Information Sciences 180 (2010) 3837–3862

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

Assessing the influence of import-coupling on OCL expression maintainability: A cognitive theory-based perspective

Luis Reynoso [a,1], Esperanza Manso [b,2], Marcela Genero [c,*], Mario Piattini [c,3]

[a] Facultad de Informática, University of Comahue, Buenos Aires 1400, 8300 Neuquén, Argentina
[b] GIRO Research Group, Department of Computer Science, University of Valladolid, Campus Miguel Delibes, E.T.I.C. 47011, Valladolid, Spain
[c] ALARCOS Research Group, Department of Information Systems and Technologies, University of Castilla-La-Mancha, Paseo de la Universidad Nº 4, 13071 Ciudad Real, Spain

ARTICLE INFO

Article history:
Received 5 February 2009
Received in revised form 3 June 2010
Accepted 21 June 2010

Keywords:
Software measures
UML/OCL models
OCL expressions
Import-coupling
Comprehensibility
Modifiability
Maintainability
Mental models
Cognitive models
Experiments
Meta-analysis

ABSTRACT

The aim of this paper is to present the definition of the measures for the import-coupling of OCL expressions, along with the empirical validation of these as early indicators of the maintainability of OCL expressions. This empirical validation has been carried out by means of an experiment and its replica, conducted with undergraduate students of Spanish and Argentinean universities, respectively. To perform this experimental activity, we have followed a cognitive theory-based approach, since the understanding of the cognitive demands that OCL expressions place on software engineers will allow us to examine in greater depth the real influence of import-coupling on the maintainability of OCL expressions. The empirical results, obtained through the analysis of the data from the experiment and its replica, first of all separately and then together after a meta-analysis study, reveal evidence suggesting that import-coupling exerts a certain amount of influence on the comprehensibility and modifiability of OCL expressions. The measures that have most influence on OCL expression comprehensibility are those related to problem objects (Number of Navigated Relationships (NNR), Weighted Number of Navigations (WNN), Depth of Navigations (DN) and Number of Attributes referred through Navigations (NAN)), relationships between problem objects (Number of Navigated Classes (NNC) and Number of Explicit Iterator variables (NEI)), as well as reified objects (Weighted Number of Collections Operations (WNCO)). On the other hand, it is only measures related to relationships between problem objects that are the main influence on OCL expression modifiability. The influence of import-coupling on the comprehensibility and modifiability of OCL expressions was reaffirmed through the cognitive complexity (i.e. we show that import-coupling affects the cognitive complexity and that the latter influences the comprehension and modification of OCL expressions). These results may have educational implications, apart from what they might mean for practitioners in the industry, as is explained in the conclusions.

* Corresponding author. Tel.: +34 926 295300x3740; fax: +34 926 295354.
E-mail addresses: lreynoso@uncoma.edu.ar (L. Reynoso), manso@infor.uva.es (E. Manso), Marcela.Genero@uclm.es (M. Genero), Mario.Piattini@uclm.es (M. Piattini).
[1] Tel.: +54 299 4490314; fax: +54 299 4490313.
[2] Tel.: +34 983423670; fax: +34 983423671.
[3] Tel.: +34 926 295300x3740; fax: +34 926 295354.

0020-0255/$ - see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.ins.2010.06.028

1. Introduction

The growing attention given to using models in software development, such as in Model-Driven Development (MDD) [1], has brought model quality into the forefront as an area of research [63,52,57]. In particular, there is special interest in the quality of models specified with the Unified Modeling Language (UML) ,[67] since it is the current modeling standard for system development and it is being adopted increasingly throughout the world [41].

Although UML models provide a good view of software architecture, they are underspecified [36], due to the fact that not all the constraints and essential aspects of a system can be represented by using diagram-based UML notation [43]. UML models should be supplemented with the use of a textual add-on, the "Object Constraint Language" (OCL) [68], which provides the expressiveness and precision that diagram-based UML notation lacks [22,15,16]. Several authors recommended using OCL for expressing constraints in UML models, when designing object-oriented systems [27,90,87]. OCL is also recommended for specifying business rules in data base applications [28], and it has been extended to express security constraints in database models [32] and in datawarehouse models [70]. OCL was also used in a precise definition of software measures for statechart diagrams [76] and for business process models [81].

Most important software providers are recognizing the importance of OCL for developers and testers, promoting their use and incorporating OCL into their own tools [59,72].

All this serves to demonstrate that OCL is becoming an increasingly significant topic. In addition, it has been recognized that OCL could help to enhance the quality of the software produced [36,89,90,7,54], though measures which objectively quantify the quality of OCL expressions do not exist.

The gap in the field as just outlined was what triggered off the research we have been carrying out over the last 5 years. Its focus has been the definition and validation of measures with which to study the influence of import-coupling on two maintainability sub-characteristics of OCL expressions, namely comprehensibility and modifiability.

Our hypothesis is that the import-coupling of an OCL expression within a UML/OCL model may influence the cognitive complexity (i.e. the mental burden of individuals: modelers, developers, testers, etc.), and that high cognitive complexity leads to the situation where OCL expressions exhibit undesirable external qualities [48], such as lower levels of comprehensibility and modifiability. This hypothesis is based on the relationships presented in Fig. 1, which constitutes the theoretical basis for developing quantitative models [8–10].

Why are we interested in the influence of import-coupling on comprehensibility and modifiability? Our reasons are as follows: The extent to which an OCL expression depends on the rest of the UML model, i.e. the import-coupling [6] of an OCL expression, could influence its comprehensibility: The larger the UML context imported to the scenario of an OCL expression, the lower the comprehensibility of that OCL expression. Moreover, when a high number of references are used, implicit assumptions may become invalid over time, thus creating a situation where an OCL expression is more likely to have to be modified. The availability of import-coupling information of a model at early stages would be useful in deciding, for example, which classes should undergo more intensive verification or validation; design decisions can also be justified better.

- We focus on the comprehensibility of OCL expressions within a UML model, since these expressions should be comprehensible and flexible enough for any modification of their meaning to be easily incorporated into the model. Moreover, as comprehension is responsible for up to half the total cost of software maintenance [64], OCL expressions should be carefully considered, since they can be difficult to read and write [20,37,42] and OCL navigation can become complex or may be rather verbose to write [20].

We have specifically focused on OCL expressions specified on UML class diagrams, since the class diagram is one of the most commonly used diagrams in software modeling and is perceived by practitioners to be the most important diagram type [31,41].

Although the mechanism which causes the effect of measures on external quality attributes is assumed to be cognitive complexity [19] (i.e. the mental burden of individuals: modelers, developers, testers, etc.), the definition of measures and the findings of empirical studies are rarely explained using cognitive theories. Darcy and Slaughter [26] argue that the consideration of a theoretical perspective, such as human cognition, provides a solid foundation upon which to derive an integrative model relating internal and external attributes of software quality, and this is fundamental to the success of software measurement [33].

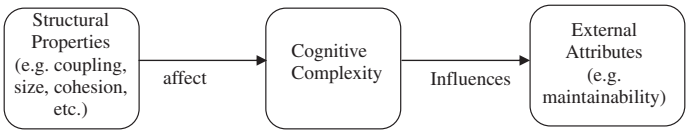


Fig. 1. Relationship between structural properties, cognitive complexity, and external quality attributes [8–10].

That being so, we believe that the introduction of cognitive theories will contribute towards the achievement of two goals:

- To provide a better and more complete explanation of the defined measures. Many traditional measures are supported by the fact that they are clearly related to cognitive limitations [5]. The understanding of the cognitive demands that OCL expressions place on software engineers will allow us to examine in greater detail the real influence of import-coupling (a structural property) on the maintainability of OCL expressions. Cognitive models have been used to attain this goal.
- To provide a better and more precise interpretation of the findings obtained through empirical studies from a cognitive point of view. The main categories of the mental models of modelers dealing with OCL expression comprehension were taken into consideration to assist in the attainment of this goal.

All that has been said so far was what led us to propose the main goals of this paper, which are the following :

- To present the cognitive theories used to explain how modelers deal with OCL expressions, providing a more solid interpretation of the empirical findings.
- To present a family of experiments, consisting of a controlled experiment and its replica, which were carried out to test our previously-defined hypothesis (Fig. 1); i.e. we want to assess the impact of import-coupling on the comprehensibility and modifiability of OCL expressions.

This paper is organized as follows: Section 2 introduces the cognitive theories used in this paper. Section 3 presents a set of measures for import-coupling, providing an explanation of these through a cognitive model. Section 4 presents a controlled experiment and its replica, which were carried out to evaluate the capability of the import-coupling measures defined as indicators of the comprehensibility and modifiability of OCL expressions. Section 5 describes the data analysis of each individual experiment. The integration of results through meta-analysis is presented in Section 6. Section 7 presents related work, and finally, Section 8 summarizes the main contributions of the paper and outlines future work.

2. Cognitive theories

In this section we briefly describe the cognitive theories used to define measures for OCL expressions, as well as to interpret the results obtained in the experimentation which validate such measures.

We believe that the selection of a cognitive theory should be dependent on the artifact measured. In our case, the artifacts are expressions in a declarative language [67]. Two cognitive theories of program comprehension were used: a cognitive model [18] and a mental model [12]. A cognitive model describes the cognitive techniques and temporary information structures in the subject's head which are used to form the mental model [86], whereas a mental model describes a subject's mental representation of a software artifact that has to be understood.

We are conscious that the application of the aforementioned theories are not straightforward, since they were defined for program comprehension, and that OCL expressions are declarative constraints which do not become embroiled in implementation details as a program does [43]. However, we consider that both models (cognitive and mental) are generally sufficient for application in the comprehension of a declarative language such as OCL. In fact, the essence of program comprehension is to identify artifacts, discover relationships, and generate abstractions [65]. We believe that, regardless of the language (imperative or declarative), the comprehension of its specifications (or products) involves different cognitive aspects.

We have therefore based the current work on the two cognitive theories mentioned previously, which will be described in more detail in the following two sections.

2.1. A cognitive model

This section briefly describes the Cognitive Complexity Model (CCM model) defined by Cant et al. [18]. The CCM gives a general cognitive theory of software complexity which elaborates on the impact of structure on comprehension. The CCM model defines two cognitive techniques that we believe are also applied by UML modelers when they comprehend OCL expressions. The two cognitive techniques (chunking and tracing) are concurrently and synergistically applied in problem solving:

- Chunking involves the recognition of a set of declarations and the extraction of information from these; this is remembered as a chunk (a single mental abstraction), whereas
- Tracing involves scanning in different directions, in order to identify pertinent chunks.

Tracing usually disrupts the process of chunking [46]. Typical examples of the application of tracing are when an OCL expression attached to a contextual type uses rolenames to refer to other classes or when an inherited property needs to be understood.

We believe that OCL expressions are key facilitators in the construction of chunks as well as compound chunks. Moreover, the constraint declared by an OCL expression is often captured by the mnemonics of the OCL expression's name, and this can help to associate high level concepts with program concepts. In our cognitive model, we argue that when a modeler is primarily chunking an OCL expression within a UML/OCL model, there are dependencies that, to be resolved, require the modeler to perform a certain amount of tracing (in different directions) to find relevant features such as rolenames, attributes, classes, etc. Having found their features, modelers will once again chunk to comprehend it. Conversely, when modelers are primarily tracing, they will need to chunk to understand the effect of the chunks identified. Moreover, as El-Eman recognizes in [30], a certain amount of coupled chunks do not affect cognitive burden, until a limit is exceeded and short-term memory overflows. On many occasions the modelers have to switch constantly between textual specification (OCL expressions) and graphical specification (UML diagram). The effects of the difficulty of chunking and tracing on complexity can be modeled graphically using a landscape model. While reading an upper-level chunk, a dependency requires the modeler to suspend reading the original OCL expression because of the need to undertake tracing if the chunk being analyzed at that moment is to be fully understood. This is exemplified in the following. The upper part of Fig. 2 shows a UML/OCL combined model where an OCL expression had been defined for the flight class, meaning that the quantity of passengers of a flight must be lower or equal to the capacity of the type of plane for that flight. The landscape model for the OCL expression, named "flight capacity", is shown in the bottom part of Fig. 2. At the top-level of the diagram there is a single chunk visible, the OCL expression, delineated by the two markers (*f*,*g*). This chunk is interrupted by three lower-level chunks. The first interruption is common to every OCL expression where the modeler traces the context of the expression (the Classifier written after the context keyword) to locate within the UML diagram. The second interruption, depicted as the "vertical drop" x1P represents in visual form the work required in tracing the relevant features in the UML diagram. In this case, that implies following a navigation from the Flight class to another class where its opposite-end rolename is defined as "plane". Having found this class, the modeler must chunk it, as well as chunking the cardinality associated with the rolename mentioned. Then the modeler should follow a new navigation from Flight to Type_of_plane using the "planetype" rolename, and after chunking the meaning of the latter class, the modeler should chunk one of its attributes, namely "capacity". The third and last interruption during the comprehension of the flight_capacity OCL expression is during the navigation (drop x4P) to the Passenger class, to obtain the size of the set of passengers. As described above, the modeler should switch between textual and graphical information (between ASCII declarations – the OCL expressions – and UML diagrammatic notations) in order to fully capture the meaning of an OCL expression. For instance, while reading a navigation in an OCL expression, the modeler should follow the rolenames used in the UML diagram. In their cognitive model, Cant et al [18] consider the spatial distance of tracing as a factor of difficulty in this cognitive technique. This factor is seen as the distance, measured in program comprehension and possibly through lines of code, between two chunks for which there is a dependency.

The spatial distance of tracing an OCL expression is captured by the distance of the contextual instance from the most distant coupled object through the DN measure (Depth of Navigations). We should also take into account that, in general, OCL expressions are not graphically attached to UML models. They are, rather, included in the underlying model repository [90], so the automated tool used to storage the model repository should facilitate the visualization of both the OCL expression and UML diagram at the same time, otherwise the spatial distance could negatively influence the cognitive load.

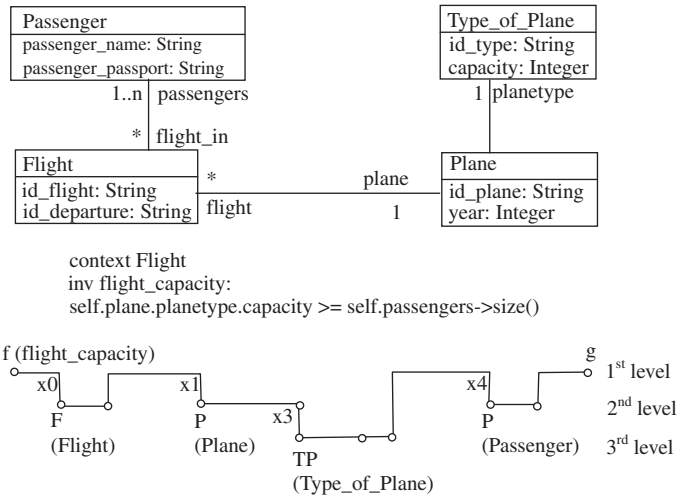


Fig. 2. An OCL expression attached to an UML class.

2.2. A mental model

This section introduces the concept of a mental model. People create mental representations to reason about and explain the objects and information in the world, behavior of systems, etc. A predictive representation of real world systems is called a mental model [66]. Mental models play an important role in software comprehension and correspondingly in comprehension-related tasks, such as modification and debugging [74,53]. Mental models are part of the dimension of Direction of Comprehension that refers to approaches regarding how programmers comprehend software. Three well-known strategic approaches are the bottom-up, top-down and opportunistic cognitive models. The main categories of mental models employed by modelers in comprehending OCL expressions are:

- *Problem objects:* The objects (main concepts) of the problem domain to which the OCL expressions are attached.
- *Relationships between problem objects:* Association, composition and inheritance relationships between objects.
- *Reified objects:* These are not problem domain objects per se, but are represented to complete the representation of relationships between problem objects, e.g. OCL collections.

These categories are based on a work of Burkhardt et al. [12].

Using the previous example of Fig. 2, the chunking of the OCL expression involves the chunking of four problem objects: the flight, the plane, the planetype and the passenger; the comprehension of three relationships between objects: passengers of a flight, the plane used in the flight and the type of plane associated with the plane. An example of a reified object is the set of passengers (a collection is obtained through navigation) which is represented by the subexpression self.passengers.

3. A proposal of measures for OCL expressions

This section presents a summary of the measures we have defined for measuring import-coupling in OCL expressions. These measures were defined following a rigorous and methodological process, proposed in [17], originally ,and refined and extended in [75]. The informal definition using natural language and theoretical validation of these measures was presented in [35]. Their definition was later formalized by using OCL, in order to avoid ambiguities and misconceptions [78].

In addition, we will explain the definition of measures using the CCM model proposed in [18] which, as was commented in Section 2, is based on two main cognitive techniques: chunking and tracing. These techniques provide the basis for giving a cognitive explanation of the definition of import-coupling measures for OCL expressions. We believe that it is relevant to have a cognitive model of the concepts (or chunks) that may be imported or captured within an OCL expression, and of the OCL mechanisms that allow modelers to import them.

As Cant et al. commented, it is difficult to determine what constitutes a "chunk", since it is a product of semantic knowledge [18]. Examples of chunks are classes, methods, attributes, etc. Within a UML/OCL model, an OCL expression constitutes a suitable "chunk" unit which modelers should comprehend as a whole declaration that constrains an aspect of the system being modeled. Other kinds of chunks are the attributes and operations (object properties) which are mentioned within an OCL expression. To understand an OCL expression as a chunk, UML modelers must carry out tracing (of the UML diagram to which the expression is attached) to find pertinent object properties imported to the OCL expression scenario and then chunk them. The understanding of an OCL expression as a chunk therefore involves a strong intertwining of tracing and chunking techniques. The tracing activity within the comprehension of source code is rather different from the cognitive techniques of tracing within an OCL expression. In the former, the scanning of the relevant chunks is not guided, whereas in the latter the tracing is mostly traced by OCL navigations.

Our first attempt to define the measures implied analyzing which OCL concepts, specified in its metamodel [67], were relevant to both of the cognitive techniques of chunking and tracing. These cognitive techniques are exemplified through the following examples:

- The contextual instance, i.e. self, provides a point of reference for the interpretation of an OCL expression. Whenever we use the self keyword, we are chunking the main object to which the expression refers.
- OCL expressions can be read and evaluated from left to right, combining different properties, similarly to the composition of functions. However, each time a navigation is used, the modeler should stop comprehending the OCL expression and should trace the UML diagram through the rolenames (properties) mentioned in the expression. Although the meaning of the navigation is subsumed in the meaning of the expression, tracing the UML diagram disrupts the left to right understanding of the expression. The evaluation of the navigation will involve the understanding of the rolenames (or class names if rolenames are omitted in the diagram) mentioned in the navigation, in order to trace these in the class diagram, as well as to look for the classes and attributes (or operations) mentioned through the navigation, and eventually to understand the OCL expressions associated with them if they have OCL expressions attached.

For the sake of brevity, we shall show only the complete definition of two measures: the Number of Explicit Self (NES) and the Number of Navigated Relationships (NNR).

NES is defined as follows:

- *Definition:* This measure counts the number of times "self" is used in an explicit form in an OCL expression.
- *Example:* The upper part of Fig. 2 shows part of a UML class diagram in which an OCL expression, denominated as "flight_capacity", has been defined in the context of the Flight class (the contextual type), meaning that the number of passengers on a flight must be lower than, or equal to, the capacity of the type of plane for that flight. The value of NES for the expression of Fig. 2 is 2, since "self" was used twice.
- *Goal:* By using the contextual instance, it is possible to access different object properties (attributes, operations, association-end) of the contextual type. So each of the property accessed through "self" involves tracing from the OCL expression to the contextual type in the UML model and from that UML contextual type in the model to another UML artifact (classes if the property used is an association-end or attribute or method of the contextual type). A high number of references of the contextual instance will increase the knowledge of the contextual type and will consequently also increase the extent of the OCL expression. The greater the number of times "self" is used may indicate that the context is that much harder to understand. For that reason, we believe that the number of references to the contextual instance should be kept as low as possible.
- *Cognitive explanation:* The first step in constructing a conceptual model is to identify a set of fundamental concepts with which to describe the domain; these concepts appear in the model as classes or types. Cant et al. [18] argue that classes are typical examples of chunks during software comprehension. Self represents an object of the contextual type (usually a class diagram). Self is the target object that should be chunked by the modelers in order to understand or modify an OCL expression. The greater the amount of times that the contextual instance appears within an OCL expression, the greater the context of the contextual type to be comprehended. There are many research works dealing with the upper limit of the human capacity to process, with reliable accuracy, information on concepts which are interacting simultaneously . For example, Miller [60] argues that people are able to remember about 7 chunks in short-term memory (STM) tasks, while other authors such as Cowan [24] or Broadbent [11] propose an average capacity limit of about 4 chunks. Although cognitive thresholds for the processing of information are not within the scope of this paper, we believe that the particular limit for modelers when comprehending an OCL expression depends on the different chunks involved, and also on the familiarity of the information encountered, i.e. experience of the domain comes into play [52]. Nevertheless, the value of NES for an OCL expression should be low, so as not to overload the modelers.

NNR is defined as follows:

- *Definition:* This measure counts the total number of relationships that are navigated in an expression. Two special cases are considered: (1) If a relationship is navigated twice, then this relationship is counted once (when a concept is repeatedly recalled, it becomes easier to recall once more and so can be easily assimilated [40]), (2). Whenever an association class is navigated, we will consider the association to which the association class is attached.
- *Example:* The value of NNR corresponding to the expression of Fig. 2 is 3, because three relationships were navigated: the relationship between Flight and Plane (through the plane rolename), the relationship between Plane and Planetype (using planetype) and the relationship between Flight and Passenger (using passenger).
- *Goal:* As Warmer and Kleppe [90] remark, one argument against complex navigation expressions is that the writing, reading, and understanding of OCL expressions become very difficult. The meaning of each relationship involves the understanding of how the objects are coupled to each other, the arity of the relationships, etc. The larger the set of relationships to be navigated, the greater the context to be comprehended.
- *Cognitive explanation:* From a cognitive point of view, NNR implicitly quantifies the effort of tracing by modelers when it is necessary for them to understand the association and composition relationships during navigations within an OCL expression.

As mentioned previously, the whole set of measures we have defined can be found in [35]. Due to the fact that OCL expressions are short assertions using different OCL mechanisms, it is not common to find all the import-coupling concepts at the same time in an OCL expression, as this would be cumbersome. That being the case, in this paper we have decided to consider the most relevant and core concepts from OCL literature, i.e. those which are most frequently used in common OCL expressions (Table 1). Table 1 shows each measure's acronym and name (first and second columns respectively), the cognitive techniques to which they are related (third column) and the structural property captured by the measure (fourth column, obtained from its theoretical validation, carried out following Briand et al.'s framework [6]).

NNR, NAN, NNC, WNN and DN are measures related to tracing, because they measure an aspect to do with navigations. Nevertheless, tracing and chunking are applied concurrently. So, although NNC and NAN are measures which involve tracing, they are also considered to be chunking measures. This is because, once the modelers resolve the tracing activity, they should chunk the meaning of a class or attribute, respectively. NEI measures the chunking of objects related to collections operations. NKW and NCO are seen as chunking measures, since the expression itself needs to be chunked, i.e. a constraint associated to the contextual instance.

Although the main focus of this article is to validate import-coupling measures empirically, we must take into account size measures, due to the fact that, in any OCL expression, both structural properties coexist. We have followed a recommendation provided in [30], which suggests that controlling the size would not bias findings in coupling studies during experimentation. Table 1 therefore also shows the size measures which we considered in this paper.

Table 1
Measures for OCL expressions defined within UML/OCL models.

Measure	Measure description	Cognitive technique	Theoretical validation
NNR	Number of Navigated Relationships	Tracing	IB-coupling[a]
NAN	Number of Attributes referred through Navigations	Tracing and chunking	IB-coupling
NNC	Number of Navigated Classes	Tracing and chunking	IB-coupling
WNN	Weighted Number of Navigations	Tracing	IB-coupling
DN	Depth of Navigations	Tracing	IB-coupling
WNCO	Weighted Number of Collection Operations	Tracing	IB-coupling
NEI	Number of Explicit Iterator variables	Chunking	IB-coupling
NES	Number of Explicit *Self*	Chunking	Size
NKW	Number of OCL Key Words	Chunking	Size
NCO	Number of Comparison Operators	Chunking	Size

[a] Interaction based coupling.

4. A family of experiments

The empirical validation of measures through controlled experiments is fundamental if we are to ensure that the measures are actually significant and useful in practice [51].

An experiment may be a part of a common family of studies, rather than being an isolated event [2]. Common families of experiments allow researchers to answer questions that are beyond the scope of individual experiments and let them generalize findings across studies. Evidence is thus provided to confirm or reject specific hypotheses. In addition, common families of studies can contribute to devising important and relevant hypotheses that may not be suggested by individual experiments. We have therefore run a family of experiments, to ascertain whether the measures defined for import-coupling can be used as early indicators of comprehensibility and modifiability in OCL expressions.

The experimental process [91] followed to carry out this family of experiments, consisting of one experiment and its replica, will be described thoroughly in the remainder of this section. The experiment and its replica are described concurrently, since their experimental phases are identical. Details are given of the differences, which basically refer to the selected subjects and data analysis.

We have carried out a strict identical replica [2], given that the replicator is the same person (the first author of the paper) as the conductor of the original experiment, and that the experiment, as well as its replica, share the goal of testing the same hypotheses. The subjects were different in each experiment.

4.1. Family goals and context definition

Using GQM template [3] for goal definition, the goal of the family is to:

Analyze import-coupling measures of OCL expressions attached to UML class diagrams
With regard to their capability of being used as comprehensibility and modifiability indicators of OCL expressions
From the point of view of researchers
In the context of the undergraduate students from two universities, a Spanish university (UPV) and an Argentinean one (UNSL).

The subjects who participated in this family of experiments have the following characteristics:

- *Experiment:* Forty-six students enrolled in a fifth-year course of Software Engineering at the Technical University of Valencia (UPV) were invited to participate in a seminar of 10 h about OCL. As an inducement to enroll on the course, the students were informed that they would be given an assessment and that its result would be taken into account in their being awarded or not an extra credit as part of their course assessment process.
- *Replica:* Thirty six students who participated in a course in the Eighth International School of Computer Science (held in San Luis, Argentina) were the subjects of the replica. The duration of the course was 20 h, and the replica experiment was run during the last 2 h. The subjects were undergraduate students from various universities, along with graduate students. The data obtained in this replication was called "UNSL".

4.2. Variables selection and hypotheses formulation

The independent variable (IV) is the import-coupling of OCL expressions. The dependent variables (DVs) are the comprehensibility (COM) and modifiability (MOD) of OCL expressions, which are two sub-characteristics of maintainability [48].

The IV is measured by using some of the measures previously presented in Table 1. We used NNR, NNC, WNN, DN, WNCO, NEI and NAN measures, since they all capture an aspect of import-coupling in their intent, through navigation or collection operation concepts.

The last three measures shown in Table 1 (NKW, NES and NCO) are size measures and, as we mentioned previously, size was studied as a confounding factor for many coupling measures. Bearing this in mind, we carefully defined the experimental object in such a way that the values of these measures do not bias our findings during experimentation. We have also attempted to keep their value as constant as possible.

The values of the measures corresponding to the UML/OCL models used as experimental objects are shown in Table 2. The content of the first column is related to the complexity of the models and is explained in Section 4.3.

The DVs variables were measured by using the following measures, whose values were obtained through the experimental tasks (Section 4.3):

- *Comprehensibility Efficiency (COM Eff):* Correct answers/COM Time.
- *Modifiability Efficiency (MOD Eff):* Correct options selected/MOD Time.

Moreover, rating tasks (Section 4.3) were used to obtain two measures (of the subject perception), called COM Subjective Complexity (COM SubComp) and MOD Subjective Complexity (MOD SubComp), respectively. These measures are essential if we are to estimate the cognitive load of subjects when dealing with UML/OCL combined models.

We formulated different hypotheses, according to distinct beliefs. These are explained below, and are summarized in Fig. 3:

- *Belief 1:* The Efficiency of the subjects will be different according to the levels of import-coupling of the UML/OCL models they have to manipulate.
  Hypotheses 1: $H_{0,1}$ The means of the (COM or MOD) Eff are the same in the different levels of import-coupling, i.e. the mean efficiency (Mean Eff) is the same for all the models. $H_{1,1} = \neg H_{0,1}$.
- *Belief 2:* The greater the import-coupling, the lower the subjects' efficiency, i.e. the import-coupling in OCL expressions influences the degree of correctness of the tasks performed per time, i.e. the subject's efficiency (COM Eff or MOD Eff).
  Hypotheses 2: $H_{0,2}$ There is correlation between the measures defined for OCL expressions, related to import-coupling and their (COM or MOD) Eff. $H_{1,2} = \neg H_{0,2}$.
- *Belief 3:* We expect that if the import-coupling increases, then the subjects will perceive the tasks as difficult to comprehend (i.e. "quite difficult to understand" or "barely understandable"), or hard to modify. The import-coupling in OCL expressions makes an impact on the subjective rate of the subjects' (COM SubComp or MOD SubComp) tasks.
  Hypotheses 3: $H_{0,3}$ There is correlation between the OCL expression measures related to import-coupling and the (COM or MOD) SubComp. $H_{1,3} = \neg H_{0,3}$.

Table 2
Values of the measures used in the family of experiments.

	Models	Measures									
		NNR	WNN	DN	NAN	NNC	NEI	NES	WNCO	NKW	NCO
LC	Model1	1	1	1	1	1	1	1	2	3	1
LC	Model2	2	2	1	0	2	0	2	2	3	1
MC	Model3	3	6	4	0	2	1	2	3	3	0
MC	Model4	3	6	4	0	2	1	2	3	3	0
HC	Model5	2	3	4	2	2	2	2	7	3	1
HC	Model6	3	6	3	1	3	3	1	8	3	1

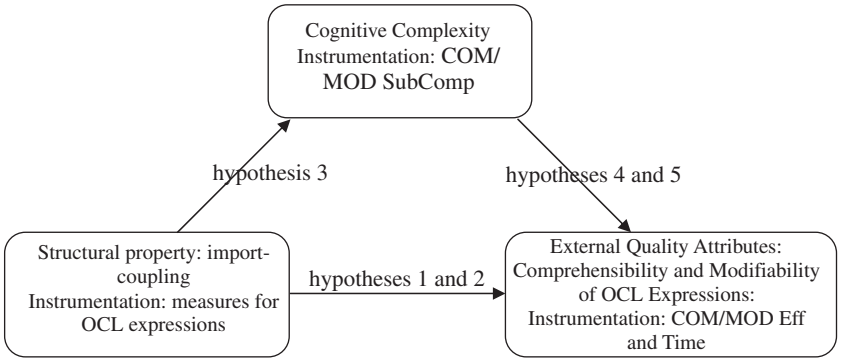


Fig. 3. Experimental hypotheses.

- *Belief 4:* The subjective criterion of subjects when they have to rate tasks is influenced by the COM (or MOD) Time. For example, we expect subjects to rate as difficult those tasks which have taken them more time to complete.
  Hypotheses 4: $H_{0,4}$ The COM or MOD SubComp are not correlated with the COM and MOD Time. $H_{1,4}$: $\neg H_{0,4}$.
- *Belief 5:* The perception of the subjects with regard to the complexity of the tasks is influenced by their efficiency when performing such tasks, i.e. we believe that the degree of correctness of the tasks performed per time (the COM Eff or MOD Eff) could be an indicator of the subjective rating given by the subjects with regard to the complexity of the required tasks.
  Hypotheses 5: $H_{0,5}$ The (COM or MOD) SubComp is not correlated with the COM and MOD Eff. $H_{1,5}$: $\neg H_{0,5}$.

4.3. Experimental design and procedure

The experimental objects were six UML/OCL combined models, each of which contained one OCL expression. They were designed by covering a wide range of the measure values (except in the case of the NES, NKW, and NCO measures). However, it is impossible to cover all the possible combinations of measure values. Fifteen models were designed initially, but we believed that some models were quite similar, and that the fact of having so many models of the same complexity could bias the experiment result. We therefore carried out a hierarchical clustering of the 15 models, to place them into three groups according to their measure values: Low, Medium or High Coupling (each level of import-coupling is identified by using the acronyms LC, MC, HC respectively) (see first column of Table 2). Finally, we obtained two models for each group, i.e. six models (see example (model 2) in Appendix A at the end of the paper). The remaining material can be found in http://alarcos.esi.uclm.es/OCLExperiments. Each model included a test, which consisted of three types of tasks:

- *Comprehension tasks (COM tasks):* The subjects had to answer a questionnaire consisting of four questions, which reflected whether or not they had understood the OCL expression attached to the class diagram. The questions concerned the meaning of the OCL expression, the types used in the expressions, as well as the navigation concepts.
- *Modifiability tasks (MOD tasks):* These tasks were different from the previous family of experiments [77], in which the subjects had to write a new OCL expression according to a new requirement. In this family, two different modifications were asked for, in the form of new requirements expressed in natural language. For each modification, the subjects had to select one of three OCL expressions (a multiple choice task). The selected expression had to represent the original OCL expressions (that associated with the model) after being modified according to the new requirement. Note that the correct OCL expression that had to be selected by the subjects had the same measure values as the original expression associated with the model, i.e. it presented the same structural properties. In the previous experiment this situation was not controlled, and the subjects wrote various correct answers, i.e. the OCL expressions were semantically equivalent but their structural properties differed from each other, and the effort of modifiability tasks could not be properly measured.
- *Rating tasks:* Card et al. [19] argue that one way in which to operationalize cognitive complexity is to equate it with the ease of comprehending an object-oriented artefact. Therefore, in order to estimate the cognitive complexity of modelers when dealing with OCL expressions, we gather the subjects' perception of the complexity of comprehension tasks and modification tasks, using linguistic labels. After finishing each task (COM or MOD tasks), the subjects use a scale of five linguistic labels to rate them (e.g. for COM tasks we used the "Easily understandable", "Quite easy to understand", "Normal", "Quite difficult to understand", "Barely Understandable" labels). This rating indicates the subjects' perception of how complex it was for them to perform the COM and MOD tasks.

The six tests were assigned to each subject, so we had a factorial within-subject design. The tests were assigned to the subjects in such a way that no two subjects did the six tests in the same order. The first three tests (and the second three tests) assigned to each subject were models of different levels of coupling, i.e. HC, MC or LC models.

In this paper, C_1 is identified as the collection of the first tests performed by all the subjects, C_2 is the second collection, and so on. It is important to note that our intention was that the same number of subjects in each C_i examined the six models. This meant that the design was balanced.

5. Data analysis and interpretation

In this section, we shall summarize the main aspects of the analysis of the empirical data, carried out with SPSS [85]. We shall carry out a descriptive and exploratory study (Section 5.1) first. We shall then test the hypotheses formulated (Section 5.2). We shall later give a cognitive explanation of the empirical findings (Section 5.3) and describe the threats to validity of the experiment (Section 5.4).

As all the hypotheses formulated are concerned with the degree of dependency between two variables, a bivariate correlation analysis can be used. In Table 3we also explain the hypotheses and the correlation tests used.

The Spearman correlation coefficient works with observation pairs (X_i, Y_j) of measures with at least, ordinal scale, over n-objects (in our case 6 models), but the observations must be independent. This means, for example, that if we study a dependent variable, such as COM Eff, of the subject k in the i-diagram, we are not allowed to consider any other observation of the same k-subject. The descriptive analysis and the correlations of the formulated hypotheses are therefore tested for each C_i (the i-tests performed by all the experimental subjects), in which i ranges from 1 to 6.

Table 3
Synopsis of hypotheses and the statistical tests applied.

Relation between	Efficiency COM Eff, MOD Eff	Time COM time MOD time	Subjective Complexity COM SubComp MOD SubComp
Models	Hypotheses 1 Test: Friedman	–	–
OCL expression measures	Hypotheses 2 Test: Spearman	–	Hypotheses 3 Test: Spearman
COM SubComp MOD SubComp	Hypotheses 5 Test: Spearman	Hypotheses 4 Test: Spearman	

To explain the meaning of the non-significant results, we have obtained the test power. That is, the probability of accepting the null hypothesis when is false [82]. The test power may be obtained for correlation tests if they are parametric, and can be obtained approximately for some non-parametric tests. In our study, we have obtained the power of a Spearman test as 91% of the Pearson test.

5.1. Descriptive and exploratory studies

In this section we explain the descriptive and exploratory aspects of both studies, hereafter referred to as "UPV" and "UNSL".

5.1.1. Descriptive and exploratory studies for UPV

The top left-hand side of Fig. 4 depicts the average of the subjects' efficiency, along with the answer for each C_i. Fig. 4 shows that the subjects are more efficient in MOD tasks (dashed line) than in COM tasks (solid line). Indeed, the subjects were more efficient than in the previous family (see related work in Section 7). As explained previously, we altered the MOD tasks so that rather than when writing a new OCL expressions according to a new requirement, the subjects had to select, from the three proposed OCL expressions, the one which modeled the new requirement correctly. However, in both kinds of tasks the subjects improved their efficiency as time went by.

In relation to the median of the subjective complexity (SubComp), during the given time (see top right-hand side of Fig. 4), it would appear that the subjects rated the MOD tasks as being more difficult than the COM tasks. The collected data, grouped by the model complexity (LC, MC, and HC) is displayed at the bottom of Fig. 4, according to subject efficiency and subjective complexity, respectively.

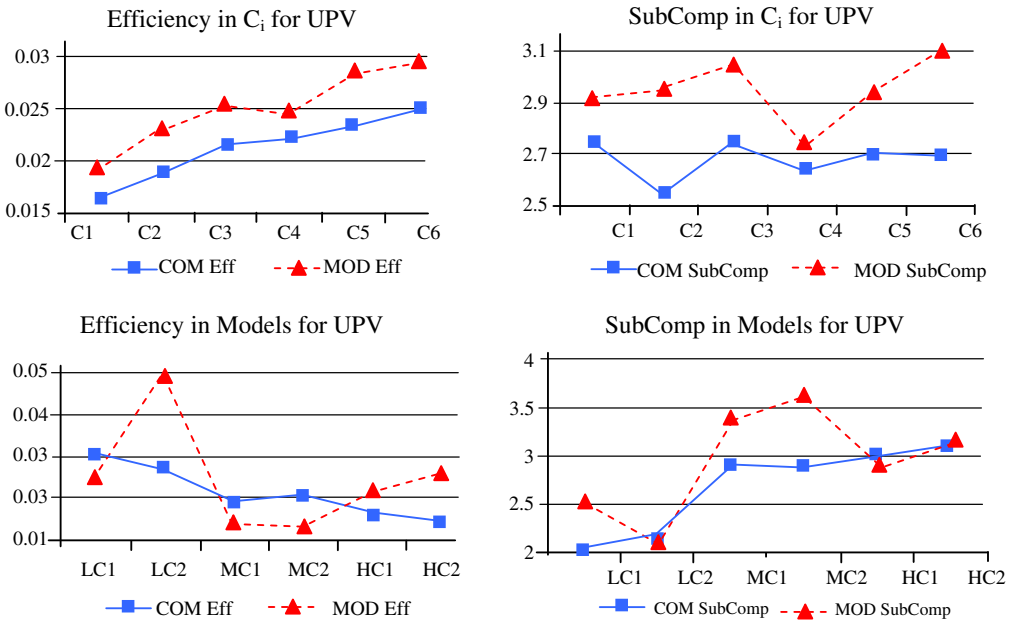


Fig. 4. Descriptive statistics for UPV.

The COM Eff decreases as the complexity of models increases (note that the horizontal axis shows the two models of each complexity, from low to high complexity). The same was expected to occur in MOD Eff. However, as we found in a previous study [61], the subjects were less efficient in performing MC models than HC models. The findings about the subjects' efficiency (grouped by model complexity) are similar to the findings of their subjective complexity. An increasing subjective complexity is observed in the COM tasks as the model complexity increases. However, in the MOD tasks, MC models are rated as being more difficult than HC models. The main difference between MC and HC models is that in the former the complexity is based mainly on combined navigations, (see the value of WNN), whereas in the latter the complexity is based mainly on an intertwining collection of operations (see the value of WNCO). We believe that it was more difficult for the subjects to identify and trace which relationships they should use (rolename, attribute name, etc.) in MOD tasks, than to identify which operation collections should be used to modify the expression. Finally, through the use of Shapiro–Wilk tests, we found that the COM/MOD Eff did not follow a normal distribution.

5.1.2. Descriptive and exploratory studies for UNSL

The top left-hand side of Fig. 5 shows the average efficiency of subjects', along with each C_i for UNSL. This figure shows that the subjects were more efficient in MOD tasks than in COM tasks (equal to UPV). On the majority of occasions, the subjects improved their efficiency as time went by (similar to UPV). With regard to the subjective complexity (SubComp) during the time given (see top right-hand side of Fig. 5), the subjects rated the MOD tasks as being more difficult than the COM tasks (equal to UPV). All of these findings are similar to UPV.

The bottom left and right-hand sides of Fig. 5 show the subject efficiency and subjective complexity, grouped by the model complexity, respectively. The COM Eff decreases as the complexity of the models increases. As far as the MOD Eff was concerned, we found the same result as in UPV: the subjects were less efficient in performing MC models than HC models.

The same results as those obtained in UPV were found with regard to subjective complexity: An increasing subjective complexity is observed in the COM tasks as model complexity increases. However, in the MOD tasks, MC models are rated as more difficult than HC models. We concluded that a similar pattern of data had been obtained in both experiments. We also found (through the use of Shapiro–Wilk tests) that the DVs did not follow a normal distribution.

5.2. Hypotheses testing

This section includes the analysis of each hypothesis for both experiments.

5.2.1. Testing hypotheses 1 for UPV and UNSL

To test hypothesis 1, we used the Friedman chi-square test (a non-parametric test for multiple related samples) which tests the null hypothesis that Mean Eff is the same in all the models considered. The results were significant (p-values were equal to 0.000, less than 0.05), i.e. the Mean Eff of subjects when performing COM and MOD tasks is different, depending on the complexity of the diagrams.

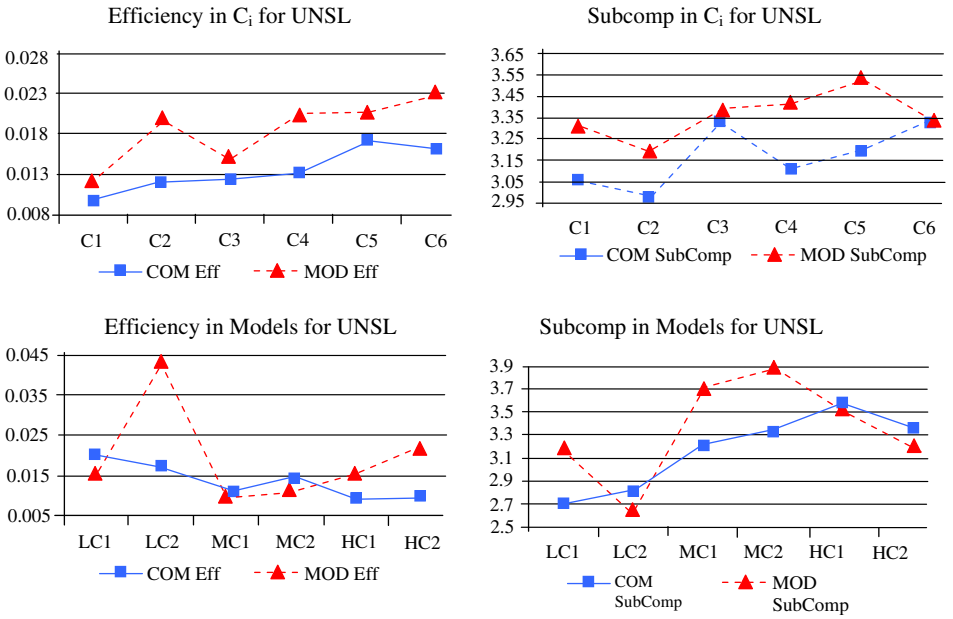


Fig. 5. Descriptive statistics for UNSL.

5.2.2. Testing hypotheses 2 and 3 for UPV and UNSL

In order to test hypotheses 2 and 3, a Spearman correlation analysis was performed, since this correlation test measures the rank-order association between two scale or ordinal variables. We used a level of significance of $\alpha = 0.05$, meaning that the level of confidence is 95% (i.e. the probability that we accept H_0 when H_0 is true is 0.95). We studied the correlation for independent observations, i.e. for each C_i, as was justified at the beginning of this section.

From Table 4, we conclude that the NNR, WNN, DN, NNC, WNCO and NEI measures are significantly correlated with the COM Efficiency in UPV. In UNSL, the same measures, with the exception of NEI, are correlated (four or five C_i) with COM Efficiency (Table 5). The NNR, WNN and DN measures are significantly correlated with the MOD Efficiency in the UPV. In UNSL, only the DN measure is correlated in more than three models (four C_i) with the MOD efficiency. These results are outlined in the first four rows of Table 6, which shows the quantity of significant coefficients found at level 0.05 between measures and DVs for the C_i.

The NNR, WNN, DN, NNC, NEI and WNCO measures are significantly correlated with the subjective complexity of the subject for COM tasks in UPV (see summarized results in the fifth row of Table 6). Almost the same set (sixth row of the same table) WNN, DN, WNCO and NEI measures, are correlated with the subjective complexity of the subject for COM tasks in more than two C_i (three or four C_i) for the UNSL. The NNR, WNN, DN and WNCO measures are significantly correlated with the MOD subjective complexity of the subjects for MOD tasks in UPV (seventh row). With regard to UNSL, only the NNR, WNN and DN measures are correlated with the subjective complexity of the subject for MOD tasks in three C_i, whereas DN is correlated in five C_i (last row).

The observed power of the non-significant results for Spearman correlation between measures and COM/MOD Eff (hypothesis 2) was:

- For COM Eff, in UPV lower than 43%, and lower than 45% in UNSL.
- For MOD Eff, in UPV and UNSL lower than 46%.

The observed power of the non-significant results for Spearman correlation between measures and COM/MOD SubComp (hypothesis 3) was:

Table 4
p-values of Spearman correlation between measures and COM/MOD Eff for UPV (significant results at level 0.05 are shown in bold font).

Direction of comprehension (mental model)	Rel. between problem objects				Problem objects		Reified objects
Measures	NNR	WNN	DN	NAN	NNC	NEI	WNCO
UPV COM Eff C_1	0.2543	**0.0928**	**0.0438**	**0.0069**	**0.0388**	**0.0002**	**0.0003**
UPV COM Eff C_2	**0.0018**	**0.0002**	**0.0001**	0.1938	**0.0006**	**0.0001**	**0.0000**
UPV COM Eff C_3	**0.0000**	**0.0000**	**0.0000**	0.4702	**0.0033**	**0.0126**	**0.0001**
UPV COM Eff C_4	**0.0158**	**0.0037**	**0.0195**	**0.0230**	**0.0001**	**0.0000**	**0.0000**
UPV COM Eff C_5	**0.0011**	**0.0004**	0.0583	0.2601	**0.0000**	**0.0002**	**0.0000**
UPV COM Eff C_6	**0.0022**	**0.0007**	**0.0036**	0.4339	**0.0000**	**0.0047**	**0.0000**
UPV MOD Eff C_1	0.0553	**0.0213**	**0.0002**	0.3639	0.3154	0.7056	0.5806
UPV MOD Eff C_2	**0.0013**	**0.0001**	**0.0000**	0.6063	0.7019	0.0864	**0.0295**
UPV MOD Eff C_3	**0.0002**	**0.0000**	**0.0000**	0.0615	0.8617	0.5163	0.2342
UPV MOD Eff C_4	**0.0253**	**0.0077**	**0.0001**	0.6771	0.9343	0.3236	0.1503
UPV MOD Eff C_5	**0.0058**	**0.0014**	**0.0000**	0.5381	0.8213	0.1745	0.1236
UPV MOD Eff C_6	**0.0000**	**0.0000**	**0.0000**	0.7194	0.1425	**0.0083**	**0.0009**

Table 5
p-values of Spearman correlation between measures and COM/MOD Eff for UNSL (significant results at level 0.05 are shown in bold font).

Direction of comprehension (mental model)	Rel. between problem objects				Problem objects		Reified objects
Measures	NNR	WNN	DN	NAN	NNC	NEI	WNCO
UNSL COM Eff C_1	0.1865	0.3052	0.7746	**0.0128**	0.6545	0.1099	0.4496
UNSL COM Eff C_2	**0.0128**	**0.0042**	**0.023**	0.1788	**0.0031**	**0.0027**	**0.0002**
UNSL COM Eff C_3	**0.1030**	**0.0054**	**0.0265**	0.9087	**0.0042**	0.0530	**0.0049**
UNSL COM Eff C_4	0.2429	0.1857	0.2973	0.5674	**0.0137**	0.2752	0.055
UNSL COM Eff C_5	**0.0308**	**0.0092**	**0.0004**	**0.1908**	**0.0350**	**0.0025**	**0.0002**
UNSL COM Eff C_6	**0.0034**	**0.0017**	**0.0112**	**0.6669**	**0.0330**	**0.0115**	**.0004**
UNSL MOD Eff C_1	0.8749	0.762	0.0959	0.0753	0.5173	0.1111	0.2952
UNSL MOD Eff C_2	0.1345	0.0501	**0.0006**	0.739	0.3825	0.1549	0.2210
UNSL MOD Eff C_3	0.0692	**0.0403**	**0.0226**	0.2019	0.6236	0.9763	0.8233
UNSL MOD Eff C_4	**0.0069**	**0.0006**	**0.0002**	0.6851	0.8556	**0.0243**	**0.0198**
UNSL MOD Eff C_5	0.1795	0.0914	**0.0075**	0.9838	0.8250	0.2916	0.2701
UNSL MOD Eff C_6	0.2109	0.1756	**0.0222**	0.3621	0.1518	0.9381	0.8913

Table 6
Time significant correlations were found between measures and DVs for the C_i.

Direction of comprehension (mental model)	Rel. between problem objects				Problem objects		Reified objects
Measures	NNR	WNN	DN	NAN	NNC	NEI	WNCO
UPV COM Eff	**5**	**5**	**5**	2	**6**	**6**	**6**
UNSL COM Eff	**4**	**4**	**4**	**3**	**5**	**3**	**4**
UPV MOD Eff	**5**	**6**	**6**	0	**0**	1	**2**
UNSL MOD Eff	1	2	**5**	0	0	1	1
UPV COM SubComp	**5**	**5**	**5**	0	**4**	**4**	**6**
UNSL COM SubComp	2	**4**	**3**	0	2	**3**	**4**
UPV MOD SubComp	**6**	**6**	**6**	1	0	1	**4**
UNSL MOD SubComp	**3**	**3**	**5**	0	0	0	1

- For COM SubComp, in UPV lower than 46%, and lower than 43% in UNSL.
- For MOD SubComp, in UPV less than 38%, and lower than 42% in UNSL.

The non-significant results do not mean that there is no correlation, because the probability that this decision (non-correlation) was correct is less than 0.46 in all cases considered for hypotheses 2 and 3.

5.2.3. Testing hypotheses 4 and 5 for UPV and UNSL

To test the hypotheses 4 and 5, we transformed the COM SubComp and MOD SubComp variables, by assigning numbers to the linguistic labels, ranging from 1 (assigned to "Easily understandable/modifiable") to 5 (which corresponded with "Barely understandable/modifiable"). We then used a Spearman coefficient to contrast the hypotheses $H_{0,4}$ and $H_{0,5}$. The results observed were:

- When we studied the correlation between COM/MOD Eff/Time and COM/MOD SubComp in UPV, the results were significant in all Ci, except for the correlations between COM/MOD Eff/Time and COM SubComp in C_1, where the results were non-significant. These non-significant results do not mean that there is no correlation, because the probability that this decision (non-correlation) was correct is less than 0.40.
- In UNSL, the results were significant for correlation between COM Time and COM SubComp in C_2 and C_3; for correlation between MOD Time and MOD SubComp in C_1, C_4 and C_5. The results for correlation between COM Eff and COM SubComp were significant in C_3 and C_4; for correlation between MOD Eff and MOD SubComp in C_1, C_3, C_4 and C_5. The power test for the non-significant results was less than 49%, so we can not believe correlation does not exist.

To sum up, it seems that in UNSL the subjective perception of subjects when they have to rate tasks is not as influenced by the COM/MOD Eff/Time as it was in UPV.

Furthermore, we must take into account that the powers of the non-significant results are below 48% in all the cases considered. For that reason we can not presume that there is no correlation in those cases either. The powers were calculated by approximation to the corresponding Pearson one.

Finally, the coefficients are negative for Efficiency variables and positive for Time variables (in both experiments), i.e. those tasks rated as difficult were time-consuming tasks and the subjects were less efficient.

5.3. Cognitive explanation of the findings obtained

In order to give a cognitive explanation, we based our reasoning on the use of the main categories of the mental model of subjects when they comprehend an OCL expression. In Table 7 we provide a mapping between these categories [12] and the proposed measures. The categories were empirically obtained in a previous work, presented in [79].

The analysis of this mapping, i.e. why each category is related to each of the measures used in the experiments is described as follows:

- *Problem objects.* The main problem objects within an OCL expression are:
  - The main object of the expression, the contextual instance, measured by NES and NIS.
  - Coupled objects of the problem domain that are obtained through navigations (measured by NNC).

Table 7
Mapping between mental model categories of subjects and measures.

	Rel. between problem objects				Problem objects		Reified objects
Measures	NNR	WNN	DN	NAN	NNC	NEI	WNCO

  - The iterators can be used as an explicit reference to problem objects. Iterators are commonly specified in collection operations which loop over the collection [90]. Attribute references in OCL can be interpreted as a client-server component of Burkhardt et al. [12] since, as Warmer and Kleppe [90] explain, any attribute reference in an OCL expression needs to be mapped to the corresponding get operation when implementing OCL expressions. Nevertheless, we consider that object properties, such as methods and attributes, are considered as part of the references of problem object properties. NAN and NAS, therefore, being attributes and operation references, belong to this category.
- *Relationship between problem objects.* Relationships between problem objects are any navigation within an OCL expression using association-ends of UML relationships (measured by NNR, WNN and DN).
- *Reified objects.* OCL collections constitute reified objects, measured by the WNCO measure.
- *Elementary operations.* Boolean and comparison operators are elementary operations (measured by NBO and NCO). OCL keywords are also part of this component (measured by NKW).

Hereafter, we will give a plausible cognitive explanation of the findings obtained through testing hypotheses 1–5.

5.3.1. Cognitive explanation of the testing of hypothesis 1

As was expected, the efficiency of the subjects is different according to the levels of import-coupling of the UML/OCL models. The cognitive complexity of the modelers dealing with models of different complexity is different and this is also evidenced by their efficiency in answering the tasks. The efficiency of the modelers is low when they have to deal with models which have high import-coupling.

5.3.2. Cognitive explanation of the testing of hypotheses 2 and 3

With regard to the conclusions and the distinct measures affecting the COM and MOD tasks, we can add the following:

- Regarding the main categories of the subjects' mental model, problem objects (NNC, NEI), relationship of problem objects (NNR, WNN, DN) and reified objects (WNCO) affect the COM Eff (hypothesis 2). However, only Relationship of problem objects (NNR, WNN, and DN) affects MOD Eff. Almost the same set of groups affects the COM and MOD SubComp (hypothesis 3). We believe that OCL comprehension demands a broad familiarity with the expression and its contextual information. However, during modification, the relationship between problem objects is the major category of the mental models that affects this activity.
- Measures related to chunking and tracing affect COM tasks, and it is mainly tracing that affects the MOD tasks.

5.3.3. Cognitive explanation of the testing of hypotheses 4 and 5

The fact that the instrumentation used for OCL expression maintainability is correlated with the subjective complexity means that the subjects' perception of the complexity of the tasks is influenced by the COM (or MOD) Time and also by their COM (or MOD) Efficiency when performing such tasks.

5.4. Threats to validity

We shall now discuss various threats to validity and the way in which we attempted to alleviate them:

- *Threats to external validity.* We have dealt with the following issues:
  - *Subjects.* In this family, we used students as experimental subjects. The tasks to be performed did not require high levels of industrial experience, so we believed that this experiment could be considered appropriate, as suggested in the literature [2,47]. Working with students also implies a set of advantages, such as the fact that the prior knowledge they bring to the experiment is rather homogeneous. We might add that there is the possible availability of a large number of subjects [88], and that there exists the chance to test experimental design and initial hypotheses [84]. An additional advantage of using novices as subjects in experiments on understandability is that the cognitive complexity of the objects under study is not hidden by the experience of the subjects.
    Nevertheless, although the students have a good background in UML, they did not have a high enough level of knowledge to perform our experiment and they needed to be given a course in OCL.
  - *Interaction of setting and treatments.* This is the effect of not having a representative experimental setting or material. In the experiment we used OCL expressions which could be considered representative of real cases. Moreover, we gave a course in OCL, using the same terminology as that appearing in its most recent version (2.0).
- *Threats to construct validity.* In this family we proposed an objective measure for the variables used in the hypothesis: (1) for the dependent variable we have used a measure of how precise the subjects were at answering tasks per time (the COM and MOD Efficiency) along with the time that the subjects spent on different tasks (the COM and MOD time); (2) for those hypotheses related to the subjects' cognitive aspects we have used a qualitative measure of the subjects' subjective opinion, and we have used linguistic labels, providing a scale to rate tasks; (3) the validity of the independent variables is guaranteed by Briand et al.'s framework ,which was used to validate them [6]. A further issue is that of:

  - *Confounding constructs and levels of constructs.* This threat is not primarily concerned with the presence or absence of a construct, but with the level of the construct that is of importance to the outcome [91]. In relation to the knowledge of the UML language, the subjects' years of experience with UML is different: UPV students have one year of experience, while UNSL students have different levels of experience.
- *Threats to internal validity.* We have dealt with the following issues:
  - *Maturation.* Subjects may react differently as time goes by. For example, the subjects are affected negatively during the experiment if they get tired or bored [91]. We dealt with this issue, running a pilot experiment to estimate the average time the subject would spend on performing the four tests. The estimated time was 1 h, and we believed that this would not produce boredom effect [49]. In fact, after running the experiment, we proved that the estimated time was correct. But a positive effect (Learning) may exist during the course of the experiment; Figs. 4 and 5 show this effect in a descriptive way for Efficiency through C_i.
  - *Selection.* This is the effect of natural variation on human performance. Depending on how the subjects are selected from a larger group, the selection effects may vary [91]. The selection of subjects was the same in both experiments. The subjects were volunteers who decided to take part in either an OCL course (UPV) or an international school course (UNSL), and as Wohlin [91] argues, the effect of using volunteers may influence the results, owing to their motivation and suitability for a new task.
  - *Persistence effects.* The subjects had never performed a similar experiment, so the persistence effect was avoided.
  - *Other factors.* Plagiarism and influence between subjects could be controlled. Two subjects who sat adjacently performed the (COM or MOD) tasks in a different order. A supervisor controlled the test and the subjects were asked not to talk to each other.
- *Threats to conclusion validity.* In the conclusion validity we wish to ensure 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 the universities that collaborate with our research. Secondly, the quantity and quality of the data collected, as well as the data analysis, were sufficient to support our conclusion, principally as described in previous sections, concerning the existence of a statistical relationship between independent and dependent variables.

6. Meta-analysis study

In families of experiments it is recommendable to synthesize the individual results using meta-analysis methods. By doing so, the results obtained from the individual data analysis can be improved when they are integrated. There are several statistical methods that allow us to integrate and interpret a set of results obtained through different experiments that are inter-related because they check similar hypotheses [38,45,61]. In the present study, we have used meta-analysis, because it allows us to extract more general conclusions, even though some of the experimental conditions are not exactly the same.

Meta-analysis is a set of statistical techniques for combining the different effect sizes of the experiments to obtain the global effect of a treatment or independent variable. There are many different types of effect size, but they fall into two main types [83]:

- standardized mean difference (e.g., Cohen's d or Hedges g) or
- correlation (e.g., Pearson's r) between continuous variables

It is possible to convert one effect size into another, so each really just offers a different scaled measure of the strength of an effect or a relationship.

As measures may come from different environments and not be homogeneous, a standardized measure of each one needs to be obtained and then those measures for estimating the global effect size of the factor must be combined.

When non-parametric statistical test are used in the report of a study, there are particular difficulties in deriving effect size estimates. In many cases this is impossible [38]. In our study we have used the Spearman non-parametric statistical test, which can be studied as the Pearson correlation test.

The Meta-Analysis v2 tool [4] was used to carry out the meta-analysis presented in this work.

For effect size we used the Spearman correlation coefficient, which measures the correlation between each measure considered in the hypotheses 2, 3, 4 and 5. For hypothesis 1 there is no need to perform a meta-analysis, given that the individual analysis was significant in both experiments (UPV, UNSL), and therefore the meta-analysis would not contribute anything new.

From Spearman correlations we obtained Hedges' g metric [45,50], which has been used to synthesize the correlation through the two groups, UNSL and UPV.

Eq. (1) show the Hedges' g metric, a weighted mean whose weights depend on sample size, where $w_i = (n_i - 3)$ and n_i is the sample size of the ith experiment.

$$\overline{Z} = \frac{\sum_i w_i z_i}{\sum_i w_i} \tag{1}$$

The higher the value of Hedges' g is, the higher the corresponding correlation. For studies in Software Engineering we can classify effect sizes into three different values: small, medium and large [50]. Once the global effect size is obtained, we can provide a confidence interval or a p-value which allows us to decide about the meta-analysis hypotheses, in a similar way to how a number of well-known empirical studies had been conducted [29,44,55,62].

6.1. Meta-analysis results

The meta-analysis results which we set out below will report the effect size measured as a Hedges'g or as a correlation coefficient [50], along with a classification obtained according to the value of the effect size. The results also include the p-value. The meanings of these concepts are explained as follows:

- The magnitude of effect size (Hedges' g) is computed based on the two correlation coefficients, that measure for instance, the overall correlation between each measure of OCL expression and COM/MOD Eff ($H_{0,2\text{-a}}$ and $H_{0,2\text{-b}}$).
- The Small, Medium, or Large classification is an indicator of the magnitude of effect size (Table 8). For studies in Software Engineering, Hedges' g effect sizes of 1.00–3.40, 0.38–1.00 and 0–0.37 are considered large, medium and small, respectively [50], the correlation for each group being 0.456–0.868, 0.193–0.456 0.00–0.193. The Hedges' g and correlation are in absolute value.
- The p-value is an indication of whether the result is statistically significant (p-value $\leqslant 0.05$) or not.

In the following paragraphs we will describe the meta-analysis results for each of the formulated hypotheses, as obtained through the Meta-Analysis v2 tool [4].

6.1.1. Meta-analysis results of hypothesis 2

Tables 9 and 10 summarize the results we obtained from hypothesis 2 in the meta-analysis.

We will explain the COM/MOD Eff results from two points of view. First of all, we want to observe the results in the first test (C_1), when learning effects did not affect the subjects. And secondly, we are interested in analyzing the number of significant results over the six tests ($C_1 \cdots C_6$), that is, what correlations are significant throughout these.

The COM Eff results in Table 9 show that:

- In the first test (C_1) performed by the subjects, the measures related with tracing (NAN and WNCO) were significantly related with the COM Eff, and the measure related with chunking (NEI) was related with the COM Eff.
- When we analyze the correlation through the six tests, the measures related to tracing, NNR, WNN, DN, NNC and WNCO, as well as NEI, a measure related to chunking, gave at least, four significant results of the six tests.
- All the significant results were associated with negative correlation: the higher the value of the measure for OCL expression, the lower its comprehensibility.

We can therefore accept the alternative hypotheses $H_{1,2}$, that there is significant correlation between some of the measures defined for OCL expressions, related to import-coupling, and their COM Eff.

MOD Eff was similarly analyzed (Table 10):

- In the first test C_1, the measure about tracing, DN, related to navigation, is significantly related to MOD Eff.
- When all the tests are considered, the tracing measures related to navigation, NNR, WNN and DN are significantly correlated to MOD Eff, at least in four of the six tests. In fact, the DN measure was significant in all the results, and the correlation is negative, i.e. the greater the depth of navigation, the more difficult it is to comprehend the OCL expression.

We can thus accept the alternative hypotheses $H_{1,2}$, which means that there is significant negative correlation between some of the measures defined for OCL expressions related to import-coupling and their MOD Eff.

6.1.2. Meta-analysis results for hypothesis 3

The results obtained in the meta-analysis for the hypothesis 3 for COM Subcomp are the following (Table 11):

Table 8
Classification of global effect size.

|Hedges' g|	Classification	|Correlation|
1.02–3.40	Large	0.456–0.868
0.38–1.02	Medium	0.193–0.456
0.00–0.38	Small	0.00–0.193

Table 9
Hedges' g values for COM Eff.

Measure	Global measure correlation with COM Eff								
	C_1			C_2			C_3		
	G	p-value	Effect-size	g	p-value	Effect-size	g	p-value	Effect-size
NNR	0.000	0.999	Small	**−0.935**	0.000	**Medium**	**−1.268**	0.000	**Large**
WNN	−0.127	0.581	Small	−1.111	0.000	Large	**−1.400**	0.000	**Large**
DN	−0.288	0.212	Small	−1.013	0.000	Large	**−1.139**	0.000	**Large**
NAN	**−0.857**	0.000	**Medium**	−0.421	0.067	Medium	0.104	0.645	Small
NNC	−0.272	0.241	Small	−1.079	0.000	Large	**−0.965**	0.000	**Medium**
NEI	**−0.865**	0.000	**Medium**	−1.150	0.000	Large	**−0.727**	0.002	**Medium**
WNCO	**−0.701**	0.004	**Medium**	−1.509	0.000	Large	**−1.115**	0.000	**Large**
	C_4			C_5			C_6		
	G	p-value	Effect-size	g	p-value	Effect-size	g	p-value	Effect-size
NNR	**−0.585**	0.013	**Medium**	**−0.903**	0.000	**Medium**	**−0.905**	0.000	**Medium**
WNN	**−0.693**	0.004	**Medium**	**−1.033**	0.000	**Large**	−0.195	0.388	Small
DN	**−0.551**	0.019	**Medium**	**−0.840**	0.001	**Medium**	0.235	0.300	Small
NAN	**−0.464**	0.046	**Medium**	−0.395	0.093	Medium	**−0.888**	0.000	**Medium**
NNC	**−1.062**	0.000	**Large**	**−1.057**	0.000	**Large**	−0.248	0.275	Small
NEI	**−0.863**	0.001	**Medium**	**−1.153**	0.000	**Large**	**−1.179**	0.000	**Large**
WNCO	**−1.126**	0.000	**Large**	**−1.439**	0.000	**Large**	**−1.354**	0.000	**Large**

Table 10
Hedges' g metric values for MOD Eff.

Measure	Global measure correlation with MOD Eff								
	C_1			C_2			C_3		
	g	p-value	Effect-size	g	p-value	Effect-size	g	p-value	Effect-size
NNR	−0.292	0.205	Small	**−0.773**	0.001	**Medium**	**−0.915**	0.000	**Medium**
WNN	−0.425	0.068	Medium	**−0.958**	0.000	**Medium**	**−1.031**	0.000	**Large**
DN	**−0.883**	0.000	**Medium**	**−1.488**	0.000	**Large**	**−1.197**	0.000	**Large**
NAN	−0.099	0.667	Small	0.037	0.869	Small	**0.510**	0.028	**Medium**
NNC	0.265	0.243	Small	0.063	0.780	Small	0.102	0.652	Small
NEI	−0.295	0.197	Small	**−0.506**	0.030	**Medium**	−0.104	0.644	Small
WNCO	−0.247	0.276	Small	**−0.554**	0.018	**Medium**	−0.232	0.307	Small
	C_4			C_5			C_6		
	g	p-value	Effect-size	g	p-value	Effect-size	g	p-value	Effect-size
NNR	**−0.800**	0.001	**Medium**	**−0.672**	0.005	**Medium**	**−0.910**	0.000	**Medium**
WNN	**−0.998**	0.000	**Medium**	**−0.806**	0.001	**Medium**	**−1.015**	0.000	**Large**
DN	**−1.400**	0.000	**Large**	**−1.166**	0.000	**Large**	**−1.276**	0.000	**Large**
NAN	0.010	0.964	Small	0.100	0.657	Small	0.195	0.390	Small
NNC	−0.041	0.857	Small	0.071	0.754	Small	−0.037	0.872	Small
NEI	**−0.496**	0.034	**Medium**	−0.387	0.091	Medium	−0.441	0.060	Medium
WNCO	**−0.592**	0.012	**Medium**	−0.426	0.064	Medium	**−0.501**	0.037	**Medium**

- In the first test C_1, all the measures, except NNR and NNC, are significantly related with COM SubComp.
- When we considered all the tests, only NAN is correlated significantly with COM SubComp for one test. The other measures were correlated significantly in more than four of the six tests.

After testing hypothesis 3, the following findings were obtained for MOD SubComp (Table 12):

- In the first test C_1, the tracing measures related to navigation, NNR, WNN and DN, are significantly related to MOD SubComp.
- When we considered all the tests, the tracing measures NNR, WNN, DN and WNCO are significantly correlated with MOD SubComp in at least four of the six tests.

6.1.3. Meta-analysis results of hypothesis 4

For hypothesis 4, the following findings were obtained:

- The time that the subjects spent doing comprehension tasks (COM Time) is correlated with the subjective perception of the comprehension complexity (COM SubComp) in all the tests, except in C_1.

Table 11
Correlation between measures and COM SubComp.

Measure	C_1			C_2			C_3		
	ρ	p-value	Effect-size	ρ	p-value	Effect-size	ρ	p-value	Effect-size
NNR	0.180	0.113	Small	**0.345**	0.002	**Medium**	**0.399**	0.000	**Medium**
WNN	**0.252**	0.024	**Medium**	**0.393**	0.000	**Medium**	**0.439**	0.000	**Medium**
DN	**0.383**	0.000	**Medium**	**0.417**	0.000	**Medium**	**0.418**	0.000	**Medium**
NAN	**0.244**	0.030	**Medium**	0.083	0.470	Small	0.013	0.907	Small
NNC	0.172	0.130	Small	**0.276**	0.013	**Medium**	**0.347**	0.002	**Medium**
NEI	**0.364**	0.001	**Medium**	**0.316**	0.004	**Medium**	**0.297**	0.008	**Medium**
WNCO	**0.378**	0.001	**Medium**	**0.400**	0.000	**Medium**	**0.412**	0.000	**Medium**
	C_4			C_5			C_6		
	ρ	p-value	Effect-size	ρ	p-value	Effect-size	ρ	p-value	Effect-size
NNR	**0.240**	0.033	**Medium**	**0.332**	0.003	**Medium**	**0.264**	0.018	**Medium**
WNN	**0.306**	0.006	**Medium**	**0.351**	0.001	**Medium**	**0.298**	0.007	**Medium**
DN	**0.334**	0.002	**Medium**	**0.235**	0.037	**Medium**	**0.306**	0.006	**Medium**
NAN	0.185	0.104	Small	0.084	0.464	Small	0.059	0.607	Small
NNC	**0.333**	0.003	**Medium**	**0.392**	0.000	**Medium**	**0.250**	0.026	**Medium**
NEI	**0.318**	0.004	**Medium**	**0.339**	0.002	**Medium**	**0.243**	0.031	**Medium**
WNCO	**0.417**	0.000	**Medium**	**0.416**	0.000	**Medium**	**0.331**	0.003	**Medium**

Table 12
Correlation between measures and MOD SubComp.

Measure	C_1			C_2			C_3		
	ρ	p-value	Effect-size	ρ	p-value	Effect-size	ρ	p-value	Effect-size
NNR	**0.266**	0.018	**Medium**	0,062	0,588	Small	**0,500**	0,000	**Large**
WNN	**0.321**	0.004	**Medium**	**0,405**	0,000	**Medium**	**0,550**	0,000	**Large**
DN	**0.454**	0.000	**Medium**	**0,470**	0,000	**Large**	**0,585**	0,000	**Large**
NAN	−0.039	0.736	Small	0,049	0,669	Small	−0,168	0,139	Small
NNC	−0.034	0.764	Small	0,055	0,631	Small	0,158	0,164	Small
NEI	0.140	0.219	Small	**0,222**	0,049	**Medium**	0,188	0,098	Small
WNCO	0.164	0.148	Small	**0,257**	0,022	**Medium**	**0,301**	0,007	**Medium**
	C_4			C_5			C_6		
	ρ	p-value	Effect-size	ρ	p-value	Effect-size	ρ	p-value	Effect-size
NNR	**0.402**	0.000	**Medium**	**0.435**	0.000	**Medium**	**0.344**	0.002	**Medium**
WNN	**0.465**	0.000	**Large**	**0.459**	0.000	**Large**	**0.375**	0.001	**Medium**
DN	**0.548**	0.000	**Large**	**0.432**	0.000	**Medium**	**0.352**	0.001	**Medium**
NAN	−0.035	0.759	Small	−0.125	0.272	Small	−0.031	0.788	Small
NNC	0.153	0.177	Small	0.131	0.250	Small	0.153	0.178	Small
NEI	**0.227**	0.044	**Medium**	0.188	0.097	Small	**0.212**	0.060	**Medium**
WNCO	**0.324**	0.003	**Medium**	**0.251**	0.025	**Medium**	**0.270**	0.016	**Medium**

- The time that the subjects spent doing the modification tasks (MOD Time) is correlated with the subjective perception of the modification complexity (MOD SubComp) in all the tests (see Fig. 6).

6.1.4. Meta-analysis results of hypothesis 5

For hypothesis 5, the following findings were obtained:

- The efficiency of the subjects doing comprehension tasks (COM Eff) is correlated with their perception of the complexity of this activity (COM SubComp) in all the tests, except in C_1. The correlations are negative; that means that the more efficient the subjects are in performing COM tasks, the less difficult they rate the OCL expression' comprehension as being; i.e. they perceive the OCL expression to be less complex.
- The efficiency of the subjects doing modification tasks (MOD Eff) is correlated with the subjective perception of the modification (MOD SubComp) in all the tests (see Fig. 7). Furthermore, the correlations are negative, which means that the more efficient the subject are, the less complex they rate the modification of the OCL expression.

6.2. Conclusions of the meta-analysis

In the following lines we will summarize the main findings we obtained through meta-analysis:
For hypothesis 2, which studies the relationship between the measures we defined (Table 1) and COM/MOD Eff:

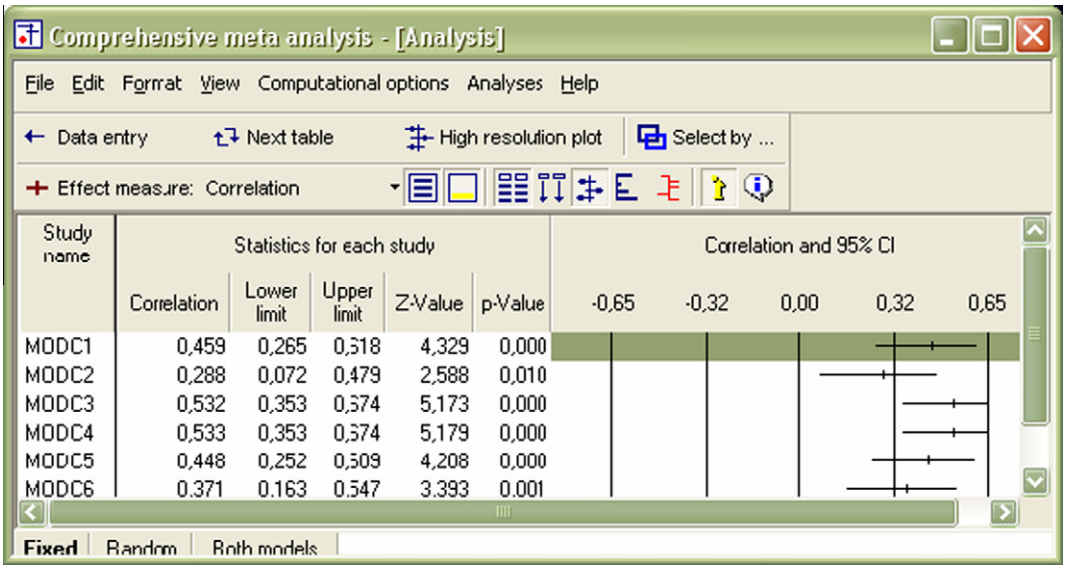


Fig. 6. Time vs. MOD SubComp meta-analysis (hypothesis 4).

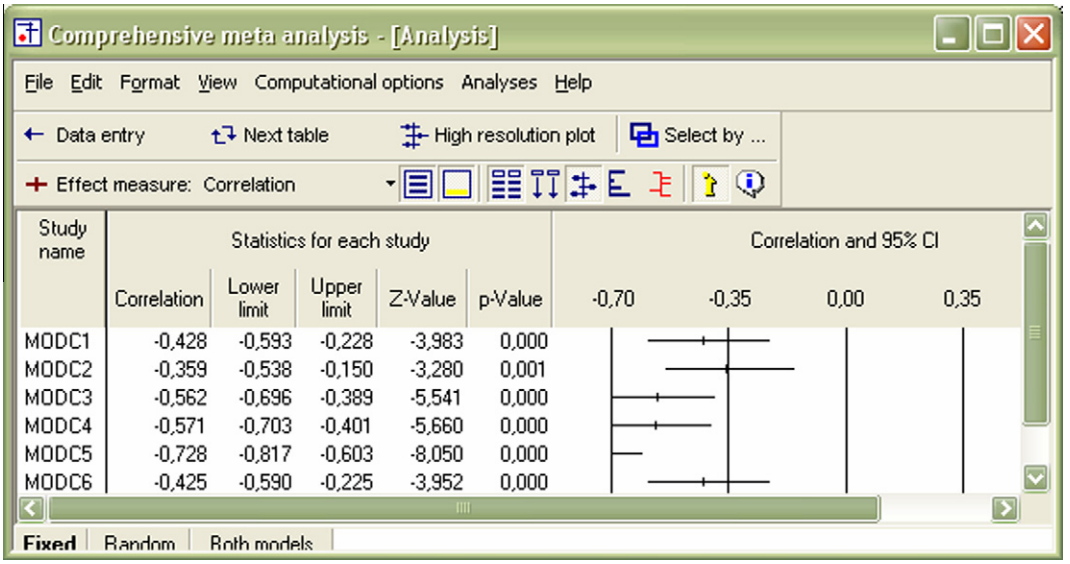


Fig. 7. Efficiency vs. MOD SubComp meta-analysis (hypothesis 5).

- The tracing measures WNCO and NEI are correlated with the COM Eff in all the tests. It seems that collection operations, as well as the number of iterator variables of collection operations, significantly influences the comprehension efficiency of OCL expressions.
- The measure DN is correlated with MOD Eff in all the tests, meaning that the depth of navigation seems to be a relevant factor of the MOD Eff. The Number of Navigated Relationships NNR, together with the Weighted Number of Navigations WNN, are important factors of the MOD Eff as well.
- The common aspect of these three measures (DN, NNR and WNN) is the navigation. From our point of view, this means that the most important factors that influence MOD tasks are that the modelers have to trace the UML diagram to identify what relationships (NNR) should be used in a modification, as well as how the relationship should be combined (WNN), and how deep the implemented navigation should be.

Meta-analysis results, when compared with those obtained in the individual analysis of each experiment reveal that:

- Essentially, the same findings were obtained when the analysis of the individual experiment was compared the meta-analysis study. NNR, NNC, WNN, DN, WNCO and NEI measures are correlated with the COM Eff.

- Regarding the MOD Eff, the same findings were obtained, given that the set of measures which have correlation are coincident.

For hypothesis 3, which studies the relation between measures and COM/MOD SubComp:

- The measures NNR, WNN, DN and WNCO are correlated with both subjective measures. This means that the import-coupling due to these navigation measures seems to be related to the subjective perception of the subjects about the complexity of the comprehensibility and modifiability of OCL expressions.

For hypothesis 4, which studies the relation between the COM/MOD Time and COM/MOD SubComp, respectively:

- The time spent by the subject in doing the tasks is correlated with both subjective perceptions (COM/MOD SubComp) in all the tests, except in C_1 for COM SubComp. The reason that C_1 is excluded is likely to be that the COM tasks were performed before MOD tasks and that C_1 represents the set of first tests performed by all the subjects.

For hypothesis 5, which studies the relation between the COM/MOD Eff and COM/MOD SubComp, respectively, the results are the same as for the COM/MOD Time.

7. Related work

Although several works concerning the comprehension and maintainability of UML models exist ([25,73,69,39,34,56,92] among others), work related to the maintainability of OCL expressions is extremely scarce. The only empirical work concerning OCL was presented in Briand et al. [7], which proved empirically that OCL has the potential to significantly improve engineers' ability to understand, inspect, and modify a system modeled with UML. These benefits are obtained only after a certain learning curve is overcome. Although the Briand el al. experiment was run with students, the authors argue [7] that they did not evaluate this fact as a threat to its external validity, because students are better equipped than professionals. Nevertheless, in some universities the students were not sufficiently prepared in OCL to perform our experiment; we found that the training of students in this topic had to be reinforced, to obtain the benefit of the language. We always gave a lecture on OCL before the experiments were run. As we have carried out all the empirical and measurement-based work related to the maintainability of OCL expressions that is in existence, all related work necessarily focuses on our own research.

We shall now briefly summarize our previous work, which focused mainly on assessing the influence of import-coupling on the maintainability of OCL expressions:

- In [77] we presented an experiment and two replicas, which were carried out in academic environments, with the goal of ascertaining whether any relationship exists between import-coupling (defined in OCL expressions through navigations and collection operations), and the comprehensibility and modifiability of OCL expressions. In this empirical study, the subjects were given three class diagrams with one OCL expression each and were asked to comprehend the expression and modify it to satisfy new requirements. The subjects were also asked to evaluate the complexity of comprehensibility and modifiability tasks in a subjective manner. Statistical analysis revealed that the measures we proposed appeared to be correlated with comprehensibility and modifiability. However, the subjects' efficiency in the modification tasks was low. We believe that the subjects' lack of experience in OCL may have been the cause of many modification task errors. This led us to believe that it was necessary to change the type of modification tasks, and that rather than asking them to carry out modifications, they should select the correct modification, i.e. they should carry out a multiple choice type of modification task. We also believe that using nine different models as an experimental object and making few observations (only three) by each subject was not appropriate. Although each subject was given three models of a different level of coupling in each observation, it was not possible to balance all the models correctly.
- The previous family of experiments [77] provided us with a basis with which to carry out another family of experiments, in which the experimental material with regard to the modification tasks was changed, the number of models was reduced, and the number of observations was incremented. This experiment was introduced, briefly, in [80].
- In [79], we described a qualitative experiment using verbal protocols, in which the subjects were given three class diagrams and were asked to think aloud in order to verbalize their thought. The aim of this experiment was to validate a categorical model of the main categories of the subjects' mental models when dealing with OCL expressions. We obtained empirical findings to indicate that the main categories applied by modelers are problem objects, relationships between problems and reified objects. We also found that the breadth of familiarity [23,71,58] with the UML diagram gained by the subjects before starting to comprehend the OCL expression comprehension activities is different. The range varies in a continuous form, which extends from those subjects who made absolutely no attempt to comprehend the diagram to those who attempted to comprehend the class diagram systematically before starting to read the OCL expression.

The current paper extends the description of the experiment introduced briefly in [80] and presents its replica, in order to provide strong evidence of the practical utility of measures for the import-coupling of OCL expressions as early indicators of

their comprehensibility and modifiability. In addition, we provide an explanation of the empirical findings, taking into consideration the main categories of the modeler's mental models dealing with OCL expression comprehension. As was mentioned above, these categories were discovered through a qualitative experiment presented in [79]. In addition, the study presented in this article presents a deeper statistical analysis and synthesis, including a meta-analysis study.

8. Conclusions and future work

The main goal of this paper is to assess whether the import-coupling affects two sub-characteristics of the maintainability of OCL expressions, such as comprehensibility and modifiability. For that purpose, we have defined a set of measures for import-coupling and empirically validated these as indicators of OCL comprehensibility and modifiability. The empirical validation was done through an experiment and its replica, run in May and June (2005) with undergraduate students at Spanish and Argentinean Universities, respectively.

In the experimentation, we have considered not only the time the subjects spent on the comprehensibility and modifiability tasks which were required in the experimental material, but also their efficiency and their subjective perception of the difficulty in carrying out these tasks. We believe that both objective (COM and MOD Efficiency) and subjective (subject's rating or COM and MOD SubComp) information is important if we are to obtain more solid findings.

The influence of import-coupling on the comprehensibility and modifiability of OCL expressions was verified through hypothesis 2. We took Briand et al.'s framework [8–10] (Fig. 3) as our basis, to reaffirm this hypothesis by using a triangulation of hypotheses relating the import-coupling and cognitive complexity (hypothesis 3), as well as the cognitive complexity and the maintainability of OCL expressions (hypotheses 4 and 5) (Fig. 3).

A meta-analysis study was performed to integrate the results obtained in the experiment and its replica, so the results obtained in the individual analysis were borne out. The main findings of the meta-analysis are shown in Table 13. The results which are shown are the significant ones in at least four of the six tests.

Cognitive theories, the cognitive model of Cant et al. [18] and components of the mental model of Burkhardt et al. [12,13], helped us to obtain a better explanation of the measure definition and to interpret the empirical findings as well. We have analyzed how the two cognitive theories are related to the main concepts captured by the OCL expressions we defined. Tables 1 and 7 describe the results of this relationship, which is explained in detail in Sections 2 and 5.3.

As regards the comprehensibility and modifiability tasks, the import-coupling of an OCL expression makes an impact on the cognitive complexity in different ways (hypothesis 3), and this effect is explained as follows, by using the cognitive theory:

- The subjects in the comprehensibility tasks gained a broad understanding of OCL expressions through Problem objects (NNC, NEI), Relation of problem objects (NNR, WNN, DN) and Reified objects (WNCO). However, once this breadth of familiarity with the OCL expressions was gained, the modelers concentrated mainly on Relation of Problems Objects (NNR, WNN, DN) during modifiability tasks. Regarding the cognitive techniques of Cant et al. [18], the modelers applied chunking and tracing in the comprehensibility tasks, whereas in the modifiability tasks tracing was the most relevant technique applied. The power test for the non-significant results was less than 50%, so we can not accept the null hypothesis (there is no correlation).

  The empirical findings could be used for educational purposes, giving the following recommendations:
- The number of collection operations (and the quantity of iterator variables used in collection operations) should be as low as possible in order to obtain OCL expressions that are easier to comprehend. Collection operations are frequently used with navigations, in order to operate with coupled objects and they could seriously affect the comprehension of OCL expressions.

Table 13
Summary of hypotheses and findings obtained through the meta-analysis.

Relation between	Efficiency	Time	Subjective complexity
	COM Eff, MOD Eff	COM time MOD time	COM SubComp MOD SubComp
OCL expression	Hypotheses 2 Spearman correlation	–	Hypotheses 3 Spearman correlation
Measures	The NNR, NNC, WNN, DN, WNCO and NEI measures are correlated with the COM efficiency The NNR, WNN and DN are significantly correlated with the MOD efficiency		The NNR, WNN, WNCO, NNC, NEI and DN measures are correlated with the COM subjective complexity; The NNR, WNN, DN and WNCO measures are significantly correlated with the MOD Subjective Complexity
COM SubComp	Hypotheses 5 Spearman correlation	Hypotheses 4 Spearman correlation	
MOD SubComp	The subjective ratings are influenced by the COM and MOD efficiency	The subjects' subjective ratings are influenced by the COM and MOD Time	

- Tracing cognitive activity seems to be the main factor affecting the modification of OCL expressions. Although OCL expression modifications are not as easy as OCL expression comprehension, the task of tracing the UML diagram associated to the OCL expression, to identify which relationships should be used to implement a modification, is one of the main factors influencing the efficiency. The way relationships are combined when more than one relationship should be used, along with the depth of this combination, affects this kind of task too. A high import-coupling with a high Depth of Navigations (DN) makes the contextual instance know details of distant objects. So, whenever possible, it is important to limit the knowledge of coupled objects to the immediate surroundings of the contextual instance [4].
- In general, we believe that although the cognitive techniques of chunking and tracing influence OCL expressions comprehension, the modification of an OCL expression demands that these techniques should be applied in an intertwining way. During OCL modifications, modelers should identify new objects, new relationships between objects and collection operations, to specify the modifications. The identification of these aspects during modification and how they must be combined demands that modelers should be more cognitively flexible [21] than when they are comprehending OCL comprehensions.
- Tracing cognitive process [18], an important factor that disrupts chunking, seems to greatly influence software maintainability.

The previous recommendations are also useful for practitioners, such as modelers and metamodelers in industry. As Cabot et al argue in [14], "the modeling community is continuously pushing forward the OCL", and "OCL is used in quite different application domains (domain-specific language and web semantics) as well as for various purposes (model verification and validation, code-generation, test-driven development, and transformations)". The availability of quality indicators of OCL expressions benefit those purposes and help to evaluate possible changes in maintainability [20,36].

There is great room for future work, in several directions. Firstly, further experimentation with practitioners should be performed. Secondly, case studies could be run in the industry, in order to obtain more conclusive and generalized results. We will attempt to establish a measure of overall complexity of a class complemented with a set of OCL expressions, since all the measures presented are defined at expression level and not at class level.

Furthermore, a similar empirical validation work could be carried out with OCL expressions attached to UML statechart diagrams, in which OCL is commonly applied to define guards and events.

Acknowledgments

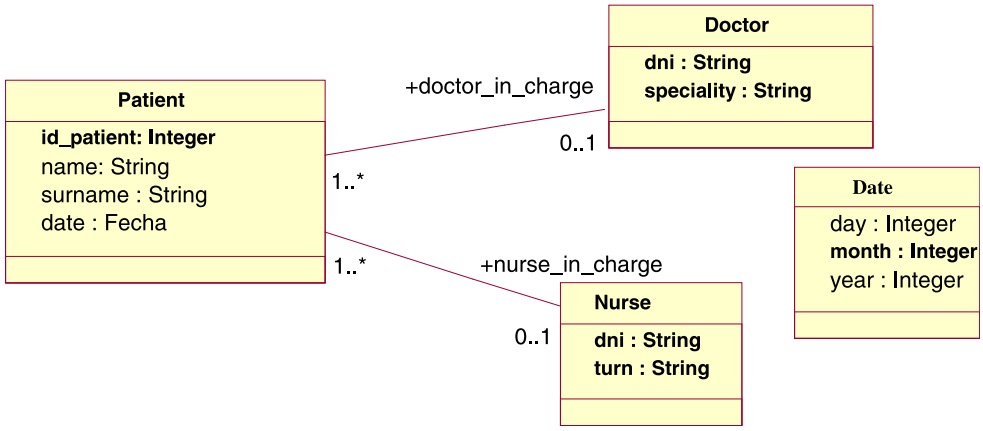
This research is part of the following projects: PEGASO/MAGO (MICINN and FEDER, TIN2009-13718-C02-01), EECCOO (MICINN TRA2009_0074), (ROADMAP MICINN TIN2008-05675), IDONEO (JCCM PAC08-0160-6141), MECCA (JCMM PII2I09-0075-8394) and the 04/e073 project financed by "Secretaría de Investigación de la Universidad Nacional del Comahue, Neuquén, Argentina".

Appendix A. Experimental materials

Below, we present an example of the experimental material (model 2) used in the experiment and replica described in this paper.

A.1. Comprehension tasks

Please write the time at which you begin the exercise: TIME (HH: MM: SS) _____



context Patient inv: (self.doctor_a_cargo->size() + self.enfermero_a_cargo->size()) = 1

With regard to the above OCL expression, please answer the following questions:

1. How many different relationships are used in navigations of relationships? Write the number and the name of the role-names used in any navigation.
2. Which of the following options (written using natural language) represents the meaning of the previous OCL expression? Choose the correct answer.
   - A patient is under the charge of both a doctor and a nurse.
   - If a patient is under the charge of a doctor, then he/she cannot be under the charge of a nurse.
   - A patient can be under the charge of a doctor or a nurse, but not both, or neither of the two.
3. How many collection operations have been used in the OCL expressions? Please mention the quantity and the name of each collection operation used.
4. Consider the navigation expressed as self.doctor. Which of the following options is true?
   - The self.doctor_a_cargo subexpression type is Integer.
   - The self.doctor_a_cargo subexpression type is Patient.
   - The self.doctor_a_cargo subexpression type is Doctor.

Please write the time at which you finish the exercise: TIME (HH: MM: SS) _____

A.2. Comprehension rating tasks

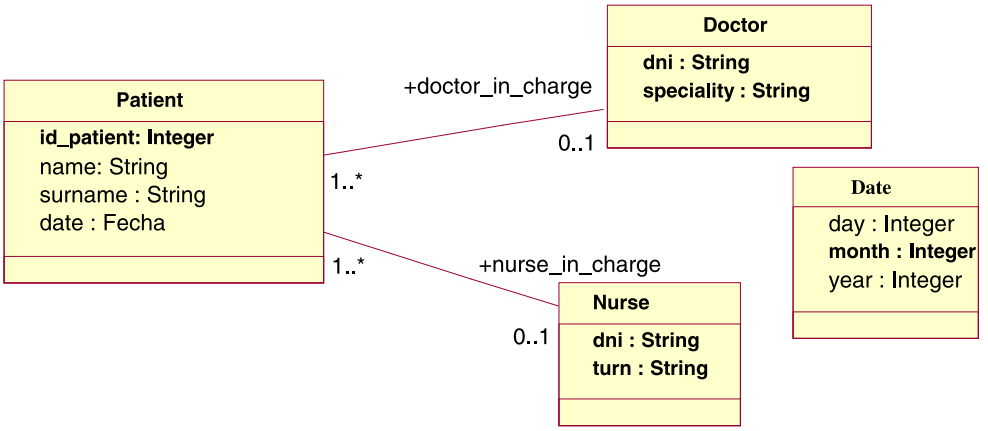
In your opinion, how understandable is the OCL expression? Choose one option only:

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

A.3. Modifications tasks

Consider the following functionality extension for the creation of a Car in the car rental system. When creating a new car, it is also necessary to indicate the class to which the car belongs (A = Luxury, B = Economy) and to buy an insurance policy from an insurance company.

Please write the time at which you begin the exercise: TIME (HH: MM: SS) _____



context Patient inv: (self.doctor_in_charge->size() + self.nurse_in_charge->size()) = 1

For each of the following requirements (which expresses a modification to the previous OCL expression written in natural language), please choose the correct OCL expression satisfying the requirement:

1. A patient is under the charge of a doctor or a nurse, or neither of them. context Patient inv:
   - (self.doctor_in_charge->size() + self.nurse_in_charge->size()) < 1
   - (self.doctor_in_charge->size() + self.nurse_in_charge - > size()) <= 1
   - (self.doctor_in_charge->size() + self.nurse_in_charge - > size()) > 1

2. A patient is under the charge of a doctor and a nurse. context Paciente inv:
   - self.doctor_in_charge>notEmpty() and self.nurse_in_chargeenfermero_a_cargo->isEmpty()
   - self.doctor_in_charge >notEmpty() and self.nurse_in_charge ->notEmpty()
   - self.doctor_in_charge ->isEmpty() and self.nurse_in_charge ->isEmpty()

Please write the time at which you finish the exercise: TIME (HH: MM: SS) _____

A.4. Modification rating tasks

In your opinion, how difficult are the modifications to the OCL expression? Choose one option only:

- Easily modifiable.
- Quite easy to modify.
- Normal.
- Quite difficult to modify.
- Barely modifiable.

References

[1] C. Atkinson, T. Kühne, Model-driven development: a metamodeling foundation, IEEE Software 20 (5) (2003) 36–41.
[2] V. Basili, F. Shull, F. Lanubile, Building knowledge through families of experiments, IEEE Transactions on Software Engineering 25 (1999) 456–473.
[3] V. Basili, A. Weiss, Methodology for collecting valid software engineering data, IEEE Transactions on Software Engineering 10 (6) (1984) 728–738.
[4] Biostat, Inc., Meta-Analysis v2, 2006, <http://www.meta-analysis.com>.
[5] D.A. Boehm-Davis, J.E. Fox, B. Philips, Techniques for exploring program comprehension, in: D.A. Boehm-Davis, W.D. Gray (Eds.), Empirical Studies of Programmers, Sixth Workshop, Norwood, Ablex, 1996, pp. 3–37.
[6] L.C. Briand, J.W. Daly, J. Wüst, A unified framework for coupling measurement in object-oriented systems, IEEE Transaction on Software Engineering 25 (1) (1999) 91–121.
[7] L.C. Briand, Y. Labiche, M. Di Penta, H. Yan-Bondoc, An experimental investigation of formality in UML-based development, IEEE Transactions on Software Engineering 31 (10) (2005) 833–884.
[8] L.C. Briand, J. Wüst, S. Ikonomovski, H. Lounis, A comprehensive investigation of quality factors in object-oriented designs: an industrial case study, Technical Report ISERN-98-29, International Software Engineering Research Network, 1998.
[9] L.C. Briand, J. Wüst, S. Ikonomovski, H. Lounis, Investigating quality factors in object-oriented designs: an industrial case-study, in: Proc. 21st Int. Conf. on Software Engineering, Los Angeles, CA, 1999, pp. 345–354.
[10] L.C. Briand, J. Wüst, H. Lounis, Replicated case studies for investigating quality factors in object-oriented designs, Empirical Software Engineering 6 (1) (2001) 11–58.
[11] D.E. Broadbent, The magic number seven after fifteen years, in: A. Kennedy, A. Wilkes (Eds.), Studies in Long-term Memory, Wiley, London, 1975, pp. 3–18.
[12] J. Burkhardt, F. Detienne, S. Wiedenbeck, Object-oriented program comprehension: effect of expertise, task and phase, Empirical Software Engineering 7 (2) (2002) 115–156.
[13] J.M. Burkhardt, F. Detienne, S. Wiedenbeck, The effect of object-oriented programming expertise in several dimensions of comprehension strategies, in: Proc. Sixth Int. Workshop on Program Comprehension, Washington, DC, USA, 1998, pp. 82–89.
[14] J. Cabot, M. Gogolla, P. Van Gorp, in: Eighth International Workshop on OCL Concepts and Tools, Models in Software Engineering: Workshops and Symposia at MODELS 2008, France, LNCS, vol. 5421, Springer, 2009, pp. 257–262.
[15] J. Cabot, E. Teniente, Transforming OCL constraints: a context change approach, in: Proc. of the 2006 ACM Symposium on Applied Computing, 2006, pp. 1196–1201.
[16] J. Cabot, E. Teniente, Transformation techniques for OCL constraints, Science of Computing Program 68 (3) (2007) 179–195.
[17] C. Calero, M. Piattini, M. Genero, Method for obtaining correct metrics, in: Proc. of the Third Int. Conf. on Enterprise and Information Systems, Setubal, Portugal, 2001, pp. 779–784.
[18] S.N. Cant, B. Henderson-Sellers, D.R. Jeffery, Application of cognitive complexity metrics to object oriented programs, Object Oriented Programming 7 (4) (1994) 52–63.
[19] D.N. Card, K. El Emam, B. Scalzo, Measurement of object-oriented software development projects, Software Productivity Consortium, Herndon, Virgina, USA, 2001.
[20] E. Cariou, R. Marvie, L. Seinturier, L. Duchien, OCL for the specification of model transformation contracts, in: Workshop OCL and Model Driven Engineering of the Seventh Int. Conf. on UML Modeling Languages and Applications, Lisbon, Portugal, 2004, pp. 69–83.
[21] V.M. Chieu, E. Milgrom, M. Frenay, Constructivist learning: operational criteria for cognitive flexibility, in: Proc. IEEE Int. Conf. on Advanced Learning Technologies, ICALT'04, Washington, DC, USA, IEEE Computer Society, 2004, pp. 221–225.
[22] S. Cook, A. Kleepe, R. Mitchell, B. Rumpe, J. Warmer, A. Wills, The Amsterdam Manifesto on OCL, Object Modeling with the OCL, The Rationale Behind the Object Constraint Language, LNCS, vol. 2263, Springer, Berlin, London, UK, 2002.
[23] C.L. Corritore, S. Wiedenbeck, Direction and scope of comprehension-related activities by procedural and object-oriented programmers: an empirical study, in: Proc. Eighth Int. Workshop on Program Comprehension, Washington, DC, USA, 2000, pp. 139–148.
[24] N. Cowan, The magical number 4 in short-term memory: a reconsideration of mental storage capacity, Behavioral and Brain Sciences 24 (1) (2001) 87–114.
[25] J.A. Cruz-Lemus, M. Genero, E. Manso, M. Piattini, Assessing the understandability of UML statechart diagrams with composite states – a family of empirical studies, Empirical Software Engineering 14 (6) (2009) 685–719.
[26] D.P. Darcy, S.A. Slaughter, The structural complexity of software, an experimental test, IEEE Transactions on Software Engineering 31 (11) (2005) 982–995.
[27] B. Demuth, H. Hussmann, Using UML/OCL constraints for relational database design, in: Proc. Second Int. Conf. on the Unified Modeling Language, 1999, pp. 598–613.
[28] B. Demuth, H. Hussmann, S. Loecher, OCL as a specification language for business rules in data base applications, in: M. Gogolla, C. Kobryn (Eds.), UML 2001 – The Unified Modeling Language, Fourth Int. Conf., Toronto, Canada, LNCS, vol. 2185, Springer, 2001, pp. 104–117.
[29] T. Dybå, E. Arisholm, D.I.K. Sjøberg, J.E. Hannay, F. Shull, Are two heads better than one? on the effectiveness of pair programming, IEEE Software 24 (6) (2007) 10–13.
[30] K. El-Emam, Object-Oriented metrics: a review of theory and practice, Technical Report NRC 44190, National Research Council Canada, Institute for Information Technology, 2001.

[31] J. Erickson, K. Siau, Theoretical and practical complexity of modeling methods, Communication of the ACM 50 (8) (2007) 46–51.
[32] E. Fernández-Medina, M. Piattini, Extending OCL for secure database development, in: Int. Conf. on UML, 2004, pp. 380–394.
[33] F. Garcia, M. Serrano, J. Lemus, F. Ruiz, M. Piattini, Managing software process measurement: a metamodel-based approach, Information Sciences 177 (2007) 2570–2586.
[34] M. Genero, M.E. Manso, A. Visaggio, M. Piattini, G. Canfora, Building measure-based prediction models for UML class diagram maintainability, Empirical Software Engineering 12 (5) (2007) 517–549.
[35] M. Genero, M. Piattini, C. Calero, Metrics for software conceptual models, Imperial College Press, 2005.
[36] M. Giese, R Heldal, From informal to formal specification in UML, in: Proc. Seventh Inter. Conf. on Unified Modelling Language, LNCS, vol. 3273, 2004, pp. 197–211.
[37] M. Giese, D. Larsson, Simplifying transformation of OCL constraints, in: L. Briand, C. Williams (Eds.), Proc. Eighth ACM/IEEE Int. Conf. on Model Driven Engineering Languages and Systems, Montego Bay, Jamaica, LNCS, vol. 3713, 2005, pp. 309–323.
[38] G.V. Glass, B. McGaw, M.L. Smith, Meta-analysis in Social Research, Sage Publications, 1981.
[39] C. Glezer, M. Last, E. Nachmany, P. Shoval, Quality and comprehension of UML interaction diagrams – an experimental comparison, Information and Software Technology 47 (10) (2005) 675–692.
[40] R. Gopal, Dynamic program slicing based on dependence relations, in: Proc. Conf. on Software Maintenance, Sorrento, Italy, 1991, pp. 191–200.
[41] M. Grossman, J.E. Aronson, R.V. McCarthy, Does UML make the grade? Insights from the software development community, Information and Software Technology 47 (6) (2005) 383–397.
[42] R. Hähnle, A. Ranta, Connecting OCL with the rest of the world, in: J. Whittle (Ed.), ETAPS Workshop on Transformations in UML, WTUML 2001, Genova, Italy, 2001.
[43] A. Hamie, F. Civello, J. Howse, S. Kent, R. Mitchell, Reflections on the object constraint language, in: Proc. First Int. Workshop on the Unified Modeling Language, LNCS, vol. 1618, 1999, pp. 162–172.
[44] W. Hayes, Research in software engineering: a case for meta-analysis, in: Proc. Sixth IEEE Int. Symposium on Software Metrics, METRICS'99, Boca Raton, USA, 1999, pp. 143–151.
[45] L.V. Hedges, I. Olkin, Statistical Methods for Meta-analysis, Academia Press, 1985.
[46] B. Henderson-Sellers, Object-oriented Metrics: Measures of Complexity, Prentice-Hall, 1996.
[47] M. Höst, B. Regnell, C. Wohlin, Using students as subjects – a comparative study of students and professionals in lead-time impact assessment, in: Proc. Fourth Conf. on Empirical Assessment and Evaluation in Software Engineering, EASE '00, Keele, UK, 2000, pp. 201–214.
[48] ISO/IEC Standard ISO 9126, Software Product Evaluation-quality Characteristics and Guidelines for Their Use, Geneva, 1991.
[49] N. Juristo, A. Moreno, Basics of Software Engineering Experimentation, Kluwer Academic Publishers., 2001.
[50] V. Kampenes, T. Dybå, J.E. Hannay, D.I.K. Sjøberg, A systematic review of effect size in software engineering experiments, Information and Software Technology 49 (11-12) (2007) 1073–1086.
[51] B. Kitchenham, S. Pflegger, N. Fenton, Towards a framework for software measurement validation, IEEE Transactions of Software Engineering 21 (12) (1995) 929–944.
[52] T.A. Klemola, A cognitive model for complexity metrics, in: H.A. Sahraoui, G. Poels, F. Brito e Abreu, H. Zuse (Eds.), Proc. of the Fourth Int. ECOOP Workshop on Quantitative Approaches in Object-Oriented Software Engineering, Sophia Antipolis and Cannes, France, 2000.
[53] C. Knight, M. Munro, Program comprehension experiences with GXL; comprehension for comprehension, in: Proc. 10th Int. Workshop on Program Comprehension, Washington, DC, USA, 2002, pp. 147–156.
[54] L. Kuzniarz, J.L. Sourouille, M. Staron, in: The First Workshop on Quality in Modeling, LNCS, vol. 4364, 2007, pp. 76–79.
[55] O. Laitenberger, K. El-Emam, T. Harbich, An internally replicated quasi-experimental comparison of checklist and perspective-based reading of code documents, IEEE Transactions on Software Engineering 27 (5) (2001) 387–421.
[56] C.F.J. Lange, M.R.V. Chaudron, Interactive views to improve the comprehension of UML models – an experimental validation, in: Proc. 15th IEEE Int. Conf. on Program Comprehension, Alberta, Canada, 2007, pp. 221–230.
[57] O.I. Lindland, G. Sindre, A. Sølvberg, Understanding quality in conceptual modeling, IEEE Software 11 (2) (1994) 42–49.
[58] D.C. Littman, J. Pinto, S. Letovsky, E. Soloway, Mental models and software maintenance, System and Software 7 (4) (1987) 341–355.
[59] D. Massey, Introduction to OCL in Together Products, 2004, <http://conferences.embarcadero.com/article/33187/images/33187/33187_10040314_P.DOC>.
[60] G.A. Miller, The magical number 7, plus or minus two: some limits on our capacity of processing information, The Psychological Review 63 (2) (1956) 81–97.
[61] J. Miller, Applying meta-analytical procedures to software engineering experiments, Systems and Software 54 (1) (2000) 29–39.
[62] J. Miller, F. McDonald., Statistical Analysis of Two Experimental Studies, EFoCS-31-98, University of Strathclyde, 1998.
[63] P. Mohagheghi, V. Dehlen, T. Neple, Definitions and approaches to model quality in model-based software development – a review of literature, Information and Software Technology 51 (12) (2009) 1646–1669.
[64] A. Mohan, N. Gold, Programming style changes in evolving source code, in: Proc. 12th IEEE Int. Workshop on Program Comprehension, Washington, DC, USA, 2004, pp. 236–240.
[65] S. Murphy, S. Tilley, S. Huang, Fourth workshop on graphical documentation: UML style guidelines, in: Proc. of the 22nd Annual Int. Conf. on Design of Communication, New York, NY, USA, 2004, pp. 118–119.
[66] D.A. Norman, Some observations on mental models, in: D. Gentner, A.L. Stevens (Eds.), Mental Models, Lawrence Erlbaum Associates Inc., Hillsdale, USA, 1983, pp. 7–14.
[67] Object Management Group, UML 2.0, OMG Document, 2005, <http://www.omg.org>.
[68] Object Management Group, UML 2.0 OCL, ptc/05-06-06, OCL FTF Report, OMG Document, 2005, <http://www.omg.org>.
[69] M.C. Otero, J.J. Dolado, Evaluation of the comprehension of the dynamic modeling in UML, Information and Software Technology 46 (1) (2004) 35–53.
[70] J. Pardillo, J. Mazón, J. Trujillo, Extending OCL for OLAP querying on conceptual multidimensional models of data warehouses, Information Sciences 180 (5) (2010) 584–601.
[71] N. Pennington, Comprehension strategies in programming, in: E. Soloway, S. Iyengar (Eds.), Empirical Studies of Programmers: Second Workshop, Ablex, Norwood, NJ, USA, 1987, pp. 100–113.
[72] G. Pollice, Formally Speaking: How to Apply OCL, 2004, Available at IBM Page <http://www.ibm.com/developerworks/rational/library/5390.html>.
[73] H.C. Purchase, L. Colpoys, M. McGill, D. Carrington, UML collaboration diagram syntax: an empirical study of comprehension, in: Proc. First Int. Workshop on Visualizing Software for Understanding and Analysis, 2002, pp. 13–22.
[74] V. Rajlich, N. Wilde, The role of concepts in program comprehension, in: Proc. 10th Int. Workshop on Program Comprehension, Washington, DC, USA, 2002, pp. 271–278.
[75] L. Reynoso, A measurement-based approach for assessing the influence of import-coupling on the maintainability of OCL expressions, Ph.D. Thesis, University of Castilla-La Mancha, Spain, 2007.
[76] L. Reynoso, J.A. Cruz-Lemus, M. Genero, M. Piattini, Formal definition of measures for UML statechart diagrams using OCL, in: Proc. 2008 ACM Symposium on Applied Computing, SAC 2008, Fortaleza, Ceara, Brazil, 2008, pp. 846–847.
[77] L. Reynoso, M. Genero, M. Piattini, Assessing the impact of coupling on the understandability and modifiability of OCL expressions within UML/OCL combined models, in: Proc. 11th IEEE Int. Software Metrics Symposium, 2005, p.14.
[78] L. Reynoso, M. Genero, M. Piattini, OCL2: using OCL in the formal definition of OCL expression measures, in: Proc. First Workshop on Quality in Modeling QIM Co-located with the ACM/IEEE Ninth Int. Conf. on Model Driven Engineering Languages and Systems, Genova, Italy, 2006, pp. 95–115.

[79] L. Reynoso, M. Genero, M. Piattini, Using verbal protocols to assess the influence of import-coupling on the comprehensibility of OCL expressions, in: Proc. Sixth IEEE Int. Conf. on Cognitive Informatics, Lake Tahoe, CA, USA, 2007, pp. 440–449.
[80] L. Reynoso, M. Genero, M. Piattini, E. Manso, Does object coupling really affect the understandability and modifiability of OCL expressions? in: Proc. 21st ACM Symposium on Applied Computing, Dijon, France, 2006, pp. 1721–1727.
[81] L. Reynoso, E. Rolón Aguilar, M. Genero, F. García, Francisco Ruiz, M. Piattini, Formal definition of measures for BPMN models, in: A. Abran, R. Braungarten, R. Dumke, J.J. Cuadrado-Gallego, J. Brunekreef (Eds.), Proc. Int. Conf. Software Process and Product Measurement, IWSM' 09, LNCS, vol. 5891, Springer-Verlag, 2009, pp. 285–306.
[82] R. Rosenthal, Meta-analytic Procedures for Social Research, Sage Publications, Newbury Park, CA, 1991.
[83] R. Rosenthal, Parametric measures of effect size, in: H. Cooper, L.V. Hedges (Eds.), The Handbook of Research Synthesis, Russell Sage Foundation, New York, 1994, pp. 231–244.
[84] D.I.K. Sjøberg, J.E. Hannay, O. Hansen, V.B. Kampenes, A. Karahasanovic, N. Liborg, A.C. Rekdal, A survey of controlled experiments in software engineering, IEEE Transactions on Software Engineering 31 (9) (2005) 733–753.
[85] SPSS 15.0. Syntax Reference Guide, Chicago, SPSS Inc., 2006.
[86] M.A. Storey, Theories, methods and tools in program comprehension: past, present and future, in: Proc. 13th Int. Workshop on Program Comprehension, IWPC '05, Washington, DC, USA, 2005, pp. 181–191.
[87] F.S.C. Tseng, C. Chen, Enriching the class diagram concepts to capture natural language semantics for database access, Data and Knowledge Engineering 67 (1) (2008) 1–29.
[88] J. Verelst, The influence of the level of abstraction on the evolvability of conceptual models of information systems, in: Proc. Third Int. Symposium on Empirical Software Engineering, ISESE 2004, Redondo Beach, CA, USA, 2004, pp. 17–26.
[89] R. Vinter, M. Loomes, R. Kornbrot, Applying software metrics to formal specifications: a cognitive approach, in: Proc. of the Fifth Int. Symposium on Software Metrics, METRICS'98, IEEE Computer Society, Washington, DC, USA, 1998, pp. 216–223.
[90] J. Warmer, A. Kleppe, The Object Constraint Language: Getting Your Models Ready for MDA, second ed., Addison-Wesley, MA, USA, 2003.
[91] C. Wohlin, P. Runeson, M. Höst, M. Ohlson, B. Regnell, A. Wesslén, Experimentation in Software Engineering: An Introduction, Kluwer Academic Publishers, Norwell, MA, USA, 2000.
[92] S. Yusuf, H. Kagdi, J.I. Maletic, Assessing the comprehension of UML class diagrams via eye tracking, in: Proc. 15th IEEE Int. Conf. on Program Comprehension, ICPC'07, Los Alamitos, CA, USA, 2007, pp. 113–122.