Editorial: ACM TOMS Replicated Computational Results Initiative

MICHAEL A. HEROUX, Sandia National Laboratories

The scientific community relies on the peer review process for assuring the quality of published material, the goal of which is to build a body of work we can trust. Computational journals such as the *ACM Transactions on Mathematical Software* (TOMS) use this process for rigorously promoting the clarity and completeness of content, and citation of prior work. At the same time, it is unusual to independently confirm computational results.

ACM TOMS has established a *Replicated Computational Results* (RCR) review process as part of the manuscript peer review process. The purpose is to provide independent confirmation that results contained in a manuscript are replicable. Successful completion of the RCR process awards a manuscript with the Replicated Computational Results Designation.

This issue of ACM TOMS contains the first [Van Zee and van de Geijn 2015] of what we anticipate to be a growing number of articles to receive the RCR designation, and the related RCR reviewer report [Willenbring 2015]. We hope that the TOMS RCR process will serve as a model for other publications and increase the confidence in and value of computational results in TOMS articles.

CCS Concepts: • **General and reference** → **Verification;** *Validation*; • **Software and its engineering** → **Formal software verfication;**

General Terms: Verification

Additional Key Words and Phrases: Replicated computational results, reproducibility, validation, publication

ACM Reference Format:
Michael A. Heroux. 2015. Editorial: ACM TOMS replicated computational results initiative. ACM Trans. Math. Softw. 41, 3, Article 13 (May 2015), 5 pages.
DOI: http://dx.doi.org/10.1145/2743015

1. INTRODUCTION

The peer review process for computational journal articles rigorously checks the clarity and completeness of content, citation of prior work and logical discussion. This process also involves scrutiny of computational results in correlation to the text and conclusions. At the same time, it is unusual to independently confirm computational results. Remarkably, we seldom rigorously probe the correctness and execution times of computational results and even more rarely ask that results be replicated, either by the author or independently.

The *ACM Transactions on Mathematical Software* (TOMS) has established a new replicated computational results (RCR) process as part of the overall peer review process. The purpose of RCR activities is to provide independent confirmation that results contained in a manuscript are correct and replicated. Successful completion of the RCR process gives the manuscript a *Replicated Computational Results* designation, which will be noted on the first page of the published article.

Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy's National Nuclear Security Administration under contract DEAC04-94AL85000.
Author's address: Sandia National Laboratories, New Mexico, P.O. Box 5800, Albuquerque, NM 87185; email: maherou@sandia.gov.
Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the Owner/Author.
2015 Copyright is held by the author/owner(s).
0098-3500/2015/05-ART13
DOI: http://dx.doi.org/10.1145/2743015

2. BACKGROUND

Reproducibility of scientific results is often equated with quality, suggesting that a published result is in some way confirmed as true, or is able to be confirmed if desired. Even so, evidence suggests that many published results are not reproducible [Begley et al. 2014; Bailey 2014; Arrowsmith 2011; Owens 2011]. An informal scan of many computational journal articles shows there is seldom enough detail in the article to enable independent confirmation of results and little evidence that the review process involved scrutiny of the results. Indeed, common personal experience shows that authors themselves can seldom reproduce their own results after a few months, if for no other reason than they are unsure of which piece of software was used to generate the original results.

Recent efforts in reproducibility [Bailey and Borwein 2014; Stodden et al. 2013] argue that openness and transparency are important ways to provide incentives to scientists. The expectation that results will be scrutinized leads to better preparation of results and an audit trail for later confirmation.

The July 2011 workshop "Reproducible Research: Tools and Strategies for Scientific Computing" in Vancouver and the December 2012 ICERM Workshop "Reproducibility in Computational and Experimental Mathematics" provided forums for developing a common understanding of reproducibility topics as well as showcases for emerging tools. The ICERM report on Reproducibility in Computational and Experimental Mathematics [Stodden et al. 2013] lays out several strategies for "Setting the Default to Reproducible." This report, and a collection of articles coming out of the Vancouver meeting [Davison 2012; Howe 2012; Vandewalle 2012; Gavish and Donoho 2012; Freire and Silva 2012; Dubois 2012], lay some of the foundation upon which the TOMS RCR effort builds. In particular, we use *replicable* as defined in the ICERM report to describe our efforts, and we formulate our review process to be consistent with ideas presented in this body of work.

3. THE TOMS REPLICATED COMPUTATIONAL RESULTS REVIEW PROCESS

The TOMS RCR process is an optional activity performed at the manuscript author's discretion. We hope that the general concern for advancing the quality of computational science results will be incentive enough for authors to engage in the RCR process. The initial response from early adopters gives us some confidence in this incentive. The RCR process starts after it becomes clear that a given manuscript is likely to be accepted, even though substantial revisions may be required. Usually, this determination is clear after the initial standard manuscript review phase.

The RCR process includes the following steps.

(1) *RCR Review Request.* When authors submit a manuscript for review, they can optionally request a replicated computational results review, which will be conducted independently from the standard review process.
(2) *Standard Reviewer Assignment.* Once the manuscript has been assigned to an associate editor (AE), the AE will assign referees for the standard review process.
(3) *RCR Suitability Review.* Concurrent with assigning standard reviewers, the Editor-in-Chief (EiC) and AE will briefly review the manuscript to determine if it is suitable for an RCR review. The decision about RCR suitability may be delayed until the first round of standard reviews is complete.
(4) *RCR Reviewer Assignment.* After determination of RCR suitability and concurrent with the standard peer review process, the AE will assign an RCR reviewer whose sole responsibility is to replicate manuscript computational results. Unlike the other reviewers, the RCR reviewer will be known to the authors and work together with the authors during the RCR process.

(5) *RCR Review Process.* Replicating computational results will require a multi-faceted approach. TOMS editors will advise the RCR reviewer on acceptable approaches, but ultimately the RCR reviewer has the responsibility to declare whether or not computational results in the manuscript are replicated. RCR reviewers will document the details of how results were replicated.
(6) *RCR Determination.* We anticipate that manuscripts submitted for RCR designation will almost surely succeed in eventually achieving this designation, at least in the introductory phase of the RCR initiative. This is important for assuring that RCR reviewers will obtain a published article as an outcome from their efforts, and to reduce authors' risk in voluntarily submitting to this review process.
(7) *RCR Review Failure.* There is some risk now and in the future that RCR efforts will fail. In this case, we must acknowledge that the manuscript is not ready for publication with the presented results. During the introductory phase, the EiC will personally manage this situation if it occurs and will work with the authors to avoid rejecting the manuscript outright. As the RCR initiative matures, we anticipate that failed RCR reviews would constitute grounds for returning the manuscript back to the authors for revision, or for rejection if concerns were serious.
(8) *Publication.* A manuscript whose computational results are successfully replicated will be published with a special RCR designation as a text-only note on the bottom of the first page. A special graphic designation may become possible in the future. The RCR referee will be acknowledged in the published paper as author of the RCR review report that will appear with the published manuscript. The RCR referee's report will be published as a TOMS article, immediately following the RCR-reviewed article. This report will also go through a light review process to assure that it is well written and contains required report elements.

3.1. Methods for Replicating Results

In the early phases of this initiative, we are being intentionally flexible about the methodologies we will accept for replicating computational results. As we gain experience, we expect to enumerate and describe acceptable methods more clearly. Presently, we have two basic approaches. The first is more desirable, but not always possible.

(1) *Independent Replication.* The authors provide the RCR reviewer access to, or sufficient description of, the computational platform used to produce the manuscript results. Access could be

—a direct transfer of software to the reviewer or a pointer to an archive of the software, and a description of a commonly available computer system the reviewer can access;
—a guest account and access to the software on the system used to produce the results;
—detailed observation of the authors replicating the results.

(2) *Review of Computational Results Artifacts.* In some situations, the authors may not be able to readily replicate computational results. Results may be from a system that is no longer available, or may be on a leadership class computing system to which access is very limited. In these situations, careful documentation of the process used to produce results could be sufficient for an RCR designation. In this case, the software should have its own substantial verification process to give the reviewer confidence that computations were performed correctly. If timing results are reported, the authors' artifacts should include validation testing of the timers used to report results.

For most computational experiments, it is very difficult to exactly replicate results. We therefore rely heavily on the judgment of the RCR reviewer in determining whether or not the replication process substantiates what the authors report in the manuscript. This is in part why we think an open review process is best for this activity. We fully expect that the authors and RCR reviewers will communicate throughout the review process, collaborating on the RCR determination.

3.2. RCR Reviewer Selection

A key element for the success of the Replicated Computational Results Initiative is the selection of RCR reviewers. It is important to note that this reviewer should be knowledgeable about the subject matter, but need not be an expert in determining if the work is novel. Instead, the RCR reviewer should be knowledgeable about the use and evaluation of the type of software under review.

For example, the direct sparse solvers community has many members, and determining whether or not an article presents novel results requires a detailed understanding of the state of research in the community. In contrast, replicating the computational results in a manuscript about direct solvers is not concerned with novelty but only correct reporting of results. The RCR reviewer could be a knowledgeable user of direct sparse solvers, a benchmarking expert or a software manager of a scientific computing project.

3.3. RCR Introduction and Transitions

The ACM TOMS Replicated Computational Results Initiative has started with this issue. We have purposely picked a weak but usable form of reproducibility: One person must determine at one point in time that computational results are replicable. The manuscript and referee report may not provide sufficient information for future replication by the authors, reviewer or the wider community. Such expectations are reserved for the future. Presently, we are simply introducing the minimal expectation that results should be replicable, which is sufficient to already change upstream behavior and increase confidence in published results.

Evaluation Phase. During the initial use of the Replicated Computational Results Designation activity, we will work closely with interested authors to assess what works and what needs improvement.

Associate Editor for Replicated Computational Results Activities. We may need to consider an editorial role focused on RCR activities. This person could be a resource for other associate editors when looking for an RCR reviewer.

Requiring Replicated Computational Results. Eventually, we want RCR designations for all TOMS publications. Because TOMS is about mathematical software, it is natural to use it as an early adopter of RCR activities and requirements. We hope that the TOMS RCR efforts will inform efforts of other publications in the future.

RCR Review Process. As the RCR process matures, we may discontinue the publication of the RCR reviewer's report since there is an obvious conflict of interest when the publication of reviewer report depends on the successful completion of the RCR process. Even so, presently this risk is worth taking.

3.4. ACM TOMS Algorithms Papers

ACM TOMS has two primary submission categories: Research and Algorithms. The RCR policy as stated here should be clear for research papers, which typically do not include a review of the discussed software, nor the software itself. However, Algorithms papers do include the submission and review of software. The RCR process is not a

replacement for the Algorithms submission software review process. Instead, the RCR process will add a replicability expectation to the existing process.

4. SUMMARY

The ACM TOMS RCR process has started its evaluation phase. The lead article in this issue of TOMS [Van Zee and van de Geijn 2015] and the associated RCR report [Willenbring 2015] are the first of what we hope will be a growing number of RCR designated publications. We view the introduction of the RCR process to be a community effort and welcome feedback and ideas.

ACKNOWLEDGMENTS

Many people have played a role in defining the criteria and process for the ACM TOMS RCR Initiative. In particular, we thank the members of the ACM publication board, especially Ronald Boisvert and Laura Lander; the members of the ad hoc committee who helped develop the details of the RCR review process, especially Ronald Boisvert (again), Sivan Toledo, George Biros, Padma Raghavan, Sherry Li and Timothy Trucano; the ACM TOMS Algorithms Editor Tim Hopkins, the ACM TOMS associate editors, especially Clément Pernet, Nick Gould and Jeremy Siek; community members Randall LeVeque, Juliana Freire and Joseph Hellerstein. I finally thank Field van Zee and Robert van de Geijn, the authors of the first RCR designated paper and James Willenbring, the first RCR reviewer, for their willingness to work through the first RCR process and give substantial advice for improving it.

REFERENCES

John Arrowsmith. 2011. Trial watch: Phase II failures: 2008-2010. *Nat. Rev. Drug Discov.* 10, 5 (May 2011), 328–329. http://dx.doi.org/10.1038/nrd3439

David H. Bailey. 2014. Fooling the masses: Reproducibility in high-performance computing. https://www.xsede.org/documents/659353/703287/xsede14_bailey.pdf.

David H. Bailey and Jonathan M. Borwein. 2014. Set the default to "Open": Reproducible science in the computer age—David H. Bailey. http://www.huffingtonpost.com/david-h-bailey/set-the-default-to-open-r_b_2635850.html.

C. Glenn Begley, George Robertson, Ken Kaitin, Khusru Asadullah, Lee Ellis, Phil Sharp, and Sharon Begley. 2014. In cancer science, many discoveries don't hold up—Reuters. http://www.reuters.com/article/2012/03/28/us-science-cancer-idUSBRE82R12P20120328.

A. P. Davison. 2012. Automated capture of experiment context for easier reproducibility in computational research. *Comput. Sci. Engin.* 14, 4 (July 2012), 48–56. DOI:http://dx.doi.org/10.1109/MCSE.2012.41

P. F. Dubois. 2012. Testing scientific programs. *Comput. Sci. Engin.* 14, 4 (July 2012), 69–73. DOI:http://dx.doi.org/10.1109/MCSE.2012.84

Juliana Freire and Claudio T. Silva. 2012. Making computations and publications reproducible with VisTrails. *Comput. Sci. Engin.* 14, 4 (July 2012), 18–25. DOI:http://dx.doi.org/10.1109/MCSE.2012.76

M. Gavish and D. Donoho. 2012. Three dream applications of verifiable computational results. *Comput. Sci. Engin.* 14, 4 (July 2012), 26–31. DOI:http://dx.doi.org/10.1109/MCSE.2012.65

Bill Howe. 2012. Virtual appliances, cloud computing, and reproducible research. *Comput. Sci. Engin.* 14, 4 (July 2012), 36–41. DOI:http://dx.doi.org/10.1109/MCSE.2012.62

Brian Owens. 2011. Reliability of new drug target claims called into question: Nature News Blog. http://blogs.nature.com/news/2011/09/reliability_of_new_drug_target.html.

V. Stodden, D. H. Bailey, J. Borwein, R. J. LeVeque, W. Rider, and W. Stein. 2013. Setting the default to reproducible: Reproducibility in computational and experimental mathematics. http://icerm.brown.edu/html/programs/topical/tw12_5_rcem/icerm_report.pdf.

Field G. Van Zee and Robert A. van de Geijn. 2015. BLIS: A framework for rapidly instantiating BLAS functionality. *ACM Trans. Math. Softw.* 41, 3 (May 2015).

P. Vandewalle. 2012. Code sharing is associated with research impact in image processing. *Comput. Science Engin.* 14, 4 (July 2012), 42–47. DOI:http://dx.doi.org/10.1109/MCSE.2012.63

James M. Willenbring. 2015. Replicated Computational Results (RCR) report for "BLIS: A Framework for Rapidly Instantiating BLAS Functionality." *ACM Trans. Math. Softw.* 41, 3 (May 2015).

Received December 2014; revised February 2015; accepted March 2015