# The Oracle Problem in Software Testing: A Survey

Earl T. Barr, Mark Harman, Phil McMinn, Muzammil Shahbaz and Shin Yoo

**Abstract**—Testing involves examining the behaviour of a system in order to discover potential faults. Given an input for a system, the challenge of distinguishing the corresponding desired, correct behaviour from potentially incorrect behavior is called the "test oracle problem". Test oracle automation is important to remove a current bottleneck that inhibits greater overall test automation. Without test oracle automation, the human has to determine whether observed behaviour is correct. The literature on test oracles has introduced techniques for oracle automation, including modelling, specifications, contract-driven development and metamorphic testing. When none of these is completely adequate, the final source of test oracle information remains the human, who may be aware of informal specifications, expectations, norms and domain specific information that provide informal oracle guidance. All forms of test oracles, even the humble human, involve challenges of reducing cost and increasing benefit. This paper provides a comprehensive survey of current approaches to the test oracle problem and an analysis of trends in this important area of software testing research and practice.

**Index Terms**—Test oracle; Automatic testing; Testing formalism.

## 1 Introduction

Much work on software testing seeks to automate as much of the test process as practical and desirable, to make testing faster, cheaper, and more reliable. To this end, we need a test oracle, a procedure that distinguishes between the correct and incorrect behaviors of the System Under Test (SUT).

However, compared to many aspects of test automation, the problem of automating the test oracle has received significantly less attention, and remains comparatively less well-solved. This current open problem represents a significant bottleneck that inhibits greater test automation and uptake of automated testing methods and tools more widely. For instance, the problem of automatically generating test inputs has been the subject of research interest for nearly four decades [46], [108]. It involves finding inputs that cause execution to reveal faults, if they are present, and to give confidence in their absence, if none are found. Automated test input generation been the subject of many significant advances in both Search-Based Testing [3], [5], [83], [127], [129] and Dynamic Symbolic Execution [75], [109], [162]; yet none of these advances address the issue of checking generated inputs with respect to expected behaviours—that is, providing an automated solution to the test oracle problem.

Of course, one might hope that the SUT has been developed under excellent design-for-test principles, so that there might be a detailed, and possibly formal, specification of intended behaviour. One might also hope that the code itself contains pre- and post- conditions that implement well-understood contract-driven development approaches [136]. In these situations, the test oracle cost problem is ameliorated by the presence of an automatable test oracle to which a testing tool can refer to check outputs, free from the need for costly human intervention.

Where no full specification of the properties of the SUT exists, one may hope to construct a

partial test oracle that can answer questions for some inputs. Such partial test oracles can be constructed using metamorphic testing (built from known relationships between desired behaviour) or by deriving oracular information from execution or documentation.

For many systems and most testing as currently practiced in industry, however, the tester does not have the luxury of formal specifications or assertions, or automated partial test oracles [91], [92]. The tester therefore faces the daunting task of manually checking the system's behaviour for all test cases. In such cases, automated software testing approaches must address the human oracle cost problem [1], [82], [131].

To achieve greater test automation and wider uptake of automated testing, we therefore need a concerted effort to find ways to address the test oracle problem and to integrate automated and partially automated test oracle solutions into testing techniques. This paper seeks to help address this challenge by providing a comprehensive review and analysis of the existing literature of the test oracle problem.

Four partial surveys of topics relating to test oracles precede this one. However, none has provided a comprehensive survey of trends and results. In 2001, Baresi and Young [17] presented a partial survey that covered four topics prevalent at the time the paper was published: assertions, specifications, state-based conformance testing, and log file analysis. While these topics remain important, they capture only a part of the overall landscape of research in test oracles, which the present paper covers. Another early work was the initial motivation for considering the test oracle problem contained in Binder's textbook on software testing [23], published in 2000. More recently, in 2009, Shahamiri et al. [165] compared six techniques from the specific category of derived test oracles. In 2011, Staats et al. [174] proposed a theoretical analysis that included test oracles in a revisitation of the fundamentals of testing. Most recently, in 2014, Pezzè et al. focus on automated test oracles for functional properties [151].

Despite this work, research into the test oracle problem remains an activity undertaken in a fragmented community of researchers and practitioners. The role of the present paper is to overcome this fragmentation in this important area of software testing by providing the first comprehensive analysis and review of work on the test oracle problem.

The rest of the paper is organised as follows: Section 2 sets out the definitions relating to test oracles that we use to compare and contrast the techniques in the literature. Section 3 relates a historical analysis of developments in the area. Here we identify key milestones and track the volume of past publications. Based on this data, we plot growth trends for four broad categories of solution to the test oracle problem, which we survey in Sections 4–7. These four categories comprise approaches to the oracle problem where:

- test oracles can be **specified** (Section 4);
- test oracles can be **derived** (Section 5);
- test oracles can be built from **implicit** information (Section 6); and
- **no automatable oracle** is available, yet it is still possible to reduce human effort (Section 7)

Finally, Section 8 concludes with closing remarks.

## 2 DEFINITIONS

This section presents definitions to establish a lingua franca in which to examine the literature on oracles. These definitions are formalised to avoid ambiguity, but the reader should find that it is also possible to read the paper using only the informal descriptions that accompany these formal definitions. We use the theory to clarify the relationship between algebraic specification, pseudo oracles, and metamorphic relations in Section 5.

To begin, we define a test activity as a stimulus or response, then test activity sequences

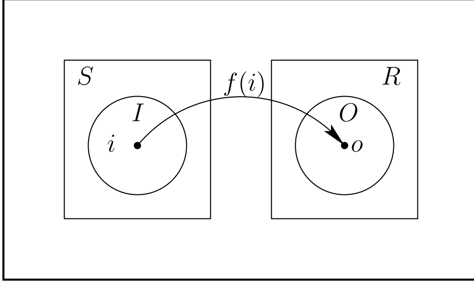


Fig. 1. Stimulus and observations: $S$ is anything that can change the observable behavior of the SUT $f$; $R$ is anything that can be observed about the system's behavior; $I$ includes $f$'s explicit inputs; $O$ is its explicit outputs; everything not in $S \cup R$ neither affects nor is affected by $f$.

that incorporate constraints over stimuli and responses. Test oracles accept or reject test activity sequences, first deterministically then probabilistically. We then define notions of soundness and completeness of test oracles.

### 2.1 Test Activities

To test is to stimulate a system and observe its response. A stimulus and a response both have values, which may coincide, as when the stimulus value and the response are both reals. A system has a set of components $C$. A stimulus and its response target a subset of components. For instance, a common pattern for constructing test oracles is to compare the output of distinct components on the same stimulus value. Thus, stimuli and responses are values that target components. Collectively, stimuli and responses are test activities:

**Definition 2.1** (Test Activities)**.** *For the SUT $p$, $S$ is the set of stimuli that trigger or constrain $p$'s computation and $R$ is the set of observable responses to a stimulus of $p$. $S$ and $R$ are disjoint.* Test activities *form the set $A = S \uplus R$.*

The use of disjoint union implicitly labels the elements of $A$, which we can flatten to the tuple $L \times C \times V$, where $L = \{\text{stimulus}, \text{response}\}$ is the set of activities labels, $C$ is the set of components, and $V$ is an arbitrary set of values. To model those aspects of the world that are independent of any component, like a clock, we set an activity's target to the empty set.

We use the terms "stimulus" and "observation" in the broadest sense possible to cater to various testing scenarios, functional and nonfunctional. As shown in Figure 1, a stimulus can be either an explicit test input from the tester, $I \subset S$, or an environmental factor that can affect the testing, $S \setminus I$. Similarly, an observation ranges from an output of the SUT, $O \subset R$, to a nonfunctional execution profile, like execution time in $R \setminus O$.

For example, stimuli include the configuration and platform settings, database table contents, device states, resource constraints, preconditions, typed values at an input device, inputs on a channel from another system, sensor inputs and so on. Notably, resetting a SUT to an initial state is a stimulus and stimulating the SUT with an input runs it. Observations include anything that can be discerned and ascribed a meaning significant to the purpose of testing — including values that appear on an output device, database state, temporal properties of the execution, heat dissipated during execution, power consumed, or any other measurable attributes of its execution. Stimuli and observations are members of different sets of test activities, but we combine them into test activities.

### 2.2 Test Activity Sequence

Testing is a sequence of stimuli and response observations. The relationship between stimuli and responses can often be captured formally; consider a simple SUT that squares its input. To compactly represent infinite relations between stimulus and response values such as $(i, o = i^2)$, we introduce a compact notation for set comprehensions:

$$x{:}[\phi] = \{x \mid \phi\},$$

where $x$ is a dummy variable over an arbitrary set.

**Definition 2.2** (Test Activity Sequence)**.** *A* test activity sequence *is an element of* $T_A = \{w \mid T \xrightarrow{*} w\}$ *over the grammar*

$$T ::= \; A\,':[' \,\phi\, ']'\, T \mid AT \mid \epsilon$$

*where $A$ is the test activity alphabet.*

Under Definition 2.2, the testing activity sequence $io{:}[o = i^2]$ denotes the stimulus of invoking $f$ on $i$, then observing the response output. It further specifies valid responses obeying $o = i^2$. Thus, it compactly represents the infinite set of test activity sequences $i_1 o_1, i_2 o_2, \cdots$ where $o_k = i_k^2$.

For practical purposes, a test activity sequence will almost always have to satisfy constraints in order to be useful. Under our formalism, these constraints differentiate the approaches to test oracle we survey. As an initial illustration, we constrain a test activity sequence to obtain a practical test sequence:

**Definition 2.3** (Practical Test Sequence)**.** *A* practical test sequence *is any test activity sequence $w$ that satisfies*

$$w = TsTrT, \textit{ for } s \in S, r \in R.$$

Thus, the test activity sequence, $w$, is *practical* iff $w$ contains at least one stimulus followed by at least one observation.

This notion of a test sequence is nothing more than a very general notion of what it means to test; we must do something to the system (the stimulus) and subsequently observe some behaviour of the system (the observation) so that we have something to check (the observation) and something upon which this observed behaviour depends (the stimulus).

A reliable reset $(\mathrm{p}, \mathfrak{r}) \in S$ is a special stimulus that returns the SUT's component $\mathrm{p}$ to its start state. The test activity sequence $(\text{stimulus}, \mathrm{p}, \mathfrak{r})(\text{stimulus}, \mathrm{p}, i)$ is therefore equivalent to the conventional application notation $\mathrm{p}(i)$. To extract the value of an activity, we write $\mathrm{v}(a)$; to extract its target component, we write $c(a)$. To specify two invocations of a single component on the different values, we must write $\mathfrak{r}_1 i_1 \mathfrak{r}_2, i_2 : [\mathfrak{r}_1, i_1, \mathfrak{r}_2, i_2 \in S, c(\mathfrak{r}_1) = c(i_1) = c(\mathfrak{r}_2) = c(i_2) \wedge v(i_1) \neq v(i_2)]$. In the sequel, we often compare different executions of a single SUT or compare the output of independently implemented components of the SUT on the same input value. For clarity, we introduce syntactic sugar to express constraints on stimulus values and components. We let $f(x)$ denote $\mathfrak{r}i : [c(i) = f \wedge \mathrm{v}(i) = x]$, for $f \in C$.

A **test oracle** is a predicate that determines whether a given test activity sequence is an acceptable behaviour of the SUT or not. We first define a "test oracle", and then relax this definition to "probabilistic test oracle".

**Definition 2.4** (Test Oracle)**.** *A* test oracle $D : T_A \nrightarrow \mathbb{B}$ *is a partial*[1] *function from a test activity sequence to true or false.*

When a test oracle is defined for a test activity, it either accepts the test activity or not. Concatenation in a test activity sequence denotes sequential activities; the test oracle $D$ permits parallel activities when it accepts different permutations of the same stimuli and response observations. We use $D$ to distinguish a deterministic test oracle from probabilistic ones. Test oracles are typically computationally expensive, so probabilistic approaches to the provision of oracle information may be desirable even where a deterministic test oracle is possible [125].

**Definition 2.5** (Probabilistic Test Oracle)**.** *A* probabilistic test oracle $\tilde{D} : T_A \nrightarrow [0, 1]$ *maps a test activity sequence into the interval* $[0, 1] \in \mathbb{R}$.

A probabilistic test oracle returns a real number in the closed interval $[0, 1]$. As with test oracles, we do not require a probabilistic test oracle to be a total function. A probabilistic test

1. Recall that a function is implicitly total: it maps every element of its domain to a single element of its range. The *partial* function $f : X \nrightarrow Y$ is the total function $f' : X' \rightarrow Y$, where $X' \subseteq X$.

oracle can model the case where the test oracle is only able to efficiently offer a probability that the test case is acceptable, or for other situations where some degree of imprecision can be tolerated in the test oracle's response.

Our formalism combines a language-theoretic view of stimulus and response activities with constraints over those activities; these constraints explicitly capture specifications. The high-level language view imposes a temporal order on the activities. Thus, our formalism is inherently temporal. The formalism of Staats et al. captures any temporal exercising of the SUT's behavior in tests, which are atomic black boxes for them [174]. Indeed, practitioners write test plans and activities, they do not often write specifications at all, let alone a formal one. This fact and the expressivity of our formalism, as evident in our capture of existing test oracle approaches, is evidence that our formalism is a good fit with practice.

### 2.3 Soundness and Completeness

We conclude this section by defining soundness and completeness of test oracles.

In order to define soundness and completeness of a test oracle, we need to define a concept of the "ground truth", $\mathcal{G}$. The ground truth is another form of oracle, a conceptual oracle, that always gives the "right answer". Of course, it cannot be known in all but the most trivial cases, but it is a useful definition that bounds test oracle behaviour.

**Definition 2.6** (Ground Truth)**.** *The* ground truth oracle*, $\mathcal{G}$, is a total test oracle that always gives the "right answer".*

We can now define soundness and completeness of a test oracle with respect to $\mathcal{G}$.

**Definition 2.7** (Soundness)**.** *The test oracle $D$ is* sound *iff*

$$D(a) \Rightarrow \mathcal{G}(a)$$

**Definition 2.8** (Completeness)**.** *The test oracle $D$ is* complete *iff*

$$\mathcal{G}(a) \Rightarrow D(a)$$

While test oracles cannot, in general, be both sound and complete, we can, nevertheless, define and use partially correct test oracles. Further, one could argue, from a purely philosophical point of view, that human oracles can be sound and complete, or correct. In this view, correctness becomes a subjective human assessment. The foregoing definitions allow for this case.

We relax our definition of soundness to cater for probabilistic test oracles:

**Definition 2.9** (Probablistic Soundness and Completeness)**.** *A probabilistic test oracle $\tilde{D}$ is* probabilistically sound *iff*

$$P(\tilde{D}(w) = 1) > \frac{1}{2} + \epsilon \Rightarrow \mathcal{G}(w)$$

*and $\tilde{D}$ is* probabilistically complete *iff*

$$\mathcal{G}(w) \Rightarrow P(\tilde{D}(w) = 1) > \frac{1}{2} + \epsilon$$

*where $\epsilon$ is non-negligible.*

The non-negligible advantage $\epsilon$ requires $\tilde{D}$ to do sufficiently better than flipping a fair coin, which for a binary classifier maximizes entropy, that we can achieve arbitrary confidence in whether the test sequence $w$ is valid by repeatedly sampling $\tilde{D}$ on $w$.

## 3 Test Oracle Research Trends

The term "test oracle" first appeared in William Howden's seminal work in 1978 [99]. In this section, we analyze the research on test oracles, and its related areas, conducted since 1978. We begin with a synopsis of the volume of publications, classified into specified, derived, implicit, and lack of automated test oracles. We then discuss when key concepts in test oracles were first introduced.

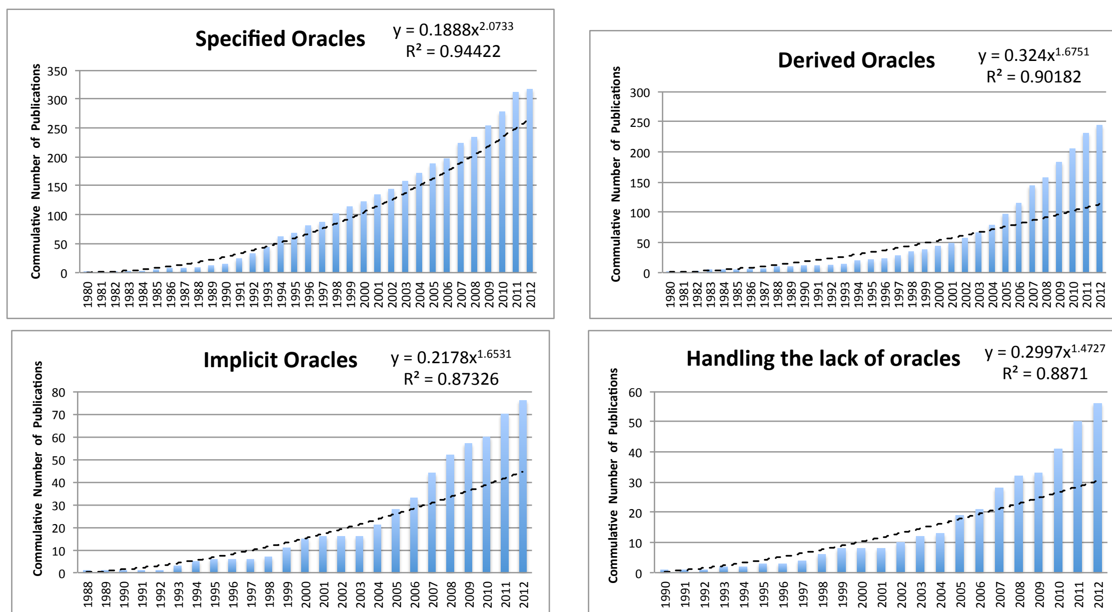


Fig. 2. Cumulative number of publications from 1978 to 2012 and research trend analysis for each type of test oracle. The $x$-axis represents years and $y$-axis the cumulative number of publications. We use a power regression model to perform the trend analysis. The regression equation and the coefficient of determination ($R^2$) indicate a upward future trend, a sign of a healthy research community.

### 3.1 Volume of Publications

We constructed a repository of 694 publications on test oracles and its related areas from 1978 to 2012 by conducting web searches for research articles on *Google Scholar* and *Microsoft Academic Search* using the queries "software + test + oracle" and "software + test oracle"[2], for each year. Although some of the queries generated in this fashion may be similar, different responses are obtained, with particular differences around more lowly-ranked results.

We classify work on test oracles into four categories: specified test oracles (317), derived test oracles (245), implicit test oracles (76), and no test oracle (56), which handles the lack of a test oracle.

Specified test oracles, discussed in detail in Section 4, judge all behavioural aspects of a system with respect to a given formal specification. For specified test oracles we searched for related articles using queries "formal + specification", "state-based specification", "model-based languages", "transition-based languages", "assertion-based languages", "algebraic specification" and "formal + conformance testing". For all queries, we appended the keywords with "test oracle" to filter the results for test oracles.

Derived test oracles (see Section 5) involve artefacts from which a test oracle may be derived — for instance, a previous version of the system. For derived test oracles, we searched for additional articles using the queries "specification inference", "specification

2. We use + to separate the keywords in a query; a phrase, not internally separated by +, like "test oracle", is a compound keyword, quoted when given to the search engine.

mining", "API mining", "metamorphic testing", "regression testing" and "program documentation".

An implicit oracle (see Section 6) refers to the detection of "obvious" faults such as a program crash. For implicit test oracles we applied the queries "implicit oracle", "null pointer + detection", "null reference + detection", "deadlock + livelock + race + detection", "memory leaks + detection", "crash + detection", "performance + load testing", "non-functional + error detection", "fuzzing + test oracle" and "anomaly detection".

There have also been papers researching strategies for handling the lack of an automated test oracle (see Section 7). Here, we applied the queries "human oracle", "test minimization", "test suite reduction" and "test data + generation + realistic + valid".

Each of the above queries were appended by the keywords "software testing". The results were filtered, removing articles that were found to have no relation to software testing and test oracles. Figure 2 shows the cumulative number of publications on each type of test oracle from 1978 onwards. We analyzed the research trend on this data by applying different regression models. The trend line, shown in Figure 2, is fitted using a power model. The high values for the four coefficients of determination ($R^2$), one for each of the four types of test oracle, confirm that our models are good fits to the trend data. The trends observed suggest a healthy growth in research volumes in these topics related to the test oracle problem in the future.

### 3.2 The Advent of Test Oracle Techniques

We classified the collected publications by techniques or concepts they proposed to (partially) solve a test oracle problem; for example, *Model Checking* [35] and *Metamorphic Testing* [36] fall into the derived test oracle and *DAISTIS* [69] is an algebraic specification system that addresses the specified test oracle problem.

For each type of test oracle and the advent of a technique or a concept, we plotted a timeline in chronological order of publications to study research trends. Figure 3 shows the timeline starting from 1978 when the term "test oracle" was first coined. Each vertical bar presents the technique or concept used to solve the problem labeled with the year of its first publication.

The timeline shows only the work that is explicit on the issue of test oracles. For example, the work on test generation using finite state machines (FSM) can be traced back to as early as 1950s. But the explicit use of finite state machines to generate test oracles can be traced back to Jard and Bochmann [103] and Howden in 1986 [98]. We record, in the timeline, the earliest available publication for a given technique or concept. We consider only published work in journals, the proceedings of conferences and workshops, or magazines. We excluded all other types of documentation, such as technical reports and manuals.

Figure 3 shows a few techniques and concepts that predate 1978. Although not explicitly on test oracles, they identify and address issues for which test oracles were later developed. For example, work on detecting concurrency issues (deadlock, livelock, and races) can be traced back to the 1960s. Since these issues require no specification, implicit test oracles can and have been built that detect them on arbitrary systems. Similarly, *Regression Testing* detects problems in the functionality a new version of a system shares with its predecessors and is a precursor of derived test oracles.

The trend analysis suggests that proposals for new techniques and concepts for the formal specification of test oracles peaked in 1990s, and has gradually diminished in the last decade. However, it remains an area of much research activity, as can be judged from the number of publications for each year in Figure 2. For derived test oracles, many solutions have been proposed throughout this period. Initially, these solutions were primarily theoretical, such as Partial/Pseudo-Oracles [196] and

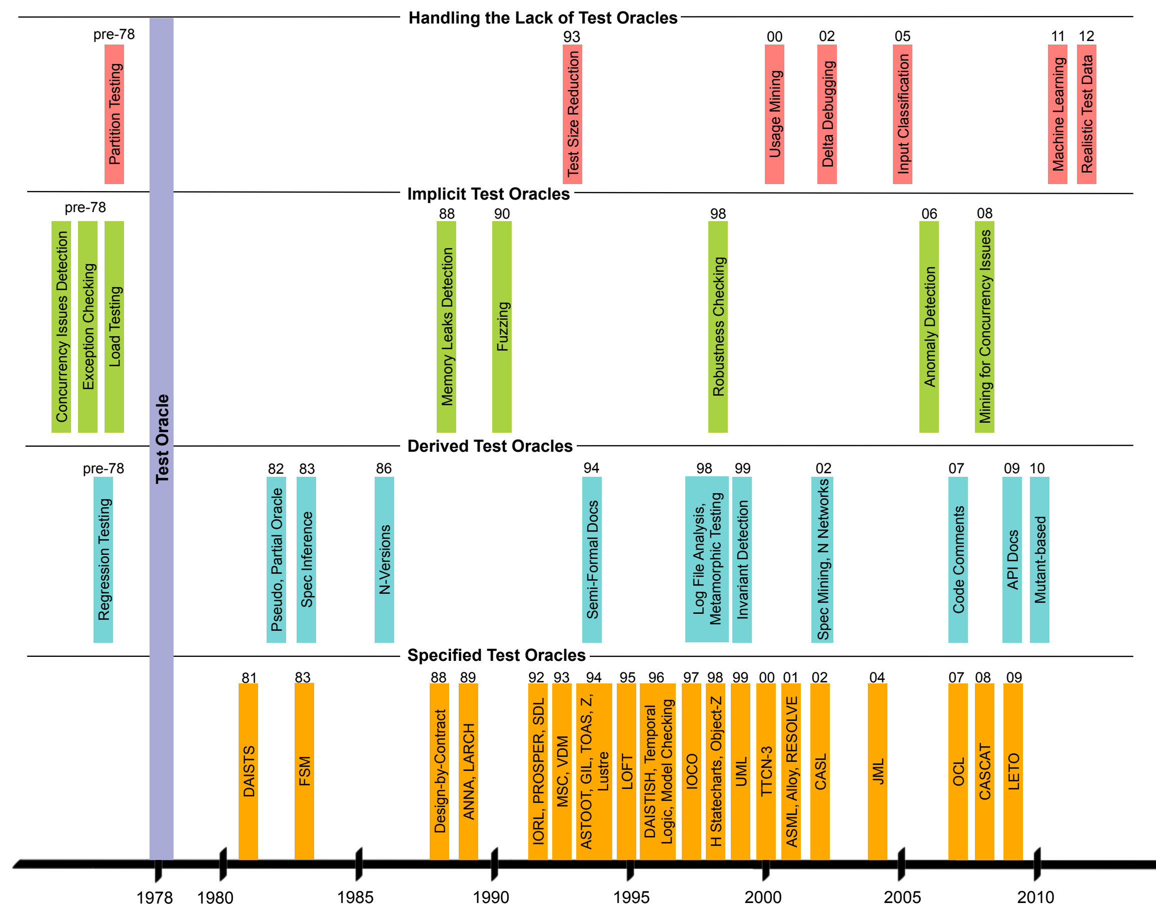


Fig. 3. Chronological introduction of test oracles techniques and concepts.

Specification Inference [194]; empirical studies however followed in late 1990s.

For implicit test oracles, research into the solutions established before 1978 has continued, but at a slower pace than the other types of test oracles. For handling the lack of an automated test oracle, *Partition Testing* is a well-known technique that helps a human test oracle select tests. The trend line suggests that only recently have new techniques and concepts for tackling this problem started to emerge, with an explicit focus on the human oracle cost problem.

## 4 Specified Test Oracles

Specification is fundamental to computer science, so it is not surprising that a vast body of research has explored its use as a source of test oracle information. This topic could merit an entire survey on its own right. In this section, we provide an overview of this work. We also include here partial specifications of system behaviour such as assertions and models.

A specification defines, if possible using mathematical logic, the test oracle for particular domain. Thus, a specification language is a notation for defining a *specified test oracle* $D$, which judges whether the behaviour of a

system conforms to a formal specification. Our formalism, defined in Section 2, is, itself, a specification language for specifying test oracles.

Over the last 30 years, many methods and formalisms for testing based on formal specification have been developed. They fall into four broad categories: model-based specification languages, state transition systems, assertions and contracts, and algebraic specifications. Model-based languages define models and a syntax that defines desired behavior in terms of its effect on the model. State transition systems focus on modeling the reaction of a system to stimuli, referred to as "transitions" in this particular formalism. Assertions and contracts are fragments of a specification language that are interleaved with statements of the implementation language and checked at runtime. Algebraic specifications define equations over a program's operations that hold when the program is correct.

## 4.1 Specification Languages

Specification languages define a mathematical model of a system's behaviour, and are equipped with a formal semantics that defines the meaning of each language construct in terms of the model. When used for testing, models do not usually fully specify the system, but seek to capture salient properties of a system so that test cases can be generated from or checked against them.

### *4.1.1 Model-Based Specification Languages*

Model-based specification languages model a system as a collection of states and operations to alter these states, and are therefore also referred to as "state-based specifications" in the literature [101], [110], [182], [183]. Preconditions and postconditions constrain the system's operations. An operation's precondition imposes a necessary condition over the input states that must hold in a correct application of the operation; a postcondition defines the (usually strongest) effect the operation has on program state [110].

A variety of model-based specification languages exist, including Z [172], B [111], UML/OCL [31], VDM/VDM-SL [62], Alloy [102], and the LARCH family [71], which includes an algebraic specification sub-language. Broadly, these languages have evolved toward being more concrete, closer to the implementation languages programmers use to solve problems. Two reasons explain this phenomenon: the first is the effort to increase their adoption in industry by making them more familiar to practitioners and the second is to establish synergies between specification and implementation that facilitate development as iterative refinement. For instance, Z models disparate entities, like predicates, sets, state properties, and operations, through a single structuring mechanism, its schema construct; the B method, Z's successor, provides a richer array of less abstract language constructs.

Börger discusses how to use the abstract state machine formalism, a very general set-theoretic specification language geared toward the definition of functions, to define high level test oracles [29]. The models underlying specification languages can be very abstract, quite far from concrete execution output. For instance, it may be difficult to compute whether a model's postcondition for a function permits an observed concrete output. If this impedance mismatch can be overcome, by abstracting a system's concrete output or by concretizing a specification model's output, and if a specification's postconditions can be evaluated in finite time, they can serve as a test oracle [4].

Model-based specification languages, such as VDM, Z, and B can express invariants, which can drive testing. Any test case that causes a program to violate an invariant has discovered an incorrect behavior; therefore, these invariants are partial test oracles.

In search of a model-based specification language accessible to domain experts, Parnas

et al. proposed TOG (Test Oracles Generator) from program documentation [143], [146], [149]. In their method, the documentation is written in fully formal tabular expressions in which the method signature, the external variables, and relation between its start and end states are specified [105]. Thus, test oracles can be automatically generated to check the outputs against the specified states of a program. The work by Parnas et al. has been developed over a considerable period of more than two decades [48], [59], [60], [145], [150], [190], [191].

#### *4.1.2 State Transition Systems*

State transition systems often present a graphical syntax, and focus on transitions between different states of the system. Here, states typically abstract sets of concrete state of the modeled system. State transition systems have been referred as visual languages in the literature [197]. A wide variety of state transition systems exist, including Finite State Machines [112], Mealy/Moore machines [112], I/O Automata [118], Labeled Transition Systems [180], SDL [54], Harel Statecharts [81], UML state machines [28], X-Machines [95], [96], Simulink/Stateflow [179] and PROMELA [97]. Mouchawrab et al. conducted a rigorous empirical evaluation of test oracle construction techniques using state transition systems [70], [138].

An important class of state transition systems have a finite set of states and are therefore particularly well-suited for automated reasoning about systems whose behaviour can be abstracted into states defined by a finite set of values [93]. State transition systems capture the behavior of a system under test as a set of states[3], with transitions representing stimuli that cause the system to change state. State transition systems model the output of a system they abstract either as a property of the states (the final state in the case of Moore machines) or the transitions traversed (as with Mealy machines).

Models approximate a SUT, so behavioral differences between the two are inevitable. Some divergences, however, are spurious and falsely report testing failure. State-transition models are especially susceptible to this problem when modeling embedded systems, for which time of occurrence is critical. Recent work model tolerates spurious differences in time by "steering" model's evaluation: when the SUT and its model differ, the model is backtracked, and a steering action, like modifying timer value or changing inputs, is applied to reduce the distance, under a similarity measure [74].

Protocol conformance testing [72] and, later, model-based testing [183] motivated much of the work applying state transition systems to testing. Given a specification $F$ as a state transition system, e.g. a finite state machine, a test case can be extracted from sequences of transitions in $F$. The transition labels of such a sequence define an input. A test oracle can then be constructed from $F$ as follows: if $F$ accepts the sequence and outputs some value, then so should the system under test; if $F$ does not accept the input, then neither should the system under test.

Challenges remain, however, as the definition of conformity comes in different flavours, depending on whether the model is deterministic or non-deterministic and whether the behaviour of the system under test on a given test case is observable and can be interpreted at the same level of abstraction as the model's. The resulting flavours of conformity have been captured in alternate notions, in terms of whether the system under test is isomorphic to, equivalent to, or quasi-equivalent to $F$. These notions of conformity were defined in the mid-1990s in the famous survey paper by Lee and Yannakakis [112] among other notable papers, including those by Bochmann et al. [26] and

3. Unfortunately, the term 'state' has different interpretation in the context of test oracles. Often, it refers to a 'snapshot' of the configuration of a system at some point during its execution; in context of state transition systems, however, 'state' typically refers to an abstraction of a set of configurations, as noted above.

Tretmans [180].

### 4.2 Assertions and Contracts

An assertion is a boolean expression that is placed at a certain point in a program to check its behaviour at runtime. When an assertion evaluates to true, the program's behaviour is regarded "as intended" at the point of the assertion, for that particular execution; when an assertion evaluates to false, an error has been found in the program for that particular execution. It is obvious to see how assertions can be used as a test oracle.

The fact that assertions are embedded in an implementation language has two implications that differentiate them from specification languages. First, assertions can directly reference and define relations over program variables, reducing the impedance mismatch between specification and implementation, for the properties an assertion can express and check. In this sense, assertions are a natural consequence of the evolution of specification languages toward supporting development through iterative refinement. Second, they are typically written along with the code whose runtime behavior they check, as opposed to preceding implementation as specification languages tend to do.

Assertions have a long pedigree dating back to Turing [181], who first identified the need to separate the tester from the developer and suggested that they should communicate by means of assertions: the developer writing them and the tester checking them. Assertions gained significant attention as a means of capturing language semantics in the seminal work of Floyd [64] and Hoare [94] and subsequently were championed as a means of increasing code quality in the development of the contract-based programming approach, notably in the language Eiffel [136].

Widely used programming languages now routinely provide assertion constructs; for instance, C, C++, and Java provide a construct called *assert* and C# provides a *Debug.Assert* method. Moreover, a variety of systems have been independently developed for embedding assertions into a host programming languages, such as Anna [117] for Ada, APP [156] and Nana [120] for C languages.

In practice, assertion approaches can check only a limited set of properties at a certain point in a program [49]. Languages based on *design by contract* principles extend the expressivity of assertions by providing means to check contracts between *client* and *supplier* objects in the form of method pre- and post-conditions and class invariants. Eiffel was the first language to offer design by contract [136], a language feature that has since found its way into other languages, such as Java in the form of Java modeling language (JML) [140].

Cheon and Leavens showed how to construct an assertion-based test oracle on top of JML [45]. For more on assertion-based test oracles, see Coppit and Haddox-Schatz's evaluation [49], and, later, a method proposed by Cheon [44]. Both assertions and contracts are enforced observation activity that are embedded into the code. Araujo et al. provide a systematic evaluation of design by contract on a large industrial system [9] and using JML in particular [8]; Briand et al. showed how to support testing by instrumenting contracts [33].

### 4.3 Algebraic Specification Languages

Algebraic specification languages define a software module in terms of its interface, a signature consisting of sorts and operation symbols. Equational axioms specify the required properties of the operations; their equivalence is often computed using term rewriting [15]. Structuring facilities, which group sorts and operations, allow the composition of interfaces. Typically, these languages employ first-order logic to prove properties of the specification, like the correctness of refinements. Abstract data types (ADT), which combine data and operations over that data, are well-suited to algebraic specification.

One of the earliest algebraic specification systems, for implementing, specifying and testing ADTs, is DAISTS [69]. In this system, equational axioms generally equate a term-rewriting expression in a restricted dialect of ALGOL 60 against a function composition in the implementation language. For example, consider this axiom used in DAISTS:

$$\begin{aligned}&\mathit{Pop2}(\mathit{Stack}\ S, \mathit{EltType}\ I):\\ &\qquad \mathit{Pop}(\mathit{Push}(S, I)) = \textbf{if}\ \mathit{Depth}(S) = \mathit{Limit}\\ &\qquad\qquad \textbf{then}\ \mathit{Pop}(S)\\ &\qquad\qquad \textbf{else}\ S;\end{aligned}$$

This axiom is taken from a specification that differentiates the accessor *Top*, which returns the top element of a stake without modifying the stack, and the mutator *Pop*, which returns a new stack lacking the previous top element. A test oracle simply executes both this axiom and its corresponding composition of implemented functions against a test suite: if they disagree, a failure has been found in the implementation or in the axiom; if they agree, we gain some assurance of their correctness.

Gaudel and her colleagues [19], [20], [72], [73] were the first to provide a general testing theory founded on algebraic specification. Their idea is that an exhaustive test suite composed only of ground terms, i.e., terms with no free variables, would be sufficient to judge program correctness. This approach faces an immediate problem: the domain of each variable in a ground term might be infinite and generate an infinite number of test cases. Test suites, however, must be finite, a practical limitation to which all forms of testing are subject. The workaround is, of course, to abandon exhaustive coverage of all bindings of values to ground terms and select a finite subset of test cases [20].

Gaudel's theory focuses on observational equivalence. Observational inequivalence is, however, equally important [210]. For this reason, Frankl and Doong extended Gaudel's theory to express inequality as well an equality [52]. They proposed a notation that is suitable for object-oriented programs and developed an algebraic specification language called LOBAS and a test harness called ASTOOT. In addition to handling object-orientation, Frankl and Doong require classes to implement the testing method `EQN` that ASTOOT uses to check the equivalence or inequivalence of two instances of a given class. From the vantage point of an observer, an object has observable and unobservable, or hidden, state. Typically, the observable state of an object is its public fields and method return values. `EQN` enhances the testability of code and enables ASTOOT to approximate the observational equivalence of two objects on a sequence of messages, or method calls. When ASTOOT checks the equivalence of an object and a specification in LOBAS, it realizes a specified test oracle.

Expanding upon ASTOOT, Chen et al. [40] [41] built TACCLE, a tool that employs a white-box heuristic to generate a relevant, finite number of test cases. Their heuristic builds a data relevance graph that connects two fields of a class if one affects the other. They use this graph to consider only that can affect an observable attributes of a class when considering the (in)equivalence of two instances. Algebraic specification has been a fruitful line of research; many algebraic specification languages and tools exist, including Daistish [100], LOFT [123], CASL [11], CASCAT [205]. The projects have been evolving toward testing a wider array of entities, from ADTS, to classes, and most recently, components; they also differ in their degree of automation of test case generation and test harness creation. Bochmann et al. used LOTOS to realise test oracle functions from algebraic specifications [184]; most recently, Zhu also considered the use of algebraic specifications as test oracles [210].

#### *4.3.1 Specified Test Oracle Challenges*

Three challenges must be overcome to build specified test oracles. The first is the lack of a formal specification. Indeed, the other classes

of test oracles, discussed in this survey, all address the problem of test oracle construction in the absence of a formal specification. Formal specifications models necessarily rely on abstraction that can lead to the second problem: imprecision, models that include infeasible behavior or that do not capture all the behavior relevant to checking a specification [68]. Finally, one must contend with the problem of interpreting model output and equating it to concrete program output.

Specified results are usually quite abstract, and the concrete test results of a program's executions may not be represented in a form that makes checking their equivalence to the specified result straightforward. Moreover, specified results can be partially represented or oversimplified. This is why Gaudel remarked that the existence of a formal specification does not guarantee the existence of a successful test driver [72]. Formulating concrete equivalence functions may be necessary to correctly interpret results [119]. In short, solutions to this problem of equivalence across abstraction levels depend largely on the degree of abstraction and, to a lesser extent, on the implementation of the system under test.

# 5 Derived Test Oracles

A *derived test oracle* distinguishes a system's correct from incorrect behavior based on information derived from various artefacts (e.g. documentation, system executions) or properties of the system under test, or other versions of it. Testers resort to derived test oracles when specified test oracles are unavailable, which is often the case, since specifications rapidly fall out of date when they exist at all. Of course, the derived test oracle might become a partial "specified test oracle", so that test oracles derived by the methods discussed in this section could migrate, over time, to become, those considered to be the "specified test oracles" of the previous section. For example, JWalk incrementally learns algebraic properties of the class under test [170]. It allows interactive confirmation from the tester, ensuring that the human is in the "'learning loop".

The following sections discuss research on deriving test oracles from development artefacts, beginning in Section 5.1 with pseudo-oracles and N-version programming, which focus on agreement among independent implementations. Section 5.2 then introduces metamorphic relations which focuses on relations that must hold among distinct executions of a single implementation. Regression testing, Section 5.3, focuses on relations that should hold across different versions of the SUT. Approaches for inferring models from system executions, including invariant inference and specification mining, are described in Section 5.4. Section 5.5 closes with a discussion of research into extracting test oracle information from textual documentation, like comments, specifications, and requirements.

## 5.1 Pseudo-Oracles

One of the earliest versions of a derived test oracle is the concept of a pseudo-oracle, introduced by Davis and Weyuker [50], as a means of addressing so-called *non-testable programs*:

> "Programs which were written in order to determine the answer in the first place. There would be no need to write such programs, if the correct answer were known." [196].

A pseudo-oracle is an alternative version of the program produced independently, e.g. by a different programming team or written in an entirely different programming language. In our formalism (Section 2), a pseudo-oracle is a test oracle $D$ that accepts test activity sequences of the form

$$f_1(x)o_1f_2(x)o_2 : [f_1 \neq f_2 \wedge o_1 = o_2], \qquad (1)$$

where $f_1, f_2 \in C$, the components of the SUT (Section 2), are alternative, independently produced, versions of the SUT on the same value. We draw the reader's attention to the similarity

between pseudo-oracles and algebraic specification systems (Section 4.3), like DIASTIS, where the function composition expression in the implementation language and the term-rewriting expression are distinct implementations whose output must agree and form a pseudo-oracle.

A similar idea exists in fault-tolerant computing, referred to as multi- or N-version programming [13], [14], where the software is implemented in multiple ways and executed in parallel. Where results differ at run-time, a "voting" mechanism decides which output to use. In our formalism, an N-version test oracle accepts test activities of the following form:

$$\begin{aligned} &f_1(x)o_1 f_2(x)o_2 \cdots f_k(x)o_k : \\ &[\forall i, j \in [1..k],\ i \neq j \Rightarrow f_i \neq f_j \\ &\wedge \arg\max_{o_i} m(o_i) \geq t] \end{aligned} \tag{2}$$

In Equation 2, the outputs form a multiset and $m$ is the multiplicity, or number of repetitions of an element in the multiset. The $\arg\max$ operator finds the argument that maximizes a function's output, here an output with greatest multiplicity. Finally, the maximum multiplicity is compared against the threshold $t$. We can now define a N-version test oracle as $D_{nv}(w, x)$ where $w$ obeys Equation 2 with $t$ bound to $x$. Then $D_{\text{maj}}(w) = D_{nv}(w, \lceil \frac{k}{2} \rceil)$ is an N-version oracle that requires a majority of the outputs to agree and $D_{pso}(w) = D_{nv}(w, k)$ generalizes pseudo oracles to agreement across $k$ implementations.

More recently, Feldt [58] investigated the possibility of automatically producing different versions using genetic programming, and McMinn [128] explored the idea of producing different software versions for testing through program transformation and the swapping of different software elements with those of a similar specification.

### 5.2 Metamorphic Relations

For the SUT `p` that implements the function $f$, a *metamorphic relation* is a relation over applications of $f$ that we expect to hold across *multiple* executions of `p`. Suppose $f(x) = e^x$, then $e^a e^{-a} = 1$ is a metamorphic relation. Under this metamorphic relation, `p(0.3) * p(-0.3) = 1` will hold if `p` is correct [43]. The key idea is that reasoning about the properties of $f$ will lead us to relations that its implementation `p` must obey.

*Metamorphic testing* is a process of exploiting metamorphic relations to generate partial test oracles for *follow-up* test cases: it checks important properties of the SUT after certain test cases are executed [36]. Although metamorphic relations are properties of the ground truth, the correct phenomenon ($f$ in the example above) that a SUT seeks to implement and could be considered a mechanism for creating specified test oracles. We have placed them with derived test oracles, because, in practice, metamorphic relations are usually manually inferred from a white-box inspection of a SUT.

Metamorphic relations differ from algebraic specifications in that a metamorphic relation relates different executions, not necessarily on the same input, of the same implementation relative to its specification, while algebraic specifications equates two distinct implementations of the specification, one written in an implementation language and the other written in formalism free of implementation details, usually term rewriting [15].

Under the formalism of Section 2, a metamorphic relation is

$$f(x_1)o_1 f(x_2)o_2 \cdots f(i_k)o_k : [expr \wedge k \geq 2],$$

where $expr$ is a constraint, usually arithmetic, over the inputs $x_i$ and $o_x$. This definition makes clear that a metamorphic relation is a constraint on the values of stimulating the single SUT $f$ at least twice, observing the responses, and imposing a constraint on how they interrelate. In contrast, algebraic specification is a type of pseudo-oracle, as specified in Equation 1, which stimulates two distinct implementations on the same value, requiring their output to be equivalent.

It is often thought that metamorphic relations need to concern numerical properties that can be captured by arithmetic equations, but metamorphic testing is, in fact, more general. For example, Zhou et al. [209] used metamorphic testing to test search engines such as Google and Yahoo!, where the relations considered are clearly non-numeric. Zhou et al. build metamorphic relations in terms of the consistency of search results. A motivating example they give is of searching for a paper in the ACM digital library: two attempts, the second quoted, using advanced search fail, but a general search identical to the first succeeds. Using this insight, the authors build metamorphic relations, like $R_{OR} : A_1 = (A_2 \cup A_3) \Rightarrow |A_2| \leq |A_1|$, where the $A_i$ are sets of web pages returned by queries. Metamorphic testing is also means of testing Weyuker's "non-testable programs", introduced in the last section.

When the SUT is nondeterministic, such as a classifier whose exact output varies from run to run, defining metamorphic relations solely in terms of output equality is usually insufficient during metamorphic testing. Murphy et al. [139], [140] investigate relations other than equality, like set intersection, to relate the output of stochastic machine learning algorithms, such as classifiers. Guderlei and Mayer introduced statistical metamorphic testing, where the relations for test output are checked using statistical analysis [80], a technique later exploited to apply metamorphic testing to stochastic optimisation algorithms [203].

The biggest challenge in metamorphic testing is automating the discovery of metamorphic relations. Some of those in the literature are mathematical [36], [37], [42] or combinatorial [139], [140], [161], [203]. Work on the discovery of algebraic specifications [88] and JWalk's lazy systematic unit testing, in which the specification is lazily, and incrementally, learned through interactions between JWalk and the developer [170] might be suitable for adaptation to the discovery metamorphic relations. For instance, the programmer's development environment might track relationships among the output of test cases run during development, and propose ones that hold across many runs to the developer as possible metamorphic relations. Work has already begun that exploits domain knowledge to formulate metamorphic relations [38], but it is still at an early stage and not yet automated.

### 5.3 Regression Test Suites

Regression testing aims to detect whether the modifications made to the new version of a SUT have disrupted existing functionality [204]. It rests on the implicit assumption that the previous version can serve as an oracle for existing functionality.

For corrective modifications, desired functionality remains the same so the test oracle for version $i$, $D_i$, can serve as the next version's test oracle, $D_{i+1}$. Corrective modifications may fail to correct the problem they seek to address or disrupt existing functionality; test oracles may be constructed for these issues by symbolically comparing the execution of the faulty version against the newer, allegedly fixed version [79]. Orstra generates assertion-based test oracles by observing the program states of the previous version while executing the regression test suite [199]. The regression test suite, now augmented with assertions, is then applied to the newer version. Similarly, *spectra*-based approaches use the program and value spectra obtained from the original version to detect regression faults in the newer versions [86], [200].

For perfective modifications, those that add new features to the SUT, $D_i$ must be modified to cater for newly added behaviours, i.e. $D_{i+1} = D_i \cup \Delta D$. Test suite augmentation techniques specialise in identifying and generating $\Delta D$ [6], [132], [202]. However, more work is required to develop these augmentation techniques so that they augment, not merely the test input, but also the expected output. In this way, test suite augmentation could be extended

to augment the existing *oracles* as well as the test data.

Changes in the specification, which is deemed to fail to meet requirements perhaps because the requirements have themselves changed, drives another class of modifications. These changes are generally regarded as "perfective" maintenance in the literature but no distinction is made between perfections that add new functionality to code (without changing requirements) and those changes that arise due to changed requirements (or incorrect specifications).

Our formalisation of test oracles in Section 2 forces a distinction of these two categories of perfective maintenance, since the two have profoundly different consequences for test oracles. We therefore refer to this new category of perfective maintenance as "changed requirements". Recall that, for the function $f : X \to Y$, $\mathrm{dom}(f) = X$. For changed requirements:

$$\exists \alpha \cdot D_{i+1}(\alpha) \neq D_i(\alpha),$$

which implies, of course, $\mathrm{dom}(D_{i+1}) \cap \mathrm{dom}(D_i) \neq \emptyset$ and the new test oracle cannot simply union the new behavior with the old test oracle. Instead, we have

$$D_{i+1} = \begin{cases} \Delta D & \text{if } \alpha \in \mathrm{dom}(\Delta D) \\ D_i & \text{otherwise.} \end{cases}$$

## 5.4 System Executions

A system execution trace can be exploited to derive test oracles or to reduce the cost of a human test oracle by aligning an incorrect execution against the expected execution, as expressed in temporal logic [51]. This section discusses the two main techniques for deriving test oracles from traces — invariant detection and specification mining. Derived test oracles can be built on both techniques to automatically check expected behaviour similar to assertion-based specification, discussed in Section 4.2.

### 5.4.1 Invariant Detection

Program behaviours can be automatically checked against invariants. Thus, invariants can serve as test oracles to help determine the correct and incorrect outputs.

When invariants are not available for a program in advance, they can be learned from the program (semi-) automatically. A well-known technique proposed by Ernst et al. [56], implemented in the Daikon tool [55], is to execute a program on a collection of inputs (test cases) against a collection of potential invariants. The invariants are instantiated by binding their variables to the program's variables. Daikon then dynamically infers *likely* invariants from those invariants not violated during the program executions over the inputs. The inferred invariants capture program behaviours, and thus can be used to check program correctness. For example, in regression testing, invariants inferred from the previous version can be checked as to whether they still hold in the new version.

In our formalism, Daikon invariant detection can define an unsound test oracle that gathers likely invariants from the prefix of a testing activity sequence, then enforces those invariants over its suffix. Let $I_j$ be the set of likely invariants at observation $j$; $I_0$ are the initial invariants; for the test activity sequence $r_1 r_2 \cdots r_n$, $I_n = \{x \in I \mid \forall i \in [1..n], r_i \models x\}$, where $\models$ is logical entailment. Thus, we take an observation to define a binding of the variables in the world under which a likely invariant either holds or does not: only those likely invariants remain that no observation invalidates. In the suffix $r_{n+1} r_{n+2} \cdots r_m$, the test oracle then changes gear and accepts only those activities whose response observations obey $I_n$, i.e. $r_i : [r_i \models I_n], i > n$.

Invariant detection can be computationally expensive, so incremental [22], [171] and light weight static analyses [39], [63] have been brought to bear. A technical report summarises various dynamic analysis techniques [158]. Model inference [90], [187] could also be re-

garded as a form of invariant generation in which the invariant is expressed as a model (typically as an FSM). Ratcliff et al. used Search-Based Software Engineering (SBSE) [84] to search for invariants, guided by mutation testing [154].

The accuracy of inferred invariants depends in part on the quality and completeness of the test cases; additional test cases might provide new data from which more accurate invariants can be inferred [56]. Nevertheless, inferring "perfect" invariants is almost impossible with the current state of the art, which tends to frequently infer incorrect or irrelevant invariants [152]. Wei et al. recently leveraged existing contracts in Eiffel code to infer postconditions on commands (as opposed to queries) involving quantification or implications whose premises are conjunctions of formulae [192], [193].

Human intervention can, of course, be used to filter the resulting invariants, i.e., retaining the correct ones and discarding the rest. However, manual filtering is error-prone and the misclassification of invariants is frequent. In a recent empirical study, Staats et al. found that half of the incorrect invariants Daikon inferred from a set of Java programs were misclassified [175]. Despite these issues, research on the dynamic inference of program invariants has exhibited strong momentum in the recent past with the primary focus on its application to test generation [10], [142], [207].

#### 5.4.2 Specification Mining

Specification mining or inference infers a formal model of program behaviour from a set of observations. In terms of our formalism, a test oracle can enforce these formal models over test activities. In her seminal work on using inference to assess test data adequacy, Weyuker connected inference and testing as inverse processes [194]. The testing process starts with a program, and looks for I/O pairs that characterise every aspect of both the intended and actual behaviours, while inference starts with a set of I/O pairs, and derives a program to fit the given behaviour. Weyuker defined this relation for assessing test adequacy which can be stated informally as follows.

A set of I/O pairs $T$ is an *inference adequate test set* for the program $P$ intended to fulfil specification $S$ iff the program $I_T$ inferred from $T$ (using some inference procedure) is equivalent to both $P$ and $S$. Any difference would imply that the inferred program is not equivalent to the actual program and, therefore, that the test set $T$ used to infer the program $P$ is inadequate.

This inference procedure mainly depends upon the set of I/O pairs used to infer behaviours. These pairs can be obtained from system executions either passively, e.g., by runtime monitoring, or actively, e.g., by querying the system [106]. However, equivalence checking is undecidable in general, and therefore inference is only possible for programs in a restricted class, such as those whose behaviour can be modelled by finite state machines [194]. With this, equivalence can be accomplished by experiment [89]. Nevertheless, serious practical limitations are associated with such experiments (see the survey by Lee and Yannakakis [112] for complete discussion).

The marriage between inference and testing has produced wealth of techniques, especially in the context of "black-box" systems, when source code/behavioural models are unavailable. Most work has applied $L^*$, a well-known learning algorithm, to learn a black-box system $B$ as a finite state machine (FSM) with $n$ states [7]. The algorithm infers an FSM by iteratively querying $B$ and observing the corresponding outputs. A string distinguishes two FSMs when only one of the two machines ends in a final state upon consuming the string. At each iteration, an inferred model $M_i$ with $i < n$ states is given. Then, the model is refined with the help of a string that distinguishes $B$ and $M_i$ to produce a new model, until the number of states reaches $n$.

Lee and Yannakakis [112] showed how to use $L^*$ for conformance testing of $B$ with a

specification $S$. Suppose $L^*$ starts by inferring a model $M_i$, then we compute a string that distinguishes $M_i$ from $S$ and refine $M_i$ through the algorithm. If, for $i = n$, $M_n$ is $S$, then we declare $B$ to be correct, otherwise faulty.

Apart from conformance testing, inference techniques have been used to guide test generation to focus on particular system behavior and to reduce the scope of analysis. For example, Li et al. applied $L^*$ to the integration testing of a system of black-box components [114]. Their analysis architecture derives a test oracle from a test suite by using $L^*$ to infer a model of the systems from dynamically observing system's behavior; this model is then searched to find incorrect behaviors, such as deadlocks, and used to verify the system's behaviour under fuzz testing (Section 6).

To find concurrency issues in asynchronous black-box systems, Groz et al. proposed an approach that extracts behavioural models from systems through active learning techniques [78] and then performs reachability analysis on the models [27] to detect issues, notably races.

Further work in this context has been compiled by Shahbaz [166] with industrial applications. Similar applications of inference can be found in system analysis [21], [78], [135], [188], [189], component interaction testing [115], [122], regression testing [200], security testing [168] and verification [53], [77], [148].

Zheng et al. [208] extract item sets from web search queries and their results, then apply association rule mining to infer rules. From these rules, they construct derived test oracles for web search engines, which had been thought to be untestable. Image segmentation delineates objects of interest in an image; implementing segmentation programs is a tedious, iterative process. Frouchni et al. successfully apply semi-supervised machine learning to create test oracles for image segmentation programs [67]. Memon et al. [133], [134], [198] introduced and developed the GUITAR tool, which has been evaluated by treating the current version of the SUT as correct, inferring the specification, and then executing the generated test inputs. Artificial Neural Networks have also been applied to learn system behaviour and detect deviations from it [163], [164].

The majority of specification mining techniques adopt Finite State Machines as the output format to capture the functional behaviour of the SUT [21], [27], [53], [77], [78], [89], [112], [114], [135], [148], [166], [168], [189], sometimes extended with temporal constraints [188] or data constraints [115], [122] which are, in turn, inferred by Daikon [56]. Büchi automata have been used to check properties against black-box systems [148]. Annotated call trees have been used to represent the program behaviour of different versions in the regression testing context [200]. GUI widgets have been directly modelled with objects and properties for testing [133], [134], [198]. Artificial Neural Nets and machine learning classifiers have been used to learn the expected behaviour of SUT [67], [163], [164]. For dynamic and fuzzy behaviours such as the result of web search engine queries, association rules between input (query) and output (search result strings) have been used as the format of an inferred oracle [208].

### 5.5 Textual Documentation

Textual documentation ranges from natural language descriptions of requirements to structured documents detailing the functionalities of APIs. These documents describe the functionalities expected from the SUT to varying degrees, and can therefore serve as a basis for generating test oracles. They are usually informal, intended for other humans, not to support formal logical or mathematical reasoning. Thus, they are often partial and ambiguous, in contrast to specification languages. Their importance for test oracle construction rests on the fact that developers are more likely to write them than formal specifications. In other words, the documentation defines the

constraints that the test oracle $D$, as defined in Section 2, enforces over testing activities.

At first sight, it may seem impossible to derive test oracles automatically because natural languages are inherently ambiguous and textual documentation is often imprecise and inconsistent. The use of textual documentation has often been limited to humans in practical testing applications [144]. However, some partial automation can assist the human in testing using documentation as a source of test oracle information.

Two approaches have been explored. The first category builds techniques to construct a formal specification out of an informal, textual artefact, such as an informal textual specification, user and developer documentation, and even source code comments. The second restricts a natural language to a semi-formal fragment amenable to automatic processing. Next, we present representative examples of each approach.

#### 5.5.1 Converting Text into Specifications

Prowell and Poore [153] introduced a sequential enumeration method for developing a formal specification from an informal one. The method systematically enumerates all sequences from the input domain and maps the corresponding outputs to produce an arguably complete, consistent, and correct specification. However, it can suffer from an exponential explosion in the number of input/output sequences. Prowell and Poore employ abstraction techniques to control this explosion. The end result is a formal specification that can be transferred into a number of notations, e.g., state transition systems. A notable benefit of this approach is that it tends to discover many inconsistent and missing requirements, making the specification more complete and precise.

#### 5.5.2 Restricting Natural Language

Restrictions on a natural language reduce complexities in its grammar and lexicon and allow the expression of requirements in a concise vocabulary with minimal ambiguity. This, in turn, eases the interpretation of documents and makes the automatic derivation of test oracles possible. The researchers who have proposed specification languages based on (semi-) formal subsets of a natural language are motivated by the fact that model-based specification languages have not seen wide-spread adoption, and believe the reason is the inaccessibility their formalism and set-theoretic underpinnings to the average programmer.

Schwitter introduced a computer-processable, restricted natural language called PENG [160]. It covers a strict subset of standard English with a restricted grammar and a domain specific lexicon for content words and predefined function words. Documents written in PENG can be translated deterministically into first-order predicate logic. Schwitter et al. [30] provided guidelines for writing test scenarios in PENG that can automatically judge the correctness of program behaviours.

## 6 Implicit Test Oracles

An *implicit test oracle* is one that relies on general, implicit knowledge to distinguish between a system's correct and incorrect behaviour. This generally true implicit knowledge includes such facts as "buffer overflows and segfaults are nearly always errors". The critical aspect of an implicit test oracle is that it requires neither domain knowledge nor a formal specification to implement, and it applies to nearly all programs.

Implicit test oracle can be built on any procedure that detects anomalies such as abnormal termination due to a crash or an execution failure [34], [167]. This is because such anomalies are *blatant faults*; that is, no more information is required to ascertain whether the program behaved correctly or not. Under our formalism, an implicit oracle defines a subset of stimulus and response relations as guaranteed failures, in some context.

Implicit test oracles are not universal. Behaviours abnormal for one system in one context may be normal for that system in a different context or normal for a different system. Even crashing *may* be considered acceptable, or even desired behaviour, as in systems designed to find crashes.

Research on implicit oracles is evident from early work in software engineering. The very first work in this context was related to deadlock, livelock and race detection to counter system concurrency issues [24] [107] [185] [16] [169]. Similarly, research on testing non-functional attributes have garnered much attention since the advent of the object-oriented paradigm. In performance testing, system throughput metrics can highlight degradation errors [121], [124], as when a server fails to respond when a number of requests are sent simultaneously. A case study by Weyuker and Vokolos showed how a process with excessive CPU usage caused service delays and disruptions [195]. Similarly, test oracles for memory leaks can be built on a profiling technique that detects dangling references during the run of a program [12], [57], [87], [211]. For example, Xie and Aiken proposed a boolean constraint system to represent the dynamically allocated objects in a program [201]. Their system raises an alarm when an object becomes unreachable but has not yet been deallocated.

Fuzzing is an effective way to find implicit anomalies, such a crashes [137]. The main idea is to generate random, or "fuzz", inputs and feed them to the system to find anomalies. This works because the implicit specification usually holds over all inputs, unlike explicit specifications which tend to relate subsets of inputs to outputs. If an anomaly is detected, the fuzz tester reports it along with the input that triggers it. Fuzzing is commonly used to detect security vulnerabilities, such as buffer overflows, memory leaks, unhandled exceptions, denial of service, etc. [18], [177].

Other work has focused on developing patterns to detect anomalies. For instance, Ricca and Tonella [155] considered a subset of the anomalies that Web applications can harbor, such as navigation problems, hyperlink inconsistencies, etc. In their empirical study, 60% of the Web applications exhibited anomalies and execution failures.

## 7 THE HUMAN ORACLE PROBLEM

The above sections give solutions to the test oracle problem when some artefact exists that can serve as the foundation for either a full or partial test oracle. In many cases, however, no such artefact exists so a human tester must verify whether software behaviour is correct given some stimuli. Despite the lack of an automated test oracle, software engineering research can still play a key role: finding ways to reduce the effort that the human tester has to expend in directly creating, or in being, the test oracle.

This effort is referred to as the Human Oracle Cost [126]. It aims to reduce the cost of human involvement along two dimensions: 1) writing test oracles and 2) evaluating test outcomes. Concerning the first dimension, the work of Staats et al. is a representative. They seek to reduce the human oracle cost by guiding human testers to those parts of the code they need to focus on when writing test oracles [173]. This reduces the cost of test oracle construction, rather than reducing the cost of a human involvement in testing in the absence of an automated test oracle. Additional recent work on test oracle construction includes Dodona, a tool that suggests oracle data to a human who then decides whether to use it to define a test oracle realized as a Java unit test [116]. Dodona infers relations among program variables during execution, using network centrality analysis and data flow.

Research that seeks to reduce the human oracle cost broadly focuses on finding a *quantitative* reduction in the amount of work the tester has to do for the same amount of test coverage or finding a *qualitative* reduction in the work needed to understand and evaluate test cases.

## 7.1 Quantitative Human Oracle Cost

Test suites can be unnecessarily large, covering few test goals in each individual test case. Additionally, the test cases themselves may be unnecessarily long — for example containing large numbers of method calls, many of which do not contribute to the overall test case. The goal of *quantitative human oracle cost reduction* is to reduce test suite and test case size so as to maximise the benefit of each test case and each component of that test case. This consequently reduces the amount of manual checking effort that is required on behalf of a human tester performing the role of a test oracle. Cast in terms of our formalism, quantitative reduction aims to partition the set of test activity sequences so the human need only consider representative sequences, while test case reduction aims to shorten test activity sequences.

### *7.1.1 Test Suite Reduction*

Traditionally, test suite reduction has been applied as a post-processing step to an existing test suite, e.g. the work of Harrold et al, [85], Offutt et al. [141] and Rothermel et al. [157]. Recent work in the search-based testing literature has sought to combine test input generation and test suite reduction into one phase to produce smaller test suites.

Harman et al. proposed a technique for generating test cases that penetrate the deepest levels of the control dependence graph for the program, in order to create test cases that exercise as many elements of the program as possible [82]. Ferrer et al. [61] attack a multi-objective version of the problem in which they sought to simultaneously maximize branch coverage and minimize test suite size; their focus was not this problem per se, but its use to compare a number of multi-objective optimisation algorithms, including the well-known Non-dominated Sorting Genetic Algorithm II (NSGA-II), Strength Pareto EA 2 (SPEA2), and MOCell. On a series of randomly-generated programs and small benchmarks, they found MOCell performed best.

Taylor et al. [178] use an inferred model as a semantic test oracle to shrink a test suite. Fraser and Arcuri [65] generate test suites for Java using their EvoSuite tool. By generating the entire suite at once, they are able to simultaneously maximize coverage and minimize test suite size, thereby aiding human oracles and alleviating the human oracle cost problem.

### *7.1.2 Test Case Reduction*

When using randomised algorithms for generating test cases for object-oriented systems, individual test cases can generate very long traces very quickly — consisting of a large number of method calls that do not actually contribute to a specific test goal (e.g. the coverage of a particular branch). Such method calls unnecessarily increase test oracle cost, so Leitner et al. remove such calls [113] using Zeller's and Hildebrandt's Delta Debugging [206]. JWalk simplifies test sequences by removing side-effect free functions from them, thereby reducing test oracle costs where the human is the test oracle [170]. Quick tests seek to efficiently spend a small test budget by building test suites whose execution is fast enough for it to be run after compilations [76]. These quick tests must be likely to trigger bugs and therefore generate short traces, which, as a result, are easier for humans to comprehend.

## 7.2 Qualitative Human Oracle Cost

Human oracle costs may also be minimised from a qualitative perspective. That is, the extent to which test cases, more generally testing activities, may be easily understood and processed by a human. The input profile of a SUT is the distribution of inputs it actually processes when running in its operational environment. Learning an input profile requires domain knowledge. If such domain knowledge is not built into the test data generation process, machine-generated test data tend to be

drawn from a different distribution over the SUT's inputs than its input profile. While this may be beneficial for trapping certain types of faults, the utility of the approach decreases when test oracle costs are taken into account, since the tester must invest time *comprehending* the scenario represented by test data in order to correctly evaluate the corresponding program output. Arbitrary inputs are much harder to understand than recognisable pieces of data, thus adding time to the checking process.

All approaches to qualitatively alleviating the human oracle cost problem incorporate human knowledge to improve the understandability of test cases. The three approaches we cover are 1) augmenting test suites designed by the developers; 2) computing "realistic" inputs from web pages, web services, and natural language; and 3) mining usage patterns to replicate them in the test cases.

In order to improve the readability of automatically-generated test cases, McMinn et al. propose the incorporation of human knowledge into the test data generation process [126]. With search-based approaches, they proposed injecting this knowledge by "seeding" the algorithm with test cases that may have originated from a human source such as a "sanity check" performed by the programmer, an already existing, partial test suite, or input–output examples generated by programming paradigms that involve the developer in computation, like prorogued programming [2].

The generation of string test data is particularly problematic for automatic test data generators, which tend to generate nonsensical strings. These nonsensical strings are, of course, a form of fuzz testing (Section 6) and good for exploring uncommon, shallow code paths and finding corner cases, but they are unlikely to exercise functionality deeper in a program's control flow. This is because string comparisons in control expressions are usually stronger than numerical comparisons, making one of a control point's branches much less likely to traverse via uniform fuzzing. We see here the seminal computer science trade-off between breadth first and depth first search in the choice between fuzz testing with nonsensical inputs and testing with realistic inputs.

Bozkurt and Harman, introduced the idea of mining web services for realistic test inputs, using the outputs of known and trusted test services as more realistic inputs to the service under test [32]. The idea is that realistic test cases are more likely to reveal faults that developers care about and yield test cases that are more readily understood. McMinn et al. also mine the web for realistic test cases. They proposed mining strings from the web to assist in the test generation process [130]. Since web page content is generally the result of human effort, the strings contained therein tend to be real words or phrases with high degrees of semantic and domain relevant context that can thus be used as sources of realistic test data.

Afshan et al. [1] combine a natural language model and metahueristics, strategies that guide a search process [25], to help generate readable strings. The language model scores how likely a string is to belong to a language based on the character combinations. Incorporating this probability score into a fitness function, a metaheuristic search can not only cover test goals, but generate string inputs that are more comprehensible than the arbitrary strings generated by the previous state of the art. Over a number of case studies, Afshan et al. found that human oracles more accurately and more quickly evaluated their test strings.

Fraser and Zeller [66] improve the familiarity of test cases by mining the software under test for common usage patterns of APIs. They then seek to replicate these patterns in generated test cases. In this way, the scenarios generated are more likely to be realistic and represent actual usages of the software under test.

### 7.3 Crowdsourcing the Test Oracle

A recent approach to handling the lack of a test oracle is to outsource the problem to an

online service to which large numbers of people can provide answers — i.e., through crowdsourcing. Pastore et al. [147] demonstrated the feasibility of the approach but noted problems in presenting the test problem to the crowd such that it could be easily understood, and the need to provide sufficient code documentation so that the crowd could determine correct outputs from incorrect ones. In these experiments, crowdsourcing was performed by submitting tasks to a generic crowdsourcing platform — Amazon's Mechanical Turk[4]. However, some dedicated crowdsourcing services now exist for the testing of mobile applications. They specifically address the problem of the exploding number of devices on which a mobile application may run, and which the developer or tester may not own, but which may be possessed by the crowd at large. Examples of these services include Mob4Hire[5], MobTest[6] and uTest[7].

## 8 Future Directions and Conclusion

This paper has provided a comprehensive survey of test oracles, covering specified, derived and implicit oracles and techniques that cater for the absence of test oracles. The paper has also analyzed publication trends in the test oracle domain. This paper has necessarily focused on the traditional approaches to the test oracle problem. Much work on test oracles remains to be done. In addition to research deepening and interconnecting these approaches, test oracle problem is open to new research directions. We close with a discussion of two of these that we find noteworthy and promising: test oracle reuse and test oracle metrics.

As this survey has shown, test oracles are difficult to construct. Oracle reuse is therefore an important problem that merits attention. Two promising approaches to oracle reuse are generalizations of reliable reset and the sharing of oracular data across software product lines (SPLs). Generalizing reliable reset to arbitrary states allows the interconnection of different versions of a program, so we can build test oracles that based on older versions of a program, using generalized reliable reset to ignore or handle new inputs and functionality. SPLs are sets of related versions of a system [47]. A product line can be thought of as a tree of related software products in which branches contain new alternative versions of the system, each of which shares some core functionality enjoyed by a base version. Research on test oracles should seek to leverage these SPL trees to define trees of test oracles that share oracular data where possible.

Work has already begun on using test oracle as the measure of how well the program has been tested (a kind of test oracle coverage) [104], [176], [186] and measures of oracles such as assessing the quality of assertions [159]. More work is needed. "Oracle metrics" is a challenge to, and an opportunity for, the "software metrics" community. In a world in which test oracles become more prevalent, it will be important for testers to be able to assess the features offered by alternative test oracles.

A repository of papers on test oracles accompanies this paper at http://crestweb.cs.ucl.ac.uk/resources/oracle_repository.

## 9 Acknowledgements

We would like to thank Bob Binder for helpful information and discussions when we began work on this paper. We would also like thank all who attended the CREST Open Workshop on the Test Oracle Problem (21–22 May 2012) at University College London, and gave feedback on an early presentation of the work. We are further indebted to the very many responses to our emails from authors cited in this survey, who provided several useful comments on an earlier draft of our paper.

4. http://www.mturk.com
5. http://www.mob4hire.com
6. http://www.mobtest.com
7. http://www.utest.com/

## REFERENCES

[1] Sheeva Afshan, Phil McMinn, and Mark Stevenson. Evolving readable string test inputs using a natural language model to reduce human oracle cost. In *International Conference on Software Testing, Verification and Validation (ICST 2013)*. IEEE, March 2013.

[2] Mehrdad Afshari, Earl T. Barr, and Zhendong Su. Liberating the programmer with prorogued programming. In *Proceedings of the ACM international symposium on New ideas, new paradigms, and reflections on programming and software*, Onward! '12, pages 11–26, New York, NY, USA, 2012. ACM. Track at OOPSLA/SPLASH'12.

[3] Wasif Afzal, Richard Torkar, and Robert Feldt. A systematic review of search-based testing for non-functional system properties. *Information and Software Technology*, 51(6):957–976, 2009.

[4] Bernhard K. Aichernig. Automated black-box testing with abstract VDM oracles. In *SAFECOMP*, pages 250–259. Springer-Verlag, 1999.

[5] Shaukat Ali, Lionel C. Briand, Hadi Hemmati, and Rajwinder Kaur Panesar-Walawege. A systematic review of the application and empirical investigation of search-based test-case generation. *IEEE Transactions on Software Engineering*, pages 742–762, 2010.

[6] Nadia Alshahwan and Mark Harman. Automated session data repair for web application regression testing. In *Proceedings of 2008 International Conference on Software Testing, Verification, and Validation*, pages 298–307. IEEE Computer Society, 2008.

[7] Dana Angluin. Learning regular sets from queries and counterexamples. *Inf. Comput.*, 75(2):87–106, 1987.

[8] W. Araujo, L.C. Briand, and Y. Labiche. Enabling the runtime assertion checking of concurrent contracts for the java modeling language. In *Software Engineering (ICSE), 2011 33rd International Conference on*, pages 786–795, 2011.

[9] W. Araujo, L.C. Briand, and Y. Labiche. On the effectiveness of contracts as test oracles in the detection and diagnosis of race conditions and deadlocks in concurrent object-oriented software. In *Empirical Software Engineering and Measurement (ESEM), 2011 International Symposium on*, pages 10–19, 2011.

[10] Shay Artzi, Michael D. Ernst, Adam Kieżun, Carlos Pacheco, and Jeff H. Perkins. Finding the needles in the haystack: Generating legal test inputs for object-oriented programs. In *1st Workshop on Model-Based Testing and Object-Oriented Systems (M-TOOS)*, Portland, OR, October 23, 2006.

[11] Egidio Astesiano, Michel Bidoit, Hélène Kirchner, Bernd Krieg-Brückner, Peter D. Mosses, Donald Sannella, and Andrzej Tarlecki. CASL: the common algebraic specification language. *Theor. Comput. Sci.*, 286(2):153–196, 2002.

[12] Todd M. Austin, Scott E. Breach, and Gurindar S. Sohi. Efficient detection of all pointer and array access errors. In *PLDI*, pages 290–301. ACM, 1994.

[13] A. Avizienis. The N-version approach to fault-tolerant software. *IEEE Transactions on Software Engineering*, 11:1491–1501, 1985.

[14] A. Avizienis and L. Chen. On the implementation of N-version programming for software fault-tolerance during execution. In *Proceedings of the First International Computer Software and Application Conference (COMPSAC '77)*, pages 149–155, 1977.

[15] Franz Baader and Tobias Nipkow. *Term Rewriting and All That*. Cambridge University Press, New York, NY, USA, 1998.

[16] A. F. Babich. Proving total correctness of parallel programs. *IEEE Trans. Softw. Eng.*, 5(6):558–574, November 1979.

[17] Luciano Baresi and Michal Young. Test oracles. Technical Report CIS-TR-01-02, University of Oregon, Dept. of Computer and Information Science, August 2001. http://www.cs.uoregon.edu/~michal/pubs/oracles.html.

[18] Sofia Bekrar, Chaouki Bekrar, Roland Groz, and Laurent Mounier. Finding software vulnerabilities by smart fuzzing. In *ICST*, pages 427–430, 2011.

[19] Gilles Bernot. Testing against formal specifications: a theoretical view. In *Proceedings of the International Joint Conference on Theory and Practice of Software Development on Advances in Distributed Computing (ADC) and Colloquium on Combining Paradigms for Software Development (CCPSD): Vol. 2*, TAPSOFT '91, pages 99–119, New York, NY, USA, 1991. Springer-Verlag New York, Inc.

[20] Gilles Bernot, Marie Claude Gaudel, and Bruno Marre. Software testing based on formal specifications: a theory and a tool. *Softw. Eng. J.*, 6(6):387–405, November 1991.

[21] Antonia Bertolino, Paola Inverardi, Patrizio Pelliccione, and Massimo Tivoli. Automatic synthesis of behavior protocols for composable web-services. In *Proceedings of ESEC/SIGSOFT FSE*, ESEC/FSE 2009, pages 141–150, 2009.

[22] D. Beyer, T. Henzinger, R. Jhala, and R. Majumdar. Checking memory safety with Blast. In M. Cerioli, editor, *Fundamental Approaches to Software Engineering, 8th International Conference, FASE 2005, Held as Part of the Joint European Conferences on Theory and Practice of Software, ETAPS 2005, Edinburgh, UK, April 4-8, 2005, Proceedings*, volume 3442 of *Lecture Notes in Computer Science*, pages 2–18. Springer, 2005.

[23] R. Binder. *Testing Object-Oriented Systems: Models, Patterns, and Tools*. Addison-Wesley, 2000.

[24] A. Blikle. Proving programs by sets of computations. In *Mathematical Foundations of Computer Science*, pages 333–358. Springer, 1975.

[25] Christian Blum and Andrea Roli. Metaheuristics in combinatorial optimization: Overview and conceptual comparison. *ACM Computing Surveys (CSUR)*, 35(3):268–308, 2003.

[26] Gregor V. Bochmann and Alexandre Petrenko. Protocol testing: review of methods and relevance for software testing. In *Proceedings of the 1994 ACM SIGSOFT international symposium on Software testing and analysis*, ISSTA '94, pages 109–124. ACM, 1994.

[27] G.V. Bochmann. Finite state description of commu-

nication protocols. *Computer Networks*, 2(4):361–372, 1978.

[28] E. Börger, A. Cavarra, and E. Riccobene. Modeling the dynamics of UML state machines. In *Abstract State Machines-Theory and Applications*, pages 167–186. Springer, 2000.

[29] Egon Börger. High level system design and analysis using abstract state machines. In *Proceedings of the International Workshop on Current Trends in Applied Formal Method: Applied Formal Methods*, FM-Trends 98, pages 1–43, London, UK, UK, 1999. Springer-Verlag.

[30] Kathrin Böttger, Rolf Schwitter, Diego Mollá, and Debbie Richards. Towards reconciling use cases via controlled language and graphical models. In *INAP*, pages 115–128, Berlin, Heidelberg, 2003. Springer-Verlag.

[31] F. Bouquet, C. Grandpierre, B. Legeard, F. Peureux, N. Vacelet, and M. Utting. A subset of precise UML for model-based testing. In *Proceedings of the 3rd International Workshop on Advances in Model-Based Testing*, A-MOST '07, pages 95–104, New York, NY, USA, 2007. ACM.

[32] Mustafa Bozkurt and Mark Harman. Automatically generating realistic test input from web services. In *IEEE 6th International Symposium on Service Oriented System Engineering (SOSE)*, pages 13–24, 2011.

[33] L. C. Briand, Y. Labiche, and H. Sun. Investigating the use of analysis contracts to improve the testability of object-oriented code. *Softw. Pract. Exper.*, 33(7):637–672, June 2003.

[34] Cristian Cadar, Vijay Ganesh, Peter M. Pawlowski, David L. Dill, and Dawson R. Engler. EXE: Automatically generating inputs of death. *ACM Trans. Inf. Syst. Secur.*, 12(2):10:1–10:38, December 2008.

[35] J. Callahan, F. Schneider, S. Easterbrook, et al. Automated software testing using model-checking. In *Proceedings 1996 SPIN workshop*, volume 353. Citeseer, 1996.

[36] F. T. Chan, T. Y. Chen, S. C. Cheung, M. F. Lau, and S. M. Yiu. Application of metamorphic testing in numerical analysis. In *Proceedings of the IASTED International Conference on Software Engineering*, pages 191–197, 1998.

[37] W.K. Chan, S.C. Cheung, and Karl R.P.H. Leung. *A metamorphic testing approach for online testing of service-oriented software applications*, chapter 7, pages 2894–2914. IGI Global, 2009.

[38] W.K. Chan, S.C. Cheung, and K.R.P.H. Leung. Towards a metamorphic testing methodology for service-oriented software applications. In *QSIC*, pages 470–476, September 2005.

[39] F. Chen, N. Tillmann, and W. Schulte. Discovering specifications. Technical Report MSR-TR-2005-146, Microsoft Research, October 2005.

[40] Huo Yan Chen, T. H. Tse, F. T. Chan, and T. Y. Chen. In black and white: an integrated approach to class-level testing of object-oriented programs. *ACM Trans. Softw. Eng. Methodol.*, 7:250–295, July 1998.

[41] Huo Yan Chen, T. H. Tse, and T. Y. Chen. TACCLE: a methodology for object-oriented software testing at the class and cluster levels. *ACM Trans. Softw. Eng. Methodol.*, 10(1):56–109, January 2001.

[42] T. Y. Chen, F.-C. Kuo, T. H. Tse, and Zhi Quan Zhou. Metamorphic testing and beyond. In *Proceedings of the International Workshop on Software Technology and Engineering Practice (STEP 2003)*, pages 94–100, September 2004.

[43] Tsong Chen, Dehao Huang, Haito Huang, Tsun-Him Tse, Zong Yang, and Zhi Zhou. Metamorphic testing and its applications. In *Proceedings of the 8th International Symposium on Future Software Technology*, ISFST 2004, pages 310–319, 2004.

[44] Yoonsik Cheon. Abstraction in assertion-based test oracles. In *Proceedings of the Seventh International Conference on Quality Software*, pages 410–414, Washington, DC, USA, 2007. IEEE Computer Society.

[45] Yoonsik Cheon and Gary T. Leavens. A simple and practical approach to unit testing: The JML and JUnit way. In *Proceedings of the 16th European Conference on Object-Oriented Programming*, ECOOP '02, pages 231–255, London, UK, 2002. Springer-Verlag.

[46] L.A. Clarke. A system to generate test data and symbolically execute programs. *Software Engineering, IEEE Transactions on*, SE-2(3):215 – 222, sept. 1976.

[47] Paul C. Clements. Managing variability for software product lines: Working with variability mechanisms. In $10^{th}$ *International Conference on Software Product Lines (SPLC 2006)*, pages 207–208, Baltimore, Maryland, USA, 2006. IEEE Computer Society.

[48] Markus Clermont and David Parnas. Using information about functions in selecting test cases. In *Proceedings of the 1st international workshop on Advances in model-based testing*, A-MOST '05, pages 1–7, New York, NY, USA, 2005. ACM.

[49] David Coppit and Jennifer M. Haddox-Schatz. On the use of specification-based assertions as test oracles. In *Proceedings of the 29th Annual IEEE/NASA on Software Engineering Workshop*, pages 305–314, Washington, DC, USA, 2005. IEEE Computer Society.

[50] M. Davies and E. Weyuker. Pseudo-oracles for non-testable programs. In *Proceedings of the ACM '81 Conference*, pages 254–257, 1981.

[51] Laura K. Dillon. Automated support for testing and debugging of real-time programs using oracles. *SIGSOFT Softw. Eng. Notes*, 25(1):45–46, January 2000.

[52] Roong-Ko Doong and Phyllis G. Frankl. The ASTOOT approach to testing object-oriented programs. *ACM Trans. Softw. Eng. Methodol.*, 3:101–130, April 1994.

[53] Edith Elkind, Blaise Genest, Doron Peled, and Hongyang Qu. Grey-box checking. In *FORTE*, pages 420–435, 2006.

[54] J. Ellsberger, D. Hogrefe, and A. Sarma. *SDL: formal object-oriented language for communicating systems*. Prentice Hall, 1997.

[55] M.D. Ernst, J.H. Perkins, P.J. Guo, S. McCamant, C. Pacheco, M.S. Tschantz, and C. Xiao. The Daikon system for dynamic detection of likely invariants. *Science of Computer Programming*, 69(1):35–45, 2007.

[56] Michael D. Ernst, Jake Cockrell, William G. Griswold, and David Notkin. Dynamically discovering

likely program invariants to support program evolution. *IEEE Trans. Software Eng.*, 27(2):99–123, 2001.
[57] R.A. Eyre-Todd. The detection of dangling references in C++ programs. *ACM Letters on Programming Languages and Systems (LOPLAS)*, 2(1-4):127–134, 1993.
[58] R. Feldt. Generating diverse software versions with genetic programming: an experimental study. *Software, IEE Proceedings*, 145, December 1998.
[59] Xin Feng, David Lorge Parnas, T. H. Tse, and Tony O'Callaghan. A comparison of tabular expression-based testing strategies. *IEEE Trans. Softw. Eng.*, 37(5):616–634, September 2011.
[60] Xin Feng, David Lorge Parnas, and T.H. Tse. Tabular expression-based testing strategies: A comparison. *Testing: Academic and Industrial Conference Practice and Research Techniques - MUTATION*, 0:134, 2007.
[61] J. Ferrer, F. Chicano, and E. Alba. Evolutionary algorithms for the multi-objective test data generation problem. *Software: Practice and Experience*, 42(11):1331–1362, 2011.
[62] John S. Fitzgerald and Peter Gorm Larsen. *Modelling Systems - Practical Tools and Techniques in Software Development (2. ed.)*. Cambridge University Press, 2009.
[63] C. Flanagan and K. R. M. Leino. Houdini, an annotation assistant for ESC/Java. *Lecture Notes in Computer Science*, 2021:500–517, 2001.
[64] Robert W. Floyd. Assigning meanings to programs. In J. T. Schwartz, editor, *Mathematical Aspects of Computer Science*, volume 19 of *Symposia in Applied Mathematics*, pages 19–32. American Mathematical Society, Providence, RI, 1967.
[65] Gordon Fraser and Andrea Arcuri. Whole test suite generation. *IEEE Transactions on Software Engineering*, 39(2):276–291, 2013.
[66] Gordon Fraser and Andreas Zeller. Exploiting common object usage in test case generation. In *Proceedings of the 2011 Fourth IEEE International Conference on Software Testing, Verification and Validation*, ICST '11, pages 80–89. IEEE Computer Society, 2011.
[67] Kambiz Frounchi, Lionel C. Briand, Leo Grady, Yvan Labiche, and Rajesh Subramanyan. Automating image segmentation verification and validation by learning test oracles. *Information and Software Technology*, 53(12):1337–1348, 2011.
[68] Pascale Gall and Agns Arnould. Formal specifications and test: Correctness and oracle. In Magne Haveraaen, Olaf Owe, and Ole-Johan Dahl, editors, *Recent Trends in Data Type Specification*, volume 1130 of *Lecture Notes in Computer Science*, pages 342–358. Springer Berlin Heidelberg, 1996.
[69] J. Gannon, P. McMullin, and R. Hamlet. Data abstraction, implementation, specification, and testing. *ACM Transactions on Programming Languages and Systems (TOPLAS)*, 3(3):211–223, 1981.
[70] Angelo Gargantini and Elvinia Riccobene. ASM-based testing: Coverage criteria and automatic test sequence. *Journal of Universal Computer Science*, 7(11):1050–1067, nov 2001.
[71] Stephen J. Garland, John V. Guttag, and James J. Horning. An overview of Larch. In *Functional Programming, Concurrency, Simulation and Automated Reasoning*, pages 329–348, 1993.
[72] Marie-Claude Gaudel. Testing from formal specifications, a generic approach. In *Proceedings of the 6th Ade-Europe International Conference Leuven on Reliable Software Technologies*, Ada Europe '01, pages 35–48, London, UK, 2001. Springer-Verlag.
[73] Marie-Claude Gaudel and Perry R. James. Testing algebraic data types and processes: A unifying theory. *Formal Asp. Comput.*, 10(5-6):436–451, 1998.
[74] Gregory Gay, Sanjai Rayadurgam, and Mats Heimdah. Improving the accuracy of oracle verdicts through automated model steering. In *Automated Software Engineering (ASE 2014)*. ACM Press, 2014.
[75] Patrice Godefroid, Nils Klarlund, and Koushik Sen. DART: directed automated random testing. In *PLDI*, pages 213–223. ACM, 2005.
[76] Alex Groce, Mohammad Amin Alipour, Chaoqiang Zhang, Yang Chen, and John Regehr. Cause reduction for quick testing. In *ICST*, pages 243–252, 2014.
[77] Alex Groce, Doron Peled, and Mihalis Yannakakis. Amc: An adaptive model checker. In *CAV*, pages 521–525, 2002.
[78] Roland Groz, Keqin Li, Alexandre Petrenko, and Muzammil Shahbaz. Modular system verification by inference, testing and reachability analysis. In *TestCom/FATES*, pages 216–233, 2008.
[79] Zhongxian Gu, Earl T. Barr, David J. Hamilton, and Zhendong Su. Has the bug really been fixed? In *Proceedings of the 2010 International Conference on Software Engineering (ICSE'10)*. IEEE Computer Society, 2010.
[80] R. Guderlei and J. Mayer. Statistical metamorphic testing testing programs with random output by means of statistical hypothesis tests and metamorphic testing. In *QSIC*, pages 404–409, October 2007.
[81] D. Harel. Statecharts: A visual formalism for complex systems. *Science of computer programming*, 8(3):231–274, 1987.
[82] Mark Harman, Sung Gon Kim, Kiran Lakhotia, Phil McMinn, and Shin Yoo. Optimizing for the number of tests generated in search based test data generation with an application to the oracle cost problem. In *International Workshop on Search-Based Software Testing (SBST 2010)*, pages 182–191. IEEE, 6 April 2010.
[83] Mark Harman, Afshin Mansouri, and Yuanyuan Zhang. Search based software engineering: A comprehensive analysis and review of trends techniques and applications. Technical Report TR-09-03, Department of Computer Science, King's College London, April 2009.
[84] Mark Harman, Phil McMinn, Jerffeson Teixeira de Souza, and Shin Yoo. Search based software engineering: Techniques, taxonomy, tutorial. In Bertrand Meyer and Martin Nordio, editors, *Empirical software engineering and verification: LASER 2009-2010*, pages 1–59. Springer, 2012. LNCS 7007.
[85] M. Jean Harrold, Rajiv Gupta, and Mary Lou Soffa. A methodology for controlling the size of a test suite. *ACM Trans. Softw. Eng. Methodol.*, 2(3):270–285, July 1993.

[86] Mary Jean Harrold, Gregg Rothermel, Kent Sayre, Rui Wu, and Liu Yi. An empirical investigation of the relationship between spectra differences and regression faults. *Software Testing, Verification, and Reliability*, 10(3):171–194, 2000.

[87] David L. Heine and Monica S. Lam. A practical flow-sensitive and context-sensitive C and C++ memory leak detector. In *PLDI*, pages 168–181. ACM, 2003.

[88] J. Henkel and A. Diwan. Discovering algebraic specifications from Java classes. *Lecture Notes in Computer Science*, 2743:431–456, 2003.

[89] F.C. Hennie. *Finite-state models for logical machines*. Wiley, 1968.

[90] MarijnJ.H. Heule and Sicco Verwer. Software model synthesis using satisfiability solvers. *Empirical Software Engineering*, pages 1–32, 2012.

[91] R.M. Hierons. Oracles for distributed testing. *Software Engineering, IEEE Transactions on*, 38(3):629–641, 2012.

[92] Robert M. Hierons. Verdict functions in testing with a fault domain or test hypotheses. *ACM Trans. Softw. Eng. Methodol.*, 18(4):14:1–14:19, July 2009.

[93] Robert M. Hierons, Kirill Bogdanov, Jonathan P. Bowen, Rance Cleaveland, John Derrick, Jeremy Dick, Marian Gheorghe, Mark Harman, Kalpesh Kapoor, Paul Krause, Gerald Lüttgen, Anthony J. H. Simons, Sergiy Vilkomir, Martin R. Woodward, and Hussein Zedan. Using formal specifications to support testing. *ACM Comput. Surv.*, 41:9:1–9:76, February 2009.

[94] Charles Anthony Richard Hoare. An Axiomatic Basis of Computer Programming. *Communications of the ACM*, 12:576–580, 1969.

[95] M. Holcombe. X-machines as a basis for dynamic system specification. *Software Engineering Journal*, 3(2):69–76, 1988.

[96] Mike Holcombe and Florentin Ipate. Correct systems: building a business process solution. *Software Testing Verification and Reliability*, 9(1):76–77, 1999.

[97] Gerard J. Holzmann. The model checker SPIN. *IEEE Trans. Softw. Eng.*, 23(5):279–295, May 1997.

[98] W E Howden. A functional approach to program testing and analysis. *IEEE Trans. Softw. Eng.*, 12(10):997–1005, October 1986.

[99] W.E. Howden. Theoretical and empirical studies of program testing. *IEEE Transactions on Software Engineering*, 4(4):293–298, July 1978.

[100] Merlin Hughes and David Stotts. Daistish: systematic algebraic testing for OO programs in the presence of side-effects. *SIGSOFT Softw. Eng. Notes*, 21(3):53–61, May 1996.

[101] Daniel Jackson. *Software Abstractions: Logic, Language, and Analysis*. The MIT Press, 2006.

[102] Daniel Jackson. *Software Abstractions: Logic, Language, and Analysis*. The MIT Press, 2006.

[103] Claude Jard and Gregor v. Bochmann. An approach to testing specifications. *Journal of Systems and Software*, 3(4):315 – 323, 1983.

[104] D. Jeffrey and R. Gupta. Test case prioritization using relevant slices. In *Computer Software and Applications Conference, 2006. COMPSAC '06. 30th Annual International*, volume 1, pages 411–420, 2006.

[105] Ying Jin and David Lorge Parnas. Defining the meaning of tabular mathematical expressions. *Sci. Comput. Program.*, 75(11):980–1000, November 2010.

[106] M.J. Kearns and U.V. Vazirani. *An Introduction to Computational Learning Theory*. MIT Press, 1994.

[107] R.M. Keller. Formal verification of parallel programs. *Communications of the ACM*, 19(7):371–384, 1976.

[108] James C. King. Symbolic execution and program testing. *Communications of the ACM*, 19(7):385–394, July 1976.

[109] Kiran Lakhotia, Phil McMinn, and Mark Harman. An empirical investigation into branch coverage for C programs using CUTE and AUSTIN. *Journal of Systems and Software*, 83(12):2379–2391, 2010.

[110] Axel van Lamsweerde. Formal specification: a roadmap. In *Proceedings of the Conference on The Future of Software Engineering*, ICSE '00, pages 147–159, New York, NY, USA, 2000. ACM.

[111] K. Lano and H. Haughton. *Specification in B: An Introduction Using the B Toolkit*. Imperial College Press, 1996.

[112] D. Lee and M. Yannakakis. Principles and methods of testing finite state machines—a survey. *Proceedings of the IEEE*, 84(8):1090–1123, aug 1996.

[113] A. Leitner, M. Oriol, A Zeller, I. Ciupa, and B. Meyer. Efficient unit test case minimization. In *Automated Software Engineering (ASE 2007)*, pages 417–420, Atlanta, Georgia, USA, 2007. ACM Press.

[114] Keqin Li, Roland Groz, and Muzammil Shahbaz. Integration testing of components guided by incremental state machine learning. In *TAIC PART*, pages 59–70, 2006.

[115] D. Lorenzoli, L. Mariani, and M. Pezzè. Automatic generation of software behavioral models. In *proceedings of the 30th International Conference on Software Engineering (ICSE)*, 2008.

[116] Pablo Loyola, Matthew Staats, In-Young Ko, and Gregg Rothermel. Dodona: Automated oracle data set selection. In *International Symposium on Software Testing and Analysis 2014*, ISSTA'04, 2014.

[117] D. Luckham and F.W. Henke. An overview of ANNA-a specification language for ADA. Technical report, Stanford University, 1984.

[118] Nancy A. Lynch and Mark R. Tuttle. An introduction to input/output automata. *CWI Quarterly*, 2:219–246, 1989.

[119] Patrícia D. L. Machado. On oracles for interpreting test results against algebraic specifications. In *AMAST*, pages 502–518. Springer-Verlag, 1999.

[120] Phil Maker. GNU Nana: improved support for assertion checking and logging in GNU C/C++, 1998. http://gnu.cs.pu.edu.tw/software/nana/.

[121] Haroon Malik, Hadi Hemmati, and Ahmed E. Hassan. Automatic detection of performance deviations in the load testing of large scale systems. In *Proceedings of International Conference on Software Engineering - Software Engineering in Practice Track*, ICSE 2013, page *to appear*, 2013.

[122] L. Mariani, F. Pastore, and M. Pezzè. Dynamic analysis for diagnosing integration faults. *IEEE Transactions on Software Engineering*, 37(4):486–508, 2011.

[123] Bruno Marre. Loft: A tool for assisting selection of test data sets from algebraic specifications. In *TAPSOFT*, pages 799–800. Springer-Verlag, 1995.

[124] A.P. Mathur. Performance, effectiveness, and reliability issues in software testing. In *COMPSAC*, pages 604–605. IEEE, 1991.

[125] Johannes Mayer, Ralph Guderlei, Abteilung Angew, Te Informationsverarbeitung, Abteilung Stochastik, and Universitt Ulm. Test oracles using statistical methods. In *In: Proceedings of the First International Workshop on Software Quality, Lecture Notes in Informatics P-58, Kllen Druck+Verlag GmbH*, pages 179–189. Springer, 2004.

[126] P. McMinn, M. Stevenson, and M. Harman. Reducing qualitative human oracle costs associated with automatically generated test data. In *1st International Workshop on Software Test Output Validation (STOV 2010), Trento, Italy, 13th July 2010*, pages 1–4, 2010.

[127] Phil McMinn. Search-based software test data generation: a survey. *Softw. Test. Verif. Reliab.*, 14(2):105–156, June 2004.

[128] Phil McMinn. Search-based failure discovery using testability transformations to generate pseudo-oracles. In *Genetic and Evolutionary Computation Conference (GECCO 2009)*, pages 1689–1696. ACM Press, 8-12 July 2009.

[129] Phil McMinn. Search-based software testing: Past, present and future. In *International Workshop on Search-Based Software Testing (SBST 2011)*, pages 153–163. IEEE, 21 March 2011.

[130] Phil McMinn, Muzammil Shahbaz, and Mark Stevenson. Search-based test input generation for string data types using the results of web queries. In *ICST*, pages 141–150, 2012.

[131] Phil McMinn, Mark Stevenson, and Mark Harman. Reducing qualitative human oracle costs associated with automatically generated test data. In *International Workshop on Software Test Output Validation (STOV 2010)*, pages 1–4. ACM, 13 July 2010.

[132] Atif M. Memon. Automatically repairing event sequence-based GUI test suites for regression testing. *ACM Transactions on Software Engineering Methodology*, 18(2):1–36, 2008.

[133] Atif M. Memon, Martha E. Pollack, and Mary Lou Soffa. Automated test oracles for GUIs. In *SIGSOFT '00/FSE-8: Proceedings of the 8th ACM SIGSOFT international symposium on Foundations of software engineering*, pages 30–39, New York, NY, USA, 2000. ACM Press.

[134] Atif M. Memon and Qing Xie. Using transient/persistent errors to develop automated test oracles for event-driven software. In *ASE '04: Proceedings of the 19th IEEE international conference on Automated software engineering*, pages 186–195, Washington, DC, USA, 2004. IEEE Computer Society.

[135] Maik Merten, Falk Howar, Bernhard Steffen, Patrizio Pellicione, and Massimo Tivoli. Automated inference of models for black box systems based on interface descriptions. In *Proceedings of the 5th international conference on Leveraging Applications of Formal Methods, Verification and Validation: technologies for mastering change - Volume Part I*, ISoLA'12, pages 79–96. Springer-Verlag, 2012.

[136] Bertrand Meyer. Eiffel: A language and environment for software engineering. *The Journal of Systems and Software*, 8(3):199–246, June 1988.

[137] B.P. Miller, L. Fredriksen, and B. So. An empirical study of the reliability of UNIX utilities. *Communications of the ACM*, 33(12):32–44, 1990.

[138] S. Mouchawrab, L.C. Briand, Y. Labiche, and M. Di Penta. Assessing, comparing, and combining state machine-based testing and structural testing: A series of experiments. *Software Engineering, IEEE Transactions on*, 37(2):161–187, 2011.

[139] Christian Murphy, Kuang Shen, and Gail Kaiser. Automatic system testing of programs without test oracles. In *ISSTA*, pages 189–200. ACM Press, 2009.

[140] Christian Murphy, Kuang Shen, and Gail Kaiser. Using JML runtime assertion checking to automate metamorphic testing in applications without test oracles. *2009 International Conference on Software Testing Verification and Validation*, pages 436–445, 2009.

[141] A. J. Offutt, J. Pan, and J. M. Voas. Procedures for reducing the size of coverage-based test sets. In *International Conference on Testing Computer Software*, pages 111–123, 1995.

[142] C. Pacheco and M. Ernst. Eclat: Automatic generation and classification of test inputs. *ECOOP 2005-Object-Oriented Programming*, pages 734–734, 2005.

[143] D L Parnas, J Madey, and M Iglewski. Precise documentation of well-structured programs. *IEEE Transactions on Software Engineering*, 20(12):948–976, 1994.

[144] David Lorge Parnas. Document based rational software development. *Journal of Knowledge Based Systems*, 22:132–141, April 2009.

[145] David Lorge Parnas. Precise documentation: The key to better software. In Sebastian Nanz, editor, *The Future of Software Engineering*, pages 125–148. Springer Berlin Heidelberg, 2011.

[146] David Lorge Parnas and Jan Madey. Functional documents for computer systems. *Sci. Comput. Program.*, 25(1):41–61, October 1995.

[147] F. Pastore, L. Mariani, and G. Fraser. Crowdoracles: Can the crowd solve the oracle problem? In *Proceedings of the International Conference on Software Testing, Verification and Validation (ICST)*, 2013.

[148] Doron Peled, Moshe Y. Vardi, and Mihalis Yannakakis. Black box checking. *Journal of Automata, Languages and Combinatorics*, 7(2), 2002.

[149] D K Peters and D L Parnas. Using test oracles generated from program documentation. *IEEE Transactions on Software Engineering*, 24(3):161–173, 1998.

[150] Dennis K. Peters and David Lorge Parnas. Requirements-based monitors for real-time systems. *IEEE Trans. Softw. Eng.*, 28(2):146–158, February 2002.

[151] Mauro Pezzè and Cheng Zhang. Automated test oracles: A survey. In Ali Hurson and Atif Memon, editors, *Advances in Computers*, volume 95, pages 1–48. Elsevier Ltd., 2014.

[152] Nadia Polikarpova, Ilinca Ciupa, and Bertrand Meyer. A comparative study of programmer-written and automatically inferred contracts. In *ISSTA*, pages 93–104. ACM, 2009.

[153] S.J. Prowell and J.H. Poore. Foundations of sequence-based software specification. *Software Engineering, IEEE Transactions on*, 29(5):417–429, 2003.

[154] Sam Ratcliff, David R. White, and John A. Clark. Searching for invariants using genetic programming and mutation testing. In *Proceedings of the 13th annual conference on Genetic and evolutionary computation*, GECCO '11, pages 1907–1914, New York, NY, USA, 2011. ACM.

[155] F. Ricca and P. Tonella. Detecting anomaly and failure in web applications. *MultiMedia, IEEE*, 13(2):44 – 51, april-june 2006.

[156] D.S. Rosenblum. A practical approach to programming with assertions. *Software Engineering, IEEE Transactions on*, 21(1):19–31, 1995.

[157] G. Rothermel, M. J. Harrold, J. von Ronne, and C. Hong. Empirical studies of test-suite reduction. *Software Testing, Verification and Reliability*, 12:219—249, 2002.

[158] N. Walkinshaw S. Ali, K. Bogdanov. A comparative study of methods for dynamic reverse-engineering of state models. Technical Report CS-07-16, The University of Sheffield, Department of Computer Science, October 2007. http://www.dcs.shef.ac.uk/intranet/research/resmes/CS0716.pdf.

[159] David Schuler and Andreas Zeller. Assessing oracle quality with checked coverage. In *Proceedings of the 2011 Fourth IEEE International Conference on Software Testing, Verification and Validation*, ICST '11, pages 90–99, Washington, DC, USA, 2011. IEEE Computer Society.

[160] R. Schwitter. English as a formal specification language. In *Database and Expert Systems Applications, 2002. Proceedings. 13th International Workshop on*, pages 228–232, sept. 2002.

[161] Sergio Segura, Robert M. Hierons, David Benavides, and Antonio Ruiz-Cortés. Automated metamorphic testing on the analyses of feature models. *Information and Software Technology*, 53(3):245 – 258, 2011.

[162] Koushik Sen, Darko Marinov, and Gul Agha. CUTE: a concolic unit testing engine for C. In *ESEC/FSE*, pages 263–272. ACM, 2005.

[163] Seyed Shahamiri, Wan Wan-Kadir, Suhaimi Ibrahim, and Siti Hashim. Artificial neural networks as multi-networks automated test oracle. *Automated Software Engineering*, 19(3):303–334, 2012.

[164] Seyed Reza Shahamiri, Wan Mohd Nasir Wan Kadir, Suhaimi Ibrahim, and Siti Zaiton Mohd Hashim. An automated framework for software test oracle. *Information and Software Technology*, 53(7):774 – 788, 2011.

[165] Seyed Reza Shahamiri, Wan Mohd Nasir Wan-Kadir, and Siti Zaiton Mohd Hashim. A comparative study on automated software test oracle methods. In *ICSEA*, pages 140–145, 2009.

[166] M. Shahbaz. *Reverse Engineering and Testing of Black-Box Software Components*. LAP Lambert Academic Publishing, 2012.

[167] K. Shrestha and M.J. Rutherford. An empirical evaluation of assertions as oracles. In *ICST*, pages 110–119. IEEE, 2011.

[168] Guoqiang Shu, Yating Hsu, and David Lee. Detecting communication protocol security flaws by formal fuzz testing and machine learning. In *FORTE*, pages 299–304, 2008.

[169] J. Sifakis. Deadlocks and livelocks in transition systems. *Mathematical Foundations of Computer Science 1980*, pages 587–600, 1980.

[170] A. J. H. Simons. JWalk: a tool for lazy systematic testing of java classes by design introspection and user interaction. *Automated Software Engineering*, 14(4):369–418, December 2007.

[171] Rishabh Singh, Dimitra Giannakopoulou, and Corina S. Pasareanu. Learning component interfaces with may and must abstractions. In Tayssir Touili, Byron Cook, and Paul Jackson, editors, *Computer Aided Verification, 22nd International Conference, CAV 2010, Edinburgh, UK, July 15-19, 2010. Proceedings*, volume 6174 of *Lecture Notes in Computer Science*, pages 527–542. Springer, 2010.

[172] J. Michael Spivey. *Z Notation - a reference manual (2. ed.)*. Prentice Hall International Series in Computer Science. Prentice Hall, 1992.

[173] M. Staats, G. Gay, and M. P. E. Heimdahl. Automated oracle creation support, or: How I learned to stop worrying about fault propagation and love mutation testing. In *Proceedings of the 34th International Conference on Software Engineering*, ICSE 2012, pages 870–880, 2012.

[174] M. Staats, M.W. Whalen, and M.P.E. Heimdahl. Programs, tests, and oracles: the foundations of testing revisited. In *ICSE*, pages 391–400. IEEE, 2011.

[175] Matt Staats, Shin Hong, Moonzoo Kim, and Gregg Rothermel. Understanding user understanding: determining correctness of generated program invariants. In *ISSTA*, pages 188–198. ACM, 2012.

[176] Matt Staats, Pablo Loyola, and Gregg Rothermel. Oracle-centric test case prioritization. In *Proceedings of the 23rd International Symposium on Software Reliability Engineering*, ISSRE 2012, pages 311–320, 2012.

[177] Ari Takanen, Jared DeMott, and Charlie Miller. *Fuzzing for Software Security Testing and Quality Assurance*. Artech House, Inc., Norwood, MA, USA, 1 edition, 2008.

[178] Ramsay Taylor, Mathew Hall, Kirill Bogdanov, and John Derrick. Using behaviour inference to optimise regression test sets. In Brian Nielsen and Carsten Weise, editors, *Testing Software and Systems - 24th IFIP WG 6.1 International Conference, ICTSS 2012, Aalborg, Denmark, November 19-21, 2012. Proceedings*, volume 7641 of *Lecture Notes in Computer Science*, pages 184–199. Springer, 2012.

[179] A. Tiwari. Formal semantics and analysis methods for Simulink Stateflow models. Technical report, SRI

International, 2002. http://www.csl.sri.com/users/tiwari/html/stateflow.html.

[180] Jan Tretmans. Test generation with inputs, outputs and repetitive quiescence. *Software - Concepts and Tools*, 17(3):103–120, 1996.

[181] Alan M. Turing. Checking a large routine. In *Report of a Conference on High Speed Automatic Calculating Machines*, pages 67–69, Cambridge, England, June 1949. University Mathematical Laboratory.

[182] Mark Utting and Bruno Legeard. *Practical Model-Based Testing: A Tools Approach*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2007.

[183] Mark Utting, Alexander Pretschner, and Bruno Legeard. A taxonomy of model-based testing approaches. *Softw. Test. Verif. Reliab.*, 22(5):297–312, August 2012.

[184] G. v. Bochmann, C. He, and D. Ouimet. Protocol testing using automatic trace analysis. In *Proceedings of Canadian Conference on Electrical and Computer Engineering*, pages 814–820, 1989.

[185] A. van Lamsweerde and M. Sintzoff. Formal derivation of strongly correct concurrent programs. *Acta Informatica*, 12(1):1–31, 1979.

[186] J.M. Voas. PIE: a dynamic failure-based technique. *Software Engineering, IEEE Transactions on*, 18(8):717–727, 1992.

[187] N. Walkinshaw, K. Bogdanov, J. Derrick, and J. Paris. Increasing functional coverage by inductive testing: A case study. In Alexandre Petrenko, Adenilso da Silva Simão, and José Carlos Maldonado, editors, *Testing Software and Systems - 22nd IFIP WG 6.1 International Conference, ICTSS 2010, Natal, Brazil, November 8-10, 2010. Proceedings*, volume 6435 of *Lecture Notes in Computer Science*, pages 126–141. Springer, 2010.

[188] Neil Walkinshaw and Kirill Bogdanov. Inferring finite-state models with temporal constraints. In *ASE*, pages 248–257, 2008.

[189] Neil Walkinshaw, John Derrick, and Qiang Guo. Iterative refinement of reverse-engineered models by model-based testing. In *FM*, pages 305–320, 2009.

[190] Yabo Wang and DavidLorge Parnas. Trace rewriting systems. In Michaël Rusinowitch and Jean-Luc Rémy, editors, *Conditional Term Rewriting Systems*, volume 656 of *Lecture Notes in Computer Science*, pages 343–356. Springer Berlin Heidelberg, 1993.

[191] Yabo Wang and D.L. Parnas. Simulating the behaviour of software modules by trace rewriting. In *Software Engineering, 1993. Proceedings., 15th International Conference on*, pages 14–23, 1993.

[192] Yi Wei, Carlo A. Furia, Nikolay Kazmin, and Bertrand Meyer. Inferring better contracts. In *Proceedings of the 33rd International Conference on Software Engineering*, ICSE '11, pages 191–200, New York, NY, USA, 2011. ACM.

[193] Yi Wei, H. Roth, C.A. Furia, Yu Pei, A. Horton, M. Steindorfer, M. Nordio, and B. Meyer. Stateful testing: Finding more errors in code and contracts. In *Automated Software Engineering (ASE), 2011 26th IEEE/ACM International Conference on*, pages 440–443, 2011.

[194] E.J. Weyuker. Assessing test data adequacy through program inference. *ACM Transactions on Programming Languages and Systems (TOPLAS)*, 5(4):641–655, 1983.

[195] E.J. Weyuker and F.I. Vokolos. Experience with performance testing of software systems: issues, an approach, and case study. *Software Engineering, IEEE Transactions on*, 26(12):1147–1156, 2000.

[196] Elaine J. Weyuker. On testing non-testable programs. *The Computer Journal*, 25(4):465–470, November 1982.

[197] Jeannette M. Wing. A specifier's introduction to formal methods. *IEEE Computer*, 23(9):8–24, 1990.

[198] Qing Xie and Atif M. Memon. Designing and comparing automated test oracles for GUI-based software applications. *ACM Transactions on Software Engineering and Methodology*, 16(1):4, 2007.

[199] Tao Xie. Augmenting automatically generated unit-test suites with regression oracle checking. In *Proc. 20th European Conference on Object-Oriented Programming (ECOOP 2006)*, pages 380–403, July 2006.

[200] Tao Xie and David Notkin. Checking inside the black box: Regression testing by comparing value spectra. *IEEE Transactions on Software Engineering*, 31(10):869–883, October 2005.

[201] Yichen Xie and Alex Aiken. Context-and path-sensitive memory leak detection. *ACM SIGSOFT Software Engineering Notes*, 30(5):115–125, 2005.

[202] Zhihong Xu, Yunho Kim, Moonzoo Kim, Gregg Rothermel, and Myra B. Cohen. Directed test suite augmentation: techniques and tradeoffs. In *Proceedings of the eighteenth ACM SIGSOFT international symposium on Foundations of software engineering*, FSE '10, pages 257–266, New York, NY, USA, 2010. ACM.

[203] Shin Yoo. Metamorphic testing of stochastic optimisation. In *Proceedings of the 2010 Third International Conference on Software Testing, Verification, and Validation Workshops*, ICSTW '10, pages 192–201. IEEE Computer Society, 2010.

[204] Shin Yoo and Mark Harman. Regression testing minimisation, selection and prioritisation: A survey. *Software Testing, Verification, and Reliability*, 22(2):67–120, March 2012.

[205] Bo Yu, Liang Kong, Yufeng Zhang, and Hong Zhu. Testing Java components based on algebraic specifications. *Software Testing, Verification, and Validation, 2008 International Conference on*, 0:190–199, 2008.

[206] A. Zeller and R Hildebrandt. Simplifying and isolating failure-inducing input. *IEEE Transactions on Software Engineering*, 28(2):183–200, 2002.

[207] S. Zhang, D. Saff, Y. Bu, and M.D. Ernst. Combined static and dynamic automated test generation. In *ISSTA*, volume 11, pages 353–363, 2011.

[208] Wujie Zheng, Hao Ma, Michael R. Lyu, Tao Xie, and Irwin King. Mining test oracles of web search engines. In *Proc. 26th IEEE/ACM International Conference on Automated Software Engineering (ASE 2011), Short Paper*, ASE 2011, pages 408–411, 2011.

[209] Zhi Quan Zhou, ShuJia Zhang, Markus Hagenbuchner, T. H. Tse, Fei-Ching Kuo, and T. Y. Chen. Automated functional testing of online search services.

*Software Testing, Verification and Reliability*, 22(4):221–243, 2012.

[210] Hong Zhu. A note on test oracles and semantics of algebraic specifications. In *Proceedings of the 3rd International Conference on Quality Software*, QSIC 2003, pages 91–98, 2003.

[211] B. Zorn and P. Hilfinger. A memory allocation profiler for C and Lisp programs. In *Proceedings of the Summer USENIX Conference*, pages 223–237, 1988.