Reproducibility: Evaluating the Evaluations

Daniel Lopresti[1](✉) and George Nagy[2]

[1] Lehigh University, Bethlehem, PA 18015, USA
lopresti@cse.lehigh.edu
[2] Rensselaer Polytechnic Institute, Troy, NY 12180, USA
nagy@ecse.rpi.edu

Abstract. Evaluation is at the heart of reproducibility in research, and the related but distinct concept of replicability. The difference between the two is whether the determination is based on the original author's source code (replicability), or is independent of the code and based purely on a written description of the method (reproducibility). A recent study of published machine learning experiments concluded that only two-thirds were reproducible, and that paradoxically, having access to the source code did not help with reproducibility, even though it obviously provides for replicability. Reproducibility depends critically, then, on the quality and completeness of both internal and external documentation. The growing popularity of competitions at pattern recognition conferences presents an opportunity to develop and disseminate new best practices for evaluating reproducibility. As an initial step forward, we collected the final reports and reviewed the competition websites associated with recent ICPR and ICDAR conferences. We used this data from 42 competitions to assess current practices and posit ways to extend evaluations from replicability (already checked by some competitions) to reproducibility on application-oriented data. We recommend empirical standards, monitoring competitions, and modified code testing to be considered and discussed by the research community as we all work together to advance the desirable goals of conducting and publishing research that achieves higher degrees of reproducibility. Competitions can play a special role in this regard, but only if certain changes are made in the way they are formulated, run, and documented.

Keywords: Pattern recognition research · Performance evaluation · Reproducibility

1 Background

Reproducibility and replicability are critical criteria for evaluating reports of experimental research. While the precise definitions of these two terms is open to debate, and different scientific disciplines have developed different preferences, here we will adopt the terminology attributed to the field of Computer Science in a recent National Academies consensus study [1], i.e., "reproducibility" refers to independent researchers arriving at the same results using their own data and methods, while "replicability" refers to a different team arriving at the same results using the original author's artifacts.

B. Kerautret et al. (Eds.): RRPR 2021, LNCS 12636, pp. 12–23, 2021.
https://doi.org/10.1007/978-3-030-76423-4_2

Interestingly, these definitions are the opposite of those specified for obtaining the "Reproducible Research" label at the RRPR 2021 workshop [2], an apparent contradiction anticipated in the National Academics consensus study, which notes that the fields of Signal Processing and Scientific Computing tend to use this other, flipped set of definitions. While a seemingly minor detail, this observation can, at times, take on major significance.

Determining if two sets of experiments obtained the "same" results, mostly similar results, or significantly different results hinges, of course, on the way the experiments are evaluated. And while it would be convenient to assume that evaluation is a mechanical process that is itself practiced uniformly by everyone working in research, this is most certainly not the case. This has implications for reproducibility as well.

Best practices in reproducibility in the pattern recognition community can be summarized based on papers presented at conferences like ICPR, and also in the competitions that are now becoming common at conferences. For the most part, published work contains hints of this when the authors of Paper A write that they have used published code from the authors of Paper B for comparison purposes, or, alternatively, that they were "forced" to reimplement an algorithm because the code was not available. We note that this is a commentary not on the first paper in question (Paper A), but rather on the paper that it references (Paper B); i.e., it is an indirect measure of reproducibility that, so far as we know, no one has attempted to study or quantify. Instead of viewing this as a criticism of the authors of Paper B (for not publicizing their code), it can be considered a complement (publishing a paper that is clear enough that the idea can be reimplemented by others). There is also a tacit assumption that the authors of Paper A have done a good enough job reimplementing the algorithm from Paper B to make for a "fair" comparison, although, of course, there is an inherent conflict of interest, and doubts have sometimes been raised about this, often by the (outraged) authors of Paper B. Sometimes authors point out that they are not able to achieve the same level of performance as was previously published by the original authors; this can be seen as a positive (operating in the interests of full disclosure), or as a complaint (a suggestion that the authors of Paper B did not do a good enough job making their work reproducible). We should also note that simply providing code and data online in a public repository does not by itself satisfy the definition of reproducibility because, as we have noted, this requires "independent researchers arriving at the same results using their own data and methods" (more on this later), although it may very well satisfy the definition of replicability ("a different team arriving at the same results using the original author's artifacts").

2 A Relevant Experiment on Reproducibility

The issue of reproducibility (and the lack thereof) was the focus of a recent study by Edward Raff [3]. Raff uses "reproducible" and "replicable" interchangeably in his writing, but performed his initial analysis of 255 published machine learning papers without looking at the original source code, which satisfies the definition we have adopted for "reproducible." Hence, this is the terminology we will use in summarizing his conclusions here.

Raff found that reported results could be successfully reproduced in only 63.5% of the cases, a somewhat disturbing outcome for those of us working as researchers in

pattern recognition. He used 26 different features to characterize the selected papers, broken into three different broad categories: unambiguous features (e.g., features that are well defined and can be easily counted, like the number of authors, the number of references, or the publication type: book, journal, conference, workshop, or tech report), mildly subjective features (e.g., the total number of tables in the paper, which as we know can be somewhat difficult to count, or whether all of the hyperparameters are completely specified), and subjective features (e.g., the number of "conceptualization" figures, the algorithm's difficulty, or the paper's readability). Raff found 10 features to be important at a level of statistical significance for predicting reproducibility: readability had the largest impact, but also significant were rigor vs. empirical (whether a paper is more theoretical or more practical), algorithm difficulty, the presence of pseudo code, the broad subject area of the paper (e.g., the specific branch of machine learning), the number of tables (positively correlated with reproducibility) and the number of equations (negatively correlated with reproducibility), and the computing environment (higher reproducible rates for work run on a GPU, and lower reproducible rates for work run on a cluster). The responsiveness of authors to email queries was also significant in predicting the reproducibility of work reported in their paper.

With some additional work, these observations could form the basis for new evaluation paradigms for reproducibility, a topic worthy of discussion within the pattern recognition research community, and one of our primary recommendations. But perhaps the most counter-intuitive conclusion from Raff's work is his discovery that whether or not a paper's authors released their code had no significant relationship to the paper's independent reproducibility. He posits that perhaps such authors include less detail in their papers because they assume readers will find it in their code. It might seem like authors who release code are signaling that they care more about reproducibility, which makes the lack of a correlation especially surprising. Our conclusion for efforts such as RRPR's "Reproducible Label" initiative is that access to and confirmation of a paper's source code demonstrates replicability, but cannot make claims regarding a work's reproducibility. This same point is almost certainly true of the competitions that have proliferated in the pattern recognition research community. Acknowledging this, we believe it would be useful for the community to have an ongoing discussion regarding the relative importance of reproducibility and replicability, and to take actions designed to take quantifiable steps in the direction of improving both of these measures in our work.

3 Examining Community Practices via Competitions

When it comes to measuring the "status quo" in the community, we have made the decision, for now, to focus on the competitions (sometimes called "contests") that take place at major conferences rather than on individual published papers. Competitions can play an important role in fostering reproducibility, but to do so, certain changes must be made in the way they are organized, run, and documented. We survey current practices and make recommendations for adapting them. This is based on the premise that extra care is exercised by authors and by evaluators in setting up and running competitions, so in some sense they should reflect the "best case" scenario, and also because there are far fewer competitions than published papers to survey, reducing the need to find an unbiased way to sample a very large population.

Competitions do not have as their traditional focus reproducibility, but they may insist on replicability by, for example, requiring submission of working code in order to participate. It could be argued there is little reason to replicate or reproduce a method that yielded poor results on the tasks set by the competition. This is a different scenario from *publication*, where the presumption is that the proposed method is, in some sense, the best known so far, at least according to the proposers. We see, however, no reason why competitions could not more explicitly encourage and measure reproducibility in pattern recognition research, and this is another of the suggestions we offer. In doing so, competitions could lead the way in establishing best practices that will be more broadly followed once understood and accepted by the community.

We have gathered data on competitions and contests organized at two of the largest, most important conferences in our field: The International Conference on Pattern Recognition (ICPR) and The International Conference on Document Analysis and Recognition (ICDAR). Both take place every two years, in alternating years (under normal circumstances). Interestingly, while ICPR is the larger conference often drawing over 1,000 attendees, ICDAR, with around 500 attendees, fields many more competitions: for example, ICDAR 2019 had 27 competitions vs. 4 for ICPR 2018, and ICDAR 2017 had 25 competitions vs. 7 for ICPR 2016. These large differences can, of course, be explained by the traditions of the two conference series.[1]

Here we report what we have discovered about the practices of ICPR and ICDAR competitions. Our focus is on issues relating to reproducibility and replicability, and opportunities to inject more of these two considerations into competitions. Prior to doing our survey, we expected that we would find two common models. In one model, all of the experiments are performed by the competitors. There is a training set released well in advance, and a test set that is distributed with a limited amount of time to report results back to the competition organizers. The other model is that competitors must submit their code to the conference organizers who will then run it on new, previously unseen data. Even the latter is, at best, evidence of replicability, not reproducibility. The former tells us nothing, really, about either. However, it is possible the competition organizers will have made a separate effort to read and analyze an accompanying paper to try to verify whether the reported experimental results are "plausible." We looked for evidence of this in the published competition reports and on associated websites.

Since it seems unlikely competition organizers would independently implement methods published in a paper due to the amount of work that would be involved, what might it mean when they take code provided by a competitor and run in on new data, obtaining results that appear consistent with what the competitor has demonstrated through their own experiments? Could this be called a weak form of "reproducibility"? Perhaps a better term would be "robustness" or "generalizability." This makes a case for competitions that push algorithms to the edge of breaking, otherwise what is learned from the competition does not substantially improve on what already appears in published work by the competitors or other authors.

[1] While ICPR and ICDAR seemed to us to be two obvious candidates to study, as noted by one of the reviewers there are, of course, many other relevant examples that may be instructive to consider, including Kaggle, the KITTI Vision Benchmark Suite, ImageNet, and reproducedpapers.org, among others.

It is also possible to imagine competitions that truly measure reproducibility. While requiring more work, this would likely provide much more value to the community. Much of the programming required by participation in a competition already devolves to Ph.D. students. Attempting to reproduce published work seems like another ideal task for students and early career researchers, many of whom are already doing this anyway as part of their entry into the field (perhaps such efforts could also include retired experts who still wish to remain connected to the community, as suggested by one of the reviewers of this paper). As of now, important work like this receives little credit within the research community, but new forms of recognition seem possible, maybe even publication credit or awards, for those who support competitions by evaluating and reporting on the reproducibility of published methods. This is another one of our suggestions for advancing the field.

Our survey results for the ICPR and ICDAR competitions are presented in Table 1. We evaluated the following features in each case:

- Announced Competitions: as per the main conference website.
- Held Competitions: competitions that actually took place, as evidenced by a written report. As can be seen, of the announced competitions, only 67% were actually held. The remainder were cancelled. In most cases the cancellations are indicated explicitly, often on the competition webpage, but sometimes even this minimal information is missing and it is only the lack of a website or a report that indicates it was cancelled.

All the rest of the percentages below are calculated relative to the competitions that were held, not the number that was originally announced:

- Competition Website - Active: whether the competition website still responds with valid information about the competition. This is true for 78% of the competitions.
- Competition Website - Competition Reproducible: whether (in our judgement) the competition protocol could be reproduced based purely on information present on the website. This is true for 71% of the competitions (clearly if the website is no longer responsive, the competition cannot be reproduced from the website).
- Final Report - In Proceedings: whether the final report was published in the conference proceedings. This is true for 93% of the competitions. While clearly it is desirable for 100% of the reports to be published in the proceedings, whether it is feasible depends on the interplay between publication deadlines and the timing of the competitions.
- Final Report - On Website: even with the report appearing in the conference proceedings, it would seem to be valuable to also include it on the website, but we found only 14% of the final reports are on the competition websites. (Sometimes there are graphs of the final results on the website, but no written analysis – we counted this only if the final written report, or something close to it, was on the website.)
- Final Report - Competition Reproducible: a companion to the information appearing on the website, this answers the question whether the competition can be reproduced by what appears in the written report. This was true for 83% of the competitions. We found that sometimes the written report does a better job in this regard, and other times the website does a better job.

Table 1. Survey of ICDAR and ICPR competitions.

	Conference	ICDAR		ICPR		Total
		2019	2017	2018	2016	
Status	Announced Competitions	27	25	4	7	**63**
	Held Competitions	**18**	**17**	**3**	**4**	**42**
Competition Website	Active	16	11	2	4	**33**
	Competition Reproducible	15	9	2	4	**30**
Final Report	In Proceedings	17	17	3	2	**39**
	On Website	3	3	0	0	**6**
	Competition Reproducible	17	14	2	2	**35**
Entries Reproducible	Some	16	11	2	2	**31**
	All	1	0	0	1	**2**
Tests Run By?	Participants	16	15	3	2	**36**
	Organizers	2	2	0	2	**6**
	Code Required?	5	4	0	2	**11**
Public Data?	Some Data	8	6	1	3	**18**
	All Data	5	6	1	3	**15**
	With Registration	10	7	1	0	**18**
Public Source Code?	Some Code	6	5	0	0	**11**
	All Code	0	0	0	0	**0**

- Entries Reproducible - Some: whether, in our judgement, sufficient information is included for at least some of the contest entries to be able to reproduce them. This will often mean a reference to a full-length published paper describing the method, sometimes supplemented by the source code on a public repository. Entries judged not to be reproducible are described only very briefly – often in a short paragraph using only very general terms (e.g., "we used a Recurrent Neural Network"). In a few cases, there is no description of a method whatsoever. We determined that for 74% of the competitions, at least some of the entries were reproducible. If just one entry in a competition was judged reproducible, we would count it here.
- Entries Reproducible - All: using the same criteria as above, but requiring all of the entries to be described in sufficient detail to be able to reproduce each one of them. Of the 42 competitions we studied, only two of them provided enough information to reproduce all of the entries (one each from ICDAR 2019 and ICPR 2016). This is a major area for improvement we recommend moving forward.
- Tests Run By? Participants: the competition participants ran their own code and submitted results to the organizers to be judged.
- Tests Run By? Organizers: in a few notable cases, the competition required entrants to submit code in one of a number of standard formats to the organizers which they then ran. This was true for 14% of the competitions.

- Tests Run By? Code Required?: in this case, the competition rules state that entrants must provide their code to the organizers. While this might seem to be identical to the previous measure, some competitions require participants to run their own code and submit results for judging, and also to submit their code for "verification." This is a good idea. However, we were unable to find explicit statements in the reports or on the websites for those competitions suggesting that the code had actually been verified. We believe it would be improper to take the lack of any sort of comment as confirmation the verification had been performed and passed. We suggest an explicit affirmation should always be included.
- Public Data? Some Data: at least some of the data used for the competition is public via links on the website or in the report. In some cases, competitions claim to be using public data, and this may in fact be true, but if a link was not provided to confirm accessibility, we did not call it public.
- Public Data? All Data: all of the data used for the competition is public via links on the website or in the report. Only 36% of the competitions make all of their data public, at least at the time of this writing.
- Public Data? With Registration: a number of the competition websites state that the data can be obtained only after registering for the competition. This may or may not be accompanied by a claim that public data is being used (e.g., it is data from a digital library), but if registration is required, then the data is hidden behind a "wall" and not truly public.
- Public Source Code? Some Code: source code for at least some of the entrants is available on a public website (e.g., GitHub) linked by the competition. This is true for 26% of the competitions. Note that just because code is required by the competition, the code is not necessarily made public. (In some cases, the organizers took executable code and not source code.)
- Public Source Code? All Code: source code from all of the competitors is available on a public website linked by the competition. This was not the case for any of the 42 competitions. There is, of course, a tension here between allowing companies to participate in competitions while protecting their intellectual property on the one hand, and encouraging the open sharing of ideas which is the hallmark of reproducible research on the other.

In conducting our survey, we experienced a number of frustrations that can also be seen as negatively impacting reproducibility. This included links to competition websites that no longer work, and links that work but now point to new, completely different activities with no hint of the previous competition. We found one published report that described the setup of the competition, but did not provide any of the results. Another final report was published in an unrelated journal, but not in the associated conference proceedings. We saw reports that only vaguely identified the contest participants, let alone provide sufficient details for reproducibility. Finally, there were a couple competitions that failed to generate outside interest for some of the tasks, so there was no real "contest," but the organizers still produced results to include in the report by running their own code.

While all the competitions did a good job describing their evaluation metrics, many of them were not completely clear on the data they used, often only generally referencing

drawing it from a larger collection (e.g., a digital library) and sometimes including a small set of sample images on the website. Surprisingly, it also sometimes required some digging to determine that participants ran their own code – this seems to be a tacit assumption that was not always explained clearly.

Our survey was admittedly simple and limited in its scope to what was publicly visible at a time well after the competitions took place. We suspect some important information was only conveyed via email exchanges between the organizers and the competitors, and never recorded anywhere else. This seems like a reasonable expediency, but it hurts later reproducibility; all of the details ought to be fully documented.

Competitions are becoming increasingly popular and play a valuable role in conferences such as ICPR and ICDAR. They have the power to focus attention on problems the community considers important, and to drive the field forward. Our colleagues who devote substantial time and effort to organizing these activities rarely receive credit commensurate with the workload they take on. Nothing in our analysis is meant to be critical of the contributions that have been made so far, but rather an attempt to rally the community around developing best practices for reproducibility. Competitions could play a valuable role in this regard, if more attention is focused on the details.

4 Empirical Standards Favoring Reproducibility

As we have noted, readability was found to be the most critical factor by Raff in his work on reproducibility. Building on this, we might imagine developing a "best practices checklist" to be used by authors when they writeup their results, and by reviewers when evaluating submissions for conferences and journals.

In our search for additional clarity, we find it instructive to turn to a series of recent best practices discussions that have taken place in the programming language community (ACM SIGPLAN), resulting in a set of Empirical Evaluation Guidelines which are formulated as a single page (poster-format) "checklist" proposed for use when evaluating papers for publication [4]. Similar discussions having been taking place in the fields of software engineering [5] and computer graphics [6].

One community exercise that is interesting to ponder would be building on the SIGPLAN checklist, keeping aspects considered useful in pattern recognition research, and deleting or modifying those which do not apply as currently stated. The goal would be to advance reproducibility beyond its current levels, without creating too great an added burden on already busy reviewers, conference chairs, and editors.

5 Program Integrity and External Dependencies

Program bugs can, of course, introduce security risks as well as hamper reproducibility. They are seldom revealed by replication on the same data. Among the many tools available for detecting bugs, fuzz testing with random inputs is attracting much current attention [7]. These tools are not specific to document image analysis, where more targeted variations in input are desirable. For example, egregious paragraphs consisting of only a few words, mathematical formulas or unusual page-breaks may affect segmentation and layout analysis. Some scanners exhibit ambient light leaks resulting in border

noise. (Even the same scanner generates different bitmaps on successive scans of the same page.)

Classification results may depend on language-specific libraries like equation solvers. Therefore, reproducibility studies should include, beside broad test data, diverse transducers (scanners or cameras) and platforms (languages and compilers). Should they also require directions for sampling a new data source, dividing the sample into training, validation, and test sets, and rebuilding the classifier from scratch? Experiments involving human-computer interaction add further dimensions of subject, training, and protocol variability. Similar observations apply beyond the field of document analysis, of course, extending across the broad domain of pattern recognition research.

6 Suggestions for Further Action

Note that these recommendations are intended as jumping off points for further discussion, not set-in-stone policies for changing the community's current practices. Suggestions like this can serve as a basis for organizing competitions and reviewing papers, as well as for conducting experiments and writing them up for publication. As noted by the members of the SIGPLAN community [4]: "The goal of the checklist is to help authors produce stronger scholarship, and to help reviewers evaluate such scholarship more consistently. Importantly, the checklist is … meant to support informed judgment, not supplant it. The committee's hope is that this checklist can put all members of the community literally on the same page." We would echo the same goal for any similar effort within our own community.

We also note that not all attempts at building evaluation check-lists are equally helpful: if verbosity is not carefully managed, then simply attempting to read, understand, and apply the checklist becomes a chore in itself; this is one reason the authors of the SIGPLAN effort strove to fit their checklist on a single (albeit dense) page (for comparison, contrast this with the 59-page SIGSOFT effort [6]). At its heart, evaluation for reproducibility is a human factors activity, and consideration for the reviewer must be front and center.

The concept of reproducibility overlaps that of *generalizability*. Will a method that gives satisfactory results on selected data also work well enough on hitherto unseen application streams? All test data samples that we have seen are basically convenience samples, not population samples. The multitude of digital images, even when considering only document images in a specific category, discourages credible sampling. Web crawls collect huge samples, but cannot yet formulate descriptors accurate enough for reliable evaluations of generalizability beyond tiny and arbitrary test sets. Is it time to design and develop a web-scale census?

Finally, as we noted in our competition survey, evaluations for evaluating reproducibility must themselves be reproducible via archival publication of clear, complete, objective protocols. As much experimental science proceeds without theoretical foundations, the evaluation of any such process (the original experiment, assessment of its reproducibility, evaluation of this assessment …) must be open to external scrutiny. Empirical standards for evaluating the evaluations, *ad libitum*?

With these various points in mind, we offer the following specific suggestions for strengthening the role that competitions play in fostering reproducibility:

- Organizers should commit to completely and openly documenting all aspects of their competition in sufficient detail that the competition itself can be reproduced at an arbitrary later date. This means using URL's that are active indefinitely, recording all aspects of the competition protocol (including important details that may have been conveyed only via emails during the competition), and making sure the evaluation measures are both reproducible and replicable; as we have noted, these are two fundamentally different concepts. Full written reports should be included both in the conference proceedings and also on the competition website.
- Assurance of the reproducibility of an entry in the competition should be a requirement for participation. Brief one-paragraph descriptions written in general terms are not enough, nor is the release of course code by itself necessarily sufficient. Presumably competitions come toward the end of a line of research, so there should already be publications to assure authors receive proper credit for their work; secrecy and proprietary interests have no place in competitions conduced in an open research community.
- Public data should be used for competitions whenever possible. In any case, the precise data that is used, both the raw inputs and the associated ground truth, should be carefully recorded and documented. Long-lived URL's or DOI's and/or the use of public repositories should be encouraged. If the actual data cannot be made public due to usage restrictions, then sufficient meta-data should be provided that someone without privileged access to the actual data can still fully understand the nature of the competition, including the target population, the set of inputs chosen from it, and the manner in which the selection was made (i.e., which of the possible inputs were included, which were excluded, and why).
- The public release of source code is admittedly problematic when there is a desire to protect intellectual property rights. Moreover, as Raff has noted, access to source code by itself does not make a work reproducible. Competitions that do not require the public release of source code should demand a complete written description of the method sufficient to make it independently reproducible. When source code is released, a separate review should be performed to confirm that the method can still be reproduced in the event the code cannot be run for whatever reason (i.e., assuring that the code is self-documenting, or that there is a separately published paper that describes the method in sufficient detail).
- Conference organizers (competition chairs) should enforce strong documentation requirements, both as a first step toward approval and also as a final check before a competition's results are "accepted" for publication. If a competition falls short at any point in the process, it should be removed from the conference website and treated as a rejected paper is treated.

These suggestions are offered as feasible improvements; following them does not guarantee reproducibility. Nor do we suggest mandatory standards for every entry. There is no doubt, however, that better documentation enhances the value of competitions.

In this analysis we have ignored several important issues that can be problematic, including whether competitors should be allowed to run their own code (especially when methods are not reproducible), and the value in a "winner takes all" approach when differences between top-ranked entries may be slight and a simple resampling of

the test data could easily change the results. Organizers should carefully consider these questions and justify their answers to the community.

It is also important to step back from time-to-time and try to answer the question "why?" Reproducibility is a mantra in the laboratory sciences, but more difficult in the social sciences. Strains of white mice are far more homogenous than samples of students, shoppers, or voters. Information sciences occupy an intermediate position in this spectrum. Control groups of mice or cultures of bacteria are relatively easy to prepare, but control groups for most interesting pattern recognition problems are not.

For example, in the field of document analysis we do not yet have methods of collecting random samples of documents that are representative of any significant population. Yet reproducibility depends critically on a test set "similar in all relevant aspects" to the data used in the competition (and, in some competitions, also on the training and validation sets). What would this even mean? Has any publicly available dataset been claimed to be comparable to the now-famous UW-1 or MNIST datasets? What is a random sample of historical documents, or even of 17th Century English literature? *Naked Statistics* by Whelan lists a dozen subtle yet well-known possible sources of biased samples and inappropriate metrics [8].

Nor is reproducibility important for every entry in a competition when there is knowledge to be gained purely through participating. Why discourage a pair of students with a clever idea from trying their luck against large teams of professional researchers? The million-dollar Netflix Prize was won by just such a team, but the thousands of participants in the contest undoubtedly learned valuable lessons along the way.

Competitions at ICDAR extend back at least as early as 2001 [9], and as least as early as 2000 at ICPR [10]. ICDAR 2021 promises to offer three "long-term" competitions and 10 "short-term" competitions [11]. The former is described as "open for a longer time period and address challenges which could continue over the next years." Best practices are evolving, albeit slowly. It would benefit all in the community if the same degree of rigor now applied to research paper submissions was carried over to proposed competitions and their implementation.

We expect topics like these will remain an ongoing, productive discussion within the pattern recognition research community, as reflected by the RRPR workshop.

Acknowledgements. We thank the reviewers for their carefully considered feedback and helpful comments, many of which we have included in the present version of this paper.

References

1. National Academies of Sciences, Engineering, and Medicine: Reproducibility and Replicability in Science. The National Academies Press, Washington, DC (2019). https://doi.org/10.17226/25303
2. Third Workshop on Reproducible Research in Pattern Recognition (RRPR 2020). Reproducible Label. https://rrpr2020.sciencesconf.org/resource/page/id/5. Accessed 16 Oct 2020
3. Raff, E.: A step toward quantifying independently reproducible machine learning research. In: Proceedings of the 33rd Conference on Neural Information Processing Systems (NeurIPS), pp. 5,485–5,495. Curran Associates, Inc., Vancouver (2019). https://papers.nips.cc/paper/8787-a-step-toward-quantifying-independently-reproducible-machine-learning-research. Accessed 16 Oct 2020

4. Berger, E.D., et al.: ACM SIGPLAN Empirical Evaluation Guidelines (2018). https://www.sigplan.org/Resources/EmpiricalEvaluation/. Accessed 13 Oct 2020
5. Ralph, P., et al.: ACM SIGSOFT Empirical Standards (2020). https://github.com/acmsigsoft/EmpiricalStandards. Accessed 1 Nov 2020
6. Bonneel, N., et al.: Code replicability in computer graphics. ACM Trans. Graph. **39**, 4. https://replicability.graphics/. (Proceedings of SIGGRAPH 2020)
7. Klees, G., et al.: Evaluating fuzz testing. In: Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security, Association for Computing Machinery (2018). https://dl.acm.org/doi/proceedings/10.1145/3243734. Accessed 1 Nov 2020
8. Wheelan, C.: Naked Statistics. W. W. Norton & Company, New York/London (2013)
9. Competitions, University of Salford Manchester. https://www.primaresearch.org/competitions. Accessed 17 Jan 2021
10. Aksoy, S., et al.: Algorithm performance contest. In: Proceedings of the 15th International Conference on Pattern Recognition, vol. 4, pp. 4870–4876, September 2020
11. ICDAR 2021 Competitions. https://icdar2021.org/competitions/. Accessed 17 Jan 2021