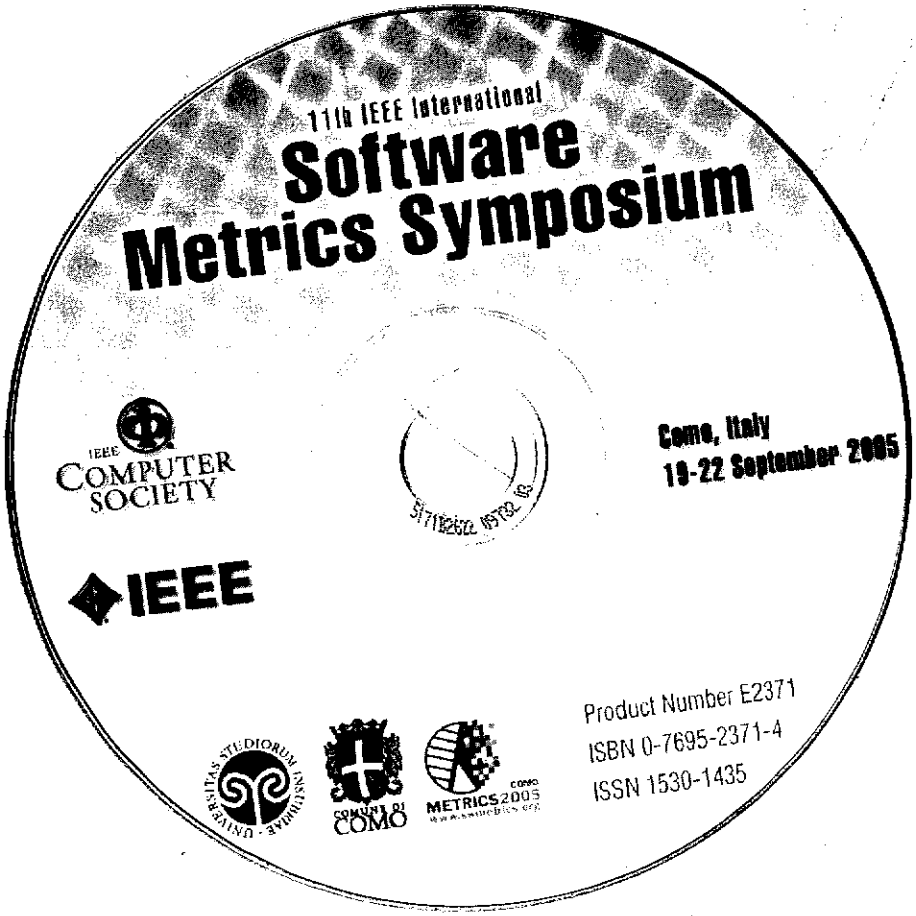

11th IEEE International
Software
Metrics Symposium
IEEE COMPUTER SOCIETY
IEEE
Como, Italy
19-22 September 2005
Product Number E2371
ISBN 0-7695-2371-4
ISSN 1530-1435
METRICS2005

Assessing the Impact of Coupling on the Understandability and Modifiability of OCL Expressions within UML/OCL Combined Models

Luis Reynoso
National University of Comahue, Neuquén, Argentina
lreynoso@uncoma.edu.ar

Marcela Genero, Mario Piattini
Castilla La Mancha University, Ciudad Real, Spain
{Marcela.Genero, Mario.Piattini@uclm.es}

Esperanza Manso
Valladolid University, Valladolid, Spain
manso@infor.uva.es

Abstract

Diagram-based UML notation is limited in its expressiveness thus producing a model that would be severely underspecified. The flaws in the limitation of the UML diagrams are solved by specifying UML/OCL combined models, OCL being an essential add-on to the UML diagrams. Aware of the importance of building precise models, the main goal of this paper is to carefully describe a family of experiments we have undertaken to ascertain whether any relationship exists between object coupling (defined through metrics related to navigations and collection operations) and two maintainability sub-characteristics: understandability and modifiability of OCL expressions. If such a relationship exists, we will have found early indicators of the understandability and modifiability of OCL expressions. Even though the results obtained show empirical evidence that such a relationship exists, they must be considered as preliminaries. Further validation is needed to be performed to strengthen the conclusions and external validity.

1. Introduction

Within the Object Oriented (OO) software development process, the importance of models is gradually becoming an essential aspect. This fact is corroborated by many recent initiatives such as Model-Driven Development (MDD)[1] and the Model-Driven Architecture (MDA) [24], which are based on the assumption that models are the basis of the software development (WK), and they constitute its primary focus and products [28]. Currently, the Unified Modeling Language (UML) [23] is the standard language in software development. However UML models only provide a good view of the software architecture [16] and they are imprecise because diagram-based notation is not expressive enough [12] . The expressiveness of the modeling technique used (for example the notation, etc.) affects one of the most important characteristics of a model, its understandability [28].

Moreover, models are used in very different ways and some authors have pointed out that some order and transparency in working with models is needed [32]. Recently, model maturity levels (MMLs) have been proposed in order to give a classification of building better models [32]. Prior to this, many researchers have defined in the literature many metrics as measures of the quality aspects of UML models [20], [15]. Common to both approaches is the intent behind them: the quality of models has a relevant repercussion in the software product that is finally delivered [30] [32].

Modelers can only obtain models of a high level of maturity using the combination of UML and the Object Constraint Language (OCL) [22], otherwise their models would be severely underspecified [32]. Due the importance of OCL, and aware that formal specification can greatly enhance the quality of produced software [16] [31], we have started to study OCL expressions as a crucial add-on to the UML diagrams. It was empirically proved that OCL has the potential to significantly improve UML-based model comprehension and maintainability [6]. We focused on assuring the quality of UML/OCL combined models, defining a set of metrics for OCL expressions [22] in a methodological way. We follows a process consisting of three main steps [7], [10]: metric definition, theoretical validation and empirical validation. As many authors have mentioned [2]; [14]; [18], [27] empirical validation of metrics, through experiments is

11th IEEE International Software Metrics Symposium (METRICS 2005)
1530-1435/05 $20.00 © 2005 IEEE

fundamental to assure that the metrics are really significant and useful in practice. Therefore, the goal of this paper is to carefully describe a family of experiments we have undertaken to ascertain if any relationship exists between the object coupling (defined through navigations and collection operations), and two maintainability sub-characteristics [17]: understandability and modifiability of OCL expressions. We decided to empirically validate object coupling because coupling is the most complex software attribute in object oriented systems [5] and a high quality software design should obey the principle of low coupling. We believe that a UML/OCL model reveals more coupling information than a model specified using UML only, due to the fact that with OCL it is possible to define OCL expressions constraining different objects through the use of a core concept of OCL: navigation. A navigation defines coupling between the objects involved [32], and the coupled objects are usually manipulated in an OCL expression through collections and its collection operations (to handle its elements).

This paper starts with a description of the metric definition for OCL expressions. Following that, we describe the purpose of the previous empirical work. In section 4 a description of a family of experiments is presented. Section 5 provides the data analysis and interpretation. The experimental threats are discussed in section 6. Finally the last section presents some concluding remarks and outlines directions for future research activities.

2. OCL Expressions Metrics

This work is part of a project we have been developing for the last three years with the aim of looking for early indicators of UML/OCL models' understandability and modifiability. These indicators will allow modelers to make better decisions early in the OO software development life cycle, contributing to the development of better quality OO software.

Because our intention is that the metric definition and traditional metrics can be supported by the fact they are clearly related to cognitive limitations [19], [4] we have considered the cognitive techniques applied by modelers during OCL comprehension and modification in the metric definition. In this way, we have taken into account the cognitive complexity (the mental burden of a person when he/she deals with artifacts) of modelers when they use OCL expressions. Our hypothesis is that structural properties of an OCL expression within an UML/OCL model (artifacts) have an impact on the cognitive complexity of modelers (subjects), and high cognitive complexity leads the OCL expression to exhibit undesirable external qualities on the final software product [17], such as less understandability or a reduced maintainability [13]. We have also hypothesed that during the comprehension of an OCL expression the modelers concurrently and synergistically apply two cognitive techniques [8], [9] : "chunking" and "tracing". The former involves the recognition of a set of declarations and the extraction of information from them, which is remembered as a chunk (a single mental abstraction), whereas the latter involves scanning, either forward or backwards, in order to identify pertinent chunks . So, we have defined a set of metrics considering the OCL concepts related to these cognitive techniques. Analysis of each of these techniques in turn leads to identification of structural properties which can be measured. In order to identify the broad set of OCL concepts, and not omit any of them, we have studied the OCL metamodel. A suite of metrics which measures the structural properties of OCL expressions can be found in [26], Table 1 shows the name of the metrics we used in the experiment presented in this paper and the cognitive technique they are related to. The metrics were theoretically validated using Briand et al. frameworks [26]. In the fourth column of Table 1 we partially show the result of the theoretical validation (only for the metrics used in this experiment).

3. Previous Experimental Work

In [25] we presented a family of experiments to ascertain whether any relation exists between the navigation depth (measured by DN) and the quantity of different object coupled (NNC) of an OCL expression and its understandability and maintainability. We obtained that OCL expressions understandability and modifiability are more dependent on how far objects coupled to the contextual instance are (DN) rather than how many different objects are coupled to the contextual instance (NNC).

We believe that the coupling defined in an OCL expression is significantly correlated with the understandability and modifiability of OCL expressions, and we still need to focus our efforts on the empirical proof of new results. For that reason, we have run new experiments evaluating a set of metrics related to coupling between objects which are defined through navigations, collections and collection operations.

Table 1: Metrics for OCL expressions of UML/OCL models

Metric	Cognitive technique	Metric Description	Theoretical Validation		
			IBC*	S*	L*
NNR	Tracing	Number of Navigated Relationships	Yes		
NAN	Tracing	Number of Attributes referred through Navigations	Yes		
NNC	Tracing	Number of Navigated Classes	Yes		
WNCO	Tracing	Weighted Number of Collection Operations	Yes		
DN	Tracing	Depth of Navigations			Yes
WNN	Tracing	Weighted Number of Navigations		Yes	
NEI	Chunking	Number of Explicit Iterator variables		Yes	
NKW	Chunking	Number of OCL KeyWords		Yes	
NES	Chunking	Number of Explicit *Self*		Yes	
NCO	Chunking	Number of Comparison Operators		Yes	
* IBC stands for *Interaction Based for Coupling*, S stands for *Size* and L stands for *Length*					

4. Family of Experiments

Relevant results can only be obtained by families of experiments rather than individual experiments. In other words, simple studies rarely provide definite answers [21] [3]. So, in order to fulfill the experiment goal previously defined in the introduction, we ran a family of experiments, consisting of three experiments. Although the experiment process follows the proposed format of Ciolkowski et al. [11] and Wohlin [33] for the sake of brevity we will show its main aspects in this section.

The first experiment took place in April 2004 and it was replicated twice in October and November 2004 respectively. Now, we describe them in details.

In the first experiment we invited the third-year students of Computer Science at the University of Alicante (UA, Spain) to do a short seminar about OCL (only 5 hours) and to do an experiment as part of the seminar. Sixty undergraduate students agreed to take part in a course. They were motivated to participate in the experiment because they would be able to obtain an extra point in the final score of the Software Engineering course if and only if they completed a test. The extra point we gave them was only dependent on finishing the exercise, not on how the exercise was done. The collected data was called "UAE".

In the first replica, twenty six students who participate in a course of the Eighth International School of Computer Science (celebrated in La Matanza University, Argentina) were the subjects of the first replica. The duration of the course was 20 hours and during the last two hours we ran the experiment replica. The subjects were undergraduate students of different universities, graduate students and teachers. The data obtained in this replication, was called "ULME" data.

In the second replica twenty nine students of fifth year enrolled on a Software Engineering course of the Austral University of Chile participated in a course of 20 hours about OCL. As an inducement to do the course, students were informed that they would do a test and its result would be considered as a point of a the course of Software Engineering. The collected data was called "UAChE".

The training sessions of the experimental subjects, seminar or courses, were conducted by the same teacher. The three experiments were carried out with supervision in a laboratory. Table 2 shows a brief description of the profile of the subjects, all the quantities are in years. We think that the course duration, the inducement for students to take the course, and their profile could have affected the experimental results.

Table 2: Subject Profile

Subject	UAE	ULME	UAChE
Average age	22	24	21
Average experience in programming	2	3	2
Average experience in modeling with UML class diagrams	1	1	Half a year

4.1. Common Aspects of the Family

In this section we will summarize the main experimental process steps common to the three experiments.

4.1.1. Independent and dependent variables: The independent variable (IV) is the object coupling defined in OCL expressions. It was measured through the metrics shown in Table 1. We used NNR , NNC, WNN, DN, WNCO, NES and NAN metrics, because in all of them an aspect of the navigation concept is captured in its intent [26]. We also use the NEI metric which is related to the collection operation iterator variables, and allows us to define the context inside the collection operations. The rest of the metrics NWK

(number of keywords) and NCO (number of comparison operators) were not related to collection operations but they are needed to define simple OCL expressions. Because we are not interested in studying the last two metrics we try to keep their value as constant as possible. For example all the OCL expressions used in experimental object are defined with three OCL keywords.

The dependent variables (DVs) are two maintainability sub-characteristics: understandability and modifiability.

4.1.2. Experimental Object: The experimental objects were nine UML/OCL combined models, each model having an OCL expression. Since we wanted to have objects of different complexity we designed them covering a wide range of the metric values (except in the case of NES, NWK, and NCO). But in reality, it is impossible to cover all of the possible combination of metrics values. Fifteen model were initially designed, but we thought that some models were quite similar, and the fact of having many models of the same complexity could bias the experiment result. For that reason we carried out a hierarchical clustering of the 15 models to group them into three groups: Low, Medium or High Complexity (we identify each complexity by using the LC, MC, HC acronyms respectively). This clustering was run using the metric values. Finally, we obtained three models of each group (see Table 3). The clustering provided us with an objective classification of the UML/OCL models, which we called ObjClass.

4.1.3. Tasks: During the test each subject had to perform three tests. The tests have the following required tasks:

- Understandability Tasks (UND-Tasks): The subjects had to answer a questionnaire consisting of 4 questions that reflected whether or not they had understood the OCL expression attached to the class diagram.
- Modifiability tasks (MOD-Tasks): The subjects had to modify the OCL expressions according to a new requirement expressed in natural language.
- Rating Tasks: After finishing any task (UND or MOD Tasks) the subject uses a scale of five linguistic labels (Table 4 shows the labels used in the UND Tasks) to rate them. This rate indicates the perception of the subjects of how complex it was for them to do UND-Tasks or MOD-Tasks. The collected data was called Understandability Subjective Complexity (UND SubComp) or Modifiability Subjective Complexity (MOD SubComp). This information is vital to estimate the cognitive load of subjects dealing with artifacts.

All three tests assigned to any subject had three different complexities, i.e. HC, MC or LC, which means there is no subject doing two tests of the same complexity. However, the tests were randomly assigned to the subjects. In this paper we identify as Ci the collection of the first tests performed by all the subjects, C2 the second collection, and so on.

We think that the time each subject spent doing each required tasks (i.e., UND Time and MOD Time) is not the most accurate measure to study the DVs. We use the understandability efficiency (UND Eff) and the modifiability efficiency (MOD Eff), being defined as:

- UND Eff = correct answers/UND Time
- MOD Eff = correct answers/MOD Time

4.1.4. Experiment Hypotheses: We formulated different hypotheses along with distinct beliefs:

- Belief 1: The structural properties related to object coupling in OCL expressions influences the degree of correctness of the performed Tasks per time, i.e. the subject's efficiency (UND Eff or MOD Eff).
- Hypotheses 1: $H_{0,1}$ There is no significant correlation between the OCL expression metrics related to object coupling and their UND Eff /MOD Eff. $H_{1,1} = \neg H_{0,1}$
- Belief 2: The structural properties related to object coupling in OCL expressions influences the subjective rate provided by subjects (UND SubComp or MOD SubComp) tasks. If so, we will be able to find an early indicator of the subject's cognitive load.
- Hypotheses 2: $H_{0,2}$ There is no significant correlation between the OCL expression metrics related to object coupling and the SubComp Eff. $H_{1,2} = \neg H_{0,2}$
- Belief 3: The UND (or MOD) time is a valued factor that influences the subjective criteria of subjects when they have to rate tasks. For example, we expect subjects to rate time-consuming understandability tasks as "quite difficult to understand" or "barely understable".
- Hypotheses 3: $H_{0,3}$ The subjective complexity (SubComp) is not correlated with the UND and MOD Time; otherwise $H_{1,3}$: $\neg H_{0,3}$
- Belief 4: We believe the degree of correctness of the tasks performed per time, i.e. the UND Eff or MOD Eff, could be an indicator of the subjective complexity of subjects when they have to rate tasks.
- Hypotheses 4: $H_{0,4}$ The subjective complexity (UND or MOD SubComp) is not correlated with the UND and MOD efficiency; otherwise $H_{1,4}$: $\neg H_{0,4}$

Table 3. Metric values for each OCL expression

object	Tracing						Chunking				Obj-class
	NNR	NNC	WNN	DN	WNCO	NAN	NEI	NES	NWK	NCO	
Model1	1	1	1	1	2	1	1	1	3	0	LC
Model2	1	1	1	1	2	1	1	1	3	1	LC
Model3	2	2	2	1	2	0	0	2	3	1	LC
Model4	3	2	6	4	3	0	1	2	3	0	MC
Model5	3	2	5	4	1	0	1	2	3	1	MC
Model6	3	2	6	4	3	0	1	2	3	0	MC
Model7	2	2	3	4	7	2	2	2	3	1	HC
Model8	3	3	3	3	5	2	2	1	3	1	HC
Model9	3	3	6	3	8	1	3	1	3	1	HC

Table 4. Linguistic labels for the Understandability Tasks (UND-Tasks)

Easily understandable	Quite easy to understand	Normal	Quite difficult to understand	Barely Understandable

5. Data Analysis and Interpretation

In this section we will summarize the main aspects of the analysis we carried out by means of SPSS [29].

As previously mentioned we have three different observations for each subject, these three observations for each subject corresponds to three models of different complexity (HC, MC or LC). C_i represents the collection of the i-tests performed by all the experimental subjects. Now we will describe which statistic test we used. Because all the hypothesis defined in the last section are concerned with dependency degree between two variables, a correlation coefficient can be used. Coefficients such as Spearman or Tau of Kendall, work with pairs of observation, (Xi, Yj), over n-objects (in our case 9 diagrams), but observations must be independent. That means for example, if we study a dependent variable, said UND Eff, of the subject "j" in the i-diagram we are not allowed to consider any other observation of the same j-subject. So, the correlations of the formulated hypothesis are tested for each C_i. In same way, studying the correlation for each C_i will indicate whether our hypotheses are dependent on the learning curve of subjects during the experiment.

The analysis of the empirical data is laid out as follows. First we will make a descriptive and exploratory study (section 5.1). In section 5.2, we will study the correlation between the proposed metrics and the dependent variable, in order to discover whether the former could predict the latter. In this section we also study the correlation between the cognitive aspects of subjects (SubComp) and the dependent variable. Afterwards (section 5.3) we analyze whether the time has influenced the students to rate the OCL expressions within UML/OCL modes, or if their efficiency has a correlation with the SubComp.

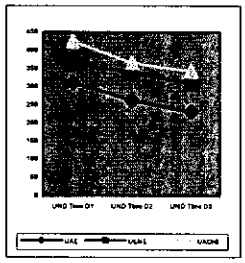
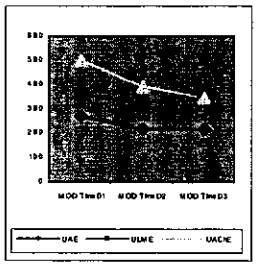
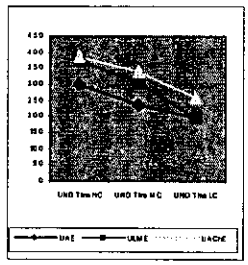
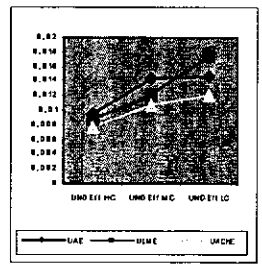
5.1. Descriptive and Exploratory Studies

The fact that the dependent variables do not follow a normal distribution was corroborated using the Shapiro Wilk tests. Table 5 shows some descriptive statistics: the Intraquartile Range (IQR), the Mean and Standard Error of the Mean (SE) for the DVs in each C_i. As previously described, the set of C_i represents the order of the performed tasks, which allows us to show how the time spent on each task decreases as new tasks are solved by subjects. In Table 6.1 and 6.2 we depict the UND and MOD Time as time passed, see that the UND and MOD time decrease during the experiment's execution. In the case of UND Eff and MOD Eff, we expected the subject rump up efficiency but it does not improve as time goes on, except in the UA experiment for UND Eff Time. However if we arrange the collected data according to the objective classification (see 6.3 and 6.4) the UND Time and UND Efficiency improves as we diminish the complexity. This is not the case for MOD Time and MOD Efficiency because the Medium Complexity (MC) tasks were more difficult to modify than the tasks corresponding to High Complexity. This situation occurs in the three experiments. The main difference between MC and HC models is that in the former the complexity is mainly based on combined navigations, (see the value of WNN) whereas in the latter the complexity is mainly based on an intertwining collection operations (see the value of WNCO).

Table 5. Mean UND/MOD Eff and UND/MOD Time during the time

	UAE			ULME			UAChE		
	IQR	Mean	SE	IQR	Mean	SE	IQR	Mean	SE
UND Eff C_1	0.035	0.012	0.007	0.026	0.013	0.008	0.05	0.011	0.01
UND Eff C_2	0.027	0.014	0.007	0.045	0.015	0.012	0.059	0.012	0.01
UND Eff C_3	0.059	0.017	0.01	0.034	0.014	0.008	0.037	0.012	0.008
MOD Eff C_1	0.03	0.007	0.009	0.021	0.006	0.006	0.03	0.007	0.008
MOD Eff C_2	0.05	0.006	0.011	0.05	0.008	0.011	0.05	0.006	0.01
MOD Eff C_3	0.033	0.006	0.008	0.033	0.007	0.009	0.031	0.006	0.009
UND Time C_1	822	311.883	152.92	567	361.576	188.848	1038	425.538	255.15
UND Time C_2	455	263.083	103.416	644	340.884	182.292	956	365.897	208.027
UND Time C_3	703	232.15	112.086	505	308.73	150.703	871	343.282	180.649
MOD Time C_1	749	266.1	162.932	998	361.615	222.793	1775	497.256	415.228
MOD Time C_2	611	213.6	125.743	807	324.23	210.228	1198	396.105	306.539
MOD Time C_3	843	214.965	130.361	729	366.923	210.58	1823	356.833	336.007

Table 6. Mean UND and MOD as time passes (C_i) and according to the objective classification

6.1 Mean UND Time during the Time	6.2 Mean MOD Time during the time	6.3 UND Time according to the Obj. Classification	6.4 UND Eff according to the Obj. Classification
			

We believe that for the subjects it was more difficult to identify and trace which relationships they should use (its rolename, attribute name, etc) in MOD Tasks, instead of identifying which operation collections should be used to modify the expression.

The descriptive statistic for mean UND Time and mean MOD Time have higher values in UAChE compared with ULME and UAE, and between the last two, the smallest mean values are from UAE. Chilean students have low experience in UML, so they required a certain amount of necessary extra time to undertake any task. Although UAE presents a higher mean UND Time than ULME their UND Efficiency are similar, if we compare the Ci.

5. 2. Testing hypotheses 1 and 2

To test the first two hypotheses, a correlation analysis was performed using Spearman's correlation coefficient with a level of significance $\alpha = 0.05$, which means the level of confidence is 95% (i.e. the probability that we accept H_0 when H_0 is true is 0.095). Tables 7 shows the Spearman significant coefficient between metrics and efficiency' DVs. The results are:

Hypotheses 1: All the metrics present a negative correlation coefficient, except several metrics as NAN and NCO in MOD Eff and NES and NCO in UND Eff in same observations within subjects. A negative coefficient means that the subjects are less efficient when the values of a metrics a high, otherwise they are more efficient.

- The NNC, WNCO and NEI metrics have several correlations with the UND Eff in the UAE and UAChE. This is logical, meaning that the number of

classes involved in the OCL expressions (NNC), the number of collection operation (WNCO) and the number of collection operation's iterator variables (NEI) influences the subjects' efficiency. This influence seems to be independent of the order of the tasks performed for UAE because we find a correlation for most of the Ci.

- The length of the navigation (DN) has correlations with the MOD Eff in the three experiments. NNR, NAN, NES and WNN have also correlations with the MOD Eff, but not for the three experiments. NAN, NES and NCO have a positive correlation coefficient, i.e. the subjects are more efficient when the values of the metrics are higher.

Hypotheses 2: All the metrics present a positive correlation coefficient except several values of NAN and NCO in MOD SubComp and NES and NCO in UND SubComp.

- We found few correlations between metrics and the UND SubComp. From the set of metrics that present a correlation just one of them is correlated twice. The significance levels were between 0.002 and 0.038.
- DN, WNN and NNR are correlated with the MOD SubComp in the three experiments. The significance levels were between 0.000 and 0.041. DN has the stronger correlation in UAE, independently of the order of the tasks. However in this experiment, the correlation of NNR and WNN is stronger as time goes on.

5. 3. Testing hypotheses 3 and 4

In order to test the 3rd and 4th hypotheses, we study the correlation using measures for ordinal data. We transformed the variables UND SubComp and MOD SubComp, assigning numbers to the linguistic labels: ranging from 1 (assigned to "Easily understandable/modifiable") to 5 (which correspond with "Barely understandable/ modifiable"). After the data was transformed we used a Kendall's Tau coefficient to analyze the correlation of $H_{0,3}$ and $H_{0,4}$. The statistics for ordinal measures are summarized in Table 8 which allow us to conclude the following:

- UND SubComp and UND Time: In the UAE and UAChE there is a statistically significant relationship between the SubComp variable and the UND Time. However in the ULME we only found correlation in one trial (C2).
- MOD SubComp and MOD Time: Regarding the MOD Time, the same results as the previous case are obtained.
- UND/MOD SubComp and UND/MOD Eff: there is a statistically significant relationship between UND SubComp and UND Eff and, between MOD SubComp and MOD Eff, in the case of UAE and UAChE experiments. In the ULME we found that MOD SubComp is correlated with the MOD Eff.

Table 7. Spearman´s correlation coefficient between Metrics and UND/MOD Eff (significant coefficients at level 0.05 are shown in bold font)

	NNR	NNC	WNN	DN	WNCO	NAN	NEI	NES	NCO
UAE UND Eff C_1	0.250	**0.021**	0.517	0.263	**0.028**	0.124	0.252	0.360	0.903
UAE UND Eff C_2	**0.035**	**0.042**	**0.027**	0.430	**0.026**	0.641	**0.029**	0.194	**0.047**
UAE UND Eff C_3	0.446	**0.002**	0.810	0.843	**0.000**	**0.000**	**0.000**	**0.019**	0.051
UAChE UND Eff C_1	0.152	**0.001**	0.938	0.590	0.057	**0.011**	**0.005**	**0.037**	**0.005**
UAChE UND Eff C_2	0.175	0.154	0.072	0.201	**0.030**	0.808	0.099	0.911	0.710
UAChE UND Eff C_3	0.404	0.696	0.769	0.488	0.585	0.674	0.670	0.765	0.747
ULME UND Eff C_1	0.278	0.150	0.279	0.484	0.066	0.147	0.053	0.350	0.698
ULME UND Eff C_2	0.440	0.993	0.677	0.982	0.748	0.762	0.970	0.456	0.132
ULME UND Eff C_3	0.987	0.338	0.760	0.311	0.126	**0.048**	0.083	**0.017**	0.296
UAE MOD Eff C_1	0.201	0.403	0.061	**0.000**	0.329	0.061	0.316	**0.000**	**0.015**
UAE MOD Eff C_2	0.479	0.851	0.794	0.689	0.072	**0.049**	0.059	0.118	0.584
UAE MOD Eff C_3	0.335	0.230	0.052	**0.001**	0.273	**0.011**	0.264	**0.000**	**0.004**
UAChE MOD Eff C_1	0.117	0.364	0.685	0.413	0.532	0.907	0.953	0.954	0.751
UAChE MOD Eff C_2	**0.031**	0.810	**0.029**	**0.010**	0.545	0.381	0.400	0.557	**0.037**
UAChE MOD Eff C_3	**0.005**	0.084	0.130	0.116	0.824	0.694	0.617	0.857	0.270
ULME MOD Eff C_1	0.166	0.374	0.479	0.057	0.903	0.680	0.977	0.241	0.831
ULME MOD Eff C_2	**0.028**	0.485	0.081	**0.010**	0.485	**0.035**	0.395	**0.021**	0.181
ULME MOD Eff C_3	0.353	0.825	0.241	0.638	0.471	**0.032**	0.186	0.312	0.543

Table 8. Correlation between subjective complexity (SubComp) and UND/MOD Time, UND/MOD Eff, significant at level 0.05 are shown in bold font

H1: τ Kendall Analysis	UAE			ULME			UAChE		
Variable	Coef.	p-value	size	Coef.	p-value	size	Coef.	p-value	size
UND SubComp - UND Time C_1	**.243**	**.015**	**60**	.136	.406	26	**.439**	**.000**	**39**
UND SubComp - UND Time C_2	**.269**	**.007**	**60**	**.401**	**.010**	**26**	**.430**	**.001**	**39**
UND SubComp - UND Time C_3	**.376**	**.000**	**60**	.060	.702	26	**.471**	**.000**	**39**
MOD SubComp - MOD Time C_1	**.366**	**.000**	**60**	**.438**	**.005**	**26**	**.289**	**.018**	**39**
MOD SubComp - MOD Time C_2	**.277**	**.005**	**60**	.251	.103	26	**.296**	**.018**	**38**
MOD SubComp - MOD Time C_3	.172	.086	58	.105	.503	26	.281	.025	37
UND SubComp - UND Eff C_1	**-.317**	**.001**	**60**	-.111	.497	26	**-.458**	**.000**	**39**
UND SubComp - UND Eff C_2	**-.300**	**.002**	**60**	**-.401**	**.010**	**26**	**-.519**	**.000**	**39**
UND SubComp - UND Eff C_3	**-.411**	**.000**	**60**	-.165	.293	26	**-.452**	**.000**	**39**
MOD SubComp - MOD Eff C_1	**-.423**	**.000**	**60**	**.436**	**.007**	**26**	**-.544**	**.000**	**39**
MOD SubComp - MOD Eff C_2	**-.439**	**.000**	**60**	**-.544**	**.001**	**26**	**-.428**	**.002**	**38**
MOD SubComp - MOD Eff C_3	**-413**	**.000**	**58**	**-.355**	**.030**	**26**	**-5.12**	**.000**	**37**

6. Validity Evaluation

A fundamental question concerning any experimental results is how valid they are. We had considered a number of validity issues inherent to this family of experiments. However for the sake of brevity we only describe the more important threats.

- **Threats to the External Validity:** Threats to this validity concern the ability to generalize experimental results outside the experiment setting. The more important threat affecting this validity is the experimental subjects. We are aware that experiments with practitioners and professionals should be carried out in order to be able to generalize the results. However, in this case, the tasks to be performed do not require high levels of experience, so, experiments with students could be appropriate [3], [18]. Moreover, it is difficult to obtain professionals having industrial experience in OCL. Probably as Briand et al. argue in [6] the chosen students (as well as those of their experiments) are better trained in modeling with UML and OCL than most software professionals.
  Nevertheless, we believe that subjects of the ULM experiment were not homogeneous. In fact, they were students coming from different universities, professionals and teachers also participated in the course. The subjects' heterogeneity could explain why the results obtained in ULME were quite different from the other two experiments in most of the hypothesis.
  We have carefully considered other factors such as the knowledge of the universe of discourse among the material used, learning effects as well as subject motivation and other factors (plagiarism, fatigue effects).
- **Threats to the Conclusion Validity:** In the conclusion validity we want to make sure that our conclusions are statistically valid. Two threats can be described. Firstly, it was not possible for us to plan the selection of a population sample by using any of the common sampling techniques, so we decided to take the whole population of the available classes in software engineering courses of universities that collaborate with our research. Secondly, the quantity and quality of the data collected and the data analysis were enough to support our conclusion, mainly as described in previous sections, concerning the existence of a statistical relationship between independent and dependent variables.
- **Threats to Construct Validity:** This validity is concerned with the relationship between theory and observation. It defines the extent to which the variables successfully measure the theoretical constructs in the hypothesis. We proposed an objective measure for the variables used in the hypothesis: (1) for the dependent variable we use a measure of how precise the subjects answering tasks per time are (the UND and MOD efficiency) as well as the time the subject spent on different tasks (the UND and MOD time). (2) for those hypotheses related to cognitive aspects of the subjects we have used a qualitative and objective measure of the subject's subjective opinion, and we use linguistic labels, providing a scale to rate tasks. (3) for the independent variables, their validity is guaranteed by Briand et al.'s framework which was used to validate them (see [26]) .

- **Threats to Internal Validity:** The internal validity is the degree of confidence in a cause-effect relationship between factors of interest and the observed results. We had alleviated some issues: knowledge of the universe of discourse among the material used, accuracy of response, learning effects as well as subject motivation and other factors (plagiarism, fatigue effects).

7. Conclusions

The lack of metrics which capture quality aspects of UML/OCL models motivated us to define a set of metrics for measuring the structural properties of OCL expressions, considered cognitive aspects of modelers in the process of metric definition [26].

We launched a family of experiments in order to analyze the coupling aspects of OCL expressions, being coupling the more complex attribute of OO system. The experiment goal was to ascertain whether any relationship exists between the object coupling (defined in OCL expressions through navigations and collection operations), and the understandability and modifiability of OCL expressions. The experiment was run at the University of Alicante (UA) with undergraduate students, and was replicated twice at the University of La Matanza (ULM) and Austral University of Chile (UACh).

In order to study the understandability and modifiability of the OCL expressions we have considered not only the time subjects spent on tasks related to this activities, but also their efficiency and their subjective perception of their activities. We think that quantitative (understandability and modifiability efficiency) and qualitative (subject's rating of their cognitive load) information is important to obtain an empirical validation. Through a thorough analysis of the collected data of the experiment and its two replicas we can summarize the obtained results as follows:

- There seems to be a statistically significant correlation between many metrics, especially those related to tracing, and the Understandability Efficiency and Modifiability Efficiency. Moreover, coupling affects in different way on the understandability and modifiability of OCL expressions. Regarding the UND or MOD Eff: collection operations, their iterators and the number of classes seem to affect the UND Eff meanwhile the length of navigations and number of relationships influences MOD Eff. The MOD SubComp (the cognitive load when subjects rate MOD Tasks) seems to be affected by the length of navigations, the number of relationships and how the navigations are combined in collection operations.
- In the UA and UACh experiments the subjects' subjective ratings (understandability or modifiability rating) are influenced by the time they used to understand or modify the OCL expressions, i.e. both times seems to affect their appreciation of the level of complexity of an OCL expression. In these two experiments the UND or MOD Eff are also correlated with UND and MOD SubComp, in stronger way. The reason the same results are not obtained in ULME could be the subjects' heterogeneity, they were students of different universities.

As the results reveals there is empirical evidence that object coupling defined in an OCL expression through navigations and collection operations is significant correlated with the maintainability of OCL expression. However, we think that the findings are preliminaries and we need to strengthen the conclusion and external validity. So, we have planned to replicate this experiment with students at the Technical University of Madrid where we will change the kind of modifiability tasks in order to improve the MOD efficiency, and we expect to reconfirm the results obtained in this family of experiments.

We are also aware that another next step is to obtain a multivariate regression analysis in order to continue interpreting the collected data obtained. Furthermore, the empirical validation of the rest of the metrics is also pending. We will work in a generalization of the benefits of the set of metrics defined for OCL expressions, trying to obtain a global complexity of UML/OCL models (note that all the proposed metrics are defined in terms of a single OCL expression).

8. Acknowledgments

This research is part of the MESSENGER project (PCC-03-003-1) financed by "Consejería de Ciencia y Tecnología de la Junta de Comunidades de Castilla-La Mancha (Spain)", the CALIPO project supported by "Dirección General de Investigación del Ministerio de Ciencia y Tecnología (Spain)" (TIC2003-07804-C05-03), the network VII-J-RITOS2 financed by CYTED. Luis Reynoso has a postgraduate grant from the agreement between the Government of Neuquén Province (Argentina) and YPF-Repsol.

9. References

[1] Atkinson, C., Kühne, T.: Model-Driven Development: A Metamodeling Foundation, IEEE Software, 20(5), 2003, pp. 36- 41.

[2] Basili, V. R., Rombach, H. D.: The TAME project: towards improvement-oriented software environments. IEEE Transactions on Software Engineering. 14 (6), 1998, pp. 758-773.

[3] Basili, V. R., Shull, F., Lanubile, F.: Building knowledge through families of experiments. IEEE Transactions on Software Engineering, 25 (4), July 1999, pp. 456-473.

[4] Boehm-Davis, D.A., Fox, J. E., Philips, B.: Techniques for Exploring Program Comprehension. Empirical Studies of Programmers, Sixth Workshop. Eds. W. Gray and D. Boehm-Davis. Norwood, NJ: Ablex, 1996, pp. 3-37.

[5] Briand, L. C., Bunse, L. C., Daly, J. W.:. A Controlled Experiment for evaluating Quality Guidelines on the Maintainability of Object-Oriented Designs. IEEE Transactions on Software Engineering, 27 (6), June 2001, pp. 513-530.

[6] Briand, L. C. ,Labiche, Y.,Yan, H. D., Di Penta, M.: A controlled Experiment on the Impact of the Object Constraint Language in UML-based Maintenance. IEEE Int. Conference on Software Maintenance, 2004.

[7] Calero, C., Piattini, M., Genero, M.: Method for Obtaining Correct Metrics. Proc. of the 3rd International Conference on Enterprise and Information Systems (ICEIS'2001), 2001, pp. 779-784.

[8] Cant, S.N., Henderson-Sellers, B., Jeffery, D.R.: Application of Cognitive Complexity Metrics to Object-Oriented Programs. Journal of Object-Oriented Programming, 7 (4), 1994, pp. 52-63.

[9] Cant, S. N., Jeffery, D. R., Henderson-Seller, B.: A Conceptual Model of Cognitive Complexity of Elements of the Programming Process. Information and Software Technology, 37 (7), 1995, pp. 351-362.

[10] Cantone, G., Donzelli, P.: Production and maintenance of software measurement models. Journal of Software Engineering and Knowledge Engineering, Vol. 5, 2000, pp. 605-626.

[11] Ciolkowski, M., Shull, F. and Biffl, S. A Family of Experiments to Investigate the Influence of Context on The Effect of Inspection Techniques. Proceedings of the 6th Int. Conference on Empirical Assessment in Software Engineering (EASE 2002), Keeke (UK), pp. 48-60 (2002).

[12] Cook, S., Kleepe, A., Mitchell, R., Rumpe, B., Warmer, J., Wills, A.: The Amsterdam Manifiesto on OCL. Advances in Object Modelling with the OCL, Springer, Berlin, LNCS 2263, 2001, pp. 115-149.

[13] El-Eman, K.: Object-Oriented Metrics: A Review of Theory and Practice. National Research Council Canada. Institute for Information Technology. March 2001.

[14] Fenton, N. E., Pfleeger, S.L: Software Metrics: A Rigorous and Practical Approach. Chapman & Hall, London, 2nd Edition. International Thomson Publishing Inc. 1997.

[15] Genero, M.: Defining and Validating Metrics for Conceptual Models, PhD Thesis, University of Castilla-La Mancha. 2002.

[16] Giese, M., Heldal, R.: From Informal to Formal Specification in UML. UML 2004, LNCS 3273, pp. 197-211, 2004.

[17] ISO/IEC 9126. Software Product Evaluation-Quality Characteristics and Guidelines for their Use. Geneva.

[18] Kitchenham, B., Pflegger, S., Fenton, N.: Towards a Framework for Software Measurement Validation. IEEE Transactions of Software Engineering, 21 (12), 1995, pp. 929-944.

[19] Klemola, T.: A Cognitive Model for Complexity Metrics. 4th International ECOOP Workshop on Quantitative Approaches in Object-Oriented Software Engineering. Sophia Antipolis and Cannes, France, 2000.

[20] Marchesi, M.: OOA Metrics for the Unified Modeling Language. In 2nd Euromicro Conference on Software Maintenance and Reengineering, pp. 67-73, 1998.

[21] Miller, J.: "Applying Meta-Analytical Procedures to Software Engineering Experiments", Journal of Systems and Software, 54, 2000, pp. 29-39.

[22] Object Management Group. UML 2.0 OCL 2nd revised submission. OMG Document ad/2003-01-07. [On-line] Available: http://www.omg.org/cgi-bin/doc?ad/2003-01-07.

[23] Object Management Group. UML Specification Version 1.5, OMG Document formal/03-03-01. [On-line] Available: http://www.omg.org/cgi-bin/doc?formal/03-03-01.

[24] Object Management Group. MDA-The OMG Model Driven Architecture. Available: http://www.omg.org./mda/, August 1st, 2002.

[25] Reynoso, L., Genero, M., Piattini, M.: A Controlled Experiment for Validating Metrics for OCL Expressions. ACM-IEEE International Symposium on Empirical Software Engineering. ISESE 2004. 19-20 August 2004 Redondo Beach CA, USA., 2004.

[26]. Reynoso, L., Genero, M., Piattini, M.: Measuring OCL Expressions: An approach based on Cognitive Techniques. Piattini M., Genero M. and Calero C. Editors, Imperial College Press, UK. 2005.

[27] Schneidewind, N. F.: Methodology For Validating Software Metrics. IEEE Transactions of Software Engineering, 18 (5), May 1992, pp. 410-422.

[28]. Selic, B.: The Pragmatics of Model-Driven Development. IEEE Software. 20 (5), pp 19–25. 2003.

[29] SPSS, 2002 SPSS 11.5. "Syntax Reference Guide". Chicago. SPSS Inc. 2002.

[30] Van Solingen, R., Berghout, E.: The Goal/Question/Metric Method: A practical guide for quality improvement of software development. McGraw-Hill, 1999.

[31] Vinter, R., Loomes, M., Kornbrot R.: Applying Software Metrics to Formal Specifications: A Cognitive Approach. 5th. International Symposium on Software Metrics. March 20 - 21, 1998, pp 216-223. Bethesda, Maryland

[32] Warmer, J., Kleppe, A.: The Object Constraint Language. Second Edition. Getting Your Models Ready for MDA. Object Technology Series. Addison-Wesley, Massachusetts, August 2003.

[33] Wohlin, C., Runeson, P., Höst, M., Ohlson, M., Regnell, B., Wesslén, A.: Experimentation in Software Engineering: An Introduction, Kluwer Academic Publishers, March 2000.