Establishing a Baseline for Measuring Advancement in the Science of Security – an Analysis of the 2015 IEEE Security & Privacy Proceedings

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

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

Sedef Akinli Kocak
Ryerson University
sedef.akinlikocak@ryerson.ca

Ayse Bener
Ryerson University
ayse.bener@ryerson.ca

Michael Felderer
University of Innsbruck
michael.felderer@uibk.ac.at

Matthias Gander
University of Innsbruck
matthias.gander@uibk.ac.at

Jason King
NC State University
jtking@ncsu.edu

Jouni Markkula
University of Oulu
jouni.markkula@oulu.fi

Markku Oivo
University of Oulu
markku.oivo@oulu.fi

Clemens Sauerwein
University of Innsbruck
clemens.sauerwein@uibk.ac.at

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

ABSTRACT

To help establish a more scientific basis for security science, which will enable the development of fundamental theories and move the field from being primarily reactive to primarily proactive, it is important for research results to be reported in a scientifically rigorous manner. Such reporting will allow for the standard pillars of science, namely replication, meta-analysis, and theory building. In this paper we aim to establish a baseline of the state of scientific work in security through the analysis of indicators of scientific research as reported in the papers from the 2015 IEEE Symposium on Security and Privacy. To conduct this analysis, we developed a series of rubrics to determine the completeness of the papers relative to the type of evaluation used (e.g. case study, experiment, proof). Our findings showed that while papers are generally easy to read, they often do not explicitly document some key information like the research objectives, the process for choosing the cases to include in the studies, and the threats to validity. 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

1. INTRODUCTION

HotSoS '16, April 19-21, 2016, Pittsburgh, PA, USA

DOI: http://dx.doi.org/10.1145/2898375.2898380

Cybersecurity breaches are in the news every day. Cyber-espionage targets intellectual property across corporate and national boundaries. Cyber-theft costs the global economy more than $375 billion every year according to the Center for Strategic and International Studies [25]. Cyber-disruption also remains a looming threat for everything from financial markets to the power grid.

However, the practice of cybersecurity today is generally reactive rather than proactive. That is, rather than improving their defenses in advance, organizations often react to attacks once they have occurred by patching the individual vulnerabilities that led to those attacks. What we need instead are scientifically founded design principles for building in security mechanisms from the beginning, giving protection against broad classes of known and unknown attacks.

Recognizing this need, government agencies have supported large research centers focused on building a science of security. We now summarize a sampling of large science of security research programs around the world.

- The Team for Research in Ubiquitous Secure Technology (TRUST) is a US National Science Foundation Science and Technology based out of the University of California at Berkeley[1];
- Four Science of Security Lablets supported by the US National Security Agency based out of Carnegie Mellon University, North Carolina State University, the University of Illinois – Urbana Champaign and the University of Maryland[2];
- The MURI project sponsored by the US Air Force Office of Scientific Research based out of Carnegie Mellon University, Cornell University, Stanford University, and the University of California at Berkeley, and the

[1] https://www.trustst c.org/about/
[2] http://cps-vo.org/node/5253

University of Pennsylvania[3]; and

- The Research Institute in Science of Cyber Security based out of the University College of London[4].

While researchers may *perform* high-quality scientific work, they may not thoroughly report their work in a manner that contains enough information to enable replication, theory building, and meta-analysis among the members of the security research community. These activities, which are key to the advancement of science, are inhibited by incomplete research reports. In other words, "If you cannot measure it, you cannot improve it" – Lord Kelvin. *Our goal is to aid security researchers in establishing a baseline in the state of scientific work in security through an analysis of indicators of scientific research as reported in top security conferences.*

Five of the authors developed and evolved a rubric of indicators of scientific research based upon the literature. Six other authors applied this rubric to evaluate the 55 papers in the *2015 IEEE Security and Privacy conference proceedings*. We chose this conference for our study because it is long established as a highly selective security conference with high-quality security papers. To drive the literature review, we established the following five research questions:

- *RQ1: What types of artifacts are being evaluated (e.g. algorithm, language, model, process, protocol, or tool)?*
- *RQ2: How are the artifacts being evaluated (case study, experiment, survey, proof, discussion)?*
- *RQ3: Are there trends in the type of artifact and the evaluation method used to evaluate it?*
- *RQ4: How frequently are replications of prior work reported?*
- *RQ5: Do the papers contain all the recommended reporting components?*

The analysis described in this paper will establish a baseline measurement for progress in the advancement of scientific research in security. It will also provide a vehicle for that measurement to allow this type of analysis to be replicated with other conferences (such as USENIX Security and ACM Computer and Communications Security) and in future years. Through establishing this baseline, we hope to additionally advance community knowledge about thorough scientific reporting so future research can build upon current results, as we work together to build a science of security. This paper contributes the following:

- A rubric of indicators of the completeness of scientific security research.
- Baseline measurements of scientific security research from the 2015 IEEE Security and Privacy conference

The rest of this paper is structured as follows. Section 2 describes the relevant background information. Section 3 introduces the rubrics used for paper analysis. Section 4 explains the methodology we used to analyze the papers. Section 5 contains the results of the analysis. Section 6 provides some overall observations across the whole set of papers. Section 7 describes the lessons we learned in conducting this analysis. Section 8 enumerates our threats to validity. Section 9 summarizes the paper.

[3]https://sites.google.com/site/sosmuri/
[4]http://www.riscs.org.uk

2. BACKGROUND

A scientific approach research process starts with observation of the world, development of theories or models to represent those observations, analysis of those theories or models, followed by validation of the theories or models via hypothesis-driven research. Finally, replication of studies helps to provide data necessary to build formalized theories that codify empirically established understandings of the world.

Because of the opportunistic nature of much of the current security research, we are not able to fully investigate some of the fundamental relationships among key variables. Therefore use of a scientific approach is critical to understand these relationships and eliminate alternative explanations or solutions. This scientific rigor requires researchers focus on empirically validating research, searching for deeper insights, establishing causality among variables, answering important questions, formulating useful hypotheses, and reporting conclusions from their data.

This section describes some of the key concepts that are important for the advancement of science: replications, theory building, and meta-analysis. Then it discusses some examples of guidelines that have been developed to support this endeavor. Finally, it discusses some previous literature reviews in security.

2.1 Replications

Replication is repetition or reproduction of a research study in different contexts to determine if the causal relationships and results/ findings of the original study can be generalized to ensure the external validity of the study [5]. A key tenet of science is reproducibility. Hence, researchers are expected to be scientifically rigorous in their research design, analysis, and reporting. This rigor enables the replication of research so that results can be confirmed or refuted in different settings.

Researchers in medicine have been debating on the validity of their published results for some time [1, 20, 15]. A July 2015 article in MedlinePlus [2] reported that researchers could not reproduce half of the 100 publications in premier psychology journals. The August issue of Science [14] also reported similar outcomes and concluded that credibility of results depends on the repeatability of the research. The outcome of these findings led journal boards to modify their acceptance and evaluation criteria to include repeatability of results.

Most of the research findings in the literature are isolated in the papers published by a particular research group or laboratory. In theory, if there is enough detail about the original study, the results can be validated independently by other researchers [26]. Replication studies enable researchers and practitioners to disseminate novel approaches and methods into different contexts. Replication of previous research is also valuable for industry. The domain, scope and process-related variables may be different for each organization. No solution provides a "silver bullet" that is applicable under all circumstances. Identifying the conditions under which a particular result is valid might help organizations identify relevant research findings [6]. Replication studies in security science are important to build a consistent body of knowledge for researchers and practitioners.

2.2 Theory Building

Researchers, regardless of domain, are interested in understanding phenomena to build knowledge as well as to find a better solution to a problem. A theory is a belief that there is a pattern in phenomena [8]. A scientific theory is a well-substantiated explanation of some aspect of the world that is acquired via the scientific method and confirmed through multiple observations and experiments[5]. Scientific theories are therefore testable and make falsifiable predictions [22].

In order to be scientific, a theory must be submitted to and survive various kinds of tests [7], including: scrutiny by a critical peer group seeks to identify flaws in the justification of the theory, and empirical testing in observational or experimental studies. In security science, as an applied discipline, in addition to providing scientific knowledge, theories should also be useful to practitioners to support decision-making. For instance, they are helpful in choosing among implemented security mechanisms and in understanding security phenomena and their impact. Therefore, theory building is critical in security science to support the communication of research knowledge, develop common research agendas, and disseminate established findings to industry.

2.3 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]. It involves a thorough search of digital libraries to identify relevant studies, a clear and objective criteria for choosing which studies to include in the analysis, a sensitivity analysis to properly interpret the findings, and an objective way of calculating the study effect. Meta analyses aim to determine the existence, size and variability of an overall effect. The results of a meta-analysis can improve the precision of estimates of effect size, answer questions not posed by the individual studies, settle controversies arising from apparently conflicting studies, and generate new hypotheses [12]. Meta-analysis has intensively been discussed and successfully been applied in medical research [29, 9] to gain insights about whether the effect of a treatment is statistically significant compared to other treatments or not. Security science could also benefit from using meta-analysis to systematically analyze the effects of specific security solutions. Such an analysis requires a body of well-performed and reported empirical studies.

2.4 Guidelines for Applying and Reporting Research

Once a community is moving in the direction of increased consistency in empirical design and reporting of results, it is easier for the members to increasingly adopt the most appropriate approaches [30]. As a research domain matures researchers develop and follow clear guidelines for applying research methods and for reporting research results as for instance provided for medical research [3, 27], psychology [28, 13] or social sciences [4, 23]. Such guidelines also enable researchers to conduct study replications (see Section 2.2) and meta analysis (see Section 2.3).

In reviewing the research methods literature in other domains such as medical research, psychology and social sciences, we observe key lessons learned in balancing scientific rigor and industrial relevance. First, it is important to report the study design process in a structured and systematic way so that it is easier to understand and compare the results. Second, it is important to discuss design alternatives ensuring that the best choice is made. Third, it is important to elaborate on how to interpret results for practical insights and use. As these fields matured they began considering the trade-offs among candidate study designs to choose the most appropriate validation for a new model, protocol, process or tool and to identify threats to validity.

Finally, the way of reporting research results itself is a critical issue. As mentioned in the previous sections, reproducibility of research heavily relies on the level of detail in the study dissemination. Recognizing the need for additional support in this area, researchers at North Carolina State University, as part of the NSA's Science of Security Lablets (one of the groups mentioned in the Introduction) are developing research planning and publication guidelines[6] covering three types of research: analytical, empirical, and solution. The guidelines provide researchers with information tailored to the type of research about what should go in a research plan to help ensure that it is scientifically defensible. The guidelines also provide guidance on how to document the plan and results in a paper to support the advancement of science of security.

2.5 Previous Literature Reviews

We have not found any similar studies in the security literature that analyze the rigor of the evaluation methods reported in papers. The mapping studies (i.e. studies providing classifications of the type of research reports and results published in a specific field) that do exist in security engineering [21, 10, 11] do not focus on identifying the evaluation methods. Therefore, it is important for such a paper to be written. This paper is the first attempt to analyze the completeness of the evaluation methods in papers of the *IEEE Symposium on Security and Privacy.*

3. PAPER EVALUATION RUBRIC

To achieve our goal we analyzed papers to characterize the completeness of the infomation included with respect to enabling the advancement of the Science of Security. To accomplish this goal, we identified four types of information about each paper.

- **Evaluation Subject Type** (Section 3.1) – The type of artifact (i.e. the Evaluation Subject) proposed and/or evaluated in the paper (i.e. the subject).
- **Is New** (Section 3.2) – A designation of whether the Evaluation Subject(s) are novel or built upon a prior solution. This characteristic helps with understanding the prevalence of replications.
- **Evaluation Approach** (Section 3.3) – The approach(es) used for evaluation in the paper (e.g. a case study, an experiment, or a proof), that is *how* did the authors evaluate the Evaluation Subject(s).
- **Completion Rubrics** (Section 3.4) – A guide for determining the completeness of the information provided in the paper relative to the goals of science stated

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

[6]http://research.csc.ncsu.edu/security/lablet/research-planning-and-publication-guidelines

earlier. We have different rubrics for each Evaluation Approach.

The following subsections provide a detailed description of each of these types of information.

3.1 Evaluation Subject Type

In a field as diverse as security, a variety of solution types exist, for example: algorithms, processes, and tools. In this paper, we refer to these solutions as **Evaluation Subjects** and the type of the subject as the **Evaluation Subject Type**. For the sake of this paper, we are not concerned with the specific Evaluation Subject itself. Rather, we classify the Evaluation Subjects into more general Evaluation Subject Types.

The primary reason for identifying different Evaluation Subject Types is that we hypothesized that researchers would likely use different Evaluation Approaches based upon the Evaluation Type. For example, an *experiment* might be an appropriate approach to evaluate a tool while a *proof* would be more appropriate for an algorithm. Therefore, we determined it was important to separate out the various Evaluation Subject Types to better understand which ones were more prevalent as well as to understand trends in how researchers evaluated each type.

Prior to embarking on our detailed analysis of the papers published in the 2015 *IEEE Symposium on Security & Privacy* proceedings, we reviewed papers from previous years of the symposium as well as papers from other security conferences, including *USENIX Security* and *ACM Computers and Communication Security (CCS)* to identify the Evaluation Subject Types of solutions proposed. (Section 4.2 describes this process in more detail.) Based on that analysis, we identified six primary **Evaluation Subject Types** in the security literature. Realizing that different researchers may use different names for the same item, we provide a concrete definition for each subject to reduce the chances of misinterpretation.

The **Evaluation Subject Types** that we identified, along with their definitions, are as follows:

- **Algorithm/Theory (AL)** - a proposal of a new algorithm/theory or an update to an existing algorithm/theory.
- **Model (M)** - a graphical or mathematical description of a system and/or its properties;
- **Language (L)** - a new programming language;
- **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 thing into something else;
- **Tool (T)** - an implementation of a process; and

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

3.2 Is New

The second key piece of information is whether the **Evaluation Subject** is novel or built upon prior solutions. In some cases, a paper will both define an Evaluation Subject and also provide an evaluation of that subject. In other cases, the Evaluation Subjects are defined elsewhere, with the goal of the new paper being more focused on performing an evaluation of one or more Evaluation Subjects. The motivation for including this piece of information in our analysis is that one of the key components of scientific advancement is the ability to replicate previous research to validate the results. Without such replication, a field will have more difficulty building a solid scientific understanding of the phenomenon that are being studied.

Therefore, to provide some insight into this factor, for each Evaluation Subject identified in a paper, we determined where it was first described. This factor can assume one of two values as follows:

- **New (N)** - The Evaluation Subject is first proposed (described) in this paper;
- **Existing (E)** - The Evaluation Subject is proposed (described) in another paper.

3.3 Evaluation Approaches

The third key piece of information is the approach used by the authors to evaluate each **Evaluation Subject**. As described in Section 2, a researcher has the choice of various approaches to evaluate the claims of his or her research. Each of these approaches has their own strengths and weakness that must be taken into account when choosing the most appropriate one for a given situation. The primary reasons for identifying which **Evaluation Approach(es)** are used in a paper are to (1) characterize the prevalence of each approach; and (2) identify any patterns in which specific Evaluation Approaches are often used for a given Evaluation Subject Type.

For the five most common Evaluation Approaches found in our previous analysis of the literature, we provide a short description along the strengths and weaknesses of that approach. Section 3.4 provides a more detailed description of what type of information should be included in a paper for each of these Evaluation Approaches.

3.3.1 Experiment (EX)

An experiment is an orderly process that seeks to test a hypothesis to establish a causal relationship. It usually has one or more Independent Variables that each have one or more treatments (e.g. experimental, control, or baseline). The primary goal of an experiment is to measure the impact of the Independent Variable(s) on the Dependent Variable(s). In an experiment, the researcher exerts some level of control over other variables that could potentially confound the results. The level of control exerted over those variables affects whether the experiment is classified as a *Controlled Experiment* or as a *Quasi-Experiment*. For the sake of this paper, we do not differentiate between these two types of experiments.

The primary benefit of an experiment is that it provides researchers with the ability to conduct statistical analyses to establish a strong causal link between the Independent Variable(s) and the Dependent Variable(s). The primary weaknesses of an experiment are that it often takes much effort to design and execute and that the level of control required to establish causality often results in an unrealistic setting. Therefore, experiments are often useful to help establish strong causality by eliminating many of the con-

founding factors that may be present in a more realistic situation.

3.3.2 Case Study (CS)

A case study is an evaluation conducted in a more realistic setting that can be provided by an experiment. Where an experiment focuses on exerting control over the environment to establish a better causal link, a case study relaxes this control to conduct an evaluation in a more realistic setting. Rather than searching for a strong statistical result, a reseacher focuses more on understanding how well the new technology works in a realistic setting and identifying factors that may affect its usefulness. In a case study, researchers can use either real systems or example systems, that is systems that were developed specifically for research purposes but contain characteristics similar to real systems. In our work, we do not differentiate between these two types of systems.

The primary benefit of a case study is that it provides the researcher with information about how well the Evaluation Subject will work inside the constraints of a real environment, with all of the other distractions and interference that may impact effectiveness. The primary weakness of a case study is that the lack of control by the researcher means that there is less confidence in the true causal nature of the result. Case studies are often valuable to test usefulness in a realistic environment once an experiment has established a causal relationship in a more controlled setting.

3.3.3 Survey (Q)

In this category, we group together a number of qualitative research methods including surveys, interviews, focus groups, and opinion polls. These approaches all have a lot in common and are each used infrequently enough that it did not make sense to analyze each one separately. In general, these research approaches focus on gathering qualitative data (along with some quantitative data) directly from a set of subjects via a series of questions or guided discussions. A researcher must then take the information gathered from these interactions and analyze it to extract the meaningful information relative to the research question at hand.

The primary benefits of surveys is that they allow a researcher to gather much more in-depth information from research participants that might be allowed by experiments or case studies. The primary weaknesses of surveys is that they are often time-consuming to conduct and analyze and that the results provided often contain some level of subjectivity.

3.3.4 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 an algorithm.

The primary strength of a proof is that its formality allows a researcher to make a direct link between existing theory and the conclusion drawn. The primary weakness of a proof is that it is only applicable in certain situations.

3.3.5 Discussion/Argument (D)

This category covers evaluation that does not contain any empirical data. 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, that is, 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.

3.4 Completion Rubrics

Each Evaluation Approach described in Section 3.3 has different characteristics. Therefore, to assess whether papers provided all of the necessary information to enable the advancement of the science of security, we designed a rubric for each Evaluation Approach. 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 results and to determine 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 Evaluation Approach used. For each Evaluation Approach, we examined the literature [17, 16, 18, 19, 24], and used our own experience, to define specifically which items should be present and what information they should contain. Based on that analysis, we defined a rubric for each Evaluation Type (as shown in Table 2 to Table 5 in Appendix 9).

Each rubric consists of a series of questions that help to determine whether all relevant information has been completely reported. To help standardize the scoring for each rubric item, we define three possible answers:

- **Yes** - the information is present in the paper and easy to find (i.e. well-formatted),

- **Partial** - the information is present in the paper, but may not be easy to find, and
- **No** - the information is omitted from the paper.

Each rubric has specific, concrete definition for these answers.

4. METHODOLOGY

In this section, we describe our research methodology, including the analysis process and the pilot studies used to validate the approach.

4.1 Steps in Analysis Process

The research team consisted of the 11 authors on this paper, which includes 5 faculty members, 5 PhD students, and 1 Postdoctoral researcher from five universities in four countries. The PhD students and the Postdoctoral researcher performed the paper analysis under the supervision of their respective faculty members. The remainder of this section refers to the PhD students and the Postdoctoral researcher collectively as *reviewers*. We followed a six-step process.

1. One faculty member author randomly assigned each paper to two reviewers. To reduce any bias, we ensured that each reviewer was paired with each of the other reviewers across the whole set of papers.
2. Using the types and rubrics defined in the Section 3, each reviewer independently analyzed each assigned paper to identify (a) the Evaluation Subject Types(s), (b) whether the Evaluation Subjects were new, (c) the Evaluation Approach(es) used, and (d) the completeness of the reporting of the Evaluation Approach(es).
3. To make this process transparent, each reviewer marked-up a PDF version of the paper to label the rubric items and whether each was fully or partially present. This step helped reviewers be more objective and to ease the process of resolving discrepancies (Step 6).
4. To prevent bias from the other reviewer's scoring, reviewers recorded their analysis in their own spreadsheet.
5. Once both reviewers had independently analyzed each paper, we ran a program to analyze the results and flag any cells for which the two reviewers disagreed.
6. For any papers that had a disagreement, the reviewers discussed these discrepancies and arrived at a final characterization for each paper.

4.2 Pilot Studies

Prior to conducting the full analysis of the 2015 papers, we tested and evolved the rubric through a series of smaller pilot studies. First, to ensure that we had well-defined *Evaluation Subject Types* and *Evaluation Approaches*, the five faculty member authors reviewed the proceedings from an earlier year of the *IEEE Security & Privacy Proceedings* while we were together at a conference. This co-location allowed us to meet frequently to arrive at the final list. Second, to ensure that we had a good set of rubric questions, all authors conducted further pilot studies by reviewing papers from the *IEEE Symposium on Security & Privacy*, *USENIX Security*, and *ACM CCS* to ensure that the questions were clear and objective. Based on these pilot reviews, we made some slight adjustments to arrive at the current versions of the Evaluation Subject Types, the Evaluation Approaches, and the Completion Rubrics. These adjustments were primarily focused on refining and clarifying the definitions of the Evaluation Subject Types, the Evaluation Approaches, and the Completion Rubric questions. These pilot studies were very helpful in ensuring that we had a collective understanding of the process.

5. RESULTS

This section reports the results of the paper analysis using the Evaluation Subject Types, Evaluation Approaches, and Rubrics defined in the previous section. After providing a description of the data preparation in Section 5.1, the remaining subsections summarize the key results, organized around the research questions. Section 6 discusses the implications of these results.

5.1 Data Preparation

The results of tool analysis of the reviews (Step 5 in Section 4.1) indicated that the reviewers agreed on 223 items and disagreed on 99 items. To resolve these discrepancies, the reviewers consulted their highlighted PDF versions of the papers (from Step 3 in Section 4.1) to recall the rationale behind their rubric answers. The reviewers then discussed the location of each rubric item (e.g. page number or paragraph based on highlighting) and why they chose a particular answer to the rubric item. We considered a discrepancy to be resolved when both reviewers agreed. We were able to resolve all discrepancies from all papers in this manner. After discrepancy resolution, we performed a manual analysis of the data to answer the following questions.

5.2 RQ1: Type of artifact evaluated

The goal of this question was to provide a general overview of the Evaluation Subject Types on which the *IEEE Security & Privacy* authors focused. This overview provides some insight into the types of contributions that are seen as being most important within the community, by virtue of their acceptance in the conference. Figure 1 shows that *Tool* and *Process* were the two most common Evaluation Subject Types, with *Language* being the least popular.

Another part of our analysis process was to determine how complex papers were in regards to the number of Evaluation Subjects within an individual paper. As shown in Figure 2, the majority of papers contained only one Evaluation Subject. Only a few papers contained three Evaluation Subjects. Overall, we can observe that the papers were not overly complex by containing multiple Evaluation Subjects.

5.3 RQ2: Method of evaluation

Similar to the Evaluation Subject Types described in the previous section, we also analyzed whether there were any Evaluation Approaches that were more prevalent than others. This analysis provides some insight into the choice of evaluation methods used by the security community and can provide guidance to other researchers about the use of those methods. Figure 3 shows that *Case Studies* were the overwhelming choice for evaluation in these papers. Interestingly *Experiments* and *Questionnaires* were used very rarely.

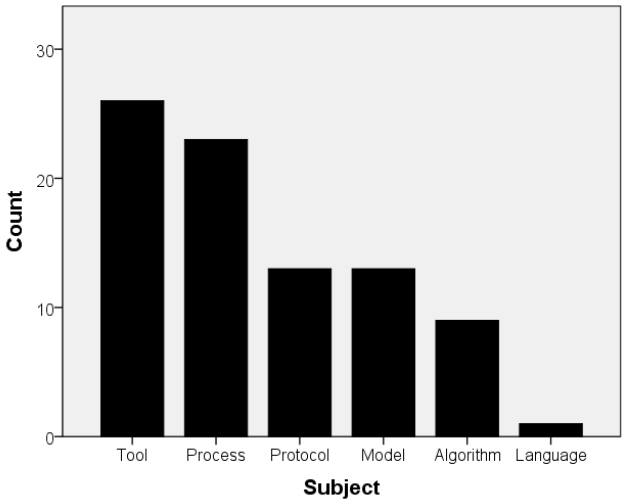

Figure 1: Frequency of Evaluation Subjects

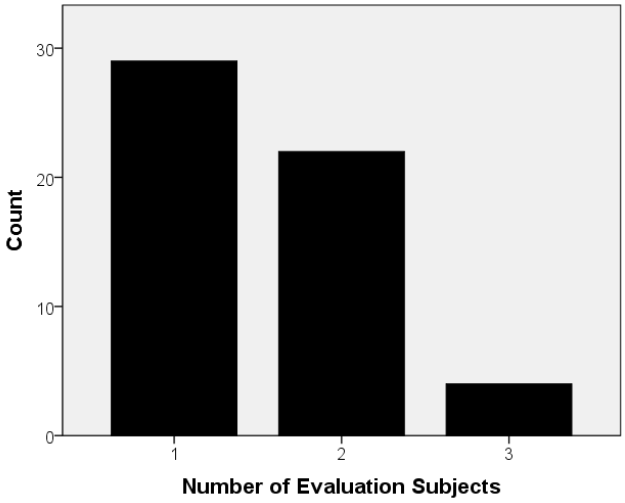

Figure 2: Number of Evaluation Subjects

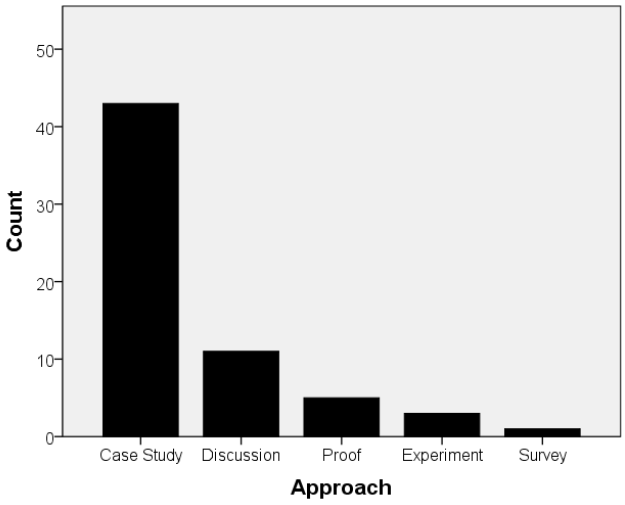

Figure 3: Frequency of Evaluation Approaches

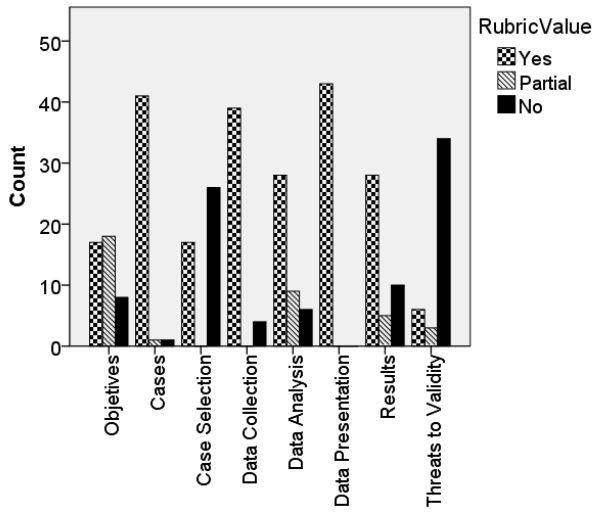

Figure 4: Case Study Rubric

5.4 RQ3: Trends in evaluation subjects and methods

To better understand whether researchers tended to use the same Evaluation Approach for a given Evaluation Subject Type, we analyzed whether there were any patterns in mapping of Evaluation Approaches to Evaluation Subject Types. Table 1 shows this mapping. As *Case Studies* were the dominant type of evaluation, they are also the most prevalent for each type of subject. Analyzing the data in a bit more detail indicates that the *Discussion* approach is the second most popular for many of the Evaluation Subject Types.

5.5 RQ4: Frequency of replication

One indication of the presence of replications is whether the Evaluation Subjects were introduced in the paper in which they are evaluated or are introduced elsewhere. The results show that a New Evaluation Subject occurred most frequently with a total of 76 new subjects in our sample while an Existing Evaluation Subject only occurred 9 times.

5.6 RQ5: Completeness of papers

The final part of our analysis process was to determine how the reviewers answered the completion rubric questions for each of the Evaluation Approaches. In this analysis, we specifically focus on how frequently the reviewers answered *Yes* to the rubric questions, because a *Yes* answer indicates that the paper was well-documented and thorough with respect to that rubric item. Because Case Studies and Proofs were the only Evaluation Approaches that had more than a few uses, we only provide the detailed results for those two Evaluation Approaches in Figure 4 and Figure 5, respectively.

One interesting observation about the Case Study results is that two items CS3 and CS8 had more *No* answers than *Yes* and *Partial* answers combined. CS3 refers to whether the methodology for choosing the cases was clearly described. CS8 refers to whether threats to validity were thoroughly discussed. Conversely, the Proofs were well-described, in general, with the *Yes* answers always being more prevalent than the *No* answers.

Table 1: Evaluation Approach Type by Evaluation Subject

	Case Study	Experiment	Discussion	Survey	Proof
Process	20	1	2	0	0
Tool	20	2	3	1	0
Model	9	0	1	0	4
Protocol	9	0	1	0	5
Algorithm	7	0	0	0	2
Language	1	0	0	0	0

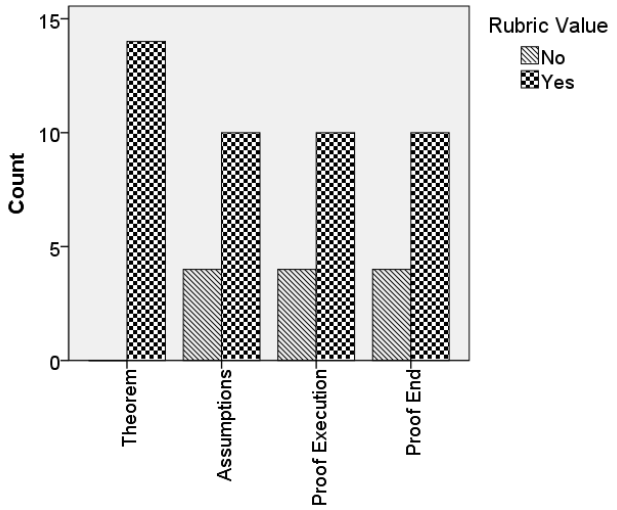

Figure 5: Proof Rubric

6. 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.

6.1 Overall Observations

In keeping with the fact that the *IEEE Symposium on Security & Privacy* is one of the top security conferences, we were pleased to observe that in general the papers were readable and easy to follow. We can observe that because papers tended to focus on only one Evaluation Subject, they have the ability to be more thorough than if they were describing multiple evaluation subjects. However, there was still some important information that was often missing from the papers or was included implicitly rather than explicitly. Without this information, it is more difficult for readers to interpret the results and for other researchers to replicate studies. Many of the initial disagreements between reviewers resulted from this lack of explicitness.

First, the papers often lacked a specific, concrete description of the *Research Objectives*. Research objectives describe the overall goal of the research. In most cases, these objectives could be inferred from the description of the paper's contribution, but they were not explicitly stated. From a scientific perspective this information is important because it helps readers have the proper context within which to understand the results presented.

Second, papers often lacked a detailed discussion of the motivation or process for selecting the specific cases included in the case studies. The cases refer to the specific people, organizations, or artifacts chosen to be the subjects of the studies. Without knowing how or why the cases were chosen, a reader cannot be sure whether the results reported in the paper will be relevant to their local context.

Third, most papers lacked a discussion of the threats to validity. Threats to validity help explain the areas in which the study design was not able to account for all potential confounding factors. It is important for researchers to identify and report these confounding factors and explain what actions they took to reduce their presence or severity. Other researchers can then decide how important those threats are in their context and choose to replicate the study to eliminate some of the threats.

6.2 Evaluation Subjects

An interesting observation from Figure 1 is that *Tools* and *Processes* were the most frequent type of Evaluation Subject. Security professionals aim to make systems as secure as possible, so it makes sense that tools and processes would be formulated as defenses. New languages may be considered less of a defensive strategy because there are already safe programming practices, but human error still causes problems. Furthermore, automation through tools and processes is more feasible and applicable across a system, while a language may be more domain specific and technical (thus more difficult to implement).

6.3 New vs. Existing Evaluation Subjects

In regards to the novelty of the Evaluation Subjects, we found that it was much more common for researchers to evaluate *New* subjects rather than *Existing* subjects. This observation suggests that there is lack of replication of studies. This lack of replication means that meta-analysis and theory building, two key aspects of the scientific approach, are not possible in many cases.

In addition, we observed that in most cases researchers used case studies rather than experiments to evaluate these new Evaluation Subjects. Given the discussion of the strengths and weakness of these approaches in Section 3, one would expect to see a higher percentage of experiments to allow for a better understanding of causality. Conversely, since the field of security is constantly evolving to address new and emerging threats, the prevalence of New subjects is not surprising.

As important as replications are to building science, researchers can be dissuaded from performing replications if reviewers of top venues do not respect the value of replications and expect novel artifacts to be created and evaluated. This phenomenon occurs in many scientific fields and could be at least a partial source of this observation.

6.4 Experiments vs. Case Studies

Finally, we make some comparisons of *Experiments* and *Case Studies*. During our reviews we found that the way papers described the research, often made it difficult to distinguish between Case Studies and Experiments. This problem is complicated by the fact that authors often use inconsistent and incorrect terminology. That is, in some cases there is no label for the Evaluation Approach. In other cases, authors use the term *Experiment* when in reality they performed a *Case Study*. This ambiguous reporting makes it more challenging to understand and interpret the results.

Overall, as shown in Figure 1, *Case Study* was the dominant Evaluation Approach. As mentioned in Section 6.3, we expected to see more *Experiments* that would be useful in establishing stronger causal relationships about the Evaluation Subjects and their intended effects. Upon reflection on the other results, it is not completely surprising that *Case Studies* were so dominant. First, the fact that most Evaluation Subjects were new rather than existing, it is reasonable for a small sample to be chosen as an initial testbed. Second, in the security domain, it is likely that demonstrating success on specific cases (i.e. that represent large or important problems) will be more effective for convincing people of the value of the new approach compared with a more controlled experimental setting. For example, a new security threat may breach the security of a specific tool. Therefore, given the assurance that the Evaluation Subject defends against the threat on a specific set of previously high-risk cases raises acceptance of that particular Evaluation Subject.

7. LESSONS LEARNED

This section provides our lessons learned in performing this review because we aim for others to replicate this work over the years to examine changes and trends in scientific reporting of security research. Overall the reviewers thought that the rubrics were clearly labeled, easy to use, and helpful for analyzing the papers. Even given the overall positive impression and given the extensive effort devoted to developing and debugging the Completion Rubrics, we still had a number of discrepancies in the initial evaluation. Through this process, we realized that there is a need to make our rubric definitions even more concrete to help reviewers more easily differentiate between the *Yes* and *Partial* answers to rubric questions.

We also realized that there were some key aspects of the papers not covered by our rubrics. For example, we did not analyze whether the related work section was included and whether it was complete. We also did not analyze whether the results tied directly back to the research objectives. We would have had to make some subjective judgments to do these analyses, which we avoided in this paper. In a future review, we will consider these factors more carefully.

As described earlier, even though there were a large number of discrepancies after the initial review, we were able to resolve them all by the end. The reviewers found this discrepancy resolution process to be the most difficult part of the study. The review team consisted of people from different countries and different time zones. These differences resulted in delays during the dispute resolution stage. In future reviews, we will establish pre-defined reviewing times in which reviewers can meet online and resolve the discrepancies in real-time.

8. THREATS TO VALIDITY

To help readers properly interpret our paper, we offer this section describing the threats to validity. A threat to validity is any information or variables that might reduce the validity of the study findings.

First, with regards to **Internal Validity**, while we made a strong effort to create an unbiased set of rubrics for evaluation, some of the feedback from our review team indicated that there was still some amount of subjectivity present. It is possible that if the same analysis was undertaken by a different set of researchers, slightly different conclusions might result. That said, we did a thorough job of piloting and validating the rubrics before we begun the process, so we think this threat is minimal.

Second, with regards to **External Validity**, we only analyzed papers from one year of one conference. It is possible that 2015 was not representative of the prior years of the *IEEE Symposium on Security & Privacy*. It is also possible that the papers published in the *IEEE Symposium on Security & Privacy* are not representative of the security research community at large. To combat these threats, we (or other researchers) will need to replicate this analysis with a more diverse set of papers.

9. SUMMARY

In this paper we have analyzed the papers from the 2015 *IEEE Symposium on Security and Privacy* to establish a baseline of the scientific rigor in the reporting of research results. The overall motivation was to determine whether security research is being documented in a way that allows for replication, meta-analysis, and theory building, three key pillars of scientific research. To perform this analysis, we established a set of **Evaluation Subject Types** and **Evaluation Approaches**. For each **Evaluation Approach** we developed a **Completeness Rubric** to help determine whether all important information was explicitly reported in each paper. Using that rubric, we determined that, while most papers are well-written, they are often lacking key information. To provide more insights, we plan to conduct a more extensive review containing the papers from multiple conferences. In addition, this baseline can serve as a comparison point for a similar review that will be conducted in five or ten years.

Acknowledgments

We acknowledge the support of the NSA Science of Security Lablets, QE LaB - Living Models for Open Systems (FFG 822740) and MOBSTECO (FWF P 26194-N15). We would also like to thank Amiangshu Bosu and Chris Corley for their contribution to the paper reviewing rubrics.

10. REFERENCES

[1] Freely associating. *Nature Genetics*, 22, 1999.
[2] How reliable are medical studies? half of findings couldn't be replicated. *MedlinePlus*, 2015.
[3] J. Abramson and Z. Abramson. *Research methods in community medicine: surveys, epidemiological research, programme evaluation, clinical trials*. John Wiley & Sons, 2011.
[4] E. Babbie. *The practice of social research*. Cengage Learning, 2015.

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

[6] B. Caglayan, B. Turhan, A. Bener, M. Habayeb, A. Miranskyy, and E. Cialini. Merits of organizational metrics in defect prediction: An industrial application. In *International Conference on Software Engineering - Software Engineering in Practice Track*, may 2015.

[7] A. Cournand and M. Meyer. The scientist's code. *Minerva*, 14(1):79–96, 1976.

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

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

[10] M. Felderer and E. Fourneret. A systematic classification of security regression testing approaches. *International Journal on Software Tools for Technology Transfer*, 17(3):305–319, 2015.

[11] M. Felderer, P. Zech, R. Breu, M. Büchler, and A. Pretschner. Model-based security testing: a taxonomy and systematic classification. *Software Testing, Verification and Reliability*, 2015.

[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. Heger. What is a theory? *LiveScience*, 2012.

[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. SjÃÿberg, 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 *International Symposium on Empirical Software Engineering*, page 10, Nov 2005.

[18] B. Kitchenham and S. Pfleeger. Personal opinion surveys. In F. Shull, J. Singer, and D. SjÃÿberg, 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 falseâĂŤbut a little replication goes a long way. *PLOS Medicine*, 4, 2007.

[21] H. Oueslati, M. M. Rahman, and L. b. Othmane. Literature review of the challenges of developing secure software using the agile approach. In *Availability, Reliability and Security (ARES), 2015 10th International Conference on*, pages 540–547. IEEE, 2015.

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

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

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

[25] P. Sandle. Cyber crime costs global economy $445 billion a year: report. *Reuters* 9 June 2014. Available: http://www.reuters.com/article/2014/06/09/us-cybersecurity-mcafee-csis-idUSKBN0EK0SV20140609.

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

APPENDIX

Paper Evaluation Rubrics

Table 2: Paper Rubric Items - Experiments

Rubric Item	Yes	Partial	No
EX1: Are the research objectives described? (e.g., goals, questions, hypotheses)	*Clearly defined and labeled (e.g. âĂIJResearch Question, RQ, Objective, âĂę)*	*Included in the text but not clearly labeled*	*Not present*
EX2: Are the methods for subject sampling described? (e.g., recruitment/selection process)	*Explicitly defined in the text*	N/A	*Not defined in the text*
EX3: Are the data collection procedures described (e.g. definition of the metrics/variables, operational constructs, measurement levels)	*Explicitly described in the text*	N/A	*Not described in the text*
EX4: 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 all of the following: statistical tests (by name) or other analysis method, results of statistical test (including p-value)*	*Paper includes some but not all of the above*	*Paper includes none of the above*
EX5: 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
EX6: 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, etcâĂę*	N/A	*Paper does not describe the data*
EX7: 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	The results are discussed, but not in a separate section	*The results are not discussed*
EX8: Is there a dedicated discussion of the threats to validity (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*

Table 3: Paper Rubric Items - Case Study

Rubric Item	Yes	Partial	No
CS1: Are the research objectives 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 âĂIJResearch Question, RQ, Objective, âĂę)*	*Included in the text but either in the wrong location or not clearly labeled (see Yes above)*	*Not present*
CS2: Are the case and its units of analysis described? (i.e., what is the context of the study, what is being tried)	*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*
CS3: Are the methods for subject selection described? (e.g., 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*
CS4: Are the data collection procedures (i.e., how was this completed) and research instruments (i.e. questionnaire, mining tools, performance computation) described?	*Described in the text*	N/A	*Not described in the text*
CS5: Are the analysis procedures described? (e.g., hypothesis checks, statistical tests, p-values, performance metrics, precision, recall, accuracy, False positive, False negative)	*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*
CS6: Is data presented with appropriate descriptive statistics to provide an understanding of the analysis? (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, charts, tables to describe the data etcâĂę*	N/A	*Paper does not contain a description of the data*
CS7: 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*	*The results are discussed, but not in a separate section*	*The results are not discussed*
CS8: Is there a dedicated discussion of the threats to validity (i.e., limitations or mitigations)?	*There is a separate Threats to Validity Section or Limitations Section*	*Threats to validity are discussed, but not in a separate section*	*Threats to validity are not discussed*

Table 4: Paper Rubric Items - Qualitative

Rubric Item	Yes	Partial	No
Q1: Are the research objectives described? (e.g., goals, questions, hypotheses)	*Clearly defined a labeled (e.g. âĂIJResearch Question, RQ, Objective, âĂę)*	*Included in the text but not clearly labeled*	*Not present*
Q2: Is a rationale behind the questions given? (i.e., why these questions and not others)	*Each question has a rationale provided*	*Some questions have a rationale*	*No questions have a rationale*
Q3: Are the evaluation procedures described?	*The paper explicitly links the research objectives to the survey questions.*	N/A	*The paper does not explicitly link the research questions/hypotheses to the survey questions.*
Q4: Are the respondents described? (e.g., demographics)	*The text explicitly describes the characteristics relevant to the research objectives*	N/A	*The text does not describe the relevant characteristics*
Q5: Are the sampling methods described? (i.e., why these respondents? e.g., mailing list, advertised, etc)	*The paper explicitly describes: the method for recruiting participants, how the questionnaire was advertised, why the participants are the correct ones, inclusion/exclusion criteria*	*The paper explicitly describes some, but not all of the above*	*The paper describes none of the above*
Q6: Is how the responses were processed described? (e.g., cleaning data, answer coding)	*The paper explicitly describes how the data was cleaned, how the answers were coded, triangulation of data and inter-rater reliability (only for qualitative analysis)*	*The paper describes some, but not all of the above*	*The paper describes none of the above*
Q7: Is there a dedicated discussion of the threats to validity (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*

Table 5: Paper Rubric Items - Proof & Discussion

Proof Rubric

Rubric Item	Yes	Partial	No
P1: Is the theorem being proved stated? (i.e., goal)	*Theorem is explicitly stated*	N/A	*Theorem is not explicitly stated*
P2: Are any assumptions used described?	*Assumptions are described*	N/A	*Assumptions are not described*
P3: Is informal material given to provide intuition on how the proof works?	*There is informal material, such as a proof sketch or an explanation of the proof in context.*	N/A	*There was no sketch or context*
P4: Is where the proof ends marked? (e.g., is there a clear ending of the proof before other, possibly unrelated, text begins)	*There is a clear end to the proof*	N/A	*There is no clear end to the proof*

Discussion Rubric

Rubric Item	Yes	Partial	No
D1: Is the goal of the argument described?	*The goal of the argument is explicitly described*	N/A	*The goal of the argument is not explicitly described*
D2: Are two or more premises and a conclusion given? (*Aristotle's rule*)	*Two or more premises and a conclusion are given*	*Some, but not all of the above are given*	*None of the above are given*
D3: Is the related knowledge described?	*Related knowledge is explicitly described*	N/A	*Related knowledge is not explicitly described*
D4: Is the supporting evidence described or cited?	*Supporting evidence is described or cited*	N/A	*Supporting evidence is not described or cited*