup>b</sup>			
$A$ (DIAGRAM)	Hypothesis	25091.033	2	12545.516	0.409	0.668	0.189
	Error	921127.549	30	30704.252 <sup>c</sup>			
$B$ (APPLICATION)	Hypothesis	967517.851	2	483758.926	15.755	0.000	1.000
	Error	921127.549	30	30704.252 <sup>c</sup>			
BLOCK or $AB^2$ component	Hypothesis	1394648.182	2	697324.091	2.149	0.151	0.512
	Error	4867405.967	15	324493.731 <sup>b</sup>			
$AB$ component	Hypothesis	184285.991	2	92142.996	3.001	0.065	0.671
	Error	921127.549	30	30704.252 <sup>c</sup>			
SUBJECT (BLOCK)	Hypothesis	4867405.967	15	324493.731	10.568	0.000	1.000
	Error	921127.549	30	30704.252 <sup>c</sup>			

<sup>a</sup>Computed with  $\alpha = 0.1$ .

<sup>b</sup>MS (SUBJECT(BLOCK)).

<sup>c</sup>MS (Error).

Table 4. First ANOVA for the NRESP variable.

Source		Dependent variable					
		Number of correct answers (NRESP)					
		Type III SS	Dg	MS	<i>F</i>	Sig.	Power <sup>a</sup>
Intercept	Hypothesis	718.685	1	718.685	424.606	0.000	1.000
	Error	25.389	15	1.693 <sup>b</sup>			
<i>A</i> (DIAGRAM)	Hypothesis	0.481	2	0.241	0.202	0.818	0.144
	Error	35.778	30	1.193 <sup>c</sup>			
<i>B</i> (APPLICATION)	Hypothesis	6.481	2	3.241	2.717	0.082	0.631
	Error	35.778	30	1.193 <sup>c</sup>			
BLOCK or <i>AB</i> <sup>2</sup> component	Hypothesis	<u>13.592</u>	2	6.796	4.015	0.040	0.756
	Error	25.592	15	1.693 <sup>b</sup>			
<i>AB</i> component	Hypothesis	<u>4.593</u>	2	2.296	1.925	0.163	0.501
	Error	35.778	30	1.193 <sup>c</sup>			
SUBJECT (BLOCK)	Hypothesis	25.389	15	1.693	1.419	0.201	0.796
	Error	35.778	30	1.193 <sup>c</sup>			

<sup>a</sup> Computed with  $\alpha = 0.1$ .<sup>b</sup> MS (SUBJECT(BLOCK)).<sup>c</sup> MS (Error).

diagram type and application domain is significant (Sig. of  $F = 0.026$ ). Also this analysis confirms that:

- The relation shown in (Winer et al., 1991, p. 704):

$$SS_{A \times B} = SS_{\text{BLOCK or } AB^2} + SS_{AB},$$

where the sum of the two numbers that appear underlined in Table 3 verifies that  $1394648.182 + 184285.991 = 1578934.173$ ; this result corresponds to the SS column of the DIAGRAM  $\times$  APPLICATION interaction in Table 5.

- Assuming that the previous interaction was not important, the observed power of the tests on the two factors (DIAGRAM and APPLICATION) is greater in the first ANOVA than in the second (see Tables 3 and 5).

When an interaction is statistically significant, the main effects of the variables that participate in this interaction are not analysed. In other words, it is not possible to further investigate any significant difference between the three types of dynamic models whilst it is not possible to separate their effect from the studied application domain. Therefore, it is appropriate to examine the simple effects for each independent variable (DIAGRAM and APPLICATION). It is a logical consequence derived from the discovery of a significant interaction between these factors. The results are presented in Table 6 and their meaning is analysed in Section 5.3.

Table 5. Second univariate ANOVA for TSEC.

Source	Dependent variable					
	Total time in seconds (TSEC)					
	Type III SS	Dg	MS	<i>F</i>	Sig.	Power <sup>a</sup>
Model	2571543.057	8	321422.882	2.499	0.025	0.921
Intercept	53465429.578	1	53465429.6	415.640	0.000	1.000
DIAGRAM	25091.033	2	12545.516	0.098	0.907	0.121
APPLICATION	967517.851	2	483758.926	3.761	0.031	0.771
DIAGRAM × APPLICATION	1578934.173	4	394733.543	3.069	0.026	0.857
Error	5788533.513	45	128634.078			
Total	61825506.151	54				
Correlated Total	8360076.573	53				

<sup>a</sup> Computed with  $\alpha = 0.01$ .

### 5.2.2. Dependent Variable: NRESP

According to the ANOVA of Table 7, we should reject the null hypothesis  $H_{01-NRESP}$  with respect to the total score. The analysis reveals the significant interaction between the diagram type and the application domain. Similarly, this study confirms the relationship indicated in (Winer et al., 1991, p. 704), where the SS column of the **DIAGRAM × APPLICATION** in Table 7 is equal to the sum of SS of its two components:  $18.185 = 13.592 + 4.593$  (numbers underlined in Table 4). It is also observed that when the interaction is negligible, the first analysis has achieved more contrasting power related to the main effects of the factors that are investigated (compare Tables 4 and 7).

As in the previous section, it is impossible to separate the **DIAGRAM** effect from the **APPLICATION** effect. This might be considered as a failure in the experiment (for example, the three OO design documents are not similar enough). However, a more reflexive interpretation is that each “dynamic” diagram contains some semantic depending on the complexity and the application domain of the OO designs. To verify this, an ANOVA of the simple effects is made for each factor in Table 8, as well as a comparison of these results with the previous ones in Section 5.3.

### 5.3. Interpretation of Experimental Results

The timing data collected is presented in summarised form in Table 9. The intersection of a line with a column gives the average total time and the standard deviation:  $\bar{X}_{TSEC}(S_{TSEC})$  for a specific type diagram and application domain.

Both in Table 9 and in Figure 1 we observe that the average time in answering all questions of the different diagrams is not distributed uniformly between the appli-

Table 6. Simple effects of total time (TSEC) for the APPLICATION factor within Diagram Type (DIAGRAM).

Source of variation	Tests of significance for TSEC using UNIQUE sums of squares				
	SS	DF	MS	<i>F</i>	Sig. of <i>F</i>
WITHIN + RESIDUAL	5812723.19	47	123674.96		
Application within Sequence Diagram	29624.76	2	14812.38	0.12	0.887
Application within Collaboration Diagram	1632750.73	2	816375.36	6.60	0.003
Application within State Diagram	883991.61	2	441995.80	3.57	0.036
Model	2546367.09	6	424394.52	3.43	0.007
Total	8359090.28	53	157718.68		

Table 7. Second univariate ANOVA for NRESP.

Source	Dependent variable					
	Number of correct answers (NRESP)					
	Type III SS	Dg	MS	<i>F</i>	Sig.	Power <sup>a</sup>
Model	25.148	8	3.144	2.313	0.036	0.898
Intercept	718.685	1	718.685	528.733	0.000	1.000
DIAGRAM	0.481	2	0.241	0.177	0.838	0.139
APPLICATION	6.481	2	3.241	2.384	0.104	0.589
DIAGRAM × APPLICATION	18.185	4	4.546	3.345	0.018	0.885
Error	61.167	45	1.359			
Total	805.000	54				
Corrected Total	86.315	53				

<sup>a</sup>Computed with  $\alpha = 0.01$ .

cations. An exception to this is the sequence diagram, where the speed of the answers does not depend on the complexity of OO designs. This fact can also be verified in the analysis of the simple effects of Table 6 with respect to TSEC (Sig. of  $F = 0.887$ ). On the other hand, there are significant differences between the three designs in the other two diagram types, as shown in the analysis of Table 6, the descriptive statistics of Table 9 and the profile plot of Figure 1.

Table 10 presents the collected scoring data in summarised form. The intersection of a line with a column gives the average total score and the standard deviation:  $\bar{X}_{\text{NRESP}}(S_{\text{NRESP}})$  for a specific type diagram and application domain.

In Figure 2 and Table 10 we see that the highest score is obtained with sequence and collaboration diagrams in the application of the Cellular Telephone and with state diagram in the Digital Dictaphone. In the case of the design of the Cellular

Table 8. Simple effects of total score (NRESP) for the APPLICATION factor within Diagram Type (DIAGRAM).

Source of variation	Tests of significance for NRESP using UNIQUE sums of squares				
	SS	DF	MS	<i>F</i>	Sig of <i>F</i>
WITHIN + RESIDUAL	61.65	47	1.31		
Application within Sequence Diagram	7.44	2	3.72	2.84	0.069
Application within Collaboration Diagram	10.11	2	5.06	3.85	0.028
Application within State Diagram	7.11	2	3.56	2.71	0.077
Model	24.67	6	4.11	3.13	0.011
Total	86.31	53	1.63		

Table 9. Statistical summary of the experiment total time (mm:ss).

	Sequence diagram	Collaboration diagram	State diagram
Cellular Telephone	16:31 (05:39)	09:46 (03:55)	14:14 (07:26)
Library System	15:56 (05:55)	22:02 (06:20)	14:27 (03:42)
Digital Dictaphone	17:34 (04:06)	16:29 (06:35)	22:11 (08:18)

Telephone, there are not any real differences among the interactions: (a) between objects represented sequentially in time (sequence diagram) or (b) between objects represented spatially (collaboration diagram). Something similar happens with the design of a Library System, since the total number of correct answers is similar regardless of the diagram type (see Table 10 and dotted line corresponding to Library in Figure 2).

If we look at Figures 1 and 2 in a parallel way, we can make a comparison of the design documents for each diagram type:

- *Sequence diagram*: As shown in Figure 1, the complexity of designs and application domains does not have effect in the time needed for understanding them. In Figure 2, however, there are differences in favour of the Cellular Telephone with respect to the variable NRESP (total score), due to the minor complexity of the Telephone document compared to the Library and Dictaphone designs.
- *Collaboration diagram*: At first sight the collaboration diagram seems to be the least comprehensible, because in Figure 1 the Library System obtains the highest time and in Figure 2 the Digital Dictaphone obtains the lowest score. However, a closer observation of Figure 1 indicates us that the relationship between the designs is TEL < DIC < LIB, considering the total time from the lowest to the highest. If we now consider the total score from the highest to the lowest, the

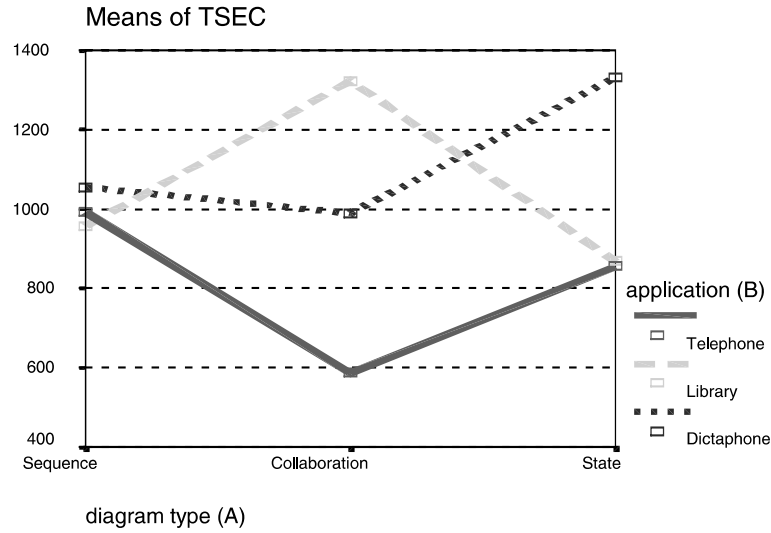


Figure 1. Profile plot of the total time (TSEC) in seconds.

Table 10. Statistical summary of the experiment total score (5 points maximum).

	Sequence diagram	Collaboration diagram	State diagram
Cellular Telephone	4.5 (0.55)	4.5 (0.84)	3.33 (1.21)
Library System	3 (1.26)	3.5 (1.38)	3.33 (1.37)
Digital Dictaphone	3.33 (1.21)	2.67 (1.63)	4.67 (0.52)

above relationship is  $TEL > LIB > DIC$ , as can be verified in Figure 2. That is, as the complexity of the application domain increases (taking into account that the Library System: LIB and the Digital Dictaphone: DIC are more complex than the Cellular Telephone: TEL), more time is invested in understanding this diagram and less correct answers are given. In addition, these statistically significant differences can be seen in Table 6 for TSEC (Sig. of  $F = 0.003$ ) and in Table 8 for NRESP (Sig. of  $F = 0.028$ ).

- *State diagram*: Examining Figures 1 and 2, total times for understanding the Cellular System and the Library System are similar, as well as the number of correct answers. However, the ANOVA reveals statistically significant differences in Table 6 for TSEC (Sig. of  $F = 0.036$ ) and in Table 8 for NRESP (Sig. of  $F = 0.077$ ). These results are due to the design of the Digital Dictaphone, because the state diagram proves to be the most appropriate for specifying the behaviour of a real-time reactive system. This diagram provides the highest number of correct answers, but it also proves to be more time consuming.

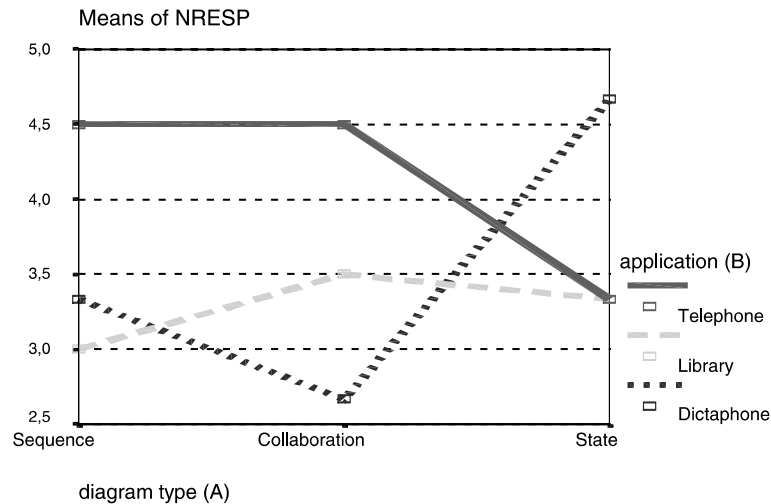


Figure 2. Profile plot of total score (NRESP).

## 6. Conclusions

Our investigation on the semantic comprehension of the dynamic models of the OO designs written in the UML notation is based in two metrics: total time and number of correct answers. The measure of time is important. Some empirical studies obtained inconclusive results, when time was not considered as a dependent variable, as analysed in (Scanlan, 1989). Scanlan concluded that “the time needed to comprehend is the most sensitive and most important measure”.

On the other hand, in this experiment the results of using time as the dependent variable are similar to the results of using total score as the dependent variable. In both cases the results show an interaction between the diagram type factor and the design document factor. This fact can be observed graphically in Figures 1 and 2, where there are not any parallel lines, meaning that the interaction is statistically significant between the factors. Therefore, the comprehension of dynamic modelling of the OO designs written in the notation UML depends on the diagram type and on the design document.

The designs, of which dynamic models are implemented using a sequence diagram, are more comprehensible. On one hand, these designs require the same amount of time for the understanding of their meaning, regardless of the complexity of the application domains. Another reason is that their comprehension implies less time and more correct answers, compared to the other diagram types.

When the dynamic behaviour is modelled in a collaboration diagram, its design document turns out to be less comprehensible than if a sequence or state diagram had been used. Concretely, this comprehension is minor in the Library System and Digital Dictaphone, which are more complex than the Cellular Telephone. In the

document of the Library System more time is needed and in the one of the Dictaphone less correct answers are achieved when a collaboration diagram appears.

With respect to the state diagram, the best results are obtained with the design of the real-time reactive system regarding the total time and the number of correct answers.

Finally, in the literature related to the experimentation in software engineering, most of the studies only deal with the main effects of the factors and do not include the effects of the interactions. However, we have shown in this experiment that investigating the interaction between factors is essential to understanding the results of the experiment. More practical work with the models is needed, in order to identify which diagrams provide the most appropriate semantics for each domain.

## **Appendix**

The experimental material and data of this research can be downloaded at the following address <http://www.sc.ehu.es/jiwdocoj/ieadm/ieadmUML.htm>. The recorded data is available for analysis for the following statistical software programs: JMP v.4, NCSS 2000 and SPSS v.9.

### ***A. Cellular Telephone***

#### **Questionnaire on the Cellular Telephone**

1. In view of the diagrams you have, it is only possible to dial:
  - a. Internal numbers of four digits inside the telephone network.
  - b. National numbers of nine digits.
  - c. There is not limit to the number of digits.
  - d. None of the previous answers.
2. If the dialled number is not valid, what happens?
  - a. Nothing, because it is not specified in these diagrams.
  - b. A new dial tone is emitted, that is, the Cellular Telephone indicates that another call can be made.
  - c. A busy tone is emitted.
  - d. An error message is emitted to inform you that the dialled number does not exist.
3. Once the telephone number is dialled:
  - a. The connection is made immediately.
  - b. You have to press the Send button.
  - c. You have to wait until the ringing tone is returned, because this tone indicates that calling is in progress.
  - d. None of the previous answers.
4. As you dial the digits corresponding to the phone number:
  - a. Each digit is visualised on a screen.

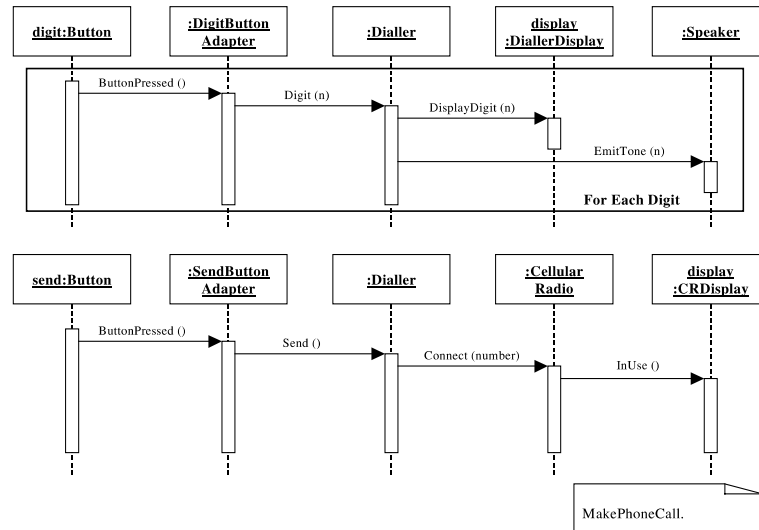


Figure 3. Sequence diagram for MakePhoneCall use case of the Telephone.

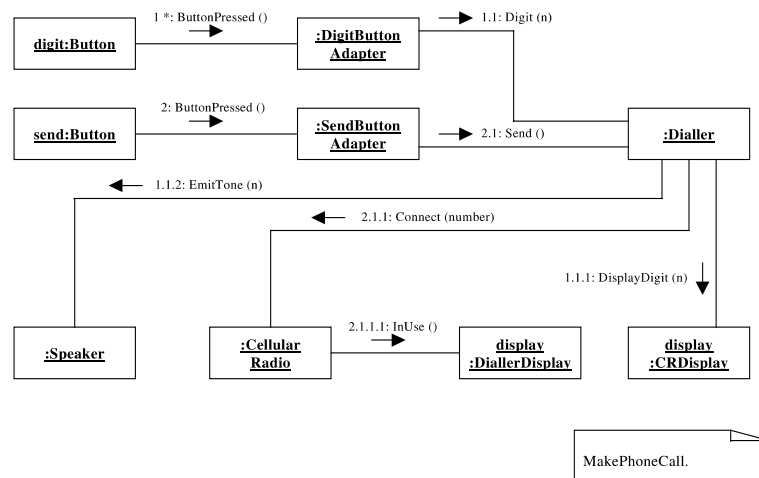


Figure 4. Collaboration diagram for MakePhoneCall use case of the Telephone.

- b. The tone corresponding to the digit pressed is emitted.
  - c. The two events described in sections (a) and (b) take place.
  - d. None of the previous answers.
5. To make a phone call, the first thing that has to happen is that the In Use display goes red.
- a. True.
  - b. False, this display is activated just after dialling all the digits of the number.

- c. False, this display is activated before pressing the Send button.
- d. None of the previous answers.

### ***B. Library System***

#### **Questionnaire on the Library System**

1. To lend an item of a book or a magazine, it is necessary to have:
  - a. A reservation apart from the title and the item.
  - b. No reservation, but the title and the item exist.
  - c. The title and the item exist, whether there is a reservation or not.
  - d. None of the previous answers, because it is not specified in these diagrams.
2. In view of the diagrams you have, which are the reasons for a reservation to be deleted?
  - a. When the time limit of 30 days has been exceeded, the elimination is automatic.
  - b. When a book or a magazine, for which a reservation existed, is loaned.
  - c. The two reasons described in sections (a) and (b).
  - d. None of the previous answers.
3. How many reservations can be made for a book or a magazine?
  - a. A maximum of 10.
  - b. Between 1 and 10.
  - c. There is no limit.
  - d. None of the previous answers.
4. When a reservation is destroyed, what is actually eliminated is the reservation of:
  - a. Title.
  - b. Item.
  - c. Title and Item.
  - d. None of the previous answers.
5. When a user, either a person or another library, wishes to borrow a book or a magazine which is still on loan, what happens in the system?
  - a. It checks whether there are other items of the book or the magazine available.
  - b. It makes a reservation for this book or magazine.
  - c. It eliminates the reservation there was for the book or magazine on loan.
  - d. Nothing, because it is not indicated in these diagrams.

### ***C. Digital Dictaphone***

#### **Questionnaire on the Digital Dictaphone**

1. In the dictaphone there are 4 recorded messages. Suppose that the message number 3 is playing now and you realise because of its content you are not interested in having it in the memory. Keeping in mind your diagrams, how is the elimination?

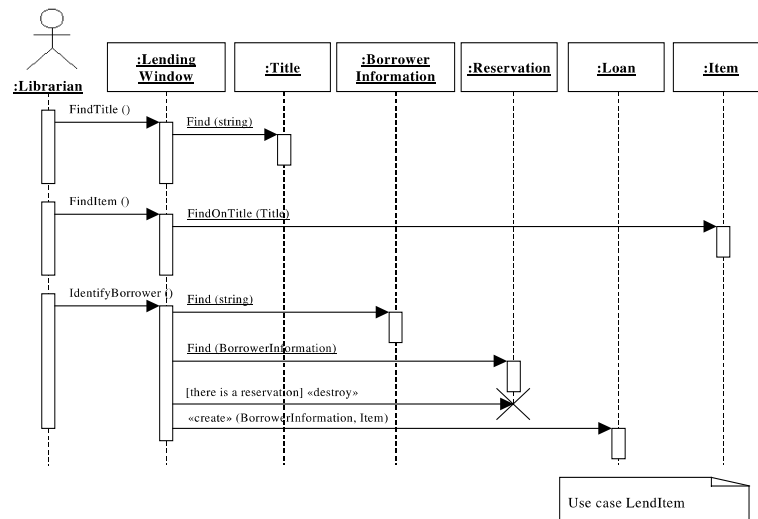


Figure 5. Sequence diagram for LendItem use case of the Library System.

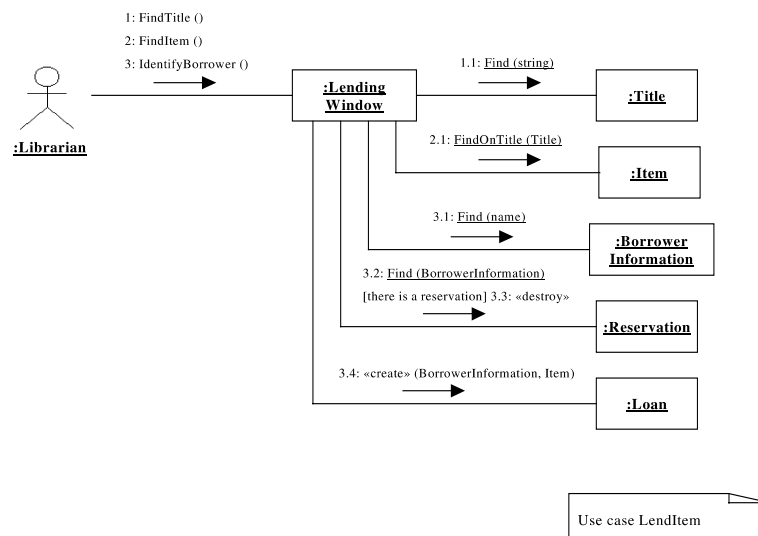


Figure 6. Collaboration diagram for LendItem use case of the Library System.

- a. This message can be erased while you are listening to it.
- b. To delete the message it is necessary to wait until it finishes.
- c. Before eliminating it from the message memory, you should always press the Stop button.

- d. None of the previous answers.
2. If the length of a message is 60 audio blocks, how long does it take to play it until the end?
  - a. 20 s.
  - b. 25 s.
  - c. 30 s.
  - d. 60 s (1 min), it is the time limit to record as well as to play a message.
3. If, while you are listening to a message, the alarm goes off, what happens?
  - a. Nothing, because this situation is not specified in these diagrams.
  - b. As there are two audio devices, one emits the alarm and the other one plays the message.
  - c. The alarm sound has priority over the message sound.
  - d. The alarm sound is not emitted until the end of the message.
4. In view of the diagrams you have, how many messages does this dictaphone allow you to play, providing you do not use the option of recording new messages?
  - a. It depends on the storage space of message memory.
  - b. It depends on the length of the messages.
  - c. It depends on the two factors expressed in sections (a) and (b).
  - d. 10 messages only.
5. Suppose the alarm is ringing at this moment, it can be interrupted when:
  - a. You press the Stop button.
  - b. A minute has gone by.
  - c. Any of the two conditions expressed in sections (a) and (b) are met.

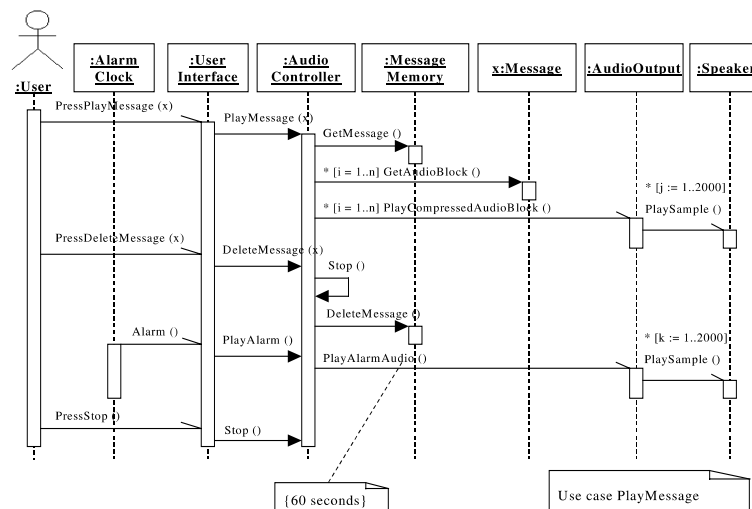


Figure 7. Sequence diagram for PlayMessage use case of the Dictaphone.



- Scanlan, D. A. 1989. Structured Flowcharts Outperform Pseudocode: An Experimental Comparison. *IEEE Software* September: 28–36.
- Solingen, van, R., and Berghout E. 1999. *The Goal/Question/Metric Method*. London, UK: McGraw-Hill.
- SPSS Inc. 1990. *SPSS Reference Guide*. Chicago: SPSS Inc.
- Travassos, G. H., Shull, F., Frederiks, M., and Basili, V. R. 1999. Detecting Defects in Object Oriented Designs: Using Reading Techniques to Increase Software Quality. *OOPSLA'99*. Denver, Colorado (USA).
- Winer, B. J., Brown, D. R., and Michels, K. M. 1991. *Statistical Principles in Experimental Design*. 3rd Ed. New York: McGraw-Hill Series in Psychology.



**Mari Carmen Otero** is Assistant Professor at the Department of Computer Languages and Systems, University of the Basque Country, Spain. She received her B.Sc. in Computer Science from the University of Deusto, Bilbao, Spain in 1988. Afterwards, she worked as a computer analyst for several local companies for 4 years. She is currently working towards her Ph.D. in the area of experimental software engineering.



**José Javier Dolado** is a Lecturer in the Department of Computer Languages and Systems at the University of the Basque Country, Spain. He received both the B.Sc. (with awards) and Ph.D. degrees in Computer Science from that university in 1985 and 1989, respectively. His current research interests are in software measurement, dynamics of the software development process, and complex systems. He is specially interested in understanding the complexity of software management and in providing tools and methods for decision support. His works have appeared in several refereed journals and he has served on the program committee of international conferences on software quality and process improvement. Dr. Dolado is a member of the ACM, ACM Sigsoft, IEEE, IEEE Computer Society and IEEE Systems, Man and Cybernetics Society.