Double-Blind Scores of an Object-Oriented Modeling Survey

by

Raymond Sze Chun Yiu

An essay
presented to the University of Waterloo
in fulfilment of the
requirement for the degree of
Master of Mathematics
in
Computer Science

Waterloo, Ontario, Canada, 1996

I hereby declare that I am the sole author of this thesis.

I authorize the University of Waterloo to lend this thesis to other institutions or individuals for the purpose of scholarly research.

I further authorize the University of Waterloo to reproduce this thesis by photocopying or by other means, in total or in part, at the request of other institutions or individuals for the purpose of scholarly research.

The University of Waterloo requires the signatures of all persons using or photocopying this thesis. Please sign below, and give address and date.

Abstract

Over the past decade, object-oriented (OO) technology has gained wide-spread acceptance in a variety of computer science fields, including software engineering, programming language design, database systems, user interfaces, operating systems and telecommunications . To support it, a new way of thinking about problems using models organized around real-world concepts, called object-oriented modeling, has emerged. Since object-oriented models are useful for understanding problems, designing and maintaining OO programs and OO databases, it is important to understand how OO concepts are used by typical object-oriented modelers.

This essay presents basic empirical observations of the OO modeling process. They help us to understand OO modelers' behavior patterns and provide a foundation for better user interface design. The observations are important because there are no empirical results prior to our work. In addition, this essay describes how we successfully used an important assessment technique – Double-Blind Scoring – for the first time in studies of programmer behavior. This technique allows a researcher to obtain quantitative data from open-ended survey questions, making it possible to investigate wide-ranging user impressions with a higher degree of objectivity.

Acknowledgements

I wish to thank my supervisor, Dr. William Cowan, who provided expert and invaluable guidance in codifying my often vague ideas, and suggested new directions of exploration througout the work. Without his encouragement, tireless readings of the working drafts and constructive criticisms, this essay would not have existed.

I would also like to acknowledge and thank to Mr. Glenn Paulley for his efforts and contributions in scoring the questionnaires for this study.

Besides, I would like to thank Ada Cheung, Peter Varlagas, Joanathan Wong and Martin Van Bommel for their help and encouragement throughout my graduate studies.

Last, but no means least, I would like to thank my parents for their encouragement and support of my graduate studies.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Over the past decade, object-oriented (OO) technology has gained wide-spread acceptance in a variety of computer science fields, including software engineering, programming language design, database systems, user interfaces, operating systems and telecommunications. To support it, a new way of thinking about problems using models organized around real-world concepts, called object-oriented modeling, has emerged. Object-oriented models are useful for understanding problems, and for designing and maintaining OO programs and OO databases. No matter how powerful the object-oriented database system (e.g. ObjectStore [6] or O_2 [2]) or how versatile the object-oriented programming language (e.g. C++ [7] or Smalltalk [11]), good OO modeling practice is always essential to exploit these tools and systems fully. Specifically, to design user interfaces that facilitate user performance on OO modeling, we need to understand how users conceive OO systems and how they accomplish the modeling task. To provide some insight into this area, a study on object-oriented modeling was launched. Its objective was to discover and understand how OO concepts are used by typical object-oriented modelers.

The study has two main results. First, it provides basic empirical observations on the OO modeling process. This is important because prior to this study, no empirical results existed. Results like these ones will help us to understand OO modelers' behavior patterns and provide a foundation for better user interface design.

Second, we have successfully used an important assessment technique – Double-Blind Scoring – for the first time in studies of programmer behavior. This technique allows a researcher to obtain quantitative data from open-ended survey questions, making it possible to investigate wide-ranging user impressions with a higher degree of objectivity.

The paper is organized as follows. A brief introduction to object-oriented terminology and basic object-oriented modeling practices is given in Chapter 2. In Chapter 3, the approach and methodology used in collecting and scoring the data for the study are presented. In particular, we focus on the Double-Blind Scoring technique, used as a method for scoring our results objectively. Survey results with statistical analysis are presented in Chapter 4. Based on the results obtained from the study, suggestions for designing user interfaces for OO modeling are given in, the concluding chapter, Chapter 5.

Chapter 2

Object-Oriented Modeling

2.1 Terminology

Object-oriented (OO) modeling is a new way of thinking about problems using *models* based on object-oriented concepts. The fundamental modeling primitive in an object-oriented model is the "object". An *object* is a combination of *data* and *behavior*, usually representing a real-world entity. An OO model or OO schema, like the one shown in Figure 2.1, represents the static structure of the objects in a system. This object-oriented model (or object model for short) is an OO schema of a database containing entities from a windowing system presented graphically using the Object Modeling Technique (OMT) graphical notation [10].

2.1.1 Class and Object

Each rectangular box in the object model represents a class. A class describes a collection of objects all of which have the same properties (ie. both data structure

and behavior). Each object is said to be an instance of its class. The first part of the rectangular box records the name of the class. The object model, shown in Figure 2.1, consists of fifteen classes, namely *Window*, *Scrolling Window*, *Canvas*, *Panel*, *Text Window*, *Scrolling Canvas*, *Shape*, *Panel Item*, *Line*, *Eillpse*, *Polygon*, *Point*, *Choice Item*, *Button*, and *Text Item* class.

The second part of the rectangular box shows the list of attribute(s) of the corresponding class. An attribute is a data value held by the objects in a class. For example, the *Shape* class has attributes *color* and *line width*; the *Ellipse* class has attributes *x*, *y*, *a*, *b*, *fill color* and *fill pattern*. Consider a blue *Shape* object with line width of 2, it will have the value *blue* for the attribute *color* and *2* for the attribute *line width*.

The third part of the rectangular box indicates the list of operation(s) (or behavoir(s)) of the corresponding class. An operation is a function that may be applied to or by objects in a class. For example, *display*, *undisplay*, *raise*, and *lower* are operations on the class *Window*; *insert* and *delete* are operations on the class *Text Window*. All objects in a class share the same operations.

2.1.2 Class Relationships

Each line in the schema represents a relationship established among objects and classes. Those relationships can be classified as generalization, aggregation and association. The cardinality of a relationship specifies how many instances of one class may relate to a single instance of a related class. A relationship can be one-to-one, one-to-many or many-to-many. A line without specific cardinality indicates a one-to-one relationship. In general, a cardinality expression is written next to the

end of the line, for example, "0+" in the relationship between the class *Panel* and *Panel Item* indicates a panel object has zero or more panel item object(s).

Generalization

Generalization (or sometimes called "is-a") is a relationship between a class and one or more refined (or inherited) versions of it. The class being refined is called the superclass of the refined class and each refined version is called the subclass of the class being refined. That is, the class *Window* is a superclass of the classes *Canvas*, *Scrolling Window* and *Panel*. The class *Canvas* is a subclass of the class *Window* and there is a generalization relationship between the class *Window* and the refined classes *Scrolling Window*, *Canvas* and *Panel*. The refinement process involves subclassing of one or more existing class(es), and refining (ie. defining additional and/or overriding inherited) properties (attributes and/or operations) inherited from its superclass(es). The OMT graphical notation for generalization is a line connecting a superclass to its subclasses with the triangle pointing upward to the superclass. Through generalization of classes, a hierarchy of class structure is formed and for simplicity, this generalization structure is often called a "generalization hierarchy".

"Inheritance" is often confused with "generalization". Generalization refers to the "is-a" relationship among classes, while inheritance is the mechanism of sharing attributes and operations using generalization. Through inheritance, a subclass inherits the properties (ie. attributes and operations) of its superclass. The subclass can define additional properties or override inherited ones on top of those inherited from its parent. For example, the class *Canvas* inherits all the attributes (*x1*, *y1*, *x2* and *y2*) and operations (*display*, *undisplay*, *raise* and *lower*) from its parent

Window. And the *Canvas* class further defines additional properties (attributes: *cx1*, *cy1*, *cx2* and *cy2*; and operations: *add-element*, and *delete-element*) of its own. Inherited properties are not graphically shown in the schema unless they are over-riden by the inherited class. As we can see, the *draw* operation is shown both in the class *Line* and the corresponding superclass *Shape*. This indicates the class *Line* overrides the operation *draw* inherited from *Shape* by defining a new version of *draw*. In general, a class can inherit from more than one existing classes, which is called "multiple inheritance". The class *Scrolling Canvas* is an example of multiple inheritance since it inherits from two superclasses *Scrolling Window* and *Canvas*.

Aggregation

Aggregation (or sometimes called "a-part-of" or "composed-of") is a relationship established by relating a composite class to a component class. For example, a triangle is defined by three vertices; a polygon is defined by three or more vertices and each vertex is an instance of the class *Point*. Hence, the class *Point* is said to be the component class of the composite class *Polygon*. The OMT graphical notation for aggregation is a line connecting a composite class to its component classes with a diamond placed at the composite class end of the line. The cardinality expresssion "3+" indicates that a polygon object is composed of three or more vertices (point objects). Through aggregation of classes, hierarchies of aggregation structure are formed and for simiplity, these aggregation structures are often called the "aggregation hierarchy".

Aggregation relates two object instances, one of them being a part of the other. Hence, when an composite object is deleted, its component objects are deleted as well. For example, when a polygon object is deleted, the corresponding component

objects, vertices, are deleted.

Association

Association relates two or more independent classes. Associations relating two independent classes are called binary associations; associations relating three independent classes are called tenary associations.

To model a canvas, which has zero or more graphical objects, an association relationship, called *Has-elements*, is created by connecting the class *Canvas* and *Shape* with a line in the schema. The *Has-elements* association relationship is estalished because line(s), ellipse(s), or polygon(s) are graphical objects and the class *Shape* is the superclass of all three classes: *Line*, *Ellipse*, and *Polygon*. The cardinality expression "0+" at the end of the *Has-elements* association indicates that this is a "zero or more" association.

Association is inherently bi-directional. Hence, if a canvas has a line as its element, that line is also called an element of that canvas. People often confuse the terms "aggregation" and "association". The main distinction between aggregation and association is that when an association instance is deleted, all participating objects, which are independent of one another, continue to exist. However, deleting a composite object results in deleting its component objects.

2.2 Object-Oriented Modeling Activities

An object-oriented (OO) modeling process is a process of thinking about problems using models based on object-oriented concepts. There are four kinds of basic OO

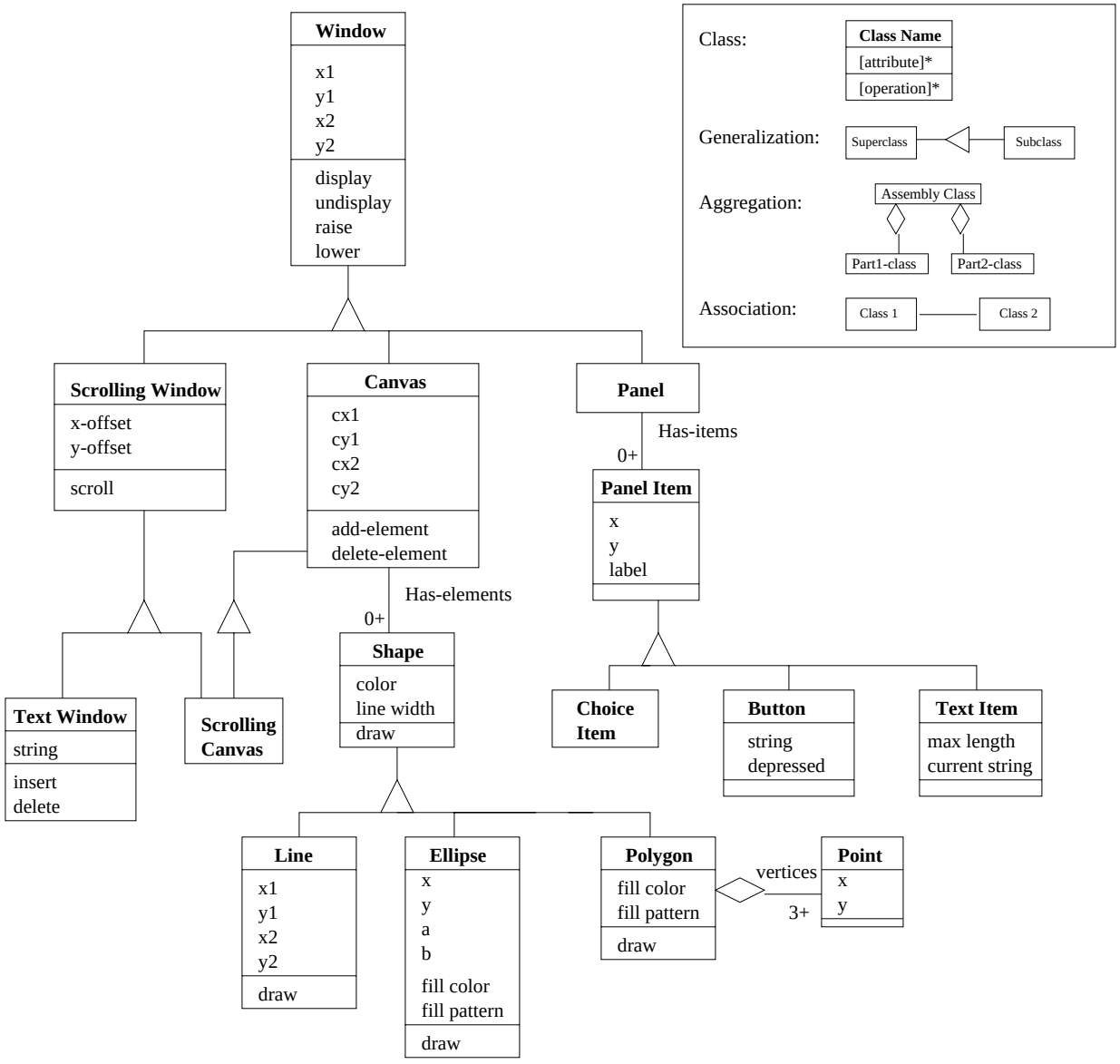

Figure 2.1: Object-Oriented Schema of Windowing System

activities which OO modelers have to go through during the modeling process. These four OO modeling activities are refinement, abstraction, composition and decomposition.

1. Refinement

   Using the generalization relationship, OO modelers can refine an existing class through the mechanism of "inheritance". The refinement process involves subclassing one or more existing class(es), and refining (ie. defining additional and/or overriding inherited) properties (attributes and/or operations) inherited from the superclass(es).

   For example, when OO modelers want to place a new class *Arc* with an attribute *arc angle* and an operation *display* into the schema shown in Figure 2.1, they place the new *Arc* class in the schema by subclassing the existing class *Ellipse*, adding the attribute *arc angle* to the refined class *Arc* and overriding the inherited operation *draw* by defining a new operation *draw*. The modified portion of the schema after the refinement process is shown in Figure 2.2.

2. Abstraction

   Sometimes, when OO modelers identify commonalities in attributes and operations among existing classes, they make use of the inheritance mechanism and abstract those classes to enhance reuse. The abstraction process usually creates abstract superclass(es), a class that has no direct instances, but with subclasses that have direct instances, by introducing an extra layer into the generalization hierarchy.

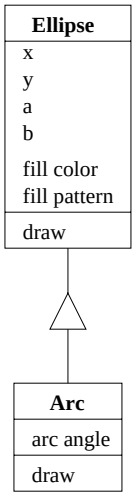

Figure 2.2: Illustration on Refinement Process

For example, when OO modelers want to create abstract classes *Closed Shape* and *Open Shape* from the existing class *Line*, *Ellipse* and *Polygon* in the schema shown in Figure 2.1, they create two new classes *Open Shape* and *Closed Shape*, both inherited from their superclass *Shape*, such that the class *Line* is inherited from the new class *Open Shape* and the classes *Ellipse* and *Polygon* are inherited from the new class *Closed Shape*. In this case, the common attributes *fill color* and *fill pattern* are abstracted as attributes of the new class *Closed Shape*. The modified portion of the schema after the abstraction process is shown in Figure 2.3.

3. Composition

   Apart from generalization, a new class can be created by composing one or more existing class(es). Composition enhances the re-use of existing classes through aggregation.

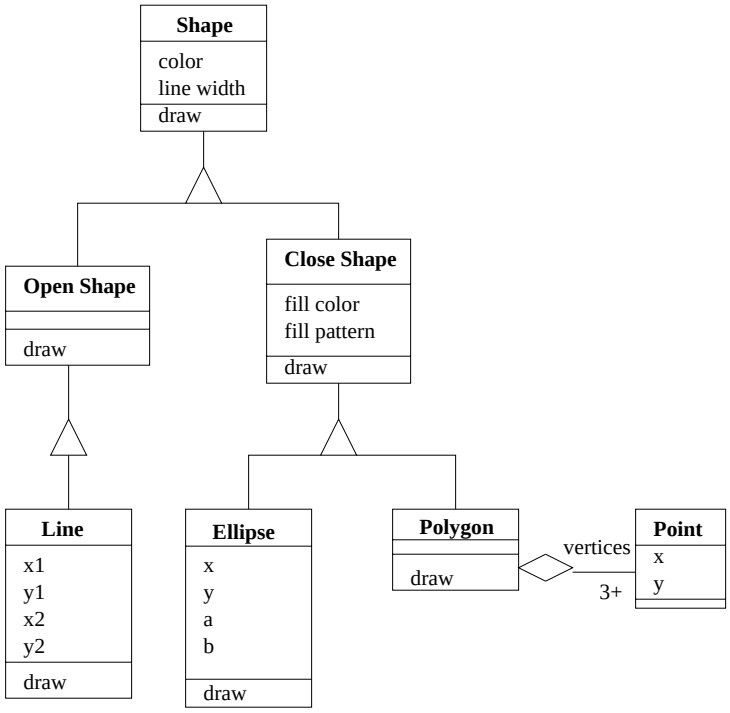

Figure 2.3: Illustration on Abstraction Process

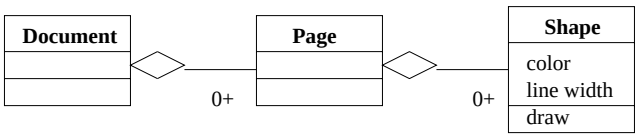

Figure 2.4: Illustration on Composition Process

For example, when OO modelers want to model the fact that 'documents are composed of pages; and pages are composed of graphical objects' starting from the schema shown in Figure 2.1, they introduce two new classes *Document* and *Page*, related to one another through the aggregation relationship, and add all necessary attributes and operations accordingly. Realizing that the class *Shape* provides all necessary properties and relationships for the class *Graphical Objects*, they would then relate the classes *Page* and *Shape* through aggregation to maximize the benefits of reuse. The modified portion of the schema after the composition process is shown in Figure 2.4.

4. Decomposition

   In order to enhance reuse, OO modelers may want to decompose an existing class by partitioning the attributes and operations of a class into new, simpler and more reusable component classes.

   For example, when OO modelers want to partition the attributes *x1*, *y1*, *x2* and *y2* of the class *Line* to form a more reusable component class based on the schema shown in Figure 2.1, they decompose the class *Line* to make it a composite class composed of the class *Point* with cardinality "2". The

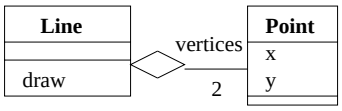

Figure 2.5: Illustration on Decomposition Process

modified portion of the schema after decomposition is shown in Figure 2.5.

Chapter 3

Survey Methodology

This chapter describes and justifies our methodology for conducting an objective and yet unbiased survey using Double Blind Scoring. By way of introduction, a brief description of survey research is given as well.

3.1 Survey Research

Experiment, survey and existing-statistics are three fundamental kinds of research technique. Experiments usually address a well-focused question, providing evidence that supports, extends, or refutes a principle or theory. Surveys usually focus on a well-defined subject with the research conducted to describe the subject in a systematic way. Existing-statistics research, on the other hand, focuses on previously researched problems and is conducted by reorganizing or combining previously conducted surveys or reports in a new way. Since object-oriented modeling is a recently established modeling methodology and no empirical results existed prior to our study, survey seems to be an obvious choice of methodology.

Survey research asks subjects questions and records their answers. Questions can be asked through a questionnaire or through an interview. In questionnaires, the questions are presented in written format, with subjects are required to write down their answers. Interviews, on the other hand, require a one-on-one verbal interaction between interviewer and subject.

3.1.1 Classifications of Survey Methods

Psychological research uses several classifications of survey methods. Three classification schemes are discussed in this section. More ways of classifying survey methods can be found in [1].

One classification divides survey methods according to content or process, the categories being *cognitive* vs. *affective*. Cognitive surveys measure the effects of mental capabilities on past and future behavior, while affective surveys measure interests, attitudes, values, motives and other non-cognitive properties.

Another classification is based on survey administration, the categories being *individual* vs. *group*. An individual survey is administered to one subject at a time, whereas a group survey is administered to many subjects simultaneously.

In terms of the method of scoring, survey methods can be classified as *objective* vs. *non-objective*. An objective survey has fixed and objective scoring standards, whereas non-objective surveys can produce different results when evaluated by different scorers.

Depending on the different purposes or objectives of survey studies, survey creators can define and categorize their surveys differently.

3.1.2 Importance of Open-Ended Survey Questions

Survey questions can be classified as open-ended or close-ended [3, 1]. An open-ended question requires the survey subject to write and construct the answer, whereas a close-ended question requires the survey subject to select the answer from a list of alternatives. There are several sub-types of close-ended question: true-false, matching and multiple-choice. A true-false question gives a binary choice of true-false or yes-no to the subject. A matching question involves a set of response options to be matched to a set of stimulus options. A multiple-choice question involves picking an answer from a list of alternatives. Guidelines on how to construct good survey questions are found in [1].

Close-ended questions have many advantages. They can be scored easily and objectively. Different scorers produce the same scores for close-ended questions because the list of alternatives for each close-ended question given to every subject is the same. These benefits are not present for open-ended questions. Scoring open-ended questions is time-consuming and subjective because different subjects give different levels of detail in their answers and choose different ways of expressing themselves. Thus, time is required to categorize the answers. Futhermore, open-ended questions can result in irrelevent or confused answers to questions. Hence, scoring open-ended questions is not as straightforward as close-ended questions.

Although close-ended surveys are often superior to open-ended ones, it is not always the case. Indeed, open-ended questions play an irreplaceable role in survey

research. First of all, unanticipated findings can be discovered through open-ended questions because answers are not limited to the alternatives created by the survey creator. Open-ended questions can yield valuable insight into what subjects are thinking and how they naturally understand their world. Close-ended questions can be used effectively only when the dimensions of the variables are well-defined. Hence, if the researcher is unfamiliar with the domain of interests or the domain of interest is not well-defined yet, open-ended surveys are an appropriate choice. Moreover, there are quite a number of disadvantages in using close-ended questions. Subjects may be frustrated when their desired answer is not available or when they are forced to make choices they would not make in the real world. Close-ended questions often force subjects to give simplistic responses to complex issues. Also, misinterpretations and random guessing of close-ended questions can go unnoticed.

Since open-ended surveys are so important and valuable in survey research, it is worthwhile to investigate ways and techniques to make subjective scoring more objective. For the next few sections of this essay, we discuss how to do this.

3.2 Overview of Double-Blind Technique

Expectancy effects [8, 12], which may be intentional or unintentional, occur when a subject or a researher produces responses that are affected by knowing the study hypothesis. Those expectations, if not eliminated, can bias the result of a study. The *Double-Blind* Technique [3] is an objective assessment technique that can eliminate expectancy effects from both subjects and reseachers.

Double-Blind is commonly used in drug research to investigate the "true" effect

of drugs on patients. Consider a drug experiment where one group of patients receives the drug under investigation and the other group does not. Now suppose that the drug group shows an improvement. We do not know whether the improvement was caused by the properties of the drug or by the patients' expectations about the effect of drug – often called a placebo effect. In order to distinguish the treatment effect from the placebo effect, a placebo group is added. Patients in the placebo group receive a "placebo" treatment in the form of pill or injection that contains harmless substances (e.g. a "sugar pill") but not the drug given to subjects in the treatment group. Then, if the improvement results from the active properties of the drug, the patients in the treatment group should show greater improvement than those in the placebo group. If the placebo group improves as much as the treatment group, the improvement is a placebo effect. If the patients are unaware of whether a "placebo" or "real" treatment is being used, we called the experiment *Single-Blind*.

Apart from the subjects' expectation, a researcher – the doctor in our example – who is aware of the purpose of the study may develop expectations about how subjects should respond. In that case, bias occurs when that researcher assesses (or scores) the subjects' behavoir. When neither the subject nor the scorer knows whether the "placebo" or "real" treatment is being used, we called the experiment *Double-Blind*. To accomplish this, the scorer must be different from the researcher.

3.3 Data Collection

This section discuss how we designed the survey and collect the data for our study. In particular, details on the pilot studies and our methodology for "blind"-ing the

subjects are presented.

3.3.1 First Blind – Survey Subjects

Our study was launched as a cognitive (c.f. section 3.1.1) survey on object-oriented modeling. Over the past two months, fifteen subjects, both graduate students and faculty members, from research groups in the Computer Science and Electrical & Computer Engineering Departments at the University of Waterloo were invited to participate the study on an individual basis. Their OO modeling or OO programming experience ranged from 1 year to 10 years. Their areas of specialty include computer graphics, database management, software engineering, symbolic computation, scientific computation, system design, computer engineering, programming language and distributed networking.

We did not formally post any announcement (or intention) of the survey nor did we encourage participants to volunteer. By choosing subjects – as opposed to advertising for them – we hoped to minimize sampling bias [4] . We believe that subjects behave most naturally when they are unaware of the variable that is being manipulated. If subjects know what we are studying, they may try to confirm the hypothesis, or they may try to look good by behaving in the most socially acceptable way. Hence, we looked for potential participants through informal verbal requests and set up convenient times with each of them individually to carry out the survey. In other words, none of the subjects knew the hypotheses of the study at the time that they participated in the survey. Hence, "blinded" (c.f. section 3.2), free-of-expectancy and unbiased answers are obtained from subjects. This "blindedness" consititutes the first "blind" of our *Double-Blind Scoring* methodology.

3.3.2 Survey Medium

Survey research can be done by questionnaire or interview (c.f. secton 3.1). We chose not to conduct the survey through interviews because results from interviews are easily biased by inconsistent survey conditions. For example, an interviewer may unknowingly emphasize certain words when reading questions to some subjects but not to others; or the interviewer may smile more when interacting with some subjects than others. Such effects can build up demand characteristics on subjects that involuntarily bias the results of the study. (A "demand characteristic" [3] is a cue that informs the subject how he or she is expected to behave. That is, an accidental emphasis of a word or smile from an interviewer can function to reinforce or discourage responses, and do so despite the best effort of the interviewer.) Hence, to make sure that same survey conditions are applied to all the subjects, we employed a questionnaire in our study.

Since we had no prior knowledge or results on object-oriented modeling and the dimensions of the variables under investigation were not well-defined, we conducted the survey using an open-ended questionnaire. As mentioned earlier, scoring open-ended questions is a subjective (c.f. section 3.1.2) and expectancy-biased (c.f. section 3.2) process. The "Double-Blind Scoring" technique is used to cope with this scoring problem. In section 3.5, we describe in detail how our "Double-Blind Scoring" methodology eliminates expectancy effects by "blind"-ing the scorer and thus enables an objective scoring mechanism by setting up scoring and criteria sheets.

The questionnaire, shown in Appendix B, is divided into two parts. The first part contains task-oriented case studies which require the subject to perform basic object-oriented modeling activities (c.f. section 2.2), such as refinement and

composition. Subjects are asked to record all the steps they go through in finding solutions to the tasks. As can be seen from the briefing sheet in Appendix A and the questionnaire in Appendix B, the survey contains a few guidelines to help subjects in structuring their answers; otherwise their responses are completely free. Those few guidelines, as can be seen, placed no restrictions on what answers could be given by the subjects. The second part of the questionnaire contains general questions about how the subjects perform OO modeling. They are expected to recall what they did in the first part while formulating their answers.

Although no time limit was enforced on any subject, the time expected to complete the questionnaire was 30-45 minutes. On average, it was completed in 40 minutes.

3.3.3 Pilot Runs

Pilot runs are small scale trials conducted prior to the actual study. They allow testing and refinement of the procedures used in the actual study.

A pilot of the survey was carried out. The pilot questionnaire was answered individually by six subjects – five graduate students and one faculty member – from the Computer Science Department at the University of Waterloo. These trial subjects were not used in the actual runs.

One major defect of the pilot questionnaire was its length. A lengthy questionnaire produces fatigue [3] in the subjects, negatively affecting their responses. On average, the pilot runs were completed in 70 minutes. The pilot subjects spent most of their time on the four case studies and lost interest in answering the second part of the questionnaire. The four case studies in the pilot questionnaire consist

of the four basic OO modeling activities (c.f. section 2.2), namely refinement, abstraction, composition and decomposition. As a remedy to the problem, the final questionnaire was shortened from four case studies to two case studies, covering only refinement and composition.

The pilot questionnaire also turned out to be too open-ended. The pilot subjects tended to provide solutions to the case studies instead of describing the sequence of steps that they went through in arriving at the solutions. Hence, the final questionnaire contained a few sub-questions to help subjects in structuring their answers. These few sub-questions, as can be seen, placed no restrictions on the content of answers given by the subjects. A briefing sheet, as shown in Appendix A, was also used in the actual study to (1) ensure that consistent information was given to all the subjects, and (2) let subjects understand that our survey is a cognitive study (c.f. section 3.1.1), recording their behavior and not quizzing them on their OO modeling ability.

3.3.4 Pilot Results

Pilot questionnaires were interpreted informally by the survey creator. These results played an important role in our "Double-Blind Scoring" methodology. We used the pilot results not only to refine the questionnaire, and also to form hypotheses for our study. Such hypotheses are essential for producing material for the scorer to use.

From the pilot data, we noticed that object-oriented (OO) modelers reduce the size of an existing OO schema by defining portion(s) that they perceive as relevant to the modeling task. How relevance is defined depends on OO experience, and on understandings of the application domain. Subjects then concentrate on only the

relevant portion(s) of the schema, ignoring the irrelevant portion(s).

Locating relevant class(es) and identifying target class(es) seem to be two significant modeling activities that OO modelers go through when browsing or exploring the organization and content of the schema in order to perform a modeling task. That is, OO modelers first try to determine the existence and location of relevant class(es) and then they try to discriminate those relevant class(es) and identify those that must be used in doing the modeling task.

When browsing the schema for relevant class(es), the OO modelers scan the generalization hierarchy 'top-down' without attending to the contents of class(es) being browsed, using the class name to evaluate relevance. On the other hand, when browsing the schema for target class(es), they search the generalization hierarchy 'bottom-up'. Class name, class relationships and class properties are used successively as discriminators for determining the validity and correctness of relevant class(es).

OO modelers are aware of the existence of the vocabulary problem [5] when doing the questionnaire. They use name similarity to locate relevant classes and identify target classes if they fail to locate the exact class name(s) specified in the task specification. Moreover, the target discrimination process works equally well whether class names are similar or identical.

3.4 Hypotheses

Results from the six pilots are generalized and turned into eleven hypotheses in five categories: relevancy, schema browsing, locating classes, identifying classes and similarity criterion, defined in more detail below.

3.4.1 Relevancy

In this essay, the word "relevancy" refers to 'a fitness for or appropriateness to the situation or occasion'. (Relevancy is determined intuitively in this study; formalizing the definition of relevancy is an important matter for future research. Double-blind scoring makes it possible to use an intuitive definition without introducing artifacts.) For example, when OO modelers want to place a new class *Arc* with an attribute *arc angle* and an operation *display* into the schema shown in Figure 2.1, the portion of the schema that includes the classes *Shape*, *Line* and *Ellipse* is said to be relevant to the modeling task. The reasons for that are (1) arc is a special kind of ellipse, and (2) according to the task specification, *Arc* could be a type of *Shape* with respect to the given OO schema. Contrarily, the portion that includes the classes *Panel Item*, *Choice Item*, *Button* and *Text Item* is said to be irrelevant with respect to the modeling task. The reasons for that are (1) arc is not a panel item, (2) arc is a not component of any panel item, and (3) there is no associative relationships between an arc and a panel item with respect to the given schema and modeling task.

The following hypothesis is based on these considerations.

H1.1. Subjects define portion(s) of relevancy with respect to each modeling task. Irrelevant portion(s) of the schema are identified and ignored throughout the modeling process.

3.4.2 Schema Browsing

To understand what schema browsing is, we need to define what browsing is. In this essay, the word "browsing" refers to 'a goal-directed process of exploring the organization and content of an information space'. Hence, the term "schema browsing" refers to the goal-directed process of exploring the organization and content of a schema.

In the survey, we specifically distinguish two time-frames, referring to them as "at the beginning" and "at the end". "At the beginning" refers to the period of time immediately after the subject looks at the modeling task and starts to browse the schema; while "at the end" refers to the period of time immediately before the subject modifies the schema.

Two other distinctions that are useful for describing schema browsing are *scan* vs. *search*, and *relevant* vs. *target*. Scan-oriented schema browsing describes browsing the schema to detect the existence of some class(es) without paying attention to the details of the classes being browsed. Contrarily, search-oriented schema browsing describes browsing the schema carefully, paying attention to details of the class(es) browsed in an effort to find or discover differences among those class(es). Relevant classes are classes that are appropriate to consider at the beginning of a modeling task; whereas the target classes are classes that are actually used in a modeling task. (Relevancy is determined inituitively for this study; formalizing the defini-

tion of relevancy is an important matter for future research.)

The following hypotheses are based on these distinctions.

H2.1. Subjects browse the schema based on the generalization hierarchy.

H2.2. At the beginning, subjects browse the schema top-down.

H2.3. At the end, subjects browse the schema bottom-up.

H2.4. When browsing the schema top-down, subjects perform scan-oriented browsing to locate (c.f. section 3.4.3) relevant class(es).

H2.5. When browsing the schema bottom-up, the subjects perform search-oriented browsing to identify (c.f. section 3.4.4) target class(es) among the relevant class(es).

3.4.3 Locating Classes

The term "locating classes" refers to the determination of the existence and location of class(es).

The following hypothesis is based on this consideration.

H3.1. While browsing, subjects rely on the class name as the means of locating relevant (c.f. section 3.4.2) class(es).

3.4.4 Identifying Classes

The term "identifying classes" refers to the process of finding class(es) that match a set of class(es) based on one or more discrimination dimensions. Class name,

class relationships and class properties are the three discriminators used to identify target class(es).

The following hypotheses are based on this consideration.

H4.1. Subjects rely on the class name as the first means for identifying target (c.f. section 3.4.2) class(es).

H4.2. If the class name is insufficient (c.f. section 3.4.2), subjects use class relationships as the second means for identification.

3.4.5 Similarity Criterion

In this essay, the word "similar" refers to 'having comparable characteristics in common'. For example, in the schema specified in Figure 2.1, we consider the class *Graphical objects* and the class *Shape* similar because of the similarity in their names.

The following hypotheses are based on similarity.

H5.1. Subjects use name similarity to locate relevant classes if they fail to locate the exact class names specified in the task specification.

H5.2. Subjects use name similarity to identify target classes. The target-class identification process (c.f. hypothesis H4.1 and H4.2) works equally well whether class names are similar or identical.

3.5 Data Scoring

In this section, we describe the details of our "Double-Blind Scoring" methodology: first, how to eliminate the expectancy effect by "blind"-ing the scorer, and second, how to create an objective scoring mechanism by setting up scoring and criteria sheets based on the hypotheses derived from the pilot study.

3.5.1 Second Blind – Scorer

After refining the questionnaire, fifteen new subjects were invited individually to fill-in the improved questionnaire. These fifteen surveys consititute the results of our study. None of these subjects participated in the pilot study.

With a survey conducted as an open-ended questionnaire, scoring these questions by the survey creator can easily produce subjective (c.f. section 3.1.2) and expectancy-biased (c.f. section 3.2) results. The problem of subjective materials scored by the survey creator is that the survey creator can easily select interpretations that are unintentionally biased in favour of the study hypotheses. If so, the results are not reliable. To eliminate expectancy effects and to ensure that scores are objectively assessed, scoring should be done by a scorer who has no knowledge of the study hypotheses. That is, a scorer who is "blind"-ed from the study hypotheses is required to score the open-ended questionnaire. The "blind"-ing of the scorer consititutes the second "blind" of our "Double-Blind Scoring" methodology.

To produce stable results, it is important to transform the subjective questionnaire answers into a form that is as objective as possible. Doing so is the job of the scorer. Thus, for the scorer, we created a close-ended scoring sheet, shown in

Appendix C, based on the hypotheses derived from the pilot study. The close-ended scoring sheet allows the scorer to transform the subjective questionnaire answers into an objective results. Furthermore, to enhance the consistency and lack of bias the scoring, a criteria sheet, shown in Appendix D, was created to document criteria while scoring questions in the scoring sheet and defining the terminology used in the scoring sheet.

Remember that because the scorer was blind to the hypotheses of the study, he could be unrestricted in responding to the questionnaire answers when deciding how they fit the categories of the scoring sheet.

A Ph.D. candidate from the department of Computer Science at the University of Waterloo, who has background knowledge on both psychological testing and object-oriented schema design, was invited to be the scorer of our study. The only interaction between the survey creator and the scorer during the study was limited in (1) the invitation to be the scorer of our study, (2) the delivery of the fifteen questionnaires, the scoring sheets (one per questionnaire) and the criteria sheet to the scorer; and (3) a one-time explanation of the scoring and criteria sheets. That is, the scorer did not know the study hypotheses when scoring the questionnaires and did not involve the questionnaire creator in interpreting any answers.

3.5.2 Scoring Sheet and Hypotheses

The use of the close-ended scoring sheet and criteria sheet, as shown in Appendix C and Appendix D respectively, enabled us to turn our scoring from a subjective scoring scheme into an objective, consistent, and unbiased scoring scheme.

In this section, we discuss how each close-ended question in the scoring sheet relate to our hypotheses, listed in the section 3.4.

Relevancy

Questions [Scoring S.1] – [Scoring S.6] were used to determine the amount of time that the subjects spent on the three different portion(s) of the schema on each of the two modeling tasks. For task 2(a) in the questionnaire (shown in Appendix B), the relevant portion of the schema is defined as Window, Canvas, Scrolling Canvas, Shape, Line, Ellipse, Polygon and Point. For task 2(b) in our questionnaire, the relevant portions of the schema are defined as (1) Window, Canvas, Scrolling Canvas, Shape, Line, Ellipse, Polygon and Point; and (2) Window, Scrolling Window, Text Window, Scrolling Canvas and Canvas. These results are used to support hypothesis H1.1.

Schema Browsing

Question [Scoring S.7] was used to determine whether the subjects follow any hierarchy to guide browsing in the schema. If so, on which hierarchy did they base their browsing? This score is a generalized score of the two modeling tasks. The result is used to support hypothesis H2.1.

Questions [Scoring S.8 – Scoring S.9] were used to determine which browsing strategy, top-down, bottom-up or random, was used by subjects when browsing the schema "at the beginning" on each of the two modeling tasks. The result is used to support hypothesis H2.2.

Questions [Scoring S.12 – S.13] were used to determine which browsing strategy, top-down, bottom-up or random, was used by subjects when browsing the schema "at the end" on each of the two modeling tasks. The result is used to support hypothesis H2.3.

Question [Scoring S.11] was used to determine whether the subjects looked at any detail of class(es) being browsed "at the beginning". This score is a generalized score of the two modeling tasks. The result is used to support hypothesis H2.4.

Question [Scoring S.14] was used to determine whether the subjects looked at any detail of class(es) being browsed "at the end". This score is a generalized score of the two modeling tasks. The result is used to support hypothesis H2.5.

Locating Classes

Question [Scoring S.10] was used to determine what information the subjects relied on when locating relevant class(es). Possibilities are class name, class property, class relationship or combinations thereof. This score is a generalized score of the two modeling tasks. The result is used to support hypothesis H3.1.

Identifying Classes

Question [Scoring S.15] was used to determine what information the subjects relied on as the first means in identifying target class(es) among the three discriminators: class name, class property, class relationship. This score is a generalized score of the two modeling tasks. The result is used to support hypothesis H4.1.

Question [Scoring S.16 – Scoring S.17] was used to determine what information the subjects relied on if the first means failed adequately to identify target class(es) among the three discriminators: class name, class property, class relationship. These scores are generalized scores of the two modeling tasks. The result is used to support hypothesis H4.2.

Question [Scoring S.18] was used to determine what information subjects used if all three discriminators failed adequately to identify target class(es). This question is included to cope with unanticipated answers.

Similarity Criterion

Question [Scoring S.19] was used to determine if subjects use name similarity if they fail to locate the exact class name(s) specified in the task specification. The result is used to support hypothesis H5.1.

Question [Scoring S.20] was used to determine if there is any difference in the target-class identification process when class(es) do not have the exact name as specified in the task specification. The result is used to support hypothesis H5.2.

3.6 Summary

In this chapter, we presented our "Double-Blind Scoring" methodology. This technique allows a researcher to obtain quantitative data from open-ended survey questions, making it possible to investigate user impressions with a higher degree of objectivity.

Here, we summarize the steps involved in our "Double-Blind Scoring" methodology.

1. Create an open-ended questionnaire for the survey study.
2. Do not announce the intention of the survey study. We want all the participanting subjects to be "blind"-ed from the purpose of the survey.
3. Choose subjects for study so as to minimize sampling bias.
4. Carry out pilot studies where all pilot subjects are "blind"-ed with respect to the intention(s) and hypotheses (if any) of the survey. Since "blind" subjects are free from expectancy, unbiased results are obtained. "Blind"-ing the subjects consititutes the first "blind" of the methodology.
5. Interprete the pilot studies informally. That is, no detailed and formal analysis is required.
6. Refine the survey based on the pilot results.
7. Create and/or refine hypotheses based on the pilot result.
8. Carry out the actual study with the improved questionnaire. "Blind"-ing the subjects from the study hypotheses is essential to obtain unbiased data. This consititutes the first "blind" of our methodology.
9. Create a close-ended scoring sheet and a criteria sheet based on the hypotheses derived from the pilots. The scoring sheet enables us to turn the subjective scoring scheme into an objective scoring scheme, while the criteria sheet further promotes the objectiveness, consistent, and unbiased nature of the scoring.

10. Use a scorer to objectively score the actual study according to survey-specific scoring and criteria sheets. The scorer cannot be the survey creator himself; otherwise the score obtained may be biased by expectancy effects. The scorer needs to be "blind"-ed from the study hypotheses to ensure the scores are objectively assessed. This consititutes the second "blind" of our methodology.

11. Summarize and analyze the "Double-Blind" scores formally and interpret the results.

Chapter 4

Survey Results

This chapter presents the survey results, as scored using the "Double-Blind Scoring" methodology. Statistical analyses and interpretations are presented to analyse and explain the validity of our study hypotheses.

4.1 Theory from Statistics

4.1.1 Confidence Interval

When estimating the population mean with a sample mean, confidence intervals provide a systematic way to generate intervals that contain the unknown population mean with probability equal to the confidence level [4]. With a 95% confidence interval using the two-tailed t-distribution at $\alpha = 0.05$, the confidence interval, CI, for the population mean μ is

$$CI_{.95} = \overline{X} \pm t_{.05/2}\frac{s}{\sqrt{N}}$$

where $\overline{X}$ is the sample mean, s is the sample standard deviation, N is the sample size.

To justify a confidence level of 95%, the computed interval contains the true population mean 95% of the time. If the confidence interval generated is very narrow, the population mean is estimated with high precision. If the confidence interval generated is wide, the population mean is estimated with low precision.

4.1.2 Testing Hypotheses

Testing a Hypothesis is an analysis procedure used to weigh the amount of evidence in the sample data against a null hypothesis, usually that a population mean μ has a specified value. The probability of rejecting a null hypothesis is called the $\alpha - level$. In general, for a given $\alpha - level$, the corresponding confidence level is $100(1-\alpha)\%$.

Testing a sample mean when the standard deviation σ of the population is not known, we use the one sample *t*-statistic [13]. To test $H_0 : \mu = \mu_0$ against, for example, $H_a : \mu \neq \mu 0$, we use the *t*-statistic computed on a *t*-distribution with N-1 degree of freedom

$$t = \frac{\overline{X} - \mu_0}{\frac{s}{\sqrt{N}}}$$

where $\overline{X}$ is the sample mean, μ_0 is the value of the population mean specified in the null hypothesis, s is the sample standard deviation, and N is the sample size.

If $\mid t \mid > t_{n-1,1-\alpha/2}$, there exists evidence for rejecting H_0 in favour of H_a at the α significance level. However, if $\mid t \mid < t_{n-1,1-\alpha/2}$, H_0 is not rejected.

To test the null hypothesis $H_0 : \mu = \mu_0$, we use the two-tailed test. To test the null hypothesis $H_0 : \mu > \mu_0$ or $\mu < \mu_0$, we use the one-tailed test.

4.1.3 Analysis of Variance

Analysis of Variance (ANOVA) [13], or F-statistic, is an analysis procedure used to compare means of several groups of data based on the assumption that each of the groups being compared has the same underlying variance. The F-statistic tests the null hypothesis $\mu_1 = \mu_2 = \ldots = \mu_p$ against the alternative that at least one mean is not equal to the others.

The F-statistic is the ratio of the treatment mean square, MS_{treat}, to the error mean square, MS_{error}. The treatment mean square is calculated by dividing the treatment sum of squares, SS_{treat}, by the number of degrees of freedom between treatments, df_{treat}. The error mean square is calculated by dividing the error sum of squares, SS_{error}, by the error degrees of freedom, df_{error}.

The term, SS_{treat}, is the sum of squared differences of the group means from the mean of all the measurements.

$$SS_{treat} = \sum_{i=1}^{groups} N_i(\overline{X_i} - \overline{\overline{X}})^2$$

where N_i is the number of observations in the i^{th} group, $\overline{X_i}$ is the mean of the i^{th} group and $\overline{\overline{X}}$ is the mean of all the measurements.

The term, SS_{error}, is a pooled sum of individual sums of squares deviations from means found within each group.

$$SS_{error} = \sum_i \sum_j (X_{ij} - \overline{X_i})^2$$

where j goes from 1 to the number of observations in the group, and i goes from 1 to the number of groups.

The total sum of squares, SS_{total}, is the sum of squared deviations of each measurement from the overall mean.

$$SS_{total} = \sum_i \sum_j (X_{ij} - \overline{\overline{X}})^2$$

where $SS_{total} = SS_{treat} + SS_{error}$.

SS_{total} is the sum of N squared deviations around one point – the grand mean. The fact that we have taken deviations around this one (estimated) point has cost us 1 *df*, leaving us with $df_{total} = N - 1$. SS_{treat} is the sum of k deviations around one point (again the grand mean), and again we have lost 1 *df* in estimating this point, leaving us with $df_{error} = k - 1$. SS_{error} represents N deviations about k points (the k treatment means), losing us $k - 1$ *df* and leaving $df_{error} = N - k$ because the mean of the means is the grand mean.

When the null hypothesis is true, both of the mean square values estimates the population variance σ^2, so the *F*-statistic is close to 1.0. The *F*-statistic has a value

Source	df	Sum of Squares (SS)	Mean Square (MS)
Factor A	$\|A\| - 1$	$SS_{treatment}$ for A (SSA)	$\frac{SSA}{\|A\| - 1}$
Factor B	$\|B\| - 1$	$SS_{treatment}$ for B (SSB)	$\frac{SSB}{\|B\| - 1}$
Error	$N - \|A\| - \|B\| + 1$	SS_{Error} (SSE)	$\frac{SSE}{N - \|A\| - \|B\| + 1}$
Total	$N - 1$	SSA + SSB + SSE	

Table 4.1: Two-Factor ANOVA Calculation Tableau

much larger than 1.0 when there is strong evidence for rejecting the null hypothesis. The probability that the null hypothesis is true owing to random flutuations is labelled as Prob by Data Desk [13], and is shown beside the *F*-statistic.

In one-factor ANOVA, the groups whose means are compared are usually thought of as different categories of a single factor or treatment. Multi-factor ANOVA introduced more factors, each specified by its own variable. Typically, the factors might affect the response either independently or jointly through an interaction between the factors. In ANOVA, interaction refers to the combined effect of two or more factors. Interaction assesses whether the response variable, as measured for one of the factors, changes at different levels of the other factors. The calculation tableau of a two-factor ANOVA without interactions is shown in Figure 4.1.

4.2 Analysis and Interpretation

We used Data Desk [13] to analyze the fifteen scored surveys, testing hypotheses in each of the five categories: relevance, schema browsing, locating classes, identifying classes and similarity criterion. Tables of scored results are in Appendix E. Statistical analyses and interpretations are presented in the following sections.

4.2.1 Relevancy

Recall that "relevancy" is defined as 'fitness for or appropriateness to the situation or occasion'. In this category is hypothesis H1.1. Results, statistical tests and interpretations are presented below.

Hypothesis H1.1

Subjects define portion(s) of relevancy with respect to each modeling task. Irrelevant portion(s) of the schema are identified and ignored in the modeling process.

Result

According to Table 4.2, subjects spent 97% and 95.7% of their time on average dealing with the relevant portion of the schema for modeling tasks 2(a) and 2(b) respectively. Only 3% and 4.3% of their time was spent on the irrelevant portions of the schema.

Analysis

From the results above,

Task / Relevancy	Task 1	Task2
Relevant Portion(s)	97%	95.7%
Irrelevant Portion(s)	3%	4.3%

Table 4.2: Summary of Scores for Hypothesis H1.1

1. Is there a difference in the percentage of time spent on the schema by different relevancy factors?

   Using the Two-Factor ANOVA, as shown in Table 4.3, the F-ratio is 3490. We conclude that there is a difference in the amount of time spent on relevant and irrelevant parts of the schema.

2. Is there a difference in the percentage of time spent on the schema by different tasks?

   No, by definition, task average is identically 50%.

3. Is there a difference in the percentage of time spent on the relevant part of the schema between the tasks?

   Using the Two-Factor ANOVA, as shown in Table 4.3, the F-ratio is 0.7. Hence, there is no evidence of interaction between relevancy factors and tasks.

   Taking the average of the two tasks, subjects spend 96.3% of their time dealing with the relevant portion of the schema and 3.6% of their time

Analysis of Variance For **TimeSpentOnSchema**

Source	df	Sum of Squares	Mean Square	F-ratio	Prob
Rly	1	128807	128807	3490.2	0.0000
Tsk	1	0	0	0	0
Rly*Tsk	1	26.6667	26.6667	0.72258	0.3989
Error	56	2066.67	36.9048		
Total	59	130900			

Table 4.3: Two-Factor ANOVA for Hypothesis H1.1

t-Tests

TimeSpentOnIrrelevancy: Test Ho: =0 vs Ha: ›0
Sample mean = 3.6667 t-statistic=3.343 with 29 d.f.
Reject Ho at alpha = 0.05 Prob <= 0.0023

Table 4.4: Irrelevancy: Test of Hypothesis on Population Mean Equals 0%

dealing with the irrelevant portion. With 95% confidence, the confidence interval for the population mean, μ, of the time spent on relevant portion(s) of the schema is

$$CI_{.95} = 96.3\% \pm 2.3\%$$

4. Could the population mean of the percentage of time spent on the irrelevant portion(s) of the schema be 0%?
   As shown in Table 4.4, the t-statistic equals 3.343. This provides strong evidence for rejecting the null hypothesis that the time spent is zero.

Interpretation

From the analysis above,

1. OO modelers spend almost all their time on relevant portions of the schema while doing modeling tasks.
2. OO modelers do not spend 100% of their time on the relevant portions of the schema because they need time to discriminate the relevant portion(s) from irrelevant ones. Once identified, irrelevant portions are ignored.
3. There is evidence of a difference between tasks.

Hence, given an O-O schema, modelers do not try to look at the whole schema while tackling modeling tasks. Instead, they define portion(s) that are relevant to the modeling task. Irrelevant portion(s) of the schema are identified and ignored.

4.2.2 Schema Browsing Strategy

Recall that schema browsing is a goal-directed process of exploring the organization and content of a schema. In the survey, we distinguish two time intervals, referring to them as "at the beginning" and "at the end". "At the beginning" refers to the period of time immediately after the subject starts a modeling task by browsing the schema; "at the end" refers to the period of time immediately before the subject modifies the schema.

Locating relevant class(es) and identifying target class(es) seem to be the two significant modeling phrases that OO modelers go through when browsing the schema in order to perform a modeling task. That is, OO modelers first try to determine

the existence and location of relevant class(es), and then they try to discriminate among the relevant class(es) to identify those that are to be used in the modeling task.

Two other distinctions that are useful for describing schema browsing are *scan* vs. *search*, and *relevant* vs. *target*. Scan-oriented schema browsing describes browsing the schema to detect the existence of class(es) without paying attention to the details of the classes being browsed. Contrarily, search-oriented schema browsing describes browsing the schema carefully, paying attention to details of the class(es) being browsed in an effort to find or discover differences among those class(es). Relevant classes are determined by scanning; target classes by searching among relevant classes.

Under this category, our hypotheses are H2.1 to H2.5. Results, statistical tests and interpretations are presented below.

Hypothesis H2.1

Subjects browse the schema following the generalization hierarchy.

Result

According to Table 4.5, 93.4% of the subjects browsed the schema following a hierarchical structure. 86.7% of the subjects used only the generalization hierarchy; 6.7% of the subjects used both the generalization and aggregation hierarchy; and 6.7% of the subjects did not follow any hierarchy. These scores are generalized scores of the two modeling tasks. The results are not differ-

	Percentage of subjects
Generalization	86.7%
Aggregation	0%
Generalization & Aggregation	6.7%
None	6.7%
Other	0%

Table 4.5: Summary of Scores for Hypothesis H2.1

entiated by task because the scoring procedure produced a single value that is the average of the two tasks.

Analysis

From the results above,

1. Is there a difference in the percentage of subjects by different hierarchy factors?
   Using the One-Factor ANOVA, as shown in Table 4.6, the F-ratio is 40.8. We conclude that there is a difference by hierarchy.

2. Could the population mean of the percentage of subjects not using hierarchy when browsing the schema be 0%?
   As shown in Table 4.7, the *t*-statistic equals 1.0 with probability of 0.1671 arising by chance. It is not possible to reject the null hypothesis that no subjects without using a hierarchy.

Analysis of Variance For **BrowsingHierarchy**

Source	**df**	**Sum of Squares**	**Mean Square**	**F-ratio**	**Prob**
Hry	4	84000.0	21000.0	40.833	0.0000
Error	70	36000.0	514.286		
Total	74	120000			

Table 4.6: One-Factor ANOVA for Hypothesis H2.1

t-Tests

NotBasedOnHierarchy: Test Ho:μ=0 vs Ha:μ>0
Sample mean = 6.6667 t-statistic=1.000 with 14 d.f.
Fail to reject Ho at alpha = 0.05 Prob <= 0.1671

Table 4.7: Schema Browsing: Test of Hypothesis on Population Mean Equals 0% in not using Hierachy when Browsing

Interpretation

From the analysis above,

1. Most OO modelers follow only the generalization hierarchy when browsing the schema.
2. There may be some use other hierarchies and some non-hierarchical browsing.

Hence, although there exists two types of hierarchy – generalization and aggregation – in an OO schema, most modelers use the generalization hierarchy when browsing.

Hypothesis H2.2

At the beginning, subjects browse the schema top-down.

Result

According to Table 4.8, 60% and 66.7% of the subjects did top-down schema browsing at the beginning on modeling tasks 2(a) and 2(b). 20% of the subjects did bottom-up schema browsing at the beginning for both modeling tasks, while 20% and 13.3% of the subjects did random schema browsing.

Analysis

From the results above,

1. Is there a difference by strategy at the beginning?
   Using the Two-Factor ANOVA, as shown in Table 4.9, the corresponding

Strategy \ Task	Task 1	Task2
Top-down	60%	66.7%
Bottom-up	20%	20%
Random	20%	13.3%

Table 4.8: Summary of Scores for Hypothesis H2.2

F-ratio is 10.8. We conclude that there is a difference in the percentage of subjects using each strategy at the beginning.

2. Is there a difference by task at the beginning?

   No, by definition, task average is identically 50%.

3. Is there a task-strategy interaction at the beginning?

   Using the Two-Factor ANOVA, as shown in Table 4.9, the F-ratio is 0.18. Hence, there is no evidence of interaction between task and browsing strategy.

   Taking the average of the two tasks, 63.3% of the subjects did top-down, 20% of the subjects did bottom-up and 16.7% of the subjects did random schema browsing at the beginning.

4. Is it possible that the percentage of subjects browsing bottom-up at the beginning is 0%?

   As shown in Table 4.10, the *t*-statistic equals 2.693. The null hypothesis

Analysis of Variance For **BrowsingAtTheBeginning**

Source	df	Sum of Squares	Mean Square	F-ratio	Prob
Tsk	1	0	0	0	0
Sty	2	40666.7	20333.3	10.765	0.0001
Tsk*Sty	2	666.667	333.333	0.17647	0.8385
Error	84	158667	1888.89		
Total	89	200000			

Table 4.9: Two-Factor ANOVA for Hypothesis H2.2

t-Tests

BottomUpAtTheBeginning: Test Ho:μ=0 vs Ha:μ>0
Sample mean = 20 t-statistic=2.693 with 29 d.f.
Reject Ho at alpha = 0.05 Prob <= 0.0058

Table 4.10: Browsing Bottom-up At The Beginning: Test of Hypothesis on Population Mean be 0%

that the percentage is 0% is rejected at the 1% level.

5. Is it possible that the percentage of subjects browsing randomly at the beginning is 0%?

   As shown in Table 4.11, the *t*-statistic equals 2.408. The null hypothesis that the percentage is 0% is rejected at the 2.5% level.

Interpretation

From the analysis above,

1. Most OO modelers browse the schema top-down at the beginning when they are tackling a modeling task.

t-Tests

RandomAtBeginning: Test Ho:μ=0 vs Ha:μ>0
Sample mean = 16.667 t-statistic=2.408 with 29 d.f.
Reject Ho at alpha = 0.05 Prob <= 0.0113

Table 4.11: Browsing Random At The Beginning: Test of Hypothesis on Population Mean be 0%

2. Some bottom-up or random schema browsing exists.

3. There is no evidence of a difference between tasks.

Hence, at the beginning of any modeling task, most OO modelers browse the schema top-down.

Hypothesis H2.3

At the end, subjects browse the schema bottom-up.

Result

According to Table 4.12, 80% and 6.7% of the subjects did bottom-up schema browsing at the end on modeling tasks 2(a) and 2(b) respectively. 6.7% and 46.7% of the subjects did top-down schema browsing at the end for tasks 2(a) and 2(b); while 13.3% and 46.7% of the subjects did random schema browsing.

Analysis

From the results above,

Strategy \ Task	Task 1	Task2
Top-down	6.7%	46.7%
Bottom-up	80%	6.7%
Random	13.3%	46.7%

Table 4.12: Summary of Scores for Hypothesis H2.3

1. Is there a difference by browsing strategy at the end?
   Using the Two-Factor ANOVA, as shown in Table 4.13, the corresponding F-ratio is 1.5. We cannot conclude that subjects favour one strategy or another at the end.

2. Is there a difference by task at the end?
   No, by definition, task average is identically 50%.

3. Is there a task-strategy interaction at the end?
   Using the Two-Factor ANOVA, as shown in Table 4.13, the F-ratio is 18.9. Hence, there is strong evidence of interaction between tasks and browsing strategies. We need to consider the two tasks separately.

4. Is there a difference by strategy at the end on modeling task 2(a)?
   Using the One-Factor ANOVA, as shown in Table 4.14, the F-ratio is 20.4. We conclude that subjects favour bottom-up browsing at the end on modeling task 2(a).

Analysis of Variance For **BrowsingAtTheEnd**

Source	df	Sum of Squares	Mean Square	F-ratio	Prob
Tsk	1	0	0	0	0
Sty	2	4666.67	2333.33	1.4554	0.2391
Tsk*Sty	2	60666.7	30333.3	18.921	0.0000
Error	84	134667	1603.17		
Total	89	200000			

Table 4.13: Two-Factor ANOVA for Hypothesis H2.3

Analysis of Variance For **BrowseAtTheEndForT2(a)**

Source	df	Sum of Squares	Mean Square	F-ratio	Prob
Sty	2	49333.3	24666.7	20.447	0.0000
Error	42	50666.7	1206.35		
Total	44	100000			

Table 4.14: One-Factor ANOVA for Hypothesis H2.3 for task 2(a)

t-Tests

TopDownAtTheEndFor2(a): Test Ho:μ=0 vs Ha:μ>0
Sample mean = 6.6667 t-statistic=1.000 with 14 d.f.
Fail to reject Ho at alpha = 0.05 Prob <= 0.1671

Table 4.15: Browsing Top-down At The End for Task 2(a): Test of Hypothesis on Population Mean Equals 0%

t-Tests

RandomAtTheEndFor2(a): Test Ho:μ=0 vs Ha:μ>0
Sample mean = 13.333 t-statistic=1.468 with 14 d.f.
Fail to reject Ho at alpha = 0.05 Prob <= 0.0822

Table 4.16: Browsing Random At The End for Task 2(a): Test of Hypothesis on Population Mean Equals 0%

5. Is it possible that no subjects browse top-down at the end for task 2(a)?
   As shown in Table 4.15, the *t*-statistic equals 1.0. It is not possible to reject the null hypothesis that no subjects browse top-down at the end of task 2(a).

6. Is it possible that no subjects browse randomly at the end for task 2(a)?
   As shown in Table 4.16, the *t*-statistic equals 1.5. There is only weak evidence for rejecting the null hypothesis that no subjects browse randomly at the end of task 2(a).

7. Do subjects exhibit a preference for any browsing strategy at the end of task 2(b)?
   Using the One-Factor ANOVA, as shown in Table 4.17, the F-ratio is

Analysis of Variance For **BrowseAtTheEndForT2(b)**

Source	df	Sum of Squares	Mean Square	F-ratio	Prob
Sty	2	16000	8000	4.0000	0.0257
Error	42	84000.0	2000.00		
Total	44	100000			

Table 4.17: One-Factor ANOVA for Hypothesis H2.3 for task 2(b)

t-Tests

BottomUpAtTheEndFor2(b): Test Ho:μ=0 vs Ha:μ>0
Sample mean = 6.6667 t-statistic=1.000 with 14 d.f.
Fail to reject Ho at alpha = 0.05 Prob <= 0.1671

Table 4.18: Browsing Bottom-up At The End for Task 2(b): Test of Hypothesis on Population Mean Equals 0%

4.0. We conclude that subjects differentiate among browsing strategies at the end of task 2(b).

8. Is it possible that no subjects browse the schema bottom-up at the end for task 2(b)?

   As shown in Table 4.18, the *t*-statistic equals 1.0. There is no evidence to reject the null hypothesis that the population mean is 0%.

Interpretation

From the analysis above,

1. OO modelers exhibit preferences among browsing strategies at the end of modeling tasks 2(a) and 2(b).

2. Most OO modelers browse the schema bottom-up at the end of modeling task 2(a).
3. Few OO modelers browse the schema bottom-up at the end of modeling task 2(b).

Hence, depending on whether the modeling task is 2(a) or 2(b), OO modelers use different schema browsing strategy at the end. Most of the subjects did bottom-up browsing at the end when tackling task 2(a), while few did bottom-up browsing at the end for task 2(b).

It is necessary to investigate more on this matter in order to generalize this result. The discrepancy may due to a carry-over effect [3]. Carry-over effects occur when the effects of one treatement are still present when the next treatment is given. A carry-over effect might occur because the given schema is small, because the same portion of the schema is used for both tasks, and/or because the subjects get familiar with the schema while doing task 2(a). Hence, while doing task 2(b), they could use their memory of task 2(a) while finding relevant and target classes.

Alternatively, the discrepancy may occur because of differences between the modeling tasks. At this point, we cannot make any general statement on this matter. More scorers, larger schemas and different schemas for different tasks are necessary to obtain more rigorous answers.

Look at Details \ Phrase	At the Beginning	At the End
Yes	33.3%	100%
No	66.7%	0%

Table 4.19: Summary of Scores for Hypotheses H2.4 and H2.5

Hypothesis H2.4 and Hypothesis H2.5

When browsing the schema top-down, subjects do scan-oriented browsing to locate (c.f. section 3.4.3) relevant class(es). Conversely, when browsing the schema bottom-up, the subjects do search-oriented browsing to identify (c.f. section 3.4.4) target class(es) among the relevant class(es).

Result

According to Table 4.19, 33.3% of the subjects looked into details of classes when they were browsing at the beginning; whereas 100% of the subjects looked at details of classes at the end. These scores are generalized scores of the two modeling tasks.

Analysis

From the results above,

Analysis of Variance For **LookAtDetails**

Source	df	Sum of Squares	Mean Square	F-ratio	Prob
TM	1	0	0	0	0
Dtl	1	16666.7	16666.7	14	0.0004
TM*Dtl	1	66666.7	66666.7	56	0.0000
Error	56	66666.7	1190.48		
Total	59	150000			

Table 4.20: Two-Factor ANOVA for Hypotheses H2.4 and H2.5

1. Is there a significant main effect showing that subjects look into details more than not?

   Using the Two-Factor ANOVA, as shown in Table 4.20, the F-ratio is 14.0. We conclude that there is a significant effect depending on inspection of details.

2. Are details more important at the end than at the beginning?

   Using the Two-Factor ANOVA, as shown in Table 4.20, the F-ratio of the interaction term is 56. Hence, there is strong evidence that details are more important at the end.

Interpretation

From the analysis above,

1. When browsing a schema, the population mean for looking at details at the beginning is different from that at the end. That is, during the two time intervals, at the beginning and at the end, OO modelers behave differently with regards to look at details of the classes being browsed.

2. At the beginning, few OO modelers look at details of the classes being browsed. However, at the end, almost all OO modelers look at details of the class(es) being browsed.

Recall our definition of scan and search browsing – At the beginning, few OO modelers do search-oriented browsing as they do not focus on class details. Conversely, all OO modelers do search-oriented browsing at the end as they focus on class details.

From the results in section 4.2.2, we know that most OO modelers browse the schema top-down at the beginning, at least for task 2(a), but not at the end. Hence, we conclude that when browsing the schema top-down, OO modelers perform scan-oriented browsing on the schema trying to locate (c.f. section 3.4.3) the relevant class(es). At the end, OO modelers perform search-oriented browsing on the schema trying to identify (c.f. section 3.4.4) the target class(es) among the relevant class(es).

4.2.3 Locating Classes

Recall that the term "locating classes" is used to refer to determining of the existence and location of class(es). Under this category is hypothesis H3.1. Results, statistical tests and interpretations are presented below.

Hypothesis H3.1

Following the browsing hierarchy, subjects rely on the class name for locating relevant (c.f. section 3.4.2) class(es).

	Percentage of subjects
Class Property (CP)	0%
Class Name (CN)	13.3%
Class Relationship (CR)	0%
CN & CP	13.3%
CN & CR	73.3%
Other	0%

Table 4.21: Summary of Scores for Hypothesis H3.1

Result

According to Table 4.21, 100% of the subjects used class name to locate relevant classes. Of that 100%, 73.3% of the subjects used both class name and class relationship, 13.3% of the subjects used only class name and 13.3% of the subjects used both class name and class property to locate the relevant classes. No subjects used only class property nor only class relationship to locate classes. These scores are generalized scores of the two modeling tasks.

Analysis

From the results above,

1. Is the difference between locating strategies statistically significant?

   Using the One-Factor ANOVA, as shown in Table 4.22, the F-ratio is

Analysis of Variance For **LocatingClasses**

Source	df	Sum of Squares	Mean Square	F-ratio	Prob
Mn	5	61000.0	12200.0	16.012	0.0000
Error	84	64000.0	761.905		
Total	89	125000			

Table 4.22: One-Factor ANOVA for Hypothesis H3.1

t-Tests

ClassNameOnlyForLocation: Test Ho:μ=0 vs Ha:μ>0
Sample mean = 13.333 t-statistic=1.468 with 14 d.f.
Fail to reject Ho at alpha = 0.05 Prob <= 0.0822

Table 4.23: Locating Classes: Test of Hypothesis on Population Mean Equals 0% using Class Name Only

16.0. We conclude that there is a difference in the percentage of subjects using different categories of locators.

2. Is it possible that no subjects use class name only while locating relevant classes?

   As shown in Table 4.23, the *t*-statistic equals 1.5. There is weak evidence for rejecting the null hypothesis that no subjects use class name only.

3. Is it possible that no subjects use both class name and class property while locating relevant classes?

   As shown in Table 4.24, the *t*-statistic equals 1.5. There is weak evidence for rejecting the null hypothesis that no subjects use class name and class property.

t-Tests

ClassName&ClassPropertiesForLocation: Test Ho:μ=0 vs Ha:μ>0
Sample mean = 13.333 t-statistic=1.468 with 14 d.f.
Fail to reject Ho at alpha = 0.05 Prob <= 0.0822

Table 4.24: Locating Classes: Test of Hypothesis on Population Mean Equals 0% using Class Name and Class Property

Interpretation

From the analysis above,

1. Class name is used by all the subjects for locating relevant classes in a schema.
2. Most OO modelers use the class name together with class relationship to locate relevant classes in a schema.
3. A few modelers use class name only, or class name together with class property for locating the relevant classes.

From the section 4.2.2, we know that most OO modelers use the generalization hierarchy as a skeleton when browsing the schema. Therefore, most OO modelers use class name and follow the generalization hierarchy while determining the existence and location of relevant classes in a schema.

4.2.4 Identifying Classes

Recall that the term "identifying classes" refers to the process of finding class(es) that match a set of class(es) based on one or more discrimination dimensions. Class

name, class relationships and class properties are the three discriminators used to identify target classes.

In this category, our hypotheses are H4.1 to H4.2. Results, statistical tests and interpretations are presented below.

Hypothesis H4.1

Subjects rely on the class name as the first means for identifying target (c.f. section 3.4.2) class(es).

Result

According to Table 4.25, 100% of the subjects used class name as the first means to identify target class(es). This score is a generalized score of the two modeling tasks.

Analysis

From the results above,

1. Are there significant differences among the fifteen different sequenced discriminators?

   Using the One-Factor ANOVA, as shown in Table 4.26, the F-ratio is 11.7. We conclude that there are significant differences among the fifteen different sequenced discriminators.

2. Are there significant differences among the five different sequenced discriminators that use class name as the first means for target identification?

Sequence of discriminators used	Percentage of subjects
CN-No-No	0%
CN-CR-No	6.7%
CN-CR-CP	60%
CN-CP-No	0%
CN-CP-CR	33.3%
CP-No-No	0%
CP-CN-No	0%
CP-CN-CR	0%
CP-CR-No	0%
CP-CR-CN	0%
CR-No-No	0%
CR-CN-No	0%
CR-CN-CP	0%
CR-CP-No	0%
CR-CP-CN	0%

Table 4.25: Summary of Scores for Hypotheses H4.1 to H4.2

Analysis of Variance For **IdentifyClasses**

Source	**df**	**Sum of Squares**	**Mean Square**	**F-ratio**	**Prob**
Mn	14	61333.3	4380.95	11.695	0
Error	210	78666.7	374.603		
Total	224	140000			

Table 4.26: One-Factor ANOVA for Hypotheses H4.1 to H4.2 for the Fifteen Sequenced Discriminators

Analysis of Variance For **IdentifyingClassUsingClassNameFirst**

Source	df	Sum of Squares	Mean Square	F-ratio	Prob
Mn	4	41333.3	10333.3	9.1949	0.0000
Error	70	78666.7	1123.81		
Total	74	120000			

Table 4.27: One-Factor ANOVA for Hypotheses H4.1 to H4.2 for the Five Sequenced Discriminators that use class name as the first means for target identification

Using the One-Factor ANOVA, as shown in Table 4.27, the corresponding F-ratio is 9.2. We conclude that there are significant differences among the five different sequenced discriminators that use class name as the first means for target identification.

Interpretation

The results and analysis show,

1. The fifteen sequenced discriminators are not all the same.
2. The five sequenced discriminators that use class name as the first means for identification are not all the same.
3. All OO modelers in the sample relied on class name as the first means when identifying target classes.

Hence, we conclude that almost all OO modelers use class name as the first means for identifying target classes.

Hypothesis H4.2

If the class name is insufficient (c.f. section 3.4.2), subjects use class relationships as the second means for identification.

Result

According to Table 4.25, 66.7% of the subjects used class relationships, and 33.3% of the subjects used class properties as the second means of identifying target class(es). These scores are generalized scores of the two modeling tasks.

Analysis

From the results above,

1. Is there significant difference between the two sequenced discriminators that use class name as the first means and class relationship as the second means for target identification?
   Using the One-Factor ANOVA, as shown in Table 4.28, the F-ratio is 13.2. We conclude that there is a significant difference between the two sequenced discriminators that use class name as the first means and class relationship as the second means for target identification.
2. Is there a significant difference between the two sequenced discriminators that use class name as the first means and class property as the second means for target identification?
   Using the One-Factor ANOVA, as shown in Table 4.29, the F-ratio is 7.0. We conclude that there is a significant difference between the two

Analysis of Variance For **IdentifyUsingCRasSecondMeans**

Source	df	Sum of Squares	Mean Square	F-ratio	Prob
Mn	1	21333.3	21333.3	13.176	0.0011
Error	28	45333.3	1619.05		
Total	29	66666.7			

Table 4.28: One-Factor ANOVA for Hypotheses H4.1 to H4.2 for the Two Sequenced Discriminators that use class name and class relationship as the first and second means respectively for target identification

Analysis of Variance For **IdentifyUsingCPasSecondMeans**

Source	df	Sum of Squares	Mean Square	F-ratio	Prob
Mn	1	8333.33	8333.33	7.0000	0.0132
Error	28	33333.3	1190.48		
Total	29	41666.7			

Table 4.29: One-Factor ANOVA for Hypotheses H4.1 to H4.2 for the Two Sequenced Discriminators that use class name and class property as the first and second means respectively for target identification

sequenced discriminators that use class name as the first means and class property as the second means for target identification.

3. Is it possible that no subjects use class relationship as the second means for identification?

   As shown in Table 4.30, the t-statistic equals 2.646. There exists strong evidence to reject the null hypothesis.

t-Tests

IdentifyUsingCPasSecondMeans: Test Ho:μ=0 vs Ha:μ>0
Sample mean = 33.333 t-statistic=2.646 with 14 d.f.
Reject Ho at alpha = 0.05 Prob <= 0.0096

Table 4.30: Identifying Classes: Test of Hypothesis on Population Mean Equals 0% in using Class Relationship as the Second Means For Identification

Interpretation

From the analysis above,

1. The two sequenced discriminators that use class name as the first means and class relationship as the second means for target identification are not the same.
2. The two sequenced discriminators that use class name as the first means and class property as the second means for target identification are not the same.
3. Most OO modelers use class relationship as the second means to identify the target classes if the class name fail to adequately identify their targets.
4. A few OO modelers use class property as the second means to identify the target classes if the class name fail to adequately identify their targets.

Hence, we conclude that most OO modelers use class relationship as the second means to identify the target classes if the first means fail to adequately identify their targets.

Locate/Identify / Use Name Similarity	Relevant Classes	Target Classes
Yes	100%	100%
No	0%	0%

Table 4.31: Summary of Scores for Hypotheses H5.1 to H5.2

4.2.5 Similarity Criterion

Recall that the word "similar" refers to 'having comparable characteristics in common'. Under this category, our hypotheses are H5.1 to H5.2. Results, statistical tests and interpretations are presented below.

Hypothesis H5.1

Subjects use name similarity to locate relevant classes if they fail to locate the exact class names specified in the task specification.

Result

According to Table 4.31, 100% of the subjects considered "Shape" to be a relevant class to deal with when they were asked to model the "Graphical Object". This score is a generalized score of the two modeling tasks.

Interpretation

From the results above,

1. All OO modelers in the sample use name similiarity to locate relevant classes.

Hence, we conclude that OO modelers use name similiarity to locate relevant classes if they fail to locate the exact class names specified in the task specification.

Hypothesis H5.2

Subjects use name similarity to identify target classes. The target- class identification process works equally well whether class names are similar or identical.

Result

According to Table 4.31, 100% of the subjects followed the same identification process in identifying the class "Shape" as target for modeling the class "Graphical Object". This score is a generalized score of the two modeling tasks.

Interpretation

From the results above,

1. All OO modelers in the sample use name similiarity to identify target classes.

2. All OO modelers in the sample use the same target identification process no matter whether class names are similar or identical.

Hence, we conclude that OO modelers use name similiarity to identify target classes. The target-class identification process works equally well whether class names are similar or identical.

Chapter 5

Conclusion and Future Work

This chapter concludes and summarizes the main findings of our study. Based on the results of the survey, suggestions for designing user interfaces to be used in OO modeling activities are given as well.

5.1 Conclusion

The study has two main results. First, it provides several empirical observations of the OO modeling process. This is important because prior to this study, no empirical results existed. Results like these ones help us to understand OO modelers' behavior patterns, and provide a starting point for future extensions.

Second, we have successfully used an important assessment technique – Double-Blind Scoring – for the first time in studies of programmer behavior. This technique allows a researcher to obtain quantitative data from open-ended survey questions, making it possible to investigate wide-ranging user impressions with a higher degree

of objectivity.

5.1.1 Empirical Observations

The following is a summary of the observations obtained from the study.

Relevancy

In our context, the word "relevancy" refers to 'a fitness for or appropriateness to the situation or occasion'. (Relevancy is determined by the survey and score sheet creator intuitively in this study; formalizing the definition of relevancy is an important matter for future research.) Given an O-O schema, modelers do not use the entire schema while doing modeling tasks. Instead, modelers define portion(s) of relevancy with respect to each modeling task. Irrelevant portion(s) of the schema are identified and ignored throughout the modeling process.

Schema Browsing

Schema browsing is a goal-directed process of exploring the organization and content of a schema. In our survey, we distinguish two time intervals and refer them as "at the beginning" and "at the end". The term "at the beginning" is used to refer to the period of time immediately after the subject looks at the modeling task and starts to browse the schema; while the term "at the end" is used to refer to the period of time immediately before the subject performs the modification of the schema.

Two other distinctions that are useful for describing schema browsing are *scan* vs.

search, and *relevant* vs. *target*. Scan-oriented schema browsing describes browsing the schema to detect the existence of class(es) without paying attention to the details of the classes being browsed. Contrarily, search-oriented schema browsing describes browsing the schema carefully, paying attention to details of the class(es) being browsed in an effort to find or discover differences among those class(es). Relevant classes are determined early in the modeling process, presumably to reduce the complexity of the task; target classes, which are used in doing the task, are then sought among the relevant classes.

Although there exist two types of hierarchy – generalization and aggregation – in an OO schema, most modelers browse the schema using the generalization hierarchy as the skeleton rather than the aggregation hierarchy or a combination of the two.

Locating relevant class(es) and identifying target class(es) seem to be two significant modeling phrases that OO modelers go through when browsing or exploring the organization and content of the schema. That is, OO modelers first try to determine the existence and location of relevant class(es) and then they try to discriminate among the relevant class(es) to identify those that are needed for the modeling task.

Most OO modelers browse the schema top-down at the beginning. However, OO modelers use different schema browsing strategies for modeling task 2(a) and 2(b) at the end. Most of the subjects did bottom-up browsing at the end when tackling task 2(a), while few did bottom-up browsing at the end for task 2(b).

Even taking into account this inter-task difference schema, browsing behavior is

different at the beginning than at the end. At the beginning, few OO modelers do search-oriented browsing as they do not focus on class details. Conversely, almost all OO modelers do search-oriented browsing at the end as they focus on class details.

Hence, we conclude that when browsing the schema top-down at the beginning, OO modelers perform scan-oriented browsing on the schema trying to locate (c.f. section 3.4.3) relevant class(es). At the end, OO modelers perform search-oriented browsing on the schema trying to identify (c.f. section 3.4.4) target class(es) among the relevant class(es).

Locating Classes

In this study, the term "locating classes" refers determining of the existence and location of class(es). In our study, we observed that most OO modelers rely on class name, and follow the generalization hierarchy while determining the existence and location of the relevant classes on a schema.

Identifying Classes

In this study, the term "identifying classes" refers to process of finding class(es) that match a set of class(es) based on one or more discrimination dimension. Class name, class relationships and class properties are possible discriminators for identifying target classes.

In our study, we observed that almost all OO modelers use class name as the first means for identifying target classes. If the class name is insufficient, most OO

modelers use class relationships; with few OO modelers using class properties as the second means to identify the target classes.

Similarity Criterion

In our context, the word "similar" refers to 'having comparable characteristics in common'. In our study, we observed that OO modelers are aware of the vocabulary problem among specifications and solve the problem using name similarity to locate relevant classes if they fail to locate the exact class names specified in the task specification.

We also observed that OO modelers use name similarity to identify target classes. The target-class identification process works equally well whether names are similar or identical.

5.1.2 Double-Blind Scoring Methodology

Open-ended questions are at least as important and valuable as close-ended questions in survey research. They are especially good for yielding insights into what the subjects are thinking, discovering unanticipated findings and investigating not-well-defined domains of interest. However, open-ended questions suffer from subjectivity. In our study, we successfully used an important assessment technique – Double-Blind Scoring. This technique allows a researcher to obtain quantitative data from open-ended survey questions, making it possible to investigate user impressions with more objectivity.

The following is the eleven-step "Double-Blind Scoring" Methodology used in our

study.

1. Create an open-ended questionnaire for the survey study.

2. Do not announce the intention of the survey study. We want all the participanting subjects to be "blind"-ed from the purpose of the survey.

3. Choose subjects for study so as to minimize sampling bias.

4. Carry out pilot studies where all the pilot subjects are "blind"-ed with respect to the intention(s) and hypotheses (if any) of the survey. Since "blind" subjects are free from expectancy, unbiased results are obtained. "Blind"-ing the subjects consititutes the first "blind" of the methodology.

5. Interprete the pilot studies in an informal way. That is, no detailed and formal analysis is required. These interpretations are not reliable because the interpretator is not blind with respect to the purpose of the study.

6. Refine the survey based on the pilot results.

7. Refine the hypotheses (if any) according to the pilot result. Possibly, create new hypotheses based on the pilot results.

8. Carry out the actual study with the improved questionnaire. "Blind"-ing the subjects from the study hypotheses essential in order to obtain unbiased data. This consititutes the first "blind" of our methodology.

9. Create a close-ended scoring sheet and a criteria sheet based on the hypotheses derived from the pilots. The scoring sheet enables us to turn the subjective scoring scheme into an objective scoring scheme, while the criteria sheet further improves the objectivity and consistency of the scoring.

10. Use a scorer to score the actual study according to survey-specific scoring and criteria sheets. The scorer cannot be the survey creator himself; otherwise the score obtained may be biased by expectancy effects. The scorer needs to be "blind"-ed from the study hypotheses to ensure the scores are objectively assessed. This consititutes the second "blind" of our methodology.

11. Summarize, analyze and interpret the "Double-Blind" scores.

5.1.3 Suggestions on Designing User Interfaces for OO Modeling

Based on the observations from the study, here are some suggestions on designing user interfaces for OO modeling.

1. Relevancy Weighting Scheme

   As we know, the first thing that OO modelers do is to define relevant portions of the schema so that they could keep the schema in a small and manageble piece. An interface that provides relevancy weighting scheme is helpful.

   Throughout our study, class name proves to provide important evidence as to whether a class is relevant with respect to the specified task. Hence, relevancy weighting schemes based on class name are likely to work well.

   In order to address the vocabulary problem, OO modelers try to use name similarity for locating relevant classes. Thus, the relevancy weighting scheme

should take into account the conceptual name-space.

The list of relevant classes returned should, by default, not reveal any class details (i.e. class relationship and class property) because most modelers do not want to see details at the beginning. At this time, they probably redefine their list of relevant classes to address the task according to the specification.

2. Filtering Mechanism

There are two types of hierarchy in a schema: generalization and aggregation. We know that most OO modelers browse the schema based on hierarchy and especially the generalization hierarchy. Hence, after OO modelers define the set of relevant classes, it should be placed in a hierarchical way based on the hierarchy (or hierarchies) that the OO modelers want to see. A good default would organize the relevant classes based on the generalization hierarchy (since most OO modelers browse based on it) and filter out all the aggregation relationships among the relevant classes chosen. Class property can be filtered out by default since most OO modelers do not want class details during the class location phrase.

3. Animation

There are three different kinds of browsing strategy: top-down, bottom-up and random. As we observed, depending on the browsing phase and the task, modelers use different browsing strategies. Hence, the interface could provide

an animation stepping through the relevant classes along a hierarchy either top-down or bottom-up. Random browsing is used once OO modelers are familiar with a portion of the schema. Hence, keeping track of the recently browsed classes or frequently browsed classes can be helpful.

4. Direct Manipulation

   The interface should provide a direct manipulation interface allowing OO modelers in the target identification phrase to pose queries like 'Is there any aggregate relationship involved with respect to this class?', 'What attributes and operations does this class have?' , or 'Show me all the superclasses of this class.'. Through graphical queries and a graphical output interface, a large number of queries can be generated dynamically in a short period of time.

5. Color

   Color is a good visual attribute for exposing differences and catching attention. Color can be used to label relationships like generalization, aggregation and association, since the amount of information displayed on a screen gets larger as the target class identificaiton phrase progresses.

5.2 Recommendations for Future Work

In this section, we propose more user interface related research on OO modeling.

1. Our discussion throughout this essay is based solely on the Object Modeling Technique (OMT) [10] graphical representation. One interesting study would compare graphical representations like OMT with the textual declarations such as C++ [7] class libraries, and see if there is a difference on users performance.

2. Up till now, no standard has been established for object-oriented data models. It would be valuable to do surveys on users' preference with respect to existing OO data models. Alternatively, it would be possible to experiment with different definitions of OO terminology (e.g. the object model, ODMG-93 [9], proposed by Object Database Management Group), comparing user performance to provide guidelines for a "user-friendly" OO data model.

3. Our study could be further extended into a survey-based or experiment-based study by refining our research and incorporating all four OO modeling activities (c.f. section 2.2) into a questionnaire or an experiment. The focus would be on investigating users' behavior and performance on each type of the four OO modeling activity (c.f. section 2.2).

4. An empirical experiment could be carried out in order to investigate our user interface design suggestions for OO modeling and see if those suggestions indeed improve user performance.

Bibliography

[1] L. R. Aiken. Psychological Testing and Assessment. Allyn and Bacon, Inc. 470 Atlantic Av., Boston, Massachusetts 02210, 1979.

[2] F. Bancilhon, C. Delobel and P. Kanellakis. Building an Object-Oriented Database System: The Story of O2. Morgan Kaufmann Publishers 1992.

[3] P. C. Cozby. Methods in Behavioral Research. Fifth edition. Mayfield Publishing Company 1993.

[4] An Introduction to Empirical Problem Solving. The Department of Statistics and Actuarial Science, University of Waterloo, Winter 1994.

[5] G. W. Furnas, T. K. Landauer, L. M. Gomez, and S. T. Dumais, "The Vocabulary Problem in Human-System Communication", Communications of the ACM, 30(11):964-971, November 1987.

[6] C. Lamb, G. Landis, J. Orenstein and D. Weinreb, "The ObjectStore Database System", CACM, 34(10), October 1991, pp. 50-63.

[7] S. B. Lippman, C++ Primer 2nd ed., AT&T Bell Laboratories, Addison Wesley June 1993.

[8] G. A. Marlatt and D. R. Rohsenow. Cognitive processes in alcohol use: Expectancy and the balanced placebo design. In N. K. Mello (Ed.), Advances in substance abuse (Vol. 1). Greenwich, CT:JAI Press 1980.

[9] R. Cattell. The Object Database Standard: ODMG-93. Morgan Kaufmann Publishers 1994.

[10] J. Runbaugh, M. Blaha, W. Premerlani, F. Eddy and W. Lorensen. Object-Oriented Modeling and Design. Prentice Hall 1991.

[11] Parcplace Systems: Objectworks - Smalltalk User's Guide. ParcPlace Systems, Mountain View, CA.

[12] R. Rosenthal. Experimenter Effects in Behavior Research. New York: Appleton-Century-Crofts 1966.

[13] P.F. Velleman and A.Y. Velleman. Data Desk Handbook. Odesta Corporation, 4048 Commercial Av., Northbrook, IL 60062, 1988.

Appendix A

Briefing Sheet

1. Please record every single step that you made before arriving at your solution to the problem.

2. You can assume that whatever information you need is available. But you must explicitly state it in your steps.

3. All the sub-questions asked can serve as guidelines. However, you can structure your own answers rather than following the sub-questions.

4. Please make a note on everything that you are unsure of regarding the question asked and the reason for the uncertainty. Then, you should continue the questionnaire by stating your assumptions.

Appendix B

Questionnaire

1. How many year(s) or month(s) of experience do you have on OO−modeling or OO−programming?

2. Please answer the followings. **You can assume whatever information you asked is available. But you need to explicitly state it out as your steps.**

   a. Place a new class 'Arc' with attribute 'arc angle' and operation 'display' into the original schema shown in Figure 2.1.

   Note: Please record every single steps that you have made before arriving your solution to the problem. The following sub-questions can be served as a guideline.

i. Which class(es), in sequence, did you focus into? Why?

ii. What superclass(es) of 2(a)(i), in sequence, did you look into? Why?

iii. What subclass(es) of 2(a)(i), in sequence, did you look into? Why?

iv. What other class(es), in sequence, did you look into? Why?

v. Which attribute(s) of which class(es), in sequence, did you look into? Why?

vi. Which operation(s) of which class(es), in sequence, did you look into? Why?

vii. What other step(s) did you go through? Why?

viii. Please indicate the above steps (i)−(vii), if any, **in sequence** that you went through.

ix. Please write down your solution to this problem.

b. Model the fact that 'Documents are composed of Pages and Pages are composed of Graphical Objects'. Please **reuse** the class(es), shown in Figure 2.1, as much as possible.

Note: Please record every single steps that you have made before arriving your solution to the problem. The following sub-questions can be served as a guideline.

i. Which class(es), in sequence, did you focus into? Why?

ii. What superclass(es) of 2(b)(i), in sequence, did you look into? Why?

iii. What subclass(es) of 2(b)(i), in sequence, did you look into? Why?

iv. What other class(es), in sequence, did you look into? Why?

v. Which attribute(s) of which class(es), in sequence, did you look into? Why?

vi. Which operation(s) of which class(es), in sequence, did you look into? Why?

vii. What other step(s) did you go through? Why?

viii. Please indicate the above steps (i)−(vii), if any, **in sequence** that you went through.

ix. Please write down your solution to this problem.

3. (a) How did you **determine the location of** the class(es) in 2(a)(i) and 2(b)(i)?

   2(a)(i)

   2(b)(i)

   (b) What kind(s) of strategy (or method) did you use to **identify** the class(es) in 2(a)(i) and 2(b)(i)? The word 'identify' means 'to determine (something) to be the same with something conceived, known, asserted, etc.'

   2(a)(i)

   2(b)(i)

4. Please list out all the **facts and assumptions** you based on in order to justify the following:

   a. The class(es) that you considered in 2(a)(i) are the correct class(es) to deal with.

   b. The solution that you considered in 2(a) are correct.

   c. The class(es) that you considered in 2(b)(i) are the correct class(es) to deal with.

   d. The solution that you considered in 2(b) are correct.

5. (a) What **portion(s) or subset(s)** of the schema, in sequence, did you look at when you were answering 2(a)–2(b)? Why?

   2(a)

   2(b)

   (b) What **particulars** inside the portion(s) or subset(s), in sequence, in 5(a) did you look at when you were answering 2(a)–2(b)?

   2(a)

   2(b)

6. Are you familiar with the application domain that was presented in the schema shown in Figure 1? If not, what are the obstacle(s) for you to answer 2(a)–2(b)?

7. Please write down any comment about this questionnaire.

8. If you would like to be informed of the results of this study, please include your name and email–address below ***or*** send me (scyiu@cgl.uwaterloo.ca) a message directly.

End of Questionnaire. Thanks for your co-operation and comments.

Appendix C

Scoring Sheet

[Scoring S.1] In tackling task 2(a), how much time (in percentage) did the subject spend in looking at the portion: Window, Canvas, Scrolling Canvas, Shape, Line, Ellipse, Polygon and Point?

⌊_____⌋ (Note: Answers for S.1, S.2, S.3 should add up to 100%)

[Scoring S.2] In tackling task 2(a), how much time (in percentage) did the subject spend in looking at the portion: Window, Scrolling Window, Text Window, Scrolling Canvas and Canvas?

⌊_____⌋ (Note: Answers for S.1, S.2, S.3 should add up to 100%)

[Scoring S.3] In tackling task 2(a), how much time (in percentage) did the subject spend in looking at the portion: Window, Panel, Panel Item, Choice Item, Button, Text Item?

⌊_____⌋ (Note: Answers for S.1, S.2, S.3 should add up to 100%)

[Scoring S.4] In tackling task 2(b), how much time (in percentage) did the subject spend in looking at the portion: Window, Canvas, Scrolling Canvas, Shape, Line, Ellipse, Polygon and Point?

⌊_____⌋ (Note: Answers for S.4, S.5, S.6 should add up to 100%)

[Scoring S.5] In tackling task 2(b), how much time (in percentage) did the subject spend in looking at the portion: Window, Scrolling Window, Text Window, Scrolling Canvas and Canvas?

⌊_____⌋ (Note: Answers for S.4, S.5, S.6 should add up to 100%)

[Scoring S.6] In tackling task 2(b), how much time (in percentage) did the subject spend in looking at the portion: Window, Panel, Panel Item, Choice Item, Button, Text Item?

⌊_____⌋ (Note: Answers for S.4, S.5, S.6 should add up to 100%)

[Scoring S.7] Which hierarchy did the subject based on as a skeleton in browsing the schema?

[ ] Generalization

[ ] Aggregation

[ ] Generalization + Aggregation

[ ] None

[ ] Other ⌊_____⌋

[Scoring S.8] In performing task 2(a), which browsing strategy did the subject used at the beginning?

[ ] Top-down

[ ] Bottom-up

[ ] Random

[Scoring S.9] In performing task 2(b), which browsing strategy did the subject used at the beginning?

[ ] Top-down

[ ] Bottom-up

[ ] Random

[Scoring S.10] Which of the following did the subject rely on in order to locate the relevant class(es)?

[ ] Class Property (Operation and/or Attribute)

[ ] Class Name

[ ] Class Relationship (Association and/or Generalization and/or Aggregation)

[ ] Other ⌊_____⌋

[Scoring S.11] At the beginning, with reference to S.8 and S.9 above, did the subject look into any particular details (eg. class name, class attribute, class operation, class relationship etc.) other than that mentioned in S.10 for classes being browsed?

[ ] Yes

[ ] No

[Scoring S.12] In performing task 2(a), which browsing strategy did the subject used at the end?

[ ] Top-down

[ ] Bottom-up

[ ] Random

[Scoring S.13] In performing task 2(b), which browsing strategy did the subject used at the end?

[ ] Top-down

[ ] Bottom-up

[ ] Random

[Scoring S.14] At the end, with reference to S.12 and S.13 above, did the subject look into any particular details (eg. class name, class attribute, class operation, class relationship etc.) other than that mentioned in S.10 for

classes being browsed?

[ ] Yes

[ ] No

[Scoring S.15] Which of the following did the subject rely on first in order to identify the target class(es)?

[ ] Class Property (Operation and/or Attribute)

[ ] Class Name

[ ] Class Relationship (Association and/or Generalization and/or Aggregation)

[ ] Other ⌊_____⌋

[Scoring S.16] Which of the following did the subject rely on after the first mean (mentioned in S.15) failed to adequately identify the target class(es)?

[ ] Class Property (Operation and/or Attribute)

[ ] Class Name

[ ] Class Relationship (Association and/or Generalization and/or Aggregation)

[ ] Other ⌊_____⌋

[Scoring S.17] Which of the following did the subject rely on after the second mean (mentioned in S.16) failed to adequately identify the target class(es)?

[ ] Class Property (Operation and/or Attribute)

[ ] Class Name

[ ] Class Relationship (Association and/or Generalization and/or Aggregation)

[ ] Other ⌊_____⌋

[Scoring S.18] Which of the following did the subject rely on after the third mean (mentioned in S.17) failed to adequately identify the target class(es)?

[ ] Class Property (Operation and/or Attribute)

[ ] Class Name

[ ] Class Relationship (Association and/or Generalization and/or Aggregation)

[ ] Other ⌊_____⌋

[Scoring S.19] For task 2(b), did the subject consider the class "Shape" to be one of the relevant class to deal with when locating the class "Graphical Objects"?

[ ] Yes

[ ] No

[Scoring S.20] For task 2(b), did the subject investigate further (according to S.15 – S.18 if any) to identify the correctness of using the class "Shape" as the class "Graphical Objects"?

[ ] Yes

[ ] No

Appendix D

Criteria Sheet

D.1 Scoring Approach

For each questionnaire, the scorer should fill-in the corresponding scoring sheet right after (s)he studied that questionnaire. The scorer should answer every questions in the scoring sheet quick. That is, it would be no good if the scorer re-thinks or re-studies the questionnaire carefully because (s)he may end up in a situation where s(he) feels uncomfortable in picking any of the choices provided.

Launching this study, we expect to get a general knowledge on object-oriented modeling. Hence, the scoring is expected to arrive at average case results. That is, we are not trying to distinguish the result of 4-4-7 from 5-5-5. Preciseness is not a major issue in the scoring. Thus, we are not expecting the scorer to re-think or re-study the questionnaire in order to fill-in the evaluations in a too-precise way. Rather, the scorer is expected to answer all the questions quick. Besides, once a questionnaire is scored, the scorer is not expected to re-consider his/her scoring

again since the scorer may give different answers at different times.

D.2 Scoring Criteria

[Criteria C.1] In answering [Scoring S.1] – [Scoring S.6] , the scorer is expected to estimate the time that the subject spent on each portion of the schema in tackling the given task in terms of the metric – percentage. The scorer is expected to answer this quick.

[Criteria C.2] For [Scoring S.1] – [Scoring S.6] , the phrase "looking at the portion" means "looking at any subset of class(es) among those listed in the portion".

[Criteria C.3] For [Scoring S.7],

(I) The browsing skeleton is considered to be "generalization" if the subject browsed the schema along the generalization hierarchy.

(II) The browsing skeleton is considered to be "aggregate" if the subject either started at the "Polygon", and immediately looked at the class "Point" before anything else; **OR** started at the "Point" class, and immediately looked at the class "Polygon" before anything else.

(III) The browsing skeleton is considered to be "generalization + aggregation" if the subject browsed the schema along the generalization hierarchy plus the class "Point".

(IV) The browsing skeleton is considered to be "None" if the subject browsed the schema by not following any imposed hierarchical struc-

ture.

[Criteria C.4] For [Scoring S.8] – [Scoring S.11], the phrase "at the beginning" refers to the period of time right after the subject looked at the modeling task and started to browse the schema.

[Criteria C.5] For [Scoring S.8] – [Scoring S.14], the browsing strategy is considered to be "Top-down" if the subject started from the root with a main direction of browsing downwards along the hierarchy.

[Criteria C.6] For [Scoring S.8] – [Scoring S.14], the browsing strategy is considered to be "Bottom-up" if the subject started from the leaf with a main direction of browsing upwards along the hierarchy.

[Criteria C.7] For [Scoring S.8] – [Scoring S.14], the browsing strategy is considered to be "Random" if the subject started from a random spot and browsed the hierarchy in a random manner (ie. not following any imposed hierarchies)

[Criteria C.8] For [Scoring S.8] – [Scoring S.14], the subject may not explicitly write down their browsing strategy nor the sequence of all the class(es) that has(have) been browsed. For example, they may say,"spot the ABC class immediately". The scorer should try his(her) best to estimate how the subject located the class(es) based on what the subject did and estimate the sequence of browsed class(es) in order to determine the strategy used by the subject.

[Criteria C.9] For [Scoring S.10] – [Scoring S.19], we use the word "relationship" to refer to "association", "aggregation" and/or "generalization". Generalization is a "is-a" relationship between a class and one or more refined versions of it. Aggregation is a relationship established by relating a composite class to a component class. Association is a relationship

established by relating two or more independent classes. For example, when we say "Professors advise students", the relationship "advise" would be an association relationship that relates the classes "Professor" and "Student".

[Criteria C.10] For [Scoring S.10] – [Scoring S.20], the word "locate" refers to "to determine the existence and location of something"; while the word "identify" refers to "to determine something to be the same with something conceived, known, asserted, etc."

[Criteria C.11] For [Scoring S.10] – [Scoring S.20], the subject may do the location of relevant class(es) and identification of target class(es) at the same time. In that case, the wordings of "locate" and "identify" used in the scoring sheet are interchangeable. Otherwise, they should be read as mentioned in [Criteria C.10].

[Criteria C.12] For [Scoring S.10], [Scoring S.15] – [Scoring S.20], relevant class refers to a class that is appropriate but not necessarily correct in addressing a modeling task; whereas the target class refers to a class which is precisely correct in addressing a modeling task.

[Criteria C.13] For [Scoring S.12] – [Scoring S.14], the phrase "at the end" refers to the period of time right before the subject performed the modification on the schema for a modeling task.

[Criteria C.14] For [Scoring S.11] and [Scoring S.14], the scorer is expected to consider "Yes" when the subject looked into any class details for more than 50% of the classes being browsed. Otherwise, the scorer is expected to give the answer "No".

[Criteria C.15] For [Scoring S.19], the scorer is expected to pick the answer "No" when the subject could not find the class "Graphical Objects" and created a new class "Graphical Objects" immediately without considering that the class "Shape" would be a possible relevant candidate to help tackling the problem. Otherwise, "Yes" is expected.

[Criteria C.16] For [Scoring S.20], the scorer is expected to pick "Yes" if the subject tried to identify the class "Shape" as the target for the class "Graphical Objects" according to the means described in [Scoring S.15] – [Scoring S.18] (if any). Otherwise, "No" is expected.

[Criteria C.17] For [Scoring S.15] – [Scoring S.18], the scorer is expected to infer or generalize the sequence of means that the subject relied on in order to identify target class(es) with respect to the modeling tasks.

Appendix E

Scored Results

Questions on Scoring Sheet: S.1 - S.6
Metric: percentage of time spent

Subject	2(a) Schema Portions			2(b) Schema Portions		
	Relevant	Irrelevant	Irrelevant	Relevant	Relevant	Irrelevant
1	100	0	0	70	30	0
2	100	0	0	80	10	10
3	90	5	5	90	5	5
4	100	0	0	50	50	0
5	100	0	0	90	10	0
6	100	0	0	70	30	0
7	100	0	0	40	60	0
8	85	10	5	100	0	0
9	100	0	0	70	30	0
10	90	5	5	25	70	5
11	100	0	0	50	40	10
12	100	0	0	90	5	5
13	100	0	0	60	15	25
14	90	5	5	60	35	5
15	100	0	0	100	0	0
Average (%)	97.0	1.7	1.3	69.7	26.0	4.3

Table E.1: Relevancy: Scored Results for Questions S.1 – S.6 in Scoring Sheet

Questions on Scoring Sheet: S.7
Metric: percentage of subjects

Subject	Generalization	Aggregation	Generalization & Aggregation	None	Other
1	1	0	0	0	0
2	1	0	0	0	0
3	1	0	0	0	0
4	0	0	0	1	0
5	1	0	0	0	0
6	1	0	0	0	0
7	0	0	1	0	0
8	1	0	0	0	0
9	1	0	0	0	0
10	1	0	0	0	0
11	1	0	0	0	0
12	1	0	0	0	0
13	1	0	0	0	0
14	1	0	0	0	0
15	1	0	0	0	0
Average (%)	86.7	0.0	6.7	6.7	0.0

Table E.2: Schema Browsing: Scored Results for Questions S.7 in Scoring Sheet

Questions on Scoring Sheet: S.8, S.9
Metric: percentage of subjects

Subject	2(a) AtTheBeginning			2(b) AtTheBeginning		
	Top-down	Bottom-up	Random	Top-down	Bottom-up	Random
1	0	1	0	1	0	0
2	1	0	0	0	0	1
3	1	0	0	1	0	0
4	0	0	1	1	0	0
5	1	0	0	1	0	0
6	1	0	0	1	0	0
7	0	0	1	1	0	0
8	1	0	0	0	0	1
9	1	0	0	0	1	0
10	0	0	1	1	0	0
11	0	1	0	1	0	0
12	1	0	0	1	0	0
13	0	1	0	0	1	0
14	1	0	0	1	0	0
15	1	0	0	0	1	0
Average (%)	60.0	20.0	20.0	66.7	20.0	13.3

Table E.3: Schema Browsing: Scored Results for Questions S.8 – S.9 in Scoring Sheet

Questions on Scoring Sheet: S.12, S.13
Metric: percentage of subjects

Subject	2(a) AtTheEnd			2(b) AtTheEnd		
	Top-down	Bottom-up	Random	Top-down	Bottom-up	Random
1	0	1	0	1	0	0
2	0	1	0	1	0	0
3	0	1	0	1	0	0
4	0	1	0	1	0	0
5	0	1	0	0	0	1
6	0	1	0	0	0	1
7	0	1	0	0	0	1
8	0	1	0	0	0	1
9	0	0	1	0	0	1
10	0	1	0	1	0	0
11	0	1	0	1	0	0
12	0	0	1	0	0	1
13	1	0	0	0	1	0
14	0	1	0	0	0	1
15	0	1	0	1	0	0
Average (%)	6.7	80.0	13.3	46.7	6.7	46.7

Table E.4: Schema Browsing: Scored Results for Questions S.12 – S.13 in Scoring Sheet

Questions on Scoring Sheet: S.11 - S.14
Metric: percentage of subjects

Subject	AtTheBeginningLookAtDetails		AtTheEndLookAtDetails	
	Yes	No	Yes	No
1	1	0	1	0
2	0	1	1	0
3	0	1	1	0
4	1	0	1	0
5	1	0	1	0
6	0	1	1	0
7	1	0	1	0
8	0	1	1	0
9	0	1	1	0
10	1	0	1	0
11	0	1	1	0
12	0	1	1	0
13	0	1	1	0
14	0	1	1	0
15	0	1	1	0
Average (%)	33.3	66.7	100.0	0.0

Table E.5: Schema Browsing: Scored Results for Questions S.11 – S.14 in Scoring Sheet

Questions on Scoring Sheet: S.10
Metric: percentage of subjects

Subject	Class Property (CP)	Class Name (CN)	Class Relationship (CR)	CN & CP	CN & CR	Other
1	0	0	0	0	1	0
2	0	1	0	0	0	0
3	0	0	0	0	1	0
4	0	0	0	1	0	0
5	0	0	0	1	0	0
6	0	0	0	0	1	0
7	0	1	0	0	0	0
8	0	0	0	0	1	0
9	0	0	0	0	1	0
10	0	0	0	0	1	0
11	0	0	0	0	1	0
12	0	0	0	0	1	0
13	0	0	0	0	1	0
14	0	0	0	0	1	0
15	0	0	0	0	1	0
Average (%)	0.0	13.3	0.0	13.3	73.3	0.0

Table E.6: Locating Classes: Scored Results for Questions S.10 in Scoring Sheet

Questions on Scoring Sheet: S.15 - S.18
Metric: percentage of subjects

1st means 2nd if 1st fails 3rd if 2nd fails Subject	CN No No	CN CR No	CN CR CP	CN CP No	CN CP CR	CP No No	CP CN No	CP CN CR	CP CR No	CP CR CN	CR No No	CR CN No	CR CN CP	CR CP No	CR CP CN
1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
2	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
3	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
4	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
5	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
6	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
7	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
8	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
9	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
10	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
11	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
12	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
13	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
14	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
15	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
Average (%)	0.00	6.7	60.0	0.00	33.3	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Table E.7: Identifying Classes: Scored Results for Questions S.15 – S.18 in Scoring Sheet

Questions on Scoring Sheet: S.19 - S.20
Metric: percentage of subjects

Subject	Use Name Similarity for Class Location		Use Name Similarity for Identification	
	Yes	No	Yes	No
1	1	0	1	0
2	1	0	1	0
3	1	0	1	0
4	1	0	1	0
5	1	0	1	0
6	1	0	1	0
7	1	0	1	0
8	1	0	1	0
9	1	0	1	0
10	1	0	1	0
11	1	0	1	0
12	1	0	1	0
13	1	0	1	0
14	1	0	1	0
15	1	0	1	0
Average (%)	100.0	0.0	100.0	0.0

Table E.8: Similarity Criterion: Scored Results for Questions S.19 – S.20 in Scoring Sheet