Reproducibility: Buy Low, Sell High

Roy Maxion
Carnegie Mellon University

Abstract—

Are papers in computer security reproducible? Amidst a general crisis of reproducibility in science, we address the health of the reproducibility culture in computer science and in security biometrics in particular. On the basis of various definitions of reproducible science, all of the published articles in the first quarter of 2020, in a subfield of security biometrics, were assessed with an eye to a range of concerns that impinge on reproducibility. While many of these papers met at least some reproducibility criteria, no papers met them all. Constructive advice is offered for raising awareness and for making research more reproducible in the interest of advancing both security and science more generally. Finally, we muse: in a culture of faster, better, cheaper, where does reproducibility fit? Reproducibility in general needn't cost much, but its benefit can be substantial. Buy low, sell high.

■ **REPRODUCIBILITY** is the ability to be copied or reproduced so that consistent results are obtained when an experiment or procedure is repeated. Reproducibility lies at the heart of the scientific method. When studies cannot be reproduced, their credibility is cast into doubt, sometimes forcing the retraction of a paper.

Perhaps one of the best-known reproducibility failures was cold fusion, a process by which energy from nuclear fusion was said to be generated at room temperatures [5]. Its 1989 announcement raised considerable controversy, but cold fusion was finally deemed unreproducible, effectively ending the careers of the original investigators.

Somewhat later, in the 2000s, the public perception of science entered what has been called “the reproducibility crisis,” in which many scien-

tific studies were found impossible to reproduce. An early call to arms was a paper by John Ioannidis of Stanford University: “Why Most Published Research Findings Are False [8].” An especially disturbing set of unreproducible findings was reported by Glenn Begley and Lee Ellis, then of Amgen, a pharmaceuticals company, in a 2012 paper in which they found that 47 out of 53 (89%) landmark preclinical studies of cancer drugs could not be reproduced [2]. This was put into stark relief by NPR journalist Richard Harris in his cleverly-titled 2017 book, “Rigor Mortis: How Sloppy Science Creates Worthless Cures, Crushes Hope, and Wastes Billions” [7], decrying the dearth of methodological rigor in empirical science.

In 2020 we have the issue of reproducibility in artificial intelligence and machine learning, especially as they are employed in computer security. AI/ML systems are being used to detect viruses and malware, malicious activity, network fraud and more [11]. A topic close to the security community is access authorization – how to present reliable credentials that enable access to a computer system. One way that access authorization can be managed is through behavioral biometrics, in which non-physical characteristics unique to an individual (e.g., typing rhythm, effected through keystroke biometrics) are presented to a system to facilitate access. Well over a thousand keystroke biometrics papers have appeared in the open literature since 1990, although the idea dates to the late 1800s when telegraph operators identified each other by their “fists” – individuals’ personal rhythms and spacings of Morse code’s dots and dashes [3]. Indeed, keystroke biometrics has found a place in commercial practice.¹

Since unreproducibility seems so rampant in other disciplines, we are motivated to audit an area of computer security with which we are familiar – behavioral keystroke biometrics – to see how we’re performing with respect to the rest of the world. If the audit finds us wanting, we should be able to suggest some simple measures to improve our performance.

In this paper we conduct a reproducibility audit to identify signs that published experiments or results may not be reproducible. The paper starts with a conceptual description and scoping rationale for our reproducibility audit, followed by definitions of reproducibility and a primer on keystroke biometrics to acquaint readers with the fundamentals of the discipline. (Un)reproducibility indicators pertaining to data collection and analysis are given, scoring the papers pass/fail on reproducibility criteria, with examples provided. Results are presented, along with final discussion points.

REPRODUCIBILITY AUDIT

Concept

A reproducibility audit of an empirical paper is a systematic examination in which the likelihood of reproducibility is assessed against a set

of criteria. The audit explores what a paper reveals about its research endeavor that’s necessary for reproducibility: what apparatus and materials were used; if human subjects were involved, how many were there, why, and what population did they come from; what was the source of data, and how was it procured and analyzed, etc?² If the details of these and other elements of the paper are not transparently provided, then reproduction will be hampered or not possible. It may seem obvious that these things should be included in a research report, but a surprising number of papers fall well short on these and other measures.

An audit goes beyond a current assessment of the field. Non-reproducibility is just the beginning of a larger problem in that readers may unwittingly accept an unreproducible and perhaps invalid result at face value, and base their own research on it. They then risk propagating unsupported claims into the future, invalidating not only their own work but also the work of others.³

Scope

The area that we choose for audit is keystroke biometrics, an authentication technique for improving computer security. In addition to being familiar to us, the field has an active community of researchers doing experimental work, and its lessons transfer easily to other domains. In addition, there are numerous extant datasets that are available for testing new analysis techniques with the objective of providing improved results. There are wide variations in experimental and analytic approaches, and in the kinds of equipment used, so reproducibility requires clear and complete documentation. The audit process examines these details to determine their impact on the experimental work being reproduced.

While our narrowing of scope limits the topics that could be addressed, it also provides a focus from which much can be learned, whether one is personally involved in keystroke research or not. Since abstraction is said to be one of the defining principles of computer science, we hope that experimental researchers in subdisciplines of the field can abstract appropriately relevant approaches and apply them in their own work.

¹www.behaviosec.com claims “over 99% accuracy.”

²Full description of (un)reproducibility indicators, page 4.

³This has already happened; see Covid19 pandemic: [4]

DEFINITIONS OF REPRODUCIBILITY

What, exactly, does “reproducibility” mean? Definitions abound, and there is modest controversy over them. Well-known institutions have weighed in: ACM with their reproducibility badges [1], the international Joint Committee on Guides for Metrology [9], and the National Academy of Sciences [10]. The relevant terms are: repeatability, replicability, and reproducibility. The National Academy adds computational reproducibility. Goodman et al. [6] argue persuasively that none of these terms is well defined, and even when fulfilled they don’t satisfy the higher goal of determining truth. They distilled sensible and comprehensible definitions,⁴ which we show here, supplemented by the Academy’s computational reproducibility [10]. The question to be considered is: Given what is reported in the paper, could someone of ordinary skill in the art to which a paper pertains re-run the experiment and, within the margins of experimental error, obtain the same results? Definitions follow.

Methods reproducibility: obtain the original study’s exact same results, in theory or in actuality, using the original procedures and data.

Results reproducibility: obtain the same results from the conduct of an independent study whose procedures are as closely matched to the original experiment as possible.

Computational reproducibility: obtain consistent computational results using the same input data, computational steps, methods, code, and conditions of analysis. These 3 terms generalize the usual concepts [1] of: Repeatability: repeated measurements being obtained with stated precision by the original team; Replicability: executing the same experiment with the same experimental setup by a different team; and Reproducibility: executing the same experiment by a different team with a different experimental setup.

KEYSTROKE BIOMETRICS: PRIMER

Here we provide an intuitive primer on keystroke biometrics, the target domain of our reproducibility audit. Keystroke biometrics, sometimes called keystroke dynamics, is the process of identifying individual users on the basis of their typing rhythms, for which it is generally

⁴Inferential reproducibility [6] is out of scope, hence omitted.

accepted that no two persons’ rhythms are the same. Typing rhythms are derived from the timestamps of key-press and key-release events on the keyboard. There are two phases to most keystroke experiments: data collection and data analysis. Each should be conducted in a reproducible way.

Data collection occurs while subjects are typing, and can take two forms: with and without intervention. For example, if an investigator wants to see if a certain emotion influences typing rhythms, then an intervention is required; pre-intervention typing is done first, then the emotion (e.g., stress) is induced in the typist, followed by post-intervention typing: an A-B-A design.

Data analysis is the process that determines how well a subject’s typing rhythm identifies him/her without mistakenly confusing him/her with some other subject. If a machine learning approach is used, there is usually a training phase in which a model (a classifier) for a subject is formed, and a testing phase to determine the extent to which confusions exist among subjects. Measures used to judge accuracy are centered on the signal-detection theory concepts of hit, miss, false alarm and correct rejection rates. If the data collection involved an intervention, the effects of the intervention are also subject to evaluation.

METHODOLOGY

We provide our survey and evaluation methodology so that others can reproduce our searches, actions, decisions, and results.

Searching for papers

Keystroke dynamics is the original term for this technology, originating over 30 years ago. The technology was later recognized as a biometric, and later still as a behavioral biometric. We searched five on-line databases, using “keystroke dynamics” and “keystroke biometrics” as search terms, limited to year = 2020; this was done on 01 April 2020, capturing papers published in the first quarter of the year. The databases (and number of papers found) were:

- Web of Science (5)
 - Scopus (20)
 - IEEEExplore (3)
 - ACM (3)
 - Google Scholar (162)
- Total: 193

Exclusion criteria

A broad search like this typically turns up suitable and unsuitable results, and the latter need to be excluded. Duplicates were eliminated, as were survey papers and papers that only tangentially mentioned keystrokes. Papers mentioning authentication were retained. Papers were limited to those whose references to keystroke-related work were no later than 2012, in an attempt to ensure that authors had ample opportunity since then to be aware of and take action against non-reproducibility. While this procedure might not have identified all of the keystroke papers published this year, it did return a suitable sample for reproducibility assessment. Google Scholar returned a large number of documents because it indexes papers' bibliographies and terms in foreign-language papers. Out of the 193 papers from the search, 170 were excluded, leaving 23 experimental papers for reproducibility assessment. Among those papers were journal papers (13), conference papers (8), one workshop paper, and one book chapter. Sponsorship was attributed to national granting agencies (8), academic institutions (3), corporations (4) and unstated (8).

Evaluation criteria

In an assessment of reproducibility there are certain questions to be asked of each paper under review. A codebook, containing evaluation criteria, provides a structured list of questions or issues to consider in an assessment, along with a set of fixed responses to these criteria (e.g., yes vs no, or present vs absent). For example, one question was: did the investigator make his/her data publicly available, for which the responses would be yes or no. The codebook used here is available on request from the author.

(UN)REPRODUCIBILITY INDICATORS

The details that indicate reproducibility are similar to the details of a recipe: a list of ingredients, amounts, equipment/utensils, and preparation instructions. If any of these is missing, the cook will be unable to faithfully (re)produce the dish. Similarly, leaving out details from data collection or analysis protocols can cripple a dependable reproduction, or render its reporting incomprehensible. Insufficient or omitted detail is regarded as a fault, and scored as "fail."

We focus on a set of methodological elements that readers need to know to be able to reproduce a study. In each of these categories we will note the things that facilitate reproduction and will help readers understand how the experiment was done. These elements are collectively referred to as the "method" by which an experiment was done. In the following subsections we will identify and comment on the dominant indicators of (un)reproducibility found in our review. These are divided into two main groups: the data collection protocol and the data analysis protocol.

Data collection protocol

To fully support reproducibility, the data collection aspects of a study should be clear and well organized. Just as a recipe is easier to follow if the entire recipe is (a) in well-defined sections, and (b) on 1-2 consecutive pages, an experiment will be easier to reproduce if the details are reported in a similar fashion using standardized, labeled sections grouped together in the paper.

When collecting keystroke data, details pertaining to exactly how the experiment was done and what was involved become important, not only for replication or reproduction, but also to provide the reader with enough context to judge whether the undertaking was sound. We start by discussing the apparatus employed in an experiment, and then move on to other aspects of the experimental method.

- **Apparatus & instrumentation (App):** Experimenters, authors and reviewers need to know enough of the equipment details to discern any factors that might impinge on the reproducibility or on the validity of the experiment itself, or on its claims. There should be enough information to allay doubts about the true cause of the reported results, or to replicate the apparatus and experimental conditions. This is true even when a dataset from another investigator is used; quoting apparatus details from the original source makes the secondary paper self contained. If the alternate dataset doesn't contain apparatus details, it may be better to find another source.

Almost all keystroke experiments make use of a range of equipment: computers, keyboards, timing devices, affect-inducing mechanisms, custom instrumentation, etc. A description of the apparatus and equipment supporting an experiment

should include all details that could affect the outcome of the experiment.

The make and model of the keyboard, whether it's wireless or wired, USB or PS/2 (which can have very different timing characteristics), uses a particular switch type (e.g., mechanical, membrane, dome, scissor, spring, optical), or has a country-specific key layout (US, European, Russian) is critical information for keystroke research. If there is something about the keyboard (e.g., Cherry MX blue/red keys, etc.) that could make a difference in an experimental outcome, then include that in the description, too. Include make, model, operating system, OS version, for all computers. Describe operating conditions including ancillary services like networks, backups, system updates, or anything that could put transient loads on the machine. Revision numbers or other identification pertaining to firmware embedded in the computer or any ancillary device should be included. Special or custom equipment (e.g., key-press and key-release timers) should be described, along with their validation and calibration procedures. Provide the keystroke timing resolution. Timing details matter; the coarser the timing is, the more likely that keystroke events will be missed or mistimed. Windows operating systems have a default tick-timer resolution of 15.625 milliseconds, although other timers are available; Linux typically times at 10 ms; custom timers have custom resolutions, usually finer.

If an intervention was used (e.g., to induce affect in a subject) and it involved special apparatus, that apparatus should be described. For example, if blood pressure plays a role in an experiment, then the BP monitor (make, model, error/variability characteristics, and calibration date) should be provided. Not all BP monitors are created equal. A substandard one can affect experimental validity and reproducibility as well.

Commercially available equipment can be briefly characterized, annotated with its name, manufacturer, and model number. Descriptions of custom apparatus may be accompanied by a photograph. Manuscripts failing a reproducibility audit in this category were 16 out of 23: **70%**.

- **Materials (Mat):** Most keystroke biometric experiments involve stimuli of some kind. These usually take the form of character strings to be typed, and are referred to as the "materials"

needed to prompt a subject's action. All such materials should be described in detail so that readers can understand them and use them in replications or follow-on experiments.

Any material presented to the subject to be typed or used to inspire the subject to type freely should be included in the paper. Readers need to know the stimulus items and how many repetitions of each item were typed. For example, if a mock password is assigned to be typed a specified number of times (either accepting or rejecting typographical errors), the mock password should be included in the paper. For free text, the prompting topic should be given and samples of the resulting text included, possibly in an appendix. If free text is inspired by pictures, the source of the pictures should be given. For all experimental materials, include the rationale for having chosen them, so that readers can judge whether or not the material is well-suited to the task. If there was no particular reason or rationale for any of these factors, then say so.

If experiments induce or measure affect, mood, or other conditions, then affect-inducing sounds, pictures, lights, videos, etc. should be reported, with their sources cited. One example is the IAPS International Affective Picture System (developed at the University of Florida), a set of pictures whose effect on people is described and accompanied by a normative rating for each picture; a measurement instrument is the NASA TLX scale for reporting cognitive load or stress.

Some investigators administer brief demographic surveys to ascertain various characteristics of subjects – gender, age, handedness, injuries, current emotional state, etc. Such surveys need not be included in a paper unless out-of-the-ordinary questions are asked; these and more-targeted questionnaires can be briefly noted in the running text or included in an appendix. The reason for stating the questions asked on a survey is that the particular phrasing of the question can influence a respondent's answer; if a question isn't explicitly stated, then readers may make wrong assumptions about it. That said, simple survey questions about gender, age, handedness, etc. are not likely to be misconstrued.

If a dataset from another investigation is used, its materials should be included in the current paper in the interest of self containment.

Many of the studies reviewed here did not report details of their experimental materials – for some, not even the typed strings – putting reproducibility of those experiments at risk. Manuscripts failing a reproducibility audit in this category were 11 out of 23: **48%**

- **Subjects (Sub):** Investigators should provide details regarding who participated in the study, how many there were, where they came from, and how they were recruited (e.g., posters, agencies, course credit, etc.). The population from which subjects were drawn should be noted (e.g., undergraduate students at the university, ambulatory patients from a Parkinson's clinic, etc.) along with an explanation of how the subjects are representative of the intended sample population. The sampling method (convenience, snowball, stratified, etc.) should be given, perhaps with an explanation for the choice of method. If subjects were compensated, then state the compensation and amount (e.g., cash, gift card, course credit). If no compensation was given, say so. A sound scientific rationale (e.g., statistical power analysis) should support the choice of number of subjects used. High-level demographic characteristics (e.g., gender, age, etc.) should be provided. Other characteristics, such as handedness, should be given if they could have an effect on the outcome.

Inclusion and exclusion criteria should be provided so that readers are aware of the homogeneity (or diversity) of the subject pool. For example, in a typing task one might reasonably exclude anyone unable to type 30 words per minute, or anyone with a condition, even temporary, that would affect the task performance. The final selection of subjects should eliminate bias; for example, seek balance in the number of professional vs novice typists.

If subjects withdraw from the study, the reason should be given. Sometimes these reasons reveal biases in the subject pool that could change the conclusion of an experiment. Unless we know at least the basic characteristics of subjects and how they were obtained, any reproduction could fail to faithfully represent the original experimental conditions. Manuscripts failing a reproducibility audit in this category were 15 out of 23: **65%**.

- **Instructions to subjects (Ins):** What, exactly, were the subjects instructed to do? Were they asked to type as they normally do, or to type

as quickly, or as accurately, as possible? Perhaps they've been asked to type with only their index finger, on a particular section of the keyboard. If readers don't know, they cannot understand, much less repeat the experiment.

Some experiments address affective states, perhaps assessing changes in typing behavior before and after an emotionally charged event. Subjects are asked to engage in some affect-inducing task and later respond to queries about their affective state. Instructions regarding these activities should be reported in the appropriate portion of the text so that readers are made aware of exactly what the subject was asked to do. Unconventional instructions should be included verbatim. Without explicit instructions, subjects are left to infer their task, risking that some subjects may perform a task other than what was intended, without experimenter cognizance.

The instructions section should provide details of what subjects were asked to do, so that readers can adequately reconstruct the experiment, and perform it in the same way that the original investigator intended, without relying on assumptions. Manuscripts failing a reproducibility audit in this category were 16 out of 23: **70%**.

- **Environment (Env):** In general, experiments are conducted either in controlled conditions (laboratory environment) or uncontrolled conditions (real-world field environment). Although the distinction is often clear from context, it's important to be explicit about the conditions under which an experiment was run. For example, a field study in a shopping mall might have modest supervision by the experimenter. A field experiment over the Internet could be unsupervised, resulting in a larger experimental variance. Even in lab environments, conditions can vary – noise, lighting, foot traffic, distractions, rumbling floors near large air handlers, etc. Because environmental states and conditions can influence experimental outcomes, they should be reported. Sometimes a particularly well controlled environment can make the difference between success and failure. Descriptions of experiment environments need not be wordy, but it's important to note unusual or unexpected circumstances, particularly in otherwise controlled conditions. Manuscripts failing a reproducibility audit in this category were 19 out of 23: **83%**.

- **Design:** This section should provide the gist of how an experiment was put together – a high-level sketch providing an account of the factors that contribute to addressing the research question. What variables were sought in data collection, and used in data analysis? Usually the dominant variables or “features” include the amount of time a key is held down and the time taken from one key-press (or key-release) to the next. Investigators sometimes explore derived features, such as several contiguous key-presses and key-releases, or even key-events from different regions of the keyboard. Whatever the features are, they should be explicitly identified and the procedure for extracting them from the raw data noted in the design (or data analysis) section. If the project compares two (or more) groups of subjects or algorithms or tasks, then these elements should be described along with the procedures for accommodating them. A consistent framework helps in reconstructing the experiment.

Although a section describing the experimental design can be very helpful in reproducing an experiment, the lack of a design section won't, in and of itself, prevent reproducibility. We decline to score the papers on this criterion, while noting that no paper under review contained the material that one would expect in a design section.

- **Procedure (Pro):** The complete experience of what the subjects did during the experiment, from the moment they started to the moment they finished, should be described here, chronologically. For example, a subject arrived in the lab or other experiment setting; was welcomed; was asked to read and sign a consent form; filled out a demographic questionnaire; confirmed the instructions spoken aloud by the experimenter (or read by the subject); was familiarized with equipment; had electrodes attached (the placement of which should be described in the apparatus section); engaged in a warm-up or practice exercise; executed the experiment; was debriefed and compensated; and left the laboratory. For on-line or field experiments, the procedure may be slightly different, but the same kinds of things should be described. After the subject leaves, any action by the experimenter, such as archiving and documenting data, should also be noted.

If there was an intervention, such as provoking an emotional response, it should be completely

described here, too, as well as any manipulation or attention checks, whether explicit or implicit.

The description of the procedure should include any details that could affect the outcome of the experiment. These details may expose idiosyncrasies or omissions that would change experiment outcomes, and alert readers to proceed with caution. Manuscripts failing a reproducibility audit in this category were 19 out of 23: **83%**.

- **Data:** There are generally two sources of keystroke data: gathered by the author, or obtained from another investigator. In either case, the data used in the experiment should be described briefly, to give the reader a high-level picture of what the data look like. For example, a dataset could be described as: each of 100 subjects typed a 10-character string (\yor6Tpsmz) 50 times a day for eight days. Whether original or imported, a data availability statement is essential. It should describe how the data can be obtained and under what conditions. If it is not available then computational reproducibility will not be possible [10]. If the experiment involves a mix of original and imported data, the availability for all sources should be given. Similar information is needed for data obtained from interventions.

Irrespective of the source of data, it should be described in detail sufficient for a reader to understand its content and structure. This includes the features contained in or derived from the data, e.g., key-press and key-release times, n-gram configurations, timing details, etc. A brief table of descriptive statistics would be a plus. Such descriptions need only convey the essence of the data used in the experiment. Manuscripts failing a reproducibility audit for data description (**DatD**) were 8 out of 23, **35%**; for data availability (**DataA**), 14 out of 23, **61%**; 7 failed both for a **30%** overall data-category fail.

Data analysis protocol

This section of a paper describes the analytical and/or statistical procedures and software used to reach the biometric results being sought, as well as for interventions and manipulation tests that ensure intervention success. For software packages, version numbers and dates should be given. Data analysis is scored across its four constituent categories, not broken down into subcategories.

Table 1. Number and percent (rounded) of papers that failed exactly N categories of a reproducibility audit.

Number of Categories Failed - N	0	1	2	3	4	5	6	7	8	9
Number of papers failing N Categories	0	0	3	1	1	3	4	3	4	4
Percent of papers failing N Categories	0%	0%	13%	4%	4%	13%	17%	13%	17%	17%

- **Analytical software:** All software (code or library), whether custom-written or contained in a package (e.g., R, Matlab, scikit-learn, etc.) should be clearly identified with a citation, a version number and a date, and made publicly available. This is to ensure that the same software is used in reproduction attempts. Software and package versions change from time to time; if a reader can't determine the version number used in the original analysis, there will be no assurance that the package will be the same in any reproduction. This admits the possibility that the original work will be declared unreproducible, when simply using the correct version would have been all right. All configuration parameters, seeds, etc., should be noted. The instructions for setting these parameters are often obscure and should be noted. If custom programs or scripts supported the analysis, they need to be made available.

- **Analytical hardware:** Some software can produce slightly different results when run on different hardware/OS platforms. For example, 32 bit and 64 bit platforms are likely to render floating point calculations differently, possibly influencing matrix inversions. Different compilers can produce different code quality and different levels of optimization, and mathematical libraries may differ across compilers. A Windows compilation may not be identical to a MacOS compilation, and the source code may differ, as well. Because the platform can influence reproducibility outcomes, it's important to report where the software was run, e.g., on a Windows PC, a MAC, a Linux box, etc., and what operating system and version were used. Stating the name of the hardware (make and model) and the name and version of the operating system is often sufficient.

- **Detector/Classifier:** It's surprising how often an author doesn't reveal the type of detector or classifier used in keystroke biometric investigations. When using a specific type of anomaly detector or classifier (K nearest neighbor, random forest, neural net, etc.) the type should be noted. If using a bespoke detector or classifier, then give it a name, along with a description of how it

works. This will alert readers to the nature of the software used, enabling apt comparisons, and allowing the use of a similar mechanism in any reproduction. Source code should, as much as possible, be made available.

- **Evaluation metric:** Most biometric system evaluation metrics have their origin in signal detection theory. They are given a range of names depending on an investigator's need, background and preference. Reproducing a study is made harder by not having certainty about the evaluation metrics used in the original paper, especially when nonstandard terminology is used. Authors should be clear about what metric is used, how it's calculated, and how it's depicted graphically in figures. Commonly used metrics are: true and false positives/negatives, false accept/rejection rates, area under the ROC curve, and equal error rate. Manuscripts failing a reproducibility audit in the overall data analysis protocol category (**Anal**) were 23 of 23: **100%**.

RESULTS

For a paper to be judged as reproducible it had to pass a yes/no test on all nine reproducibility indicators or categories, each of which was rated in terms of the percentage of papers that lacked adequate transparency; e.g., 70% of papers failed the apparatus category. A paper that failed one category may have had other aspects for which it was reproducible. If the paper failed the subjects category, for example, other parts of the paper might still be reproducible, but the omission of information in a failed category might leave readers to guess at some of the missing information, jeopardizing full reproducibility. There were nine such failure ratings,⁵ the worst of which were environment and procedure, failing at 83% each. No papers adequately described both their hardware and software platforms.

Table 1 shows the number of papers that failed the reproducibility audit in exactly 0, 1, 2 ... or 9 categories, providing a (disheartening) picture of

⁵The categories, abbreviated, are: App, Mat, Sub, Ins, Env, Pro, DatD, DatA, Anal: see pages 4-8.

the overall reproducibility of the audited papers, and perhaps a concomitant reflection on quality in general. No papers passed all categories; four papers failed all nine; another four failed eight.

Only five papers were X-reproducible, where X is one of the reproducibility types noted on page 3 – a discouraging 22% success rate. No papers fulfilled the conditions for methods reproducibility. Each of the reproducibility types requires multiple categories/indicators to pass at the same time; e.g., methods reproducibility requires a contemporaneous pass on apparatus, materials, procedure and more. The two reproducibility types that found even limited applicability were: Results reproducibility (4 papers, 17%) and Computational reproducibility (1 paper, 4%).

HOW TO IMPROVE

Our reproducibility assessment of keystroke biometrics papers in the first quarter of 2020 had a disappointing result: all 23 of the papers were found to be unreproducible with respect to one or more of the factors noted in the previous sections. Table 1 shows that 18 of the papers (78%) failed in five or more categories each. This is consistent with the Begley and Ellis paper [2], and it is an uncomfortable sign about work in our own field. Why did this happen, and what can be done to improve, both in the short and the long term?

One can only speculate as to the causal mechanisms behind the reproducibility failures. Certainly there is the oft-mentioned pressure to publish, but a number of the observed shortcomings could have been addressed without compromising publication productivity. For example, including the source and version number of an analysis package would be easy; it simply needs to be in the mind of the author (and reviewer) to include these and other critical details.

Perhaps authors are simply unaware, not realizing that omitting certain details from their papers makes it difficult for others to read, to comprehend, and to follow up on published work. Perhaps authors know that reviewers aren't harsh judges of unreproducible work, taking modest advantage. Perhaps reviewers are equally unaware, or are not required by review forms to check reproducibility, even though some reproducibility shortcomings tend to impinge on credibility, too. Indeed, there are some papers that even a gener-

ous Gedanken experiment couldn't reproduce or, for that matter, even imagine how they could have the results that they do.

If authors are concerned about being banished to the reproducibility doghouse, there are some simple short-term remedies. The community could promote awareness of the value of reproducibility, perhaps by editorials in journals or by tutorials in conferences. The community could provide better education; teaching the precepts noted here would take perhaps 2-3 lectures in a course or a research seminar. The community could provide better review forms for conferences and journals – forms whose questions to referees reflect the importance of reproducibility, which includes a nod to credibility. Publishers could insist on including a distinct methods section in experimental papers, prescribing a uniform structure reflective of the structure in the present paper; many venues in other disciplines require this. Incorporating such a structure would enhance the readability and reproducibility of the research, as well as making a method section more transparent and easier to write. Conferences could make awards for papers that excel in presenting their experimental methods. Funding agencies are under pressure to demonstrate the utility of projects that they fund; issuing directives on descriptions for supporting reproducibility could help.

In the longer term, we could develop courses in experimental methods, and teach them in graduate programs leading to degrees involving experimentation. This will take time. To make this effective in the long run, journals and conferences would need to require or reward authors for reporting research methods in a structured way. Writing is not usually part of a CS program, but there would be no harm in students being asked to pass a short writing course that is directed at reporting in the computing disciplines, with attention to reproducibility. Part of this course could include assignments to perform reproducibility audits on papers from the extant literature.

For those who feel that such reporting is too burdensome, consider the benefits, and that well-structured method sections take only about half a page in an article – a fairly light burden in exchange for a big return in transparency, quality, credibility, readability, and reproducibility. Think of it as a recipe for readers to reproduce or follow

up on your work; lay out the ingredients and the instructions so that data cooks can be successful in understanding and reproducing your work.

DISCUSSION AND CONCLUSIONS

Variation is not your friend when reproducing the work of others. Discordance between an original work and a reproduction can bring unwelcome differences in results. Papers that heed our audit criteria are more likely to beget faithful reproductions; everyone benefits.

In this paper, we focused on the reporting of the experimental process with an emphasis on the information necessary for another investigator to reproduce the reported results; we make no judgements about the reported data or results, *per se*. It is possible for even a perfectly reported experiment to produce worthless data, due to fundamental flaws in the hypothesis, underlying assumptions, instrumentation, etc., but a critical reading of the paper should reveal some or all of these flaws. Similarly, it is possible for a badly reported experiment to produce excellent data, but the reader will not be able to tell how the data were obtained and will probably not be able to repeat or confirm the work.

We performed reproducibility audits for 23 keystroke biometric papers published in the first quarter of 2020, based on a systematic review of five databases. We offered ideas for improvement that are simple and inexpensive in terms of cost and effort. We conjecture that the easiest way to improve reproducibility is for researchers to adopt a mindset focused on “what could possibly go wrong?” and taking action to avert hazard. We conclude that reporting the details of a research method, perhaps using a template based on this paper, would bring benefits not only to authors, but to the entire research community in terms of facilitating reproducibility and better science. Better science is always valuable, especially when it doesn’t cost much. Buy low, sell high.

ACKNOWLEDGMENT

The author cannot overstate the kindness and generosity of John McHugh for many helpful comments, gratefully received. This work was supported by the CMU Software Engineering Institute and by NSF grant CNS-1319117.

REFERENCES

1. Association for Computing Machinery (2016). Artifact Review and Badging. Available online at: www.acm.org/publications/policies/artifact-review-badging (Accessed 01 April, 2020).
2. Glenn C. Begley and Lee M. Ellis. “Drug development: Raise Standards for Preclinical Cancer Research.” *Nature*, Vol. 483, No. 7391, pp. 531-533, 29 March 2012.
3. William Lowe Bryan and Noble Harter. “Studies in the Physiology and Psychology of the Telegraphic Language,” *Psychological Review*, Vol. 4, No. 1, pp. 27-53, January 1897.
4. Melissa Davey. “Unreliable Data: How Doubt Snowballed Over Covid-19 Drug Research that Swept the World,” *The Guardian*, 04 June 2020. <https://www.theguardian.com/world/2020/jun/04/unreliable-data-doubt-snowballed-over-covid-19-drug-research>
5. Martin Fleischmann and Stanley Pons. “Electrochemically Induced Nuclear Fusion of Deuterium,” *Journal of Electroanalytical Chemistry*, Vol. 261, pp. 301-308, 1989.
6. Steven N. Goodman, Daniele Fanelli and John P. A. Ioannidis. “What Does Research Reproducibility Mean?” *Science Translational Medicine*, Vol. 8, Issue 341, pp. 1-6, 01 June 2016.
7. Richard F. Harris, “Rigor Mortis: How Sloppy Science Creates Worthless Cures, Crushes Hope, and Wastes Billions,” Basic Books, New York, NY, 2017.
8. John P. A. Ioannidis. “Why Most Published Research Findings Are False.” *PLoS Medicine*, Vol. 2, No. 8 (2005).
9. Joint Committee on Guides for Metrology (JCGM). *International Vocabulary of Metrology - Basic and General Concepts and Associated Terms (VIM)*, 3rd edition, JCGM 200:2012, 2012.
10. National Academies of Sciences, Engineering, and Medicine. “Reproducibility and Replicability in Science,” *The National Academies Press*, Washington, DC, 2019.
11. Tony Thomas, Athira Vijayaraghavan and Sabu Emmanuel. “Machine Learning and Cybersecurity” in *Machine Learning Approaches in Cyber Security Analytics*, Chapter 3, pp. 37-47, Springer Nature, Singapore, 2020.

Roy Maxion is an IEEE Fellow, currently with Carnegie Mellon University, Pittsburgh, PA, USA. He won the 2019 IEEE/DSN Test of Time Award, with his student, Kevin Killourhy, for their 2009 paper, “Comparing Anomaly Detection Algorithms for Keystroke Dynamics.” His research interests are keystroke biometrics, reliability, and research methods in computer science. He is on the editorial boards of the *International Journal of Machine Learning* and *IEEE Security & Privacy*. Contact him at maxion@cs.cmu.edu.