Alex Groce
School of Informatics, Computing & Cyber Systems
Northern Arizona University

September 10, 2019

Dear editor and reviewers:

We would like to thank the anonymous reviewers for their help in significantly improving the paper. Based on the comments, we corrected errors in the formalism, re-organized the paper by moving the sources of determinism discussion to earlier in the paper and providing more overview of implementation choices, including key challenges, and re-formulated the experimental section in terms of explicit research questions and answers to those questions, including a summary table.

The section numbering is changed. There is a new section 3, moved from the implementation section, on sources of nondeterminism. The delta-debugging is now section 4, implementation is section 5, the example is section 6, and section 7 is the experimental evaluation. Section 8, 9, and 10 present threats to validity, related work, and conclusions, respectively.

Below are our detailed responses to the reviewer comments.

R1:

1. Revised Formalism, fixed page 12 mistake, revised formal definition of failure nondeterminism. We now discuss the issue fo multiple executions in vertical determinism. In particular, our original formulation actually allowed for this, but this was not made clear.
2. We revised the section substantially, and added definitions and pseudo-code to make the concepts more precise and easy to follow.
3. Section 4 has been restructured and summarized more precisely.
4. Section 6 has been revised to make research questions explicit, our goals in selecting the subjects explicit, and presented as (limited) experiment rather than a series of case studies.

R2:

1. We now much more thoroughly introduce delta debugging. We explained our motivation in using mutation for vertical determinism only (it relates to an actual high-quality specification of a different kind to compare to; for redis-py, the ”spec” is largely either failure or nondeterminism represented by flaky tests, so the comparison with mutants would have very little value and be difficult to interpret. Comparison with a strong differential specification is easy to understand, and is certainly meaningful.
2. We discuss in threats and related work why we did not compare to, e.g., DeFlaker: the settings and uses are almost completely different. We were unable to devise a reasonable experiment that would compare the tools, since essentially they are orthogonal.
3. The 12 hours mentioned is rather misleading, we now elaborate what this meant, in a way that should make it possible to understand the result much more easily. Basically, 12 hours *sufficed* and any amount of time can be spent that is needed to satisfy neede certainty. This is basically independent of the flaky test suite size, though reducing those might also be a (weaker) approach to the same goal.

Sincerely,
Alex Groce (agroce@gmail.com)
Associate Professor
School of Informatics, Computing &
Cyber Systems
Northern Arizona University

Josie Holmes
Affiliated Researcher
School of Informatics, Computing &
Cyber Systems
Northern Arizona University