Characterizing Scientific Reporting in Security Literature: An analysis of ACM CCS and IEEE S&P Papers

Morgan Burcham
University of Alabama
mburcham@crimson.ua.edu

Mahran Al-Zyoud
University of Alabama
mmalzyoud@crimson.ua.edu

Jeffrey C. Carver
University of Alabama
carver@cs.ua.edu

Mohammed Alsaleh
UNC at Charlotte
malsaleh@uncc.edu

Hongying Du
NC State University
hdu2@ncsu.edu

Fida Gilani
UNC at Charlotte
sgillan4@uncc.edu

Jun Jiang
UNC at Chapel Hill
jiangcj@cs.unc.edu

Akond Rahman
NC State University
aarahman@ncsu.edu

Özgür Kafalı
NC State University
rkafali@ncsu.edu

Ehab Al-Shaer
UNC at Charlotte
ealshaer@uncc.edu

Laurie Williams
NC State University
williams@csc.ncsu.edu

ABSTRACT

Scientific advancement is fueled by solid fundamental research, followed by replication, meta-analysis, and theory building. To support such advancement, researchers and government agencies have been working towards a "science of security". As in other sciences, security science requires high-quality fundamental research addressing important problems and reporting approaches that capture the information necessary for replication, meta-analysis, and theory building. The goal of this paper is to aid security researchers in establishing a baseline of the state of scientific reporting in security through an analysis of indicators of scientific research as reported in top security conferences, specifically the 2015 ACM CCS and 2016 IEEE S&P proceedings. To conduct this analysis, we employed a series of rubrics to analyze the completeness of information reported in papers relative to the type of evaluation used (e.g. empirical study, proof, discussion). Our findings indicated some important information is often missing from papers, including explicit documentation of research objectives and the threats to validity. Our findings show a relatively small number of replications reported in the literature. We hope that this initial analysis will serve as a baseline against which we can measure the advancement of the science of security.

Keywords

Science of Security, Literature Review

HoTSoS'17, April 3-5, Hanover, Maryland
© 2017 ACM. ISBN 978-1-4503-5274-1…$15.00
DOI:10.1145/3055305.3055307

1. INTRODUCTION

Sustained scientific advancement in a field of study requires a significant level of effort from disparate members of a community. Included in these efforts are both high-quality foundational work by community members and viable methods of communicating that work to the larger community. The communication phase is important in that it allows community members to review, understand, analyze, question, replicate, and extend the published results to deepen and expand the overall knowledge of the community and the ability of research results to impact practice.

A key tenet of scientific investigation is the identification and understanding of the fundamental relationships among variables that contribute to or determine the results observed in individual studies. Often researchers cannot identify or understand these relationships on the basis of an individual study. A research community must be able to examine the results and important causal factors from multiple related studies to identify patterns that can provide the deeper insight needed to make progress in the foundational scientific understanding of a field.

However, the practice of cybersecurity today is frequently reactive rather than proactive. That is, because of the frequency and severity of constantly looming threats, organizations often operate in a mode of reacting to attacks after they occur by patching individual vulnerabilities that provided the opening for the attack. For the community to advance from reactive to proactive solutions, we need to gain a better understanding of scientifically-based design principles that allow us to build security in from the beginning. Such an approach would provide more defense against broader classes of both known and unknown attacks.

Recognizing this need, government agencies and security researchers have begun working towards a "science of security". To facilitate additional advances in the scientific underpinnings of security, the research community needs to be able to perform scientific tasks like replication, meta-analysis, and theory building. As indicated in the JASON report, “The highest priority should be assigned to establishing research protocols to enable reproducible experiments” [3].

The JASON report [3] contends, “There is every reason to believe that the traditional domains of experimental and theoretical inquiry apply to the study of cyber-security.” The advancement of science in these traditional domains involves two key requirements. First, members of a community need to be conducting high-quality research addressing relevant problems. Second, the reports describing this high-quality research need to contain the information necessary to allow for replication, meta-analysis, and theory building.

Therefore, the goal of this paper is to *aid security researchers in establishing a baseline of the state of scientific reporting in security through an analysis of the content of papers in top security conferences.* We emphasize that this only on characterizes the completeness of the information in the papers and does *not* judge the quality of the underlying work (which we assume to be of high quality). In our initial work characterizing the proceedings of the *2015 IEEE Security & Privacy* conference [8], we defined a set of rubrics based on literature on scientific evaluation [16, 17, 18, 19, 23]. The current paper expands on that original paper by answering two questions about the proceedings of the *2015 ACM CCS* and *2016 IEEE Security & Privacy* conferences:

RQ1: *What are the characteristics of the artifacts and evaluations contained in the papers?*

- **RQ1.1:** *What types of artifacts are evaluated (i.e., models, languages, protocols, processes, tools, or theories)?*
- **RQ1.2:** *What methods are used for artifact evaluation (i.e. empirical study, proof, or discussion)?*
- **RQ1.3:** *Do papers build on or extend prior work?*
- **RQ1.4:** *Are there trends in the relationship between artifact type and evaluation method?*

RQ2: *Do the papers contain all the information necessary to support the science of security?*

The analysis in this paper will help establish a baseline against which to measure progress related to the science of security. We also hope that this paper can serve as an encouragement to members of the community regarding which information should be reported in papers to support the overall advancement of the field.

The remainder of this paper is organized as follows. Section 2 describes background information. Section 3 discusses related work. Section 4 defines the rubric used for paper analysis. Section 5 explains the methodology we used to analyze the papers. Section 6 contains the results of the analysis. Section 7 provides some overall observations across the whole set of papers. Section 8 describes our lessons learned while conducting this study. Section 9 enumerates the threats to validity. Section 10 summarizes the paper.

2. BACKGROUND

This section provides background information on the key concepts that are important for the advancement of science. Then it discusses some examples of guidelines that support the application and reporting of research.

2.1 Replications

A key tenet of science is reproducibility of results. Reproducibility consists of: (1) obtaining the same results of the original study using the same method in the same environment (where possible, i.e. through a virtual machine replicating an environment); and and (2) providing enough information about the study conditions to allow colleagues to build on results and advance scientific progress [9]. Replications expand this definition by attempting to re-execute studies in different environments (driven by conscious changes to increase the robustness of the overall finding). The ability to reproduce results in various contexts allows researchers to evaluate external validity by determining the extent to which the causal relationships and results/findings of the original study can be generalized [7].

In theory, if there is enough detail about the original study, the results can be validated independently by other researchers [26]. Conversely, if this information is lacking, then research findings likely will be isolated to the original paper and scientific progress will be slower. Therefore, it is important for researchers to provide the right information in their research reports to reduce the overhead introduced when potential replicators have to solicit information from the original researchers.

The problem of replication has been discussed and addressed in different ways:

- Medical researchers have been debating on the validity of their published results for some time [1, 20, 15]. A July 2015 article in MedlinePlus [4] reported that researchers could not reproduce half of the 100 publications in premier psychology journals.
- The ACM SIGMOD community awards the Reproducible Label to database papers, which means “The experimental results of the paper were reproduced by the committee and were found to support the central results of the paper. The experiments (data, code, scripts) are made available to the community”[1].
- The Computational Science community has long recognized the need for reproducibility [24, 25, 2], but has yet to develop a comprehensive solution. Recently *ACM Transactions on Mathematical Software* introduced the Replicated Computational Results designation, awarded to papers for which the editors can obtain “independent confirmation that the results contained in the manuscript are correct and replicated” [14].

2.2 Meta-Analysis

Meta-analysis is a systematic approach to analyze the results of a set of previously conducted research studies to derive conclusions about the entire body of research [12]. This process requires researchers to follow a clearly defined process to identify all relevant studies in the literature. Based on those identified studies, meta-analysis uses various statistical approaches to determine the existence, size, and variability of an overall effect. A meta-analysis helps researchers answer new questions, resolve conflicting results, and generate new hypotheses [12]. The abundance of studies and clinical trials on various treatment protocols has provided the necessary data for medical researchers to frequently and successfully apply meta-analysis to draw general conclusions from the disparate studies [29, 11].

Similarly, as the body of studies in the security domain grows, meta-analysis will become an increasingly important

[1] http://db-reproducibility.seas.harvard.edu

method for drawing overall conclusions that can guide future research and practice. To successfully enable the use of meta-analysis in security research, it is essential that researchers provide thorough documentation of their studies.

2.3 Theory Building

One goal of research is to build the knowledge required to organize findings into coherent statements about the domain. A theory is the belief that there is a pattern in related observations [10]. A theory can help to fill in the gaps in current knowledge. Further, a scientific theory is an explanation of some phenomenon that is acquired via the scientific method and confirmed through repeated observation and experimentation[2]. Therefore, researchers can test theories and use them to make falsifiable predictions [21].

In security science, as an applied science, theories are important to guide people when making choices about the application of existing solutions to unknown problems. As a constantly evolving field, security science requires a continual growth of the body of knowledge and a deeper understanding of the underlying theories. This knowledge will allow researchers to communicate solutions to practitioners and develop common research agendas.

2.4 Guidelines for Reporting Research

As a community coalesces around accepted study design and result reporting mechanisms, it becomes easier for community members to follow appropriate methods [31]. Clear guidelines help a field mature over time, e.g. medicine [5, 27], psychology [28, 13], and social science [6, 22]. Increased maturity in these fields makes it easier to perform tasks like replication, meta-analysis, and theory building.

Analysis of literature from these fields provides insight into balancing scientific rigor and practical relevance. To have the most impact, a community must understand how to report studies (both the designs and the results), how to describe design alternatives, and how to interpret the results for practical benefit. One prime example of such a community is the Cochrane Collaboration in medicine (http://www.cochrane.org). The stated goal is to provide a world of "improved health where decisions about health and health care are informed by high-quality, relevant, and up-to-date synthesized research evidence." To achieve this goal, the community follows a set of principles that ensure: collaboration, avoiding duplicated effort, minimization of bias, relevance, quality assurance, and wide participation. Because members of the community understand how their research results will be used to further larger goals, there is an understood approach to study reporting. As a result, the Cochrane Collaboration has been able to analyze disparate research results to produce reports that transform the way health decisions are made. While we do not necessarily advocate this exact model for the security research community, the benefits that can be seen from rigorous study reporting should be informative.

3. RELATED WORK

This section describes related work about the Science of Security and previous literature reviews.

3.1 Science of Security

In 2010, JASON was tasked by the US Department of Defense to perform a study on the interplay of science and cybersecurity. The resulting report indicated that a most important attribute is "the construction of a common language and a set of basic concepts about which the security community can develop an understanding." [3]

This work is part of the U.S. National Security Agency *Science of Security Lablets*[3] which seek to develop the scientific underpinnings of security and build a body of knowledge to support rigorous design methodologies. Other similar security research programs around the world include:

- The Team for Research in Ubiquitous Secure Technology (TRUST) is a US National Science Foundation Science and Technology Center based out of the University of California at Berkeley, with the goal of developing "cyber security science and technology that will radically transform the ability of organizations to design, build, and operate trustworthy information systems for the nation's critical infrastructure"[4];
- The MURI project sponsored by the US Air Force OSR based out of Carnegie Mellon University, Cornell University, Stanford University, the University of California at Berkeley, and the University of Pennsylvania, with the goal of "advancing a science base for trustworthiness by developing concepts, relationships, and laws with predictive value"[5]; and
- The Research Institute in Science of Cyber Security based out of the University College of London, with the goal of "giving organizations more evidence, to allow them to make better decisions, aiding to the development of cybersecurity as a science"[6].

3.2 Related Studies

We previously analyzed 55 papers of the 2015 *IEEE Symposium on Security and Privacy* with a focus on the completeness of the information provided about the evaluation methods [8]. We used a set of rubrics to determine the type(s) of artifacts being evaluated, the evaluation method, and the completeness of the details for that evaluation method. Some key observations from this study include: (1) tools and processes were the most commonly evaluated artifacts; (2) most papers did not compare their results against a baseline; (3) many papers lacked a clear description of research objectives; and (4) most papers did not discuss threats to validity or study limitations. Section 7.3 compares the results of our current paper with these prior results.

The Asymmetric Resilient Cybersecurity Initiative at Pacific Northwest National Laboratory[7] developed a Science Council [30]. This organization provides insights on applying the scientific method in cybersecurity research and describes the initial impacts of applying the science practices to cybersecurity research and identified eight practices as beneficial in improving the quality of experiments and generating repeatable outcomes: Defining a Tractable Problem, Preliminary Data Assessment, Developing Falsifiable

[2] Based on a definition provided by the National Academy of Sciences (http://www.nap.edu/read/6024/chapter/2#2)

[3] http://cps-vo.org/node/5253

[4] https://www.truststc.org/about/

[5] https://sites.google.com/site/sosmuri/

[6] http://www.riscs.org.uk

[7] http://cybersecurity.pnnl.gov/arc.stm

Research Questions, Identifying Ground Truth, Documenting Assumptions, Testing Tools and Assumptions, Starting with Simple Experiments, and Assessing Progress Toward the Larger Problem.

4. PAPER EVALUATION RUBRIC

We adopted the rubric from the initial review [8] with improvements based on our experiences writing the previous paper. The following subsections provide a detailed description of the factors included in the rubric. Note that this section describes the final rubric, after updates made during the review process (See Section 5.4).

4.1 Evaluation Subject Type

In a field as diverse as security, a variety of solutions exist. We refer to those solutions as **Evaluation Subjects** because they must be evaluated in the paper. In this paper, we are interested in **Evaluation Subject Types** rather than the specific evaluation subjects. We identified the set of Evaluation Subject types through a detailed analysis of previous security literature [8]. The primary reason for identifying different Evaluation Subject Types is to characterize the frequency of each and to identify whether researchers use different Evaluation Approaches for each type. The **Evaluation Subject Types** are:

- **Model (M)** - a graphical or mathematical description of a system and/or its properties. Provides a simplified understanding of a system;
- **Language (L)** - a constructed/formal language developed as a method of communication;
- **Protocol (PL)** - a written procedural method that specifies the behavior for data exchange amongst multiple parties;
- **Process (PR)** - the computational steps required to transform one artifact into another one;
- **Tool (T)** - an implementation of a process, model, or protocol – an executable piece of software; and
- **Theory (TH)** - Proposes a new theory or update to an existing theory.

A paper can have one or more Evaluation Subjects corresponding to one or more Evaluation Subject Types.

4.2 Evaluation Subject Source

The second factor is when and by whom the **Evaluation Subject** was first introduced. In some cases, a paper will provide both the definition of an Evaluation Subject and an evaluation of that subject. In other cases, a paper will provide an evaluation of one or more Evaluation Subjects, which are defined elsewhere (possibly by different researchers). Because one of the key components of scientific advancement is building upon and replicating prior research, this information is important. Without a balance between new evaluation subjects and replication/extension of existing subjects, the literature can become unbalanced and reduce the potential for overall scientific impact. This factor can assume one of four values:

- **Authors Here (AH)** - Authors introduced the Evaluation Subject for the first time in the paper;
- **Authors Elsewhere (AE)** - Authors introduced the Evaluation Subject in a previous paper;
- **Others Modified (OM)** - Someone else introduced the Evaluation Subject and the authors modified it;
- **Others Not Modified (ON)** - Someone else introduced the Evaluation Subject and the authors used it without modification. [Note, we added this option to the original rubric.]

4.3 Evaluation Approach

The fourth factor is the approach used by the authors to evaluate each **Evaluation Subject**. A researcher has the choice of various approaches to evaluate the claims of his or her research. Each of these approaches has its own strengths and weaknesses that must be taken into account when choosing the most appropriate one. The primary reasons for identifying the **Evaluation Approach(es)** are to (1) characterize the prevalence of each approach and (2) identify any patterns in the relationship between Evaluation Subjects and Evaluation Approaches.

The remainder of this section provides a short description of the three most common Evaluation Approaches found in our previous analysis of the security literature [8]. Section 4.4 provides a more detailed description of what type of information should be included in a paper for each of these Evaluation Approaches [16, 17, 18, 19, 23].

4.3.1 Empirical Study (ES)

An empirical study is the process of collecting and analyzing data from a set of subjects (e.g. people, systems, etc...) to determine the distribution of and/or the correlation between variable(s). The primary strength of an empirical study is that they lead to conclusions based upon evidence, which can encourage replication, theory building, and meta-analysis. The primary weakness of an empirical study is the effort required to plan/execute one. Because the ES type covers a wide range of study types, we include the following additional attributes to further classify those studies[8]:

- **Participant Type** (*Simulation*, *Humans*, *Systems*)
- **Type of Study**
  - *Observational* - Study performed in a natural setting in which the researcher collects data via observation without intentionally manipulating the environment or behavior of the participants. The researcher observes the participants in a natural setting without interacting with them.
  - *Interventional* - The researcher intentionally applies treatment(s) to participants that potentially manipulate the participants' environment or behavior. When multiple treatments are considered, the participants are assigned to treatment groups and the effects of the treatments are compared across the groups.
- **Type of Data**
  - *Self-reported* - e.g. interviews, surveys
  - *Observed* - a researcher records data
  - *Automated* - e.g. via a tool/logging device
- **Comparison** - whether results from the current study are compared with other data

[8] Unlike the previous rubric in which Case Study and Experiment were separate categories, we now include both under the more general term of *empirical study*.

- *Historical* - data from a previous study
- *New* - data generated in the current study
- *None* - no comparison

4.3.2 Proof (P)

A proof is a formal approach to validate a characteristic or property of an Evaluation Subject. In a proof, a researcher provides a series of statements, typically grounded in previous theory, to establish the truth of a claim. Proofs are not useful for all Evaluation Subject Types. Rather, they only apply to those that have some type of mathematical basis that can be established via proof, for example a theory. The primary strength of a proof is that its formality allows a researcher to make a direct link between an existing theory and the conclusion. The primary weakness is that a proof can only be applied to a subset of the Evaluation Subjects.

4.3.3 Discussion/Argument (D)

This category covers evaluation that does not contain any empirical data or proof. Note that this category does not refer to papers that have a discussion of the results obtained by one of the other Evaluation Approaches. Rather, this category covers papers whose only method of validating the results is through discussion or argument. In this case, the claims of the research have not been tested empirically through observations of research participants. The primary strength of this approach is that it can be used in situations where no empirical data is available. The primary weakness is that it is based primarily, if not completely, upon the opinion of the researcher and is therefore subject to bias.

4.4 Completeness Rubrics

For other researchers to be able to understand, replicate, and build on published research, the paper needs to contain a number of key elements.

- **Research Objectives** - To help readers understand the goals of the paper and position the results, a paper should clearly state the objectives that guide the development of the research.
- **Subject/Case Selection** - Readers can better understand how to interpret the results if the authors have clearly and explicitly described the subjects of the evaluation (e.g. the system or people chosen to participate), why those subjects are appropriate, and how they were recruited or developed.
- **Description of Data Collection Procedures** - To clarify exactly what information was collected and to enable replication, a paper should provide a detailed description of the data collection procedures.
- **Description of Data Analysis Procedures** - To enable replication, a paper should provide a detailed description of the data analysis procedures, including the statistical tests chosen.
- **Threats to Validity** - A paper should include information to help a reader understand the limitations of the evaluation results and whether or not those results are applicable in his or her particular situation.

While the above items generally appear in each of the Evaluation Approaches, they may have slightly different meanings (or not even apply) depending upon the characteristics of the Evaluation Approach. For each Evaluation Approach, we combined our own experience with information from the literature [17, 16, 18, 19, 23] to define a rubric describing which items should be present and what information they should contain.

Each of the three rubrics consists of a series of questions that help determine whether all relevant information has been completely reported. For each rubric item, we define three values:

- **Labeled** - the information is present in the paper and is clearly labeled (e.g. in a specific section or prefaced by bold/italicized text);
- **Not Labeled** the information is present in the paper, but is not clearly labeled; and
- **Missing** - the information is omitted from the paper.

Clear labeling of information is critical because it eases the process of locating the information in the paper. Being able to quickly locate key information helps readers quickly identify relevant papers and helps support replication, theory building, and meta-analysis.

Each rubric has specific, concrete definitions for these answers. An example of our rubric can be found in the appendix of this paper with the full rubrics available online[9].

5. METHODOLOGY

This section describes the research methodology for analyzing paper content. After training the research team and piloting the rubrics, we first reviewed the 128 papers from the *2015 ACM CCS* conference. Then, we made some minor modifications to the process based on lessons learned. Finally, we reviewed the 55 papers from the *2016 IEEE Security & Privacy* conference. The remainder of this section explain this process in detail.

5.1 Review Team

The research team consisted of the 11 authors of this paper, drawn from three universities, including three faculty members, six graduate students, and two postdoctoral researchers. The six graduate students and one of the postdoctoral researchers performed the paper reviews under the supervision of their respective faculty members. The remainder of this section refers to the six graduate students and one postdoctoral researcher collectively as “reviewers”.

5.2 Pilot Studies

In the original paper [8] (described in Section 3.2), we extensively pilot tested the rubric by reviewing papers from earlier editions of *ACM CCS* and *IEEE S&P* and refining as necessary. The pilot tests and use in our previous study, provide confidence in the validity of the rubric and our approach. Prior to conducting the *ACM CCS* review, all reviewers went through a three-stage training process to ensure they fully understood the rubric and knew how to use the support tool (NVivo). During this process, each reviewer applied the rubric to a set of papers from the previous study for which we had agreed-upon ratings. We compared the reviewers' values against the known values to determine whether any items needed further clarification. We also asked the reviewers to provide feedback on any rubric items that needed clarification.

[9] http://carver.cs.ua.edu/Studies/SecurityReview/Rubric.html

In the first stage, each reviewer analyzed two papers. For each rubric item, the reviewer highlighted the relevant information in a PDF version of the paper and provided their rationale. Overall, the reviewers' results were somewhat consistent with the known ratings (although not perfect). We discussed the inconsistent answers with each reviewer to ensure they understood the rubric. In the second stage, each reviewer analyzed two additional papers. We again clarified the small number of discrepancies between the reviewers' results and the known results to ensure the reviewers understood the rubric. In the third stage, each reviewer analyzed one additional paper, this time using NVivo. The main goal of this phase was to ensure everyone understood how to use NVivo to code the papers. The responses from the reviewers were again largely consistent with the known values. After discussing the remaining discrepancies, the reviewers agreed that they understood the rubric and NVivo.

Based on the feedback provided during the pilots, we updated the rubric to more clearly describe the Evaluation Subject Types, the Evaluation Approaches, and the Completeness Rubric questions.

5.3 ACM CCS Review Process

To analyze the ACM CCS papers and validate our results, we followed a six-step process.

Step 1 - Assign Reviewers To maintain consistency across the set of papers, we designated two of the reviewers to be the *lead reviewers*. The lead reviewers were very familiar with the rubric and were co-located, allowing them to frequently discuss the reviewing process. Based on the reviewers' interest and knowledge of the paper topics, we assigned each paper a lead reviewer and a second reviewer.

Step 2 - Review Papers Each reviewer independently reviewed his or her assigned papers to identify: (a) The Evaluation Subject Type(s); (b) The Evaluation Subject Source; (c) The Evaluation Approach(es) Used for Each Evaluation Subject; and (d) The Completeness of the Reporting of the Evaluation Approach(es). Reviewers coded this information using a predefined set of codes in NVivo.

Step 3 - Identify Discrepancies Once each reviewer finished coding papers, we used NVivo to compare the individual codings and identify any instances where the two reviewers disagreed on their codings.

Step 4 - Resolve Discrepancies In cases where the two reviewers disagreed, they resolved the disagreement via an email discussion. Each reviewer provided an argument for his/her coding choice (along with pointers to specific locations in the paper, where relevant). The discussion continued until both reviewers agreed. If the reviewers could not agree, a third reviewer reviewed the arguments on both sides and resolved the discrepancy.

Step 5 - Author Feedback We sent the agreed upon results to the paper authors for confirmation. We asked the authors to inform us if any of our characterizations were incorrect. If the author thought we made a mistake, we asked him to indicate the specific location in the paper of the missing information. We added this step to help ensure that we were properly understanding each paper.

Step 6 - Final Update In cases where a paper author provided feedback, one of the lead reviewers analyzed each response to determine whether the feedback provided met the criteria defined in our rubric. If the reviewer determined that it did, then we updated the characterization of that paper. If the reviewer did not think the feedback matched the rubric's criteria, the other lead reviewer checked the response to confirm the decision. In this case, we kept the original characterization and noted the disagreement. We kept track of all changes and disagreements as a result of this feedback process.

5.4 Updates to Methodology

Based on the author feedback, we made the following improvements to the rubric. First, for **Evaluation Approaches** we merged *case studies* and *experiments* into the *empirical study*. Second, to provide more insight into the empirical studies, we added **Type of Study**, **Type of Data**, and **Comparison** (described in Section 4.3.1). Third, we added the **Evaluation Attribute** item (described in Section 4.3) because it is important for a readers's ability to understand the overall goals and evaluation in a paper.

To ensure the *ACM CCS* papers were consistent with the updated rubric, the reviewers extracted the new information and updated the rubric values. After making these updates, we sent a revised version to the paper authors. We followed the same process as Steps 5 and 6 in Section 5.3.

To ease the author feedback process and increase the response rate, we revised the author feedback mechanism. We posted the review rubric on the web. We then created a web page for each paper with the values for each rubric and a link directly to the description of the rubric items so the authors could easily understand the scores. We provided a button on the webpage that allowed the author to easily respond with agreement or a description of the disagreements.

5.5 IEEE S&P Review Process

Based on the lessons learned from the *ACM CCS* review process and our desire to increase study validity by having greater author involvement, we made a few modifications prior to the review of the *IEEE Security & Privacy* papers. Other than the updates to the rubric (described in Section 5.4), the major change was that we assigned only one reviewer to each paper. By modifying our process to ease the author feedback step, we decided it would be more valid to have the paper author act as the second reviewer. This choice also eased the reviewing burden on our team.

Once our reviewer had finished reviewing the paper (following Step 2 in Section 5.3), we sent the review to the author using our improved author feedback form. We integrated author feedback in the same way as with *ACM CCS*. In those cases where the author did not respond, we had a second reviewer review the paper. We followed the same process as described in Sectin 5.3 for discrepancy resolution.

6. RESULTS

This section is organized around the two research questions. Section 7 discusses the implications of these results.

6.1 RQ1: Paper Characteristics

The following subsections describe the results for each of the subquestions to RQ1.

6.1.1 RQ1.1: Type of artifact evaluated

Figure 1 shows the distribution of the **Evaluation Subject Types** for each conference. These results show that *Tools* were the most common Evaluation Subject type for both conferences. Conversely, there was only one *Languages* (occurring in an IEEE S&P Paper). Furthermore, the distribution of Evaluation Subject is different for each conference, with IEEE S&P having relatively more *Theory*, *Process*, *Model*, and *Language* and ACM having relatively more *Protocol* and *Tool*. A chi-square test shows that these distributions are significantly different (p=.003).

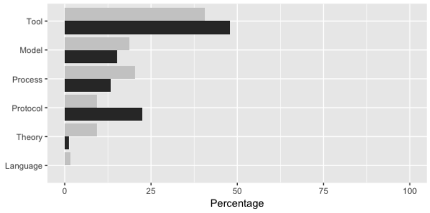

Figure 1: Frequency of Evaluation Subjects

6.1.2 RQ1.2: Method of evaluation

Figure 2 shows the distribution of the **Evaluation Approaches** for each conference. For both conferences, *Empirical Studies* were the preferred choice at over 70% in each case. While the distributions are slightly different between the two conferences, those differences are not significant.

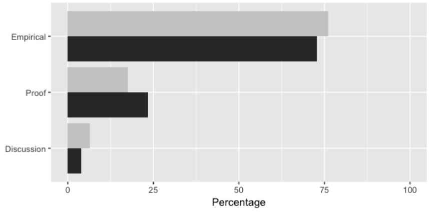

Figure 2: Frequency of Evaluation Approaches

6.1.3 RQ1.3: Building on Prior Work

Figure 3 shows the distribution of the **Evaluation Subject Sources** for each conference. The results show that papers overwhelmingly evaluate subjects that are first introduced in the current paper (more than 75% of the time in each conference). If researchers were building on prior work, we would expect to see more of the other values. For example, if a study is a replication, then the Evaluation Subject would be other than *Authors Here*.

Another indication of building on prior work is the presence of a comparison between the current study and historical data. This type of comparison is important for the research community because it helps build confidence in the study findings when researchers can confirm previous results or provide insight into the source of the differences. The majority (approximately 85%) of the papers did not contain any comparison of their results against historical data.

6.1.4 RQ1.4: Relationship between Evaluation Subjects and Evaluation Approaches

To analyze whether some Evaluation Approaches are preferred for different Evaluation Subject Types, Table 1 shows the distribution of these values. While *Empirical Studies* are the most common for all Evaluation Subjects (due to its overall dominance in the sample), its relative frequency is different for each Evaluation Subject. For example, *Theories* and *Tools* are almost 100% *Empirical*, while *Protocols* and *Models* are much closer to 50%. Due to the low expected value in some cells, it was not possible to perform a chi-square test for this distribution.

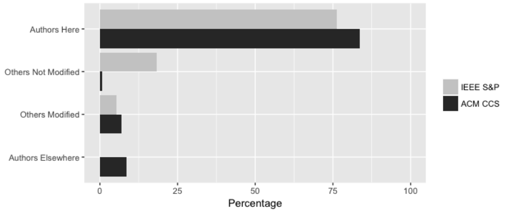

Figure 3: Frequency of Subject Sources

Table 1: Evaluation Approach Type by Evaluation Subject

	Empirical	**Proof**	**Discussion**
Process	27	9	2
Tool	104	3	3
Model	23	18	0
Protocol	21	20	5
Theory	7	0	0

6.2 RQ2: Completeness of papers

The goal of this question is to analyze whether or not the papers described all the key information related to the chosen Evaluation Approach and whether that information is clearly labeled. Because Empirical Studies and Proofs were the most dominant Evaluation Approaches, we only provide the detailed results for those two Evaluation Approaches.

Figure 4 shows the results for the **Empirical Studies**. Some interesting observations about the Empirical Study results from the figure are: (1) *Objectives* were rarely labeled. (2) *Data analysis procedures* were often not clearly defined. (3) *Threats to Validity* were missing much more often than they were present. Some additional observations from the data (not shown in Figure 4) are: (1) *Observational* studies were most often used rather than interventional ones with 70% of the IEEE and 80% of the ACM empirical studies being of type observational. (2) *Type of data gathered* was usually automated. (3) The majority of the empirical studies used systems rather than humans as participants. (4) Approximately 1/3 of the papers had no comparison of results with another 1/2 comparing to newly generated data.

Figure 5 shows the results for those studies using a Proof as the Evaluation Approach. The proofs in both conferences were generally well-documented with most studies addressing the rubric items in the paper.

6.3 Author Feedback

The author response rate to our characterizations was 51% for IEEE S&P and 33% for ACM CCS. Of those, 68% of the IEEE authors and 38% of ACM CCS authors had at least one disagreement. Based on the authors' responses, we changed an average of two items per paper for a total of 30 papers. The source of most changes were *Threats to Validity* (both conferences), *Analysis Procedures* (IEEE), *# Study Conditions* (IEEE), and *Research Objectives* (ACM). For *Threats to Validty*, the majority of the changes were

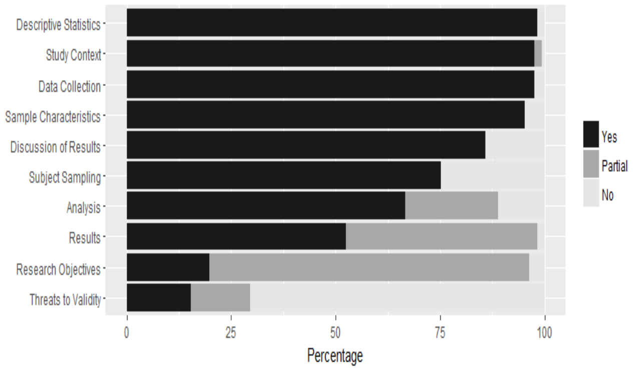

Figure 4: Empirical Study Rubric Results

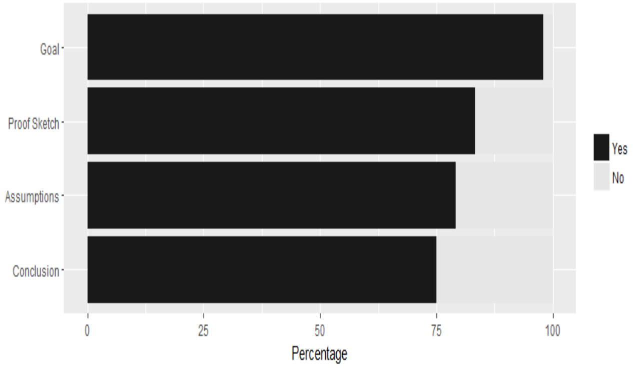

Figure 5: Proof Rubric Results

from "No" to "Partial," suggesting that, because the threats were unlabeled, they were more difficult to locate. For *Analysis Procedures* and *Research Objectives*, the majority of the changes were from "Partial" to "Yes," suggesting that the reviewers originally found these item labels unclear, but were convinced by the authors' arguments.

Based on the type of feedback provided by the authors and the changes that we made, we found this process to be valuable in ensuring the correct characterization of each paper. It is also interesting, but not surprising, to note that we made a larger percentage of changes to the IEEE papers, which were initially reviewed by only one reviewer before sending to the authors. The fact that the ACM papers had fewer changes suggests that having two reviewers per paper, prior to sending to the authors, improves accuracy. Conversely, even with two reviewers, approximately 1/3 of the papers still required changes. The large number of changes required after author interaction suggests that there is room for improvement in the clarity of paper organization (to help readers find information) and room for improvement in our review process to reduce the number of author disagreements.

7. DISCUSSION

This section provides overall observations about the results presented in the previous section. These observations were drawn both from the data as well as from our subjective impression of the papers as we reviewed them.

7.1 RQ1 - Characteristics of Evaluations

Regarding the **Evaluation Subject**, we can make some interesting observations. First, the most frequent research subject was *Tools*. Because security researchers aim to provide solutions to enable practitioners to secure their systems, it is not surprising to see the prevalence of *Tools* that automate security solutions. Second, we notice a difference between the two conferences for the second most common subject (*Protocols* for ACM CCS and *Processes* for IEEE S&P) suggesting there may be slightly different interests in the two conferences. Second, *languages* were relatively nonexistent in both proceedings with only one language being evaluated. Security researchers may see languages as an unnecessary defense mechanism, since safe programming languages and practices are already in use. Also, adoption of new languages may be avoided due to a learning curve. Furthermore, tools may be preferable due to the lower amount of effort required to start using them.

Regarding the presence of replications, we observed that the majority of **Evaluation Subjects** were frequently being evaluated by their creators, i.e. the Authors of the paper were the inventors of the artifact. Such findings suggest a lack of replication studies which hinders meta-analysis and theory building. However, as security is a constantly evolving field in which new security threats are always coming into existence, it is likely that, for some scenarios, researchers prefer to develop new solutions to mitigate emerging threats rather than studying existing artifacts. Also in regards to replications, we found that the empirical studies were either generating new data for comparison or providing no comparison. Few studies had any historical comparison.

Finally, regarding the **Evaluation Approach**, it is positive that the majority of papers used some type of *Empirical Study* for evaluation. This approach provides for more opportunities for replications than the other types. Most studies gathered data via automated means, likely because it is easier to collect and analyze large amounts of data in this manner. In addition, automation is less prone to human error. The study participants were usually systems (rather than people). Since systems are usually the threatened party in security research, this result is reasonable. The empirical studies also tended to be observational, rather than interventional, which suggests that researchers are observing security threats/practices rather than trying to manipulate those threats/practices. While the presence of observational data is positive, the primary weakness is that observational studies do not typically provide the type of control necessary to establish causation and to support theory-building and meta-analysis, key tenets of a science of security.

7.2 Discussion of RQ2 - Completeness

While performing our review, we found that most of the papers were easy to follow. Despite this quality, however, there was still some important information that was often missing from the papers or was included implicitly rather than explicitly. A lack of such information may make it difficult for readers to properly interpret the results and for other researchers to perform replications, meta-analysis, or theory building. Many of the initial discrepancies between reviewers were due to this lack of explicitness. Here we make three general observations about paper completeness.

First, many 80% of the empirical studies did not provide clearly defined and labeled *Research Objectives* to define the goals, questions, and/or hypotheses of the research. In many cases these objectives were present, for example in the description of contributions, but not clearly stated or labeled and too vague to fully understand the goal of the study itself. Because the research objective drives the study design, it is important that readers clearly understand the study goal so they can draw proper conclusions and make decisions about replication, meta-analysis, and theory building.

Second, 70% of the papers contained no discussion of the *threats to validity.* In this case, we refer to the validity threats of the Evaluation Approach, not the limitations of the Evaluation Subject (a confusion we saw in some author responses). A clear discussion of validity threats is important for researchers who want to replicate studies, because it allows the researcher to make design choices that address those threats. Furthermore, for meta-analysis and theory building, researchers need to understand the validity threats of each study in a set so they can more appropriately combine their results to gain a deeper understanding.

Third, the Proofs and Discussions from the literature were usually very well documented and thorough. However, we did note that in the IEEE S&P papers, many of the proofs had no clearly marked ending.

7.3 Comparison with Prior Results

This section compares the results of our current review with those from our previous review of the 2015 IEEE S&P papers [8]. We found some similar, although not identical, patterns between the two studies.

First, for **Evaluation Subjects**, *Tools* were the most common. We did note a difference in the second most common subject. In our current study the second most common subject was *Processes* for the IEEE S&P papers and *Protocols* for the ACM CCS papers. In the previous study (of IEEE S&P papers), *Processes* were also second most common, although their relative percentage was higher. So, it seems as though the conferences in this analysis were more heavily dominated by *Tools* than in the previous study.

Second, for the **Evaluation Approach**, the current study showed that *Empirical Studies* were the overwhelming majority. In the previous study, we had two categories, *Case Study* and *Experiment* which both map to *Empirical Study.* Together, those two types again represented the overwhelming majority of studies.

Third, in both studies, there was little evidence of *Replications.* In the current study approximately 75% of the papers evaluated a subject that was first introduced in the paper itself. In the previous study, that number was approximately 90%. While there does appear to be some increase in the number of replications, they are still relatively infrequent.

Finally, with regards to the completion rubrics, in both papers, most studies did not provide a discussion of the Threats to Validity.

8. LESSONS LEARNED

Based on the lessons learned in the original paper [8], we were able to make some adjustments for this paper. We also made some adjustments between the ACM reviews and the IEEE reviews. Because many of these changes are detailed in Section 5, we just summarize the key points here. We hope that other researchers will learn from our experiences and perform further replications of this type of review in different venues.

The first key change we made to the original process was to use a professional-grade qualitative analysis tool (NVivo) to aid in the review process, rather than using manual markups of PDF files. The use of this tool eased the review and analysis process. During the review, the reviewers could use pre-defined codes to mark sections of the paper. During analysis, NVivo made it easier to identify areas of agreement and disagreement between reviewers and to perform many of the required calculations. We recommend use of a tool like NVivo for future reviews.

The reviewer and author feedback from the ACM reviews led to a number of changes to the rubric. These changes helped clarify and simplify use of the rubric. Section 5.4 describes these changes in detail. The biggest change resulted from the fact that the distinction between an *experiment* and a *case study* turned out to be more subjective than we had planned. Therefore, we consolidated these into *empirical study* and added some additional attributes to more fully characterize each study. The reviwers indicated that this change made the rubric clearer and less subjective. Overall the reviewers found the updated rubrics clearly labeled, easy to use, and helpful for the analysis of the papers.

One of the important innovations in this iteration of the study was to solicit author feedback. We piloted an author feedback mechanism for the ACM papers, but it saw limited success. Therefore, we greatly improved that form to improve clarity, provide direct links to the rubric items, and ease the feedback process. Using the new form for both ACM and IEEE authors, we saw a much improved response rate. We recommend using a clear feedback form that allows the authors to clearly understand each characterization.

9. THREATS TO VALIDITY

To help readers properly interpret our paper, we offer this section describing the threats to validity. A threat to validity is anything that may reduce the validity of the study findings. As all studies have validity threats, we have sought to reduce them as much as possible.

The first threat is related to the review rubric. Because the results are based on the rubric, if it is not adequate, the results would be lacking. To reduce this threat, we reused an existing rubric with small additions to increase clarity. Even so, it is possible that our list of Evaluation Subjects and Evaluation Approaches is incomplete. We believe this threat is minimal because the reviewers were able to classify all papers on these two attributes. It is possible, that if this review were repeated on a different population of security papers, the rubric may be lacking. Again, this threat is minimized by the fact that the rubric has now been used to review papers from two of the largest security conferences.

Second, even with a concrete, validated rubric, reviewing a paper still has an element of subjectivity. To combat this threat, in the first round of paper reviews (ACM CCS 2015) two reviewers were assigned to independently review each paper. While there were some initial discrepancies, through discussion the reviewers resolved them. It is still not possible to eliminate all possibilities of subjectivity and bias. To further mitigate this threat, we allowed paper authors to provide feedback on the results of our analysis. Due to time limitations, we did not engage in a discrepancy resolution process with the authors (as we did between the reviewers). In the second round of reviewing papers (IEEE S&P 2016 papers and empirical papers of ACM CCS conference 2015), we emailed all authors with our findings. This had a profound impact on validating our findings.

10. SUMMARY AND FUTURE WORK

This paper analyzed the 2015 *ACM CCS* and 2016 *IEEE S&P* proceedings to establish a baseline of the state of scientific research in security through an analysis of indicators of scientific research. The overall motivation was to determine whether papers are reporting the information necessary for three key pillars of scientific research: replication, meta-analysis, and theory building. To perform this analysis, we followed published guidelines for identifying **Evaluation Subject Types** and **Evaluation Approaches**. For each **Evaluation Approach** we used a **Completeness Rubric** to help determine whether each paper described all important information regarding the evaluation. The results of this study are similar to those from the prior paper that analyzed the 2015 *IEEE Security & Privacy* proceedings. This baseline can serve as a comparison point for a similar review that will be conducted in five or ten years.

As future work, we are analyzing our dataset to select outstanding examples for each rubric item. We anticipate that this future work, combined with our previous work, will serve as a positive and clear guidance to researchers that leads them to produce better reports of their research.

11. REFERENCES

[1] Freely associating. *Nature Genetics*, 22, 1999.

[2] Reproducible research. *Computing in Science Engineering*, 12(5):8–13, Sept 2010.

[3] The science of cyber-security. Technical Report JSR-10-102, MITRE, 2010.

[4] How reliable are medical studies? half of findings couldn't be replicated. *MedlinePlus*, 2015.

[5] J. Abramson and Z. Abramson. *Research methods in community medicine: surveys, epidemiological research, programme evaluation, clinical trials*. John Wiley & Sons, 2011.

[6] E. Babbie. *The practice of social research*. Cengage Learning, 2015.

[7] A. M. T. Bobby J. Calder, Lynn W. Phillips. The concept of external validity. *Journal of Consumer Research*, 9(3):240–244, 1982.

[8] J. C. Carver, M. Burcham, S. A. Kocak, A. Bener, M. Felderer, M. Gander, J. King, J. Markkula, M. Oivo, C. Sauerwein, and L. Williams. Establishing a baseline for measuring advancement in the science of security: An analysis of the 2015 ieee security & privacy proceedings. In *Proc. of the Symposium and Bootcamp on the Science of Security*, HotSos '16, pages 38–51, 2016.

[9] C. Collberg and T. A. Proebsting. Repeatability in computer systems research. *Commun. ACM*, 59(3):62–69, Feb. 2016.

[10] C. F. Craver. Structures of scientific theories. *The Blackwell guide to the philosophy of science*, 19:55, 2008.

[11] R. DerSimonian and N. Laird. Meta-analysis in clinical trials. *Controlled clinical trials*, 7(3):177–188, 1986.

[12] A. Haidich. Meta-analysis in medical research. *Hippokratia*, 14:29–37, 2010.

[13] P. Harris. *Designing and reporting experiments in psychology*. McGraw-Hill Education (UK), 2008.

[14] M. A. Heroux. Editorial: Acm toms replicated computational results initiative. *ACM Trans. Math. Softw.*, 41(3):13:1–13:5, June 2015.

[15] J. P. Ioannidis and T. A. Trikalinos. Early extreme contradictory estimates may appear in published research: The proteus phenomenon in molecular genetics research and randomized trials. *Journal of Clinical Epidemiology*, 58(6):543 – 549, 2005.

[16] A. Jedlitschka, M. Ciolkowski, and D. Pfahl. Reporting experiments in software engineering. In F. Shull, J. Singer, and D. I. Sjøberg, editors, *Guide to Advanced Empirical Software Engineering*, pages 201–228. Springer London, 2008.

[17] A. Jedlitschka and D. Pfahl. Reporting guidelines for controlled experiments in software engineering. In *Int'l Symp. on Empir. Soft. Eng.*, page 10, Nov 2005.

[18] B. Kitchenham and S. Pfleeger. Personal opinion surveys. In F. Shull, J. Singer, and D. Sjøberg, editors, *Guide to Advanced Empirical Software Engineering*, pages 63–92. Springer London, 2008.

[19] L. Lamport. How to write a proof. *The American Mathematical Monthly*, 102(7):600–608, 1995.

[20] R. Moonesinghe, M. Khoury, and A. Janssens. Most published research findings are false–but a little replication goes a long way. *PLOS Medicine*, 4, 2007.

[21] K. Popper. *Conjectures and refutations*, volume 7. London: Routledge and Kegan Paul, 1963.

[22] H. Rahmandad and J. D. Sterman. Reporting guidelines for simulation-based research in social sciences. *System Dynamics Rev.*, 28(4):396–411, 2012.

[23] P. Runeson and M. Höst. Guidelines for conducting and reporting case study research in software engineering. *Empirical Software Engineering*, 14(2):131–164, 2009.

[24] G. Sandve, A. Nekrutenko, J. Taylor, and E. Hovig. Ten simple rules for reproducible computational research. *PLoS Comput Biol*, 9, 2013.

[25] M. Schwab, N. Karrenbach, and J. Claerbout. Making scientific computations reproducible. *Computing in Science Engineering*, 2(6):61–67, Nov 2000.

[26] F. J. Shull, J. C. Carver, S. Vegas, and N. Juristo. The role of replications in empirical software engineering. *Empirical Software Engineering*, 13(2):211–218, 2008.

[27] I. Simera, D. Moher, J. Hoey, K. Schulz, and D. Altman. A catalogue of reporting guidelines for health research. *European journal of clinical investigation*, 40(1):35–53, 2010.

[28] J. A. Smith. *Qualitative psychology: A practical guide to research methods*. Sage, 2015.

[29] A. J. Sutton, K. R. Abrams, D. R. Jones, D. R. Jones, T. A. Sheldon, and F. Song. *Methods for meta-analysis in medical research*. J. Wiley Chichester; New York, 2000.

[30] M. Tardiff. Applying the scientific method to cybersecurity research. In *IEEE Int'l Symp. on Technologies for Homeland Security*, 2016.

[31] M. V. Zelkowitz. An update to experimental models for validating computer technology. *Journal of Systems and Software*, 82(3):373–376, 2009.

Appendix: Paper Rubric Items - Empirical Study

Rubric Item	**Yes**	**Partial**	**No**
EM1: Are the research objectives of the study described? (e.g., goals, questions, hypotheses)?	*Clearly defined early in the paper (i.e. not in the results or discussion) and labeled (e.g. in bold, italics, underlined or set apart from the text with labels like Research Question, RQ, Objective)*	*Included in the text but either in the wrong location or not clearly labeled (see Yes above)*	*Not present*
EM2: Is the context of the study described? Does the paper offer details on what is being tried to solve the research problem?	*The paper explicitly defines the context of the study (i.e. the problem background or why it is important to study these particular research questions or problems) and what is being tried*	*The defines some, but not all, of the above*	*The paper defines none of the above*
EM3: Are the methods for subject sampling described? (e.g., recruitment/selection process, inclusion/exclusion criteria)?	*The paper explicitly describes how the cases were selected including the rationale for selecting the particular case(s)*	N/A	*The paper does not explicitly describe how the cases were selected*
EM4: Are the data collection procedures (e.g., how was this completed, definition of the metrics/variables, operational constructs, measurement levels) and research instruments (i.e. questionnaire, mining tools, performance computation) described??	*Described in the text*	N/A	*Not described in the text*
EM5: Are the analysis procedures described? (e.g., hypothesis checks, statistical tests, p-values, performance metrics, precision, recall, accuracy, False positive, False negative etc.)?	*Paper includes both the statistical tests (by name) or other analysis method (e.g. performance measures) and the results of statistical test (including p-value) or other analysis method*	*Paper includes one of the above*	*Paper includes none of the above*
EM6: Are the characteristics of the sample/ systems described? (e.g., demographics, specification)?	*Paper explicitly describes the characteristics of the sample*	N/A	*Paper does not explicitly describe characteristics of the sample*
EM7: Does the data presented have descriptive stats? (e.g., mean, std dev, charts or tables to describe data, etc)	*Paper contains a description of the data: e.g., mean/median, standard deviation, frequency, etc...*	N/A	*Paper does not describe the data*
EM8: Do they discuss results in relation to the research objectives? (e.g., hypotheses evaluated, questions answered, or "big picture")	*There is a separate discussion section*	*Results are discussed, but not in a separate section*	*Results are not discussed*
EM9: Do they discuss and provide reasoning for "why" the results had the given outcome?	*There is a discussion of why a particular outcome occurred in the study. Rather than presenting only the results, the authors explain "why" such results were obtained.*	N/A	*No reasoning for the outcome of the study is given.*
EM10: Is there a dedicated discussion of the threats to validity to the experiment (i.e., limitations or mitigations)?	*There is a separate Threats to Validity Section*	*Threats to validity are discussed, but not in a separate section*	*Threats to validity are not discussed*