Clarifications on the Construction and Use of the ManyBugs Benchmark

Claire Le Goues, Yuriy Brun, Senior Member, IEEE, Stephanie Forrest, Fellow, IEEE, Westley Weimer

HIGH-QUALITY research requires timely dissemination and the incorporation of feedback. Since the publication of the MANYBUGS benchmark [2] and its release on http://repairbenchmarks.cs.umass.edu/, Fan Long and Martin Rinard have provided feedback on the benchmark’s construction and use. Here, we describe that feedback and our subsequent improvements to the MANYBUGS benchmark. We thank these researchers for the feedback.

**php test harness.** The php subject is a language interpreter. Both its testing framework and the tests themselves are effectively written in php (the tests are actually written in a variant of php called phpt). To run a test, a php interpreter (the testing interpreter) runs the testing framework, which loads a second php interpreter (the tested interpreter) to interpret the test php program. The test framework (again, executed by the testing interpreter) captures the output of the tested interpreter and compares it to the expected output.

Automated repair techniques produce variant php interpreters, which should naturally serve as the tested interpreters. However, the answer to the question of what should serve as the testing interpreter is less obvious. php’s default test harness configuration uses the same version of the interpreter for both the tested and testing interpreter. However, php may be configured via a command-line argument to use a different interpreter, such as the unmodified defective version, or a separate, manually-repaired version. Testing results — whether a test passes or fails — can differ based on which version is used as the testing interpreter. For example, theoretically, automated repair could produce a variant that, when interpreting the testing framework’s code that compares a test’s output to the expected output, always deems the results identical. Such a variant would produce different test results than would a different, unmodified version of the interpreter.

Absent an established answer on how to augment or improve the developer-provided test framework with automatic repair in mind, we initially configured the php scenarios to follow the developer-provided defaults. Thus, our reported experimental results (Section 6.2 in [2]) use the variant php interpreter as both tested and testing interpreter. Here, Fig. 1 augments Fig. 4 in [2] to include the repair results on the php scenarios using an unmodified version of php (version 5.3.1) as the testing interpreter. Row “php default” is identical to that in [2]; row “php alternative” shows the newly-computed results. We kept the experimental settings the same between the runs for consistency with the previous work, with one exception: The new AE experiments are run without the “super mutant” optimization (which typically improves efficiency).

We have added the repair logs and other data associated with these experiments to the MANYBUGS release (at http://repairbenchmarks.cs.umass.edu/). We have also added support files and instructions, appending to the MANYBUGS README, to support their reproduction on the previously-released AWS virtual machines.

**Guarantees.** One of the requirements of the MANYBUGS defect scenarios was that “No part of the specification (e.g., no test case) can be satisfied by a program with explicitly degenerate behavior, such as a program that crashes immediately on launch” (Section 3.1 in [2]). We verified that the specification cannot be satisfied by the following programs:

1. `/bin/false`
2. `main(){ while(1) { }}`
3. `main(){ /*immediate segmentation fault */}`
4. an empty module.

Note that not all degenerate programs are caught by test output: timeouts on test execution catch infinite loops, and many scenarios do not compile or link trivial modules. Different approaches to constructing or configuring degenerate test programs may find different results.

**libtiff test suite quality.** The test suites used in the MANYBUGS benchmark are written by the respective project developers, and thus vary in quality. In this vein, it is worth noting that a number of the libtiff test cases...
### Table 1: Baseline results of running GenProg v2.2, TrpAutoRepair, and AE v3.0 on the 104 php defects from the MANYBUGS benchmark with the variant php interpreter to interpret both the testing framework and the test (php default) and an unrepaid php interpreter to interpret the testing framework and the variant php interpreter to interpret the test (php alternative). For each of the repair techniques, we report the number of defects repaired; the average time to repair in minutes (GenProg and TrpAutoRepair were run on 10 seeds per scenario, with each run provided a 12-hour timeout; AE is run once per scenario, with a 60-hour timeout); and the number of fitness evaluations to a repair, which serves as a compute- and scenario-independent measure of repair time (typically dominated by test suite execution time and thus varies by test suite size). Complete results, including individual log files for each defect and reproduction instructions, are available for download with the dataset.

<table>
<thead>
<tr>
<th>Program</th>
<th>Defects repaired</th>
<th>GenProg Time (min)</th>
<th>Fitness evals</th>
<th>TrpAutoRepair Time (min)</th>
<th>Fitness evals</th>
<th>AE Time (min)</th>
<th>Fitness evals</th>
</tr>
</thead>
<tbody>
<tr>
<td>php default</td>
<td>54/104</td>
<td>181</td>
<td>5.2</td>
<td>56/104</td>
<td>180</td>
<td>1.1</td>
<td>1.2</td>
</tr>
<tr>
<td>php alternative</td>
<td>19/104</td>
<td>210</td>
<td>6.1</td>
<td>27/104</td>
<td>193</td>
<td>1.3</td>
<td>1.0</td>
</tr>
</tbody>
</table>

### Discussion

The above clarifications primarily concern the test suites provided with the MANYBUGS scenario programs. All project test suites in MANYBUGS are written by the respective projects’ developers, and thus vary in quality; This is to be expected in real-world test suites [6]. However, this also means that the strength of the guarantees that such test suites provide about automatically-generated patches can vary. Moreover, different choices for test suite configuration and construction can produce different results when used to assess different patch generation techniques.

Assessing patch quality independently of tested behavior or otherwise guarding against degenerate patches that make tests pass without properly repairing the broken functionality is an unsolved research problem, and the subject of ongoing work [5], [7]. Absent a definitive understanding of how best to use developer-provided test suites to assess candidate repair quality, and to avoid interposing our own biases in benchmark construction, we sought to use these test suites in as unmodified a manner as possible. We are heartened by recent work that focuses on understanding and quantifying the quality problem, as well as on techniques that produce higher-quality patches, e.g., [1], [3], [4].

We look forward to ongoing discussion of these and related issues, with a goal of advancing collective scientific understanding of (1) what it means for a patch to be of high-quality, (2) how to measure that quality by using, augmenting, or side-stepping altogether the developer-provided test suites, and (3) how to automatically produce high-quality patches, improving software quality while decreasing the cost of developing it.

### ACKNOWLEDGMENTS

This work is supported by the National Science Foundation under grants no. CCF-1563797 and CCF-1564162.

### REFERENCES


Claire Le Goues received the BA degree in computer science from Harvard University and the MS and PhD degrees from the University of Virginia. She is an Assistant Professor in the School of Computer Science at Carnegie Mellon University, where she is primarily affiliated with the Institute for Software Research. She is interested in how to construct high-quality systems in the face of continuous software evolution, with a particular interest in automatic error repair. More information is available at: http://www.cs.cmu.edu/~clegoues.
Yuriy Brun is an Associate Professor in the College of Information and Computer Science at the University of Massachusetts, Amherst. He received the PhD degree from the University of Southern California in 2008 and completed his postdoctoral work at the University of Washington in 2012. His research focuses on software engineering, self-adaptive systems, and testing software for fairness. He received an NSF CAREER award and an IEEE TCSC Young Achiever in Scalable Computing Award. He is a senior member of the IEEE and the ACM. More information is available on his homepage: http://www.cs.umass.edu/~brun/.

Stephanie Forrest received the BA degree from St. John’s College and the MS and PhD degrees from the University of Michigan. She is currently Regents Distinguished Professor of Computer Science at the University of New Mexico and a member of the External Faculty of the Santa Fe Institute. Her research studies complex adaptive systems, including immunology, evolutionary computation, biological modeling, and computer security. She is an IEEE fellow.

Westley Weimer received the BA degree in computer science and mathematics from Cornell University and the MS and PhD degrees from the University of California, Berkeley. He is currently a Professor at the University of Michigan. His main research interests include static and dynamic analyses to improve software quality and fix defects.