Effectiveness Assessment of an Early Testing Technique using Model-Level Mutants*

M. F. Granda 1
Computer Science Department,
University of Cuenca
Ecuador
fernanda.granda@udecuenca.edu.ec

N. Condori-Fernández
University of A Coruña, Spain
n.condori.fernandez@udc.es
n.condori-fernandez@vu.nl

T. E. J. Vos, O. Pastor
PROS, Universitat Politecnica de València
Spain
opastor@pros.upv.es

ABSTRACT While modern software development technologies enhance the capabilities of model-based driven development, they introduce challenges for testers such as how to perform early testing at model level to ensure the quality of the model. In this context, we have developed an early testing technique supported by the CoTest tool to validate requirements at model level. In this paper we describe an empirical evaluation of CoTest with respect to its effectiveness in terms of its fault detection and test suite adequacy. This evaluation is carried out by model-level mutation testing using first order mutants (created by injection of a single fault) and high order mutants (containing more than one fault) with seven conceptual schemas (of different sizes) that represent the functionality of different software systems in different domains. Our findings show that the tests generated by CoTest are effective at killing a large number of mutants. However, there are also some fault types (e.g. delete the references to a class attribute or an operation call in a constraint) that our test suites were not able to detect. CoTest was more effective in terms of detecting fault types using high order mutants that first order mutants. Thus, CoTest's effectiveness is affected by the mutant type tested.

CCS CONCEPTS
- Software and its engineering → Software testing and debugging - Software and its engineering → Empirical software validation

KEYWORDS
Test Suite Effectiveness, Effectiveness Assessment, Mutation Testing, Conceptual Schemas Testing, Class Diagram Mutation

ACM Reference format:
http://dx.doi.org/10.1145/3084236.3084257

1 INTRODUCTION
Constructing software automatically from models or Conceptual Schemas (CS) is one of the current challenges in software engineering, especially in a Model-driven Engineering context [26]. A well-formed model, being an accurate representation of all the requirements for a system under construction, is a key factor in the successful development and production of the system. The development of a CS is an iterative process involving evaluation of the model, its accuracy and its improvement from the evaluation results.

Testing is a well-established technique that helps to accomplish this task and provides a level of confidence in the end product based on the coverage of the requirements achieved by the tests. In this context, we defined an early testing technique for validating Conceptual Schemas in a Model-driven environment [14][13]. This technique covers: 1) test suite generation, 2) CS under test generation, 3) test execution and report generation with the faults detected and the coverage analysis. Therefore, the technique's effectiveness and adequacy of the test suite require to be evaluated.

Effectiveness in detecting faults can be evaluated by the types and number of faults that can be detected by the technique [28]. For assessing the adequacy of a test suite, mutation testing is a method that injects artificial faults or changes into a software product (mutant) and checks whether a test suite is "good enough" to detect these artificial faults. The adequacy level of the test suite can be measured by a mutation score that is computed in terms of the number of mutants killed (detected) by the test suite [18].

Killing a mutant means that the execution is stopped because a fault was detected or because it reaches an inconsistent state and cannot continue execution. Mutants are produced by using mutation operators that describe syntactic changes to the original software product. Mutants can be classified into two types: First Order Mutants (FOM) and Higher Order Mutants (HOM) [19]. Traditional mutation testing considers FOM created by injection of a single fault: HOM contain more than one fault. Jia and
Harman claim that some HOMs are harder to kill than the FOMs [18], and so we were interested in evaluating the effectiveness of CoTest test cases in both mutant types.

Mutation testing was originally introduced by DeMi1lo et al. [5] and Hamlet [17], as a support technique for developing tests for software systems represented at the code level. However, it has also been applied to models at the design level, for example to Finite State Machines [9], State Charts [11], Activity Diagrams [19], and Network protocols [20]. However, there is no empirical evidence on the effectiveness of mutation testing in improving test suites for Conceptual Schemas.

This paper uses a mutation-testing based approach to evaluate the fault detection effectiveness of an automatically generated test suite to test a given CS. This means the CS is mutated and not the code! The goal of our study is, therefore, to analyze the test suite generated by CoTest tool for the purpose of carrying out a comparative evaluation with respect to its effectiveness in detecting faults, fault types and the adequacy of the test suite in the context of mutants (i.e. FOM and HOM) generated for seven CS.

In a previous paper [12], we proposed a set of 50 mutation operators specifically designed to generate mutants for UML Class Diagrams based CS and we evaluated the usefulness of an effective subset of mutant types of 18 mutation operators to inject defects into a CS. For this, we developed 1) the MutUML tool (Mutation for UML) [16] for the generation and parsing (i.e. syntax analysis) of first order mutants by using the set of 18 mutation operators previously defined for Conceptual Schemas based on UML Class Diagrams (CD); and, 2) the CoTest tool (Conceptual Schema Testing) [12] to support the semi-automatic generation of test cases from a requirements model, the execution of CS/CS mutants against generated tests, and reporting the results.

The main contribution of this paper is to empirically evaluate CoTest's effectiveness in detecting faults and the adequacy of the test suite, using seven CSs and mutation testing.

The paper is organized as follows. Section 2 describes the CoTest technique. Section 3 presents the experimental design. Section 4 analyses and interprets the results. Section 5 discusses the results. Section 6 summarizes the threats to validity. The conclusions and future work are given in Section 7.

2 AN EARLY TESTING TECHNIQUE: CoTest

As mentioned in Section 1, the main goal of CoTest is to automate a testing approach for Conceptual Schemas. For this, CoTest generates test cases (i.e. assertions with the expected value), transforms the conceptual schema under test into an executable CS and executes the test process for reporting the results. In this section we describe the testing environment, the steps of the CoTest technique and test cases properties.

2.1 The testing environment

The environment for testing conceptual schemas provided by CoTest is based on the Action Language for Foundational UML, or ALF [23], adopted as standard by the OMG [25]. ALF is basically a textual notation for UML behaviours that can be attached to a UML model at any point that may contain a UML behaviour, e.g. the method of an operation or the classifier behaviour of a class. Semantically, ALF maps the Foundational UML (F UML [24]) subset, then F UML provides the virtual machine for the execution of the ALF language, so that the test suite and executable model are generated and transformed into ALF language, respectively.

2.2 The CoTest Process

Fig. 1 provides the reader with a description of how CoTest operates, its phases and activities. The figure contains four main parts: CoTest artefacts, CoTest activities, software artefacts and model tester activities. As the names suggest, CoTest activities are done automatically whereas the model tester activities are done manually. CoTest encapsulates all the CoTest artefacts. The numbered ovals represent activities and the boxes represent artefacts. Arrows to/from activities represent the consumption and production of artefacts, respectively.

![Figure 1. The CoTest process](image)
Effectiveness Assessment of an Early Testing Technique using...  

A brief description of each CoSTest activity is as follows:

2.2.1 Test suite Generation

1. **Identify the input requirements:** The tester needs to select the requirements model (RM), which is based on Communication Analysis [6]. We assume that the model is syntactically well-formed.

2. **Generate the test model (TM):** CoSTest analyses the RM structure by automatically traversing all the RM nodes (event sequences) and extracting all the Test Model (TM) elements and their properties.

3. **Generate the abstract test scenarios (TS):** CoSTest computes the total number of possible test scenarios (based on event sequence) and generates the test scenarios with abstract test cases. These three steps are explained with more detail in [14].

4. **Concrete Variables:** The next step is to concretize the variables of the test cases. The tester can (i) recover a variable list from the test model and generate values automatically from the example values specified in the requirements model, or (ii) concretize manually by introducing values for each variable.

5. **Choose the test suite types:** The tester can select between two types of test cases, such as (i) partial (only positive test cases) or (ii) complete, which adds test cases with some negative conditions, such as values out of range, constraint violations, and unique value violation for class variables.

6. **Generate concretized test cases (CTC):** In this phase, CoSTest automatically transforms the abstract test cases into parameterized scripts. The output is a non-executable script for each test scenario. Scripts are not executable in the sense that they do not contain concretized variables.

CoSTest then computes and generates the total number of possible executable and concrete test cases that may be executed on the CS, including concretized variables, the test objective and an expected output (oracle) that is used to validate the CS requirements. The output of this step is a test suite formed by an executable script (ALF script) for each test scenario. The test suite for the subsequent testing process is now ready.

7. **Identify the Conceptual Schema:** The tester, which is a UML Class Diagram (CD), identifies the Conceptual Schema. We assume that the CS is syntactically well-formed.

2.2.2 CSUT Generation

8. **Generate an Executable Conceptual Schema (CSUT):** CoSTest transforms the CS into an executable format (ALF) for its execution.

2.2.3 Test Execution and Reports Generation

9. **Choose testing type:** The testing type is based on the following stop criteria: Testing should be stopped when (1) one fault is detected; or (2) all available test cases have been run.

10. **Execute Test suites:** Test cases are executed on the executable CS and the output is compared to the stored expected output (from Step 6). CoSTest generates an execution report in which the executed test cases are classified as passed, failed or inconclusive. A coverage analysis is performed and a fault report is generated.

2.3 CoSTest Test Cases

A test suite for CS is a set of one or more test scenarios. Each test scenario is a story that consists of one or more test cases. The CoSTest test cases exhibit the following properties:

- A test case consists of a fixture and one or more statements that execute one of the tests applicable to CS, such as testing assertions about the occurrence or non-occurrence of an event. The fixture is a set of statements (e.g. create an object or link, execute a method) that create a CS state and define the values of the CS variables.

- Each execution of a test case starts with the execution of the fixture. For example, if we want to test the creation of an object of the RegisterUser class in the Sudoku Game CS, a test case that corresponds to a one test scenario generated by CoSTest would be as shown in Fig. 2.

- It is assumed that the execution of each test case starts with an empty state. With this assumption, test cases of a CS are independent of each other, and the order of their execution is therefore irrelevant.

![Figure 2: A partial view of a test case](image)

- **Test case always returns a verdict which may be Pass, Fail or Inconsistent.** The execution of the test cases leads to one of the following three outputs:
  - No defects and a status of passed execution. This is considered the output expected.
  - A defect list and a status of failed execution. For example the execution of the test cases may produce an output with several defects (e.g. missing class, incorrect operation and missing operation), which is different from the expected output.
  - A defect list (optional) and "status-inconclusive" if the execution is not conclusive. For example, if the fixture has caused a fault, this leads to an inconclusive status.

In the next section, we describe the design of a controlled experiment for evaluating CoSTest by means of its effectiveness for detecting faults and test suite quality.

3 EXPERIMENTAL PLAN

Since the experiment was motivated by the need to investigate the effectiveness of CoSTest, we intended to compare the effectiveness and adequacy of the test cases when they were
applied in both first order mutants and high order mutants to detect faults in seven CS. The experiment was carried out in 2016 (from January to March) and was designed according to Wohlin et al. [29], and reported according to Juristo and Moreno [21]. This section describes the goal of the study, research questions, metrics used, the subject CS, and the experimental settings.

3.1 Goal

In line with the Goal/Question/Metric Paradigm [27], the goal of our empirical study was the following: Analyze the test suite generated by the CoTest tool for the purpose of carrying out a comparative evaluation with respect to its effectiveness in detecting faults, fault types and the adequacy of the test suite from the point of view of the testers in the context of mutants (i.e. FOM and HOM) generated for seven CS.

3.2 Research Questions

As we were interested in determining if the effectiveness was the same for both types of mutants (i.e. FOM and HOM), we posed and studied the following research questions:

- **RQ1**: How does the mutation type influence the CoTest’s effectiveness in detecting faults?
- **RQ2**: How adequate are CoTest test suites for killing both the First Order Mutants and High Order Mutants of Conceptual Schemas?

3.3 Hypotheses

We defined three hypotheses. The null hypotheses (represented by a $H_0$ in the subscript), which corresponds to the absence of an impact of the independent variables on the dependent variables. The alternative hypotheses involve the existence of such an impact and are the expected result.

$H_0$ Mutant type does not influence the effectiveness of the CoTest test cases in detecting faults in Conceptual Schemas (RQ1).

$H_0$ Mutant type does not influence the effectiveness of the CoTest test cases in detecting fault types in Conceptual Schemas (RQ2).

$H_0$ Mutant type does not influence the adequacy of the CoTest test cases (RQ3).

3.4 Variables and Metrics

3.4.1 Independent Variables

We consider one independent variables (a.k.a. factor [21]):

1. **Mutation type.** Since this study uses mutation for injecting the artificial faults into a CS, mutants can be classified into two types according to the number of mutated elements:
   - First Order Mutants (FOM), which are generated by applying mutation operators only once [18].
   - Higher Order Mutants (HOM), which are generated by applying mutation operators more than once [18].

3.4.2 Dependent Variables and Metrics

We consider the following two dependent variables (aka. response variables [21]), which are expected to be influenced to some extent by the independent variable.

1. **Fault Detection Effectiveness.** To investigate our RQ1 we need to measure the effectiveness of the CoTest tool in terms of both the number of faults found and the type (or cause) of the faults that were found [22] using the following metrics:
   - The metric **Rate of Fault Detection (FDR)** is the value calculated by dividing the number of faults detected by the tool $F_D(T)$ by the total number of faults that are expected to be identified from the CS mutants ($F_T$).
     \[ FDR(T) = \frac{F_D(T)}{F_T} \]
   - The metric **Rate of Fault Type Detection (FTDR)** is the value calculated by dividing the number of fault types detected by the tool $FT(D(T))$ by the total number of fault types that are expected to be identified from the CS mutants ($FT(T)$).
     \[ FTDR(T) = \frac{FT(D(T))}{FT(T)} \]

2. **Adaptability Test Suite.** For a test suite $T$ the adaptability score is a variable that can be used to measure the effectiveness of a test suite in terms of its ability to kill mutants because it is one outcome of the Mutation Testing process, which indicates the quality of the input test set [18]. During execution each CS mutant $M_i$ will be run against a test case suite $T$. If the result of running $M_i$ is different from the result of running CS for any test case in $T$, then the mutant $M_i$ is said to be "killed", otherwise it is said to have "survived". A CS mutant may survive either because it is equivalent to the original model (i.e. it is semantically identical to the original model although syntactically different) or the test set is inadequate to kill the mutant. Thus, the mutation score (MS) for a test suite $T$ is the ratio of the number of killed mutants $M_k(T)$ over the total number of the non-equivalent mutants $M_i$ generated for a CS, as follows:
     \[ MS(T) = \frac{M_k(T)}{M_i} \]

3.5 Experimental Context

3.5.1 Subject CS

We used seven subject CS in our study which contained a variety of characteristics that can be present in UML CD-based CS, including classes, relations (i.e. association, composite aggregation, and generalization) and different types of constraints. Table 1 summarizes the characteristics of these CS.

<table>
<thead>
<tr>
<th>Table 1: Elements of the Subject Conceptual Schemas</th>
</tr>
</thead>
<tbody>
<tr>
<td>Element</td>
</tr>
<tr>
<td>----------</td>
</tr>
<tr>
<td>Classes</td>
</tr>
<tr>
<td>Attributes</td>
</tr>
<tr>
<td>Operations</td>
</tr>
<tr>
<td>Parameters</td>
</tr>
<tr>
<td>Associations</td>
</tr>
<tr>
<td>Constraints</td>
</tr>
<tr>
<td>Generalizations</td>
</tr>
</tbody>
</table>
Effectiveness Assessment of an Early Testing Technique using...

A brief description of each CS is as follows:

1. Video Club (VC) CS represents the functionality of a chain of video stores to manage movies, partners and movie rentals.

2. Medical Treatment (MT) CS defines part of a Medical Treatment business process for a fictional hospital named University Hospital Santiago Grisolia, developed by España et al. [8].

3. Sudoku Game (SG) CS was developed by Tort and Olivé [2] as an object-oriented CS of the Sudoku Game system. This CS defines the functionality for managing different users, playing with their sudokus and generating new ones.

4. Expense Report (ER) CS defines the functionality of an information system to manage the expense-report life cycle of a business. This CS deals with several entities such as departments, employees, projects, and expense types.

5. Online Conference Review (OCR) CS, which is based on the description of the CyberChair System [4], defines the functionality of an information system to deal with members (committee chair and program committee) of a conference, as well as authors that submit papers to be evaluated for inclusion in the conference proceedings.

6. Super Stationery (SS) CS defines the information system of a company that provides stationery and office material to its clients. This CS was developed by España et al. [7].

7. Incident Management (IM) CS defines the functionality of an information system to solve the incoming incidents (reception, process, allocation process and resolution process). This CS is a real case taken from Eversi Company1, a multinational firm offering business consulting, as well as development, maintenance and improvement IT.

3.5.2 Mutation operators

In a CS, missing, unnecessary and incorrectly modelled requirements are the main causes of a CS inaccuracy that can be detected by the requirements. In a previous work [12], 50 mutation operators were defined for CS, and 18 were selected for generating only first order mutants.

In this work, in order to mutate the CSs and evaluate CoSTest’s effectiveness and the adequacy of the test suite, we used 27 mutation operators defined in [12]. 18 for FOM (see Table 2) and 9 for HOM (see Table 9). In these tables each of the 27 mutant operators is represented by a three-letter acronym and a number. The acronym consists of 3 parts: (i) one letter that corresponds to the defect type injected by the mutation operator C=unnecessary, W=wrong and M=missing; (ii) two letters that represent the modelling element (i.e. CO=constraint, GE=generalization, AS=association, CL=class, AT=attribute, OP=operation, and PA=parameter) affected by the mutation; and (iii) a sequential number within its category, for example, the "Missing Association" (MAS).

Table 2: Mutation operators for CS FOM taken from [12]

<table>
<thead>
<tr>
<th>#</th>
<th>Code</th>
<th>Mutation Operator rule</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>UFA2</td>
<td>Adds an extraneous Parameter to an Operation</td>
</tr>
<tr>
<td>2</td>
<td>WCO1</td>
<td>Changes the constraint by deleting the references to a class Attribute</td>
</tr>
<tr>
<td>3</td>
<td>WCO3</td>
<td>Change the constraint by deleting the calls to specific operation.</td>
</tr>
<tr>
<td>4</td>
<td>WCO4</td>
<td>Changes an arithmetic operator for another and supports binary operators: +, -, *</td>
</tr>
<tr>
<td>5</td>
<td>WCO5</td>
<td>Changes the constraint by adding the conditional operator &quot;not&quot;</td>
</tr>
<tr>
<td>6</td>
<td>WCO6</td>
<td>Changes a conditional operator for another and supports operators: or, and</td>
</tr>
<tr>
<td>7</td>
<td>WCO7</td>
<td>Changes the constraint by deleting the conditional operator &quot;not&quot;</td>
</tr>
<tr>
<td>8</td>
<td>WCO8</td>
<td>Changes a relational operator for another and supports operators: &lt;, &lt;=, &gt;, =&gt;, =, !=</td>
</tr>
<tr>
<td>9</td>
<td>WCO9</td>
<td>Changes a constraint by deleting a unary arithmetic operator (‘)</td>
</tr>
<tr>
<td>10</td>
<td>WAS1</td>
<td>Interchanges the members of an Association</td>
</tr>
<tr>
<td>11</td>
<td>WAS2</td>
<td>Changes the association type (e.g., normal, composite)</td>
</tr>
<tr>
<td>12</td>
<td>WAS3</td>
<td>Changes the multiplicity of an Association member (i.e., &quot;*&quot;, &quot;0..1&quot;, &quot;0..1&quot;)</td>
</tr>
<tr>
<td>13</td>
<td>WCL1</td>
<td>Changes visibility kind of the Class (e.g., private)</td>
</tr>
<tr>
<td>14</td>
<td>WOP2</td>
<td>Changes the visibility kind of an operation</td>
</tr>
<tr>
<td>15</td>
<td>WPA</td>
<td>Changes the Parameter data type (i.e., String, Integer, Boolean, Date, Real)</td>
</tr>
<tr>
<td>16</td>
<td>MCO</td>
<td>Deletes a constraint (i.e. pre-condition, post-condition constraint, body constraint)</td>
</tr>
<tr>
<td>17</td>
<td>MAS</td>
<td>Deletes an Association</td>
</tr>
<tr>
<td>18</td>
<td>MPA</td>
<td>Deletes a Parameter from an Operation</td>
</tr>
</tbody>
</table>

Table 3: Mutation operators for CS HOM taken from [12]

<table>
<thead>
<tr>
<th>#</th>
<th>Code</th>
<th>Mutation Operator rule</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>WC02</td>
<td>Changes the property (attribute) data type in the constraint</td>
</tr>
<tr>
<td>2</td>
<td>WGE</td>
<td>Changes the Generalization member end</td>
</tr>
<tr>
<td>3</td>
<td>WAT1</td>
<td>Changes the Attribute feature &quot;Is Derived&quot; to true</td>
</tr>
<tr>
<td>4</td>
<td>WAT2</td>
<td>Changes the Attribute property &quot;Is Derived&quot; to false</td>
</tr>
<tr>
<td>5</td>
<td>WAT3</td>
<td>Changes the Attribute data type</td>
</tr>
<tr>
<td>6</td>
<td>MG0</td>
<td>Deletes a Generalization relation</td>
</tr>
<tr>
<td>7</td>
<td>MCL</td>
<td>Deletes the class (i.e., normal or association class)</td>
</tr>
<tr>
<td>8</td>
<td>MAT</td>
<td>Deletes an Attribute</td>
</tr>
<tr>
<td>9</td>
<td>MOP</td>
<td>Deletes an Operation</td>
</tr>
</tbody>
</table>

Therefore, we need to add more steps to the operator (going from FOM to HOM). The HOM should delete the association together with the respective constraint. This way, the mutant will not be detected by the parser and can generate a valid mutant for testing. Our experiment was carried out under a within-subject design, all our subjects were exposed to the two treatments of our independent variable (mutation type) [3].

1www.eversi.com
3.6 Experimental Procedure
This section describes the details of the experimental setup including the subject CS used, instrumentation, data collection, and analysis. Fig. 4 summarizes the experimental process, which involved performing the following seven steps:

3.6.1 Choose CS Subjects
The selected subjects are described in Section 3.5.1. These CS were of different sizes and domains (e.g., information systems, games). The selected CS comprised an industrial case (i.e., BI), some others were found in the literature (i.e., [2], [5] and [10]), and some were added because they contained the CS elements required to inject the faults.

3.6.2 Generate Test Suites
A test suite was generated to kill CS mutants for each CS subject by following Steps 1-6 of Section 2.2, we then analyzed and recorded the information on the generated test cases in order to eliminate repeated or invalid test cases. The CoTest report was then used for this task.

3.6.3 Execute Test Suites on CS
Each test suite executed on the respective CS subject using our freely available CoTest validation tool. We assessed whether an invalid test case required a manual setting (e.g., concretize variables that require several values because they should be unique values or adjust a negative test case so that it can create a valid sequence of events to validate constraints).
We adjusted the test cases in order to get a successful testing process with the original CS and registered the invalid test cases.

3.6.4 Generate CS Mutants
As this step is quite computationally expensive and cumbersome, we used our MymUI tool [16] for generating first order mutants, in contrast to the high order mutants, which were generated manually. Both mutant types were generated by using the mutation operators introduced in Section 3.5.2. A syntax analysis was then performed by using the Alf parser to ensure that the mutants were valid and could be used in a testing process.

Figure 3. Excerpt of a UML CD-based CS and the application of five mutation operators

Figure 4. Steps taken in experimental process

---

3 https://staq.doc.pwr.wroc.pl/CoTest.html
In this study, we used all the FOMs generated by the tool for all CS subjects (see Table 8 in Appendix). In actual testing scenarios, CS do not typically contain as many faults as these numbers of mutants. The numbers of selected mutants derived by this process for our subject CSs can be found at an external source9. In the other case, since there is no tool to automatically generate HOMs, to simulate more realistic scenarios, we randomly selected 3 mutants from the pools of mutants created for each mutation operator. Our goal was 27 mutants per CS, 3 mutants by each mutation operator from Table 4, but some versions of our CSs did not have enough mutants to allow formation of so many groups. So, our random selection algorithm stopped generating mutants for each mutation operator when it could not generate any more unique mutants, resulting in several cases in which mutants numbered less than 27, i.e., for WAT2, WGB, and MGE operators (see Table 9 in Appendix).

### 3.6.5 Select and generate an executable CS mutant

Each CS mutant is transformed into an executable CS (CSUT) by using the respective CoSTest module (see Step 8 in Section 2.2).

### 3.6.6 Execute Test Suites on CS Mutants and Collect Data

We ran each test suite using CoSTest for each generated mutant. CoSTest generates automatically a report (i.e., data collection) with the status of the test suite (i.e., passing/failing/inconclusive). We then manually examined the FOM with zero kills (i.e., status=passing) and eliminated any that were semantically equivalent to the original CS. The analysis of survivor mutants in order to identify equivalent mutants is a prerequisite for calculating a mutation score. An example of an equivalent mutant is shown in Fig. 5.

![Figure 5. Excerpt of a Constraint mutated by WCQ8](image)

### Table 4: Faults and Fault Types detected by Mutant Type

<table>
<thead>
<tr>
<th>Fault Types</th>
<th>CS</th>
<th>VC</th>
<th>MT</th>
<th>SG</th>
<th>ER</th>
<th>OCR</th>
<th>FOM</th>
<th>HOM</th>
<th>FOM</th>
<th>HOM</th>
<th>FOM</th>
<th>HOM</th>
<th>FOM</th>
<th>HOM</th>
<th>FOM</th>
<th>HOM</th>
<th>FOM</th>
<th>HOM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Extraneous</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Derived</td>
<td>3</td>
<td>5</td>
<td>3</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Missing Class</td>
<td>5</td>
<td>1</td>
<td>6</td>
<td>3</td>
<td>11</td>
<td>2</td>
<td>7</td>
<td>2</td>
<td>10</td>
<td>2</td>
<td>9</td>
<td>3</td>
<td>6</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Missing Constraint</td>
<td>52</td>
<td>13</td>
<td>50</td>
<td>10</td>
<td>36</td>
<td>2</td>
<td>37</td>
<td>1</td>
<td>19</td>
<td>21</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Missing Operation</td>
<td>13</td>
<td>7</td>
<td>2</td>
<td>14</td>
<td>4</td>
<td>17</td>
<td>6</td>
<td>6</td>
<td>3</td>
<td>23</td>
<td>4</td>
<td>7</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Missing Association</td>
<td>4</td>
<td>8</td>
<td></td>
<td></td>
<td></td>
<td>13</td>
<td>12</td>
<td></td>
<td>8</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Incorrect Operation</td>
<td>1</td>
<td>6</td>
<td>9</td>
<td>9</td>
<td></td>
<td>12</td>
<td>13</td>
<td>13</td>
<td>8</td>
<td>2</td>
<td>9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Incorrect Parameter</td>
<td>3</td>
<td>27</td>
<td>21</td>
<td>29</td>
<td>58</td>
<td>2</td>
<td>16</td>
<td>1</td>
<td>82</td>
<td>20</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FDR</td>
<td>0.71</td>
<td>1.00</td>
<td>0.74</td>
<td>1.00</td>
<td>0.63</td>
<td>0.93</td>
<td>0.71</td>
<td>1.00</td>
<td>0.61</td>
<td>0.90</td>
<td>0.74</td>
<td>1.00</td>
<td>0.58</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FTDR</td>
<td>0.83</td>
<td>1.00</td>
<td>0.80</td>
<td>1.00</td>
<td>0.86</td>
<td>1.00</td>
<td>0.80</td>
<td>1.00</td>
<td>0.80</td>
<td>1.00</td>
<td>0.88</td>
<td>0.75</td>
<td>1.00</td>
<td>0.86</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

9 [https://staq.dsic.upv.es/webstaq/mutuml/experiment_data.htm](https://staq.dsic.upv.es/webstaq/mutuml/experiment_data.htm)

We used the CoSTest option to export the results (faults and coverage analysis) of the testing process of the CS subject. If there are further CS to be studied, Steps 2 to 5 are repeated with the next subject.

### 3.6.7 Analysis of Testing Results

The CoSTest effectiveness and adequacy of the test suite is calculated from the information recorded in this process (see Section 3.6.6). These results are given in the next Section.

### 4 ANALYSIS AND INTERPRETATION OF RESULTS

This section describes the analysis and interpretation of the results related to our response variables for RQ1 and RQ2. The Statistical analysis was carried out on the Statistical Package for Social Sciences (SPSS) V23.0. Since the first research question (RQ1) was aimed at evaluating CoSTest’s effectiveness at detecting faults, we compared the number and types of faults detected for mutant type (i.e., FOM and HOM) in the different CS subjects. Table 4 shows both the number of the faults and the number of fault types detected in each CS subject by mutant type (i.e., FOM and HOM). Shapiro-Wilk tests were performed to evaluate the samples normality. We used this test as our numerical means of assessing normality because it is more appropriate for small sample sizes (<50 samples).

### 4.1 Effectiveness based on Rate of Fault Detection

Since all Sig. values for Shapiro-Wilk tests were 0.165 for FOM and 0.001 for HOM, these variables do not follow a normal distribution (<0.05 for HOM). So, we considered both mutant types as independent groups. Then, the Mann-Whitney U Test was used to test our first null hypothesis (H0). Fig. 6 shows the box-plot containing data on the number of faults per mutant type and Table 5 shows the results of the Mann-Whitney U Test.
4.2 Effectiveness based on Rate of Fault Type Detection

As in the previous analysis, all Sig. values for Shapiro-Wilk tests were 0.234 for FOM and 0 for HOM, which meant these variables did not have a normal distribution (i.e., <0.05 for HOM). Considering both mutant types as independent groups, we selected the Mann-Whitney U Test (non-parametric test) to evaluate the second null hypothesis (Hα). Since the fault type detection rate is different between FOM and HOM (see Fig. 7), we rejected hypothesis Hα. In other words, the number of fault types detected is different for each mutant type; (U = 4, p=0.005<0.05).

4.3 Test Suite Adequacy

In RQ2, we aimed to verify whether the mutation score of CoSTest test suites was the same for killing the different mutant types. To do this, we compared the mutation score for HOMs and FOMs in the seven different CS subjects. Table 6 shows the mutation score summarized for each CS subject and by each mutant type. Tables 7-8 (see Appendix) show the detailed mutation scores for each CS Subject and mutant type (FOM and HOM) respectively.

Table 6: Mutation Score by Mutant Type

<table>
<thead>
<tr>
<th>Mutant Type</th>
<th>VC</th>
<th>MT</th>
<th>SG</th>
<th>ER</th>
<th>OCR</th>
<th>SS</th>
<th>IM</th>
</tr>
</thead>
<tbody>
<tr>
<td>FOM</td>
<td>0.87</td>
<td>0.60</td>
<td>0.73</td>
<td>0.90</td>
<td>0.75</td>
<td>0.82</td>
<td>0.74</td>
</tr>
<tr>
<td>HOM</td>
<td>1.00</td>
<td>1.00</td>
<td>0.89</td>
<td>1.00</td>
<td>0.96</td>
<td>1.00</td>
<td>1.00</td>
</tr>
</tbody>
</table>

Fig. 8 depicts the box-plot of our collected data for mutation score per mutant type. As the results show, the values of mutation score gave a better value for HOM than for FOM.

5 DISCUSSION

Our main results regarding CoSTest’s effectiveness (RQ1) and the adequacy of the test suites (RQ2) are the following: mutant type can influence these two variables, with better effectiveness and test suite adequacy in high order mutants than in first order mutants. So, test suites generated by CoSTest are effective at killing a large number of mutants. However, there are fault types that our test suites cannot detect, as explained below.

Thus, the mutants generated by the WAS2 mutation operator (changes the association type, i.e. normal, composite) and WAS3 mutation operator (changes the member end multiplicity of an
Effectiveness Assessment of an Early Testing Technique using...

Association, i.e. "\textsuperscript{+}, \textsuperscript{-}, 0..1, \textsuperscript{+}0..1" cannot be killed (mutation score=0) by a traditional mutation adequate test set.

Also, the fault types Incorrect Constraint and Incorrect Generalization injected by the mutation operators WC01, WC03, WC04, WC05, WC08 and WGE were hard to detect (mutation score <0.7). This showed the weakness of test cases in testing some constraints, such as derivation rules, which needed to be executed in reverse order when there was a relation between classes that affected the computed result. For example, they first calculated the total of the expense report and then the total of the expense report details. This means these test cases will have to be improved.

Additionally, we found that a lower mutation score for some mutants related with constraints (WCOx) was because the test suites only consider coverage at element level and not at constraint level (i.e. condition branch).

We therefore plan to include test cases with values to make sure that different conditions (e.g. true vs false) will be tested. However, the coverage analysis is important to detect defects when the assertions assert only return values and not side effects (see Fig. 9 in which the coverage analysis is reduced, but all tests still pass.

![Figure 9. Example of an assertion conditional](image)

In addition, we found that CoSTest test suites do not test whether the cardinalities of the association ends meet a certain limit (only creating links according to the test scenario) thereby leading to missed faults, such as an Incorrect Association injected by the WAS3 mutation operator. As well as changing a navigable association to a shared aggregation or vice versa (WAS2) generates an equivalent mutant because "aggregation=shared" has no semantic effect in an executable model using Alf. Thus, another validation technique is required to validate these elements' properties (i.e. inspection of the CS).

Finally, one of the strengths of CoSTest test cases is that it can detect types of defect about misunderstanding requirements (i.e. "Missing" and "Unnecessary" types) that are not normally detected at the CS level, by generating test cases based on user requirements. In a previous work [15] we found a tendency to report only defects related to verification, such as "Wrong" type (e.g. incorrect) rather than defects related to validation.

6 THREATS TO VALIDITY

There are several threats that potentially affect the validity of our study including threats to internal validity, threats to external validity, and threats to construct validity.

Threats to internal validity are conditions that can affect the dependent variables of the experiment without the researcher's knowledge. In our study, the selection of mutation operators is the main threat to internal validity. According to Andrews et al. [1], when using carefully selected mutation operators and after removing equivalent mutants, the mutants can provide a good indication of the fault detection ability of a test suite. Therefore,

in order to minimize this threat we used the MuJava tool [16] to inject faults systematically, by avoiding non-valid and equivalent mutants and optimizing the testing coverage. This tool implements the mutation operators defined in a previous work [12].

Threats to external validity are conditions that limit the ability to generalize the results of our experiments to industrial practice. This threat is reduced by using seven CS of different sizes (see Section 3.5.1) and domain (i.e. information systems, games).

Moreover, a CS was taken from industry, some well-documented CS were found in the literature (i.e. [8], [2] and [7]), and others (i.e. ER, OCR, and VC) were selected because they contained the relevant CS elements required to inject the faults.

Threats to construct validity refer to the suitability of our evaluation metrics. We used well-known metrics to measure the effectiveness (rate of number of faults and number of detected fault types) [28] and the adequacy of the test suites (mutation score) [20]. We therefore believe there is little threat to the construct validity.

7 CONCLUSIONS AND FUTURE WORK

Test cases are important artefacts in any software product as a support to users (e.g. modeler/tester/developer) for checking the reliability of their software product.

In this paper, we evaluated empirically the test cases generated by the CoSTest tool with respect to its effectiveness in terms of its fault detection in Conceptual Schemas and the adequacy of the test suite.

Fault detection effectiveness was measured in terms of rate of faults detection and their causes (fault type) by the test suites. Test suite adequacy was measured in terms of the mutation score value. Our evaluation included the analysis of the variables for mutant types (FOM and HOM).

The Effectiveness and adequacy of the test suites was affected by the mutant type and better results were obtained in detecting faults in HOM. These results suggest that the CoSTest technique is robust in detecting types of defects that are not normally detected at the CS level.

However, some mutation operators achieved a score lower than 0.7 in the mutation score. These results suggest that the test suite should include a test for certain characteristics of CS elements, such as associations, and improve the coverage at the constraint level in order to enhance the effectiveness of the test suites.

In future work we plan to identify features of test cases that would lead to improved effectiveness. We also intend to replicate this experiment on a wide variety of subjects to verify the results, including at least two CS (subjects) per domain.

ACKNOWLEDGMENTS

This work has been developed with the financial support by SENESCYT of the Republic of Ecuador, SHIP (SMEs and HEIs in Innovation Partnerships, ref. EACEA/A2/URB/CL 554187), PERTEST (TIN2013-46928-C3-1-R), European Commission (CaSo project) and Generalitat Valenciana (PROMETEOII/2014/039).
## A APPENDIX

### Table 7. Mutation Score of CoTest Test Suites for First Order Mutants

<table>
<thead>
<tr>
<th>CS</th>
<th>VC</th>
<th>MV</th>
<th>MT</th>
<th>SG</th>
<th>ER</th>
<th>OCR</th>
<th>SS</th>
<th>IM</th>
</tr>
</thead>
<tbody>
<tr>
<td>UPA</td>
<td>6</td>
<td>0</td>
<td>1.00</td>
<td>13</td>
<td>0</td>
<td>1.00</td>
<td>15</td>
<td>0</td>
</tr>
<tr>
<td>WC01</td>
<td>0</td>
<td>0.00</td>
<td>0</td>
<td>6</td>
<td>5</td>
<td>0.82</td>
<td>1</td>
<td>0.00</td>
</tr>
<tr>
<td>MC03</td>
<td>1</td>
<td>0.00</td>
<td>0</td>
<td>6</td>
<td>4</td>
<td>0.41</td>
<td>1</td>
<td>0.00</td>
</tr>
<tr>
<td>WC04</td>
<td>2</td>
<td>0</td>
<td>1.00</td>
<td>7</td>
<td>8</td>
<td>0.19</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>WC05</td>
<td>1</td>
<td>0.00</td>
<td>6</td>
<td>5</td>
<td>5</td>
<td>0.55</td>
<td>8</td>
<td>0.75</td>
</tr>
<tr>
<td>WC06</td>
<td>3</td>
<td>0</td>
<td>1.00</td>
<td>7</td>
<td>5</td>
<td>0.57</td>
<td>9</td>
<td>0.60</td>
</tr>
<tr>
<td>WC07</td>
<td>1</td>
<td>0</td>
<td>1.00</td>
<td>7</td>
<td>4</td>
<td>0.50</td>
<td>9</td>
<td>0.50</td>
</tr>
<tr>
<td>WC08</td>
<td>60</td>
<td>0</td>
<td>1.00</td>
<td>6</td>
<td>20</td>
<td>0.68</td>
<td>21</td>
<td>7</td>
</tr>
<tr>
<td>WC09</td>
<td>5</td>
<td>0</td>
<td>1.00</td>
<td>7</td>
<td>7</td>
<td>0.15</td>
<td>9</td>
<td>0.45</td>
</tr>
<tr>
<td>MID</td>
<td>0</td>
<td>0.00</td>
<td>12</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>13</td>
<td>0</td>
</tr>
<tr>
<td>WC11</td>
<td>5</td>
<td>0</td>
<td>1.00</td>
<td>6</td>
<td>15</td>
<td>0.73</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>WC12</td>
<td>1</td>
<td>0</td>
<td>1.00</td>
<td>3</td>
<td>3</td>
<td>0.00</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>WPA</td>
<td>1</td>
<td>0</td>
<td>1.00</td>
<td>9</td>
<td>5</td>
<td>0.75</td>
<td>15</td>
<td>0</td>
</tr>
<tr>
<td>MCD</td>
<td>25</td>
<td>0</td>
<td>1.00</td>
<td>9</td>
<td>5</td>
<td>0.75</td>
<td>15</td>
<td>0</td>
</tr>
<tr>
<td>MFA</td>
<td>1</td>
<td>0</td>
<td>1.00</td>
<td>10</td>
<td>15</td>
<td>0</td>
<td>23</td>
<td>0</td>
</tr>
<tr>
<td>AB</td>
<td>80</td>
<td>12</td>
<td>0.87</td>
<td>68</td>
<td>17</td>
<td>0.80</td>
<td>122</td>
<td>45</td>
</tr>
</tbody>
</table>

### Table 8. Mutation Score of CoTest Test Suites for High Order Mutants

<table>
<thead>
<tr>
<th>CS</th>
<th>VC</th>
<th>MV</th>
<th>MT</th>
<th>SG</th>
<th>ER</th>
<th>OCR</th>
<th>SS</th>
<th>IM</th>
</tr>
</thead>
<tbody>
<tr>
<td>WC02</td>
<td>3</td>
<td>0</td>
<td>1.00</td>
<td>3</td>
<td>0</td>
<td>1.00</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>WC1</td>
<td>3</td>
<td>0</td>
<td>1.00</td>
<td>3</td>
<td>0</td>
<td>1.00</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>WAT</td>
<td>3</td>
<td>1</td>
<td>0.00</td>
<td>3</td>
<td>0</td>
<td>1.00</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>MS</td>
<td>3</td>
<td>0</td>
<td>1.00</td>
<td>3</td>
<td>0</td>
<td>1.00</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>MA</td>
<td>3</td>
<td>0</td>
<td>1.00</td>
<td>3</td>
<td>0</td>
<td>1.00</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>MAF</td>
<td>3</td>
<td>0</td>
<td>1.00</td>
<td>3</td>
<td>0</td>
<td>1.00</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>AB</td>
<td>20</td>
<td>0</td>
<td>1.00</td>
<td>18</td>
<td>0</td>
<td>1.00</td>
<td>25</td>
<td>0</td>
</tr>
</tbody>
</table>

## REFERENCES


del-driven Architecture in Practice. Berlin Heidelberg.

