Common Pitfalls in Writing about Security and Privacy Human Subjects Experiments, and How to Avoid Them

Stuart Schechter
Microsoft Research
StuS@microsoft.com

ABSTRACT

Reviewers of papers that describe human subjects experiments of security and privacy often observe that authors are prone to a set of common mistakes. In this document I provide advice to help researchers avoid these mistakes and document their experiments properly.

Executive summary

1. State the hypothesis or hypotheses you are testing precisely.
2. If testing a security hypothesis, have a clear and defensible threat model.
3. Avoid misleading yourself or your reader in any way, especially in selling your contribution or in translating results into conclusions.
4. Carefully explain how participants' behaviors were observed, scored, and then fed into statistical tests.
5. Ask colleagues who were not involved in the research to read an early draft of your paper.
6. Disclose all limitations in your study design and results that you are aware of.
7. Label all axes in graphs and add captions to ensure figures are self explanatory.
8. Do not assume that correlation implies causation, or that the lack of statistical significance implies a hypothesis is false.
9. Ensure that your data source meets the requirements of your statistical tests. (When in doubt, use a non-parametric test.)
10. Don't be afraid to ask for help.

1. INTRODUCTION

Program committees must often reject papers with fascinating ideas or clever experimental methodologies – which we would love to see presented – because the validity of the experimental results are unclear: we cannot ascertain key experimental details from the paper, how data were collected, or whether a statistical test is indeed sufficient to support a hypothesis. Many of the mistakes that force program committees to reject papers are common and easily avoided.

I have written this document to guide researchers in how to avoid the most common pitfalls when submitting to the Symposium on Usable Privacy and Security (SOUPS) and other venues that accept human subjects studies about security and privacy. I provide a mixture of generally accepted practices for writing computer science papers, practices specific to human subjects studies in security, and less universally accepted advice based on my opinion and past experiences as an author and reviewer. This work is not intended to be a complete guide to writing a paper. Rather, it is intended to help those with a general knowledge of how to write an academic paper to adapt their skills to writing up security and privacy human subjects studies and to help all authors avoid common pitfalls.

2. YOUR CONTRIBUTION

It is important to define your contribution by explaining the general problem you are trying to solve and the specific instance of the problem that is the basis for your work, the hypotheses you wanted to test, unique features of your approach, and your results.

As you lay out your paper, and especially your contribution, you must be meticulously careful to avoid misleading yourself or your reader in any way. Prior work should not be unduly disparaged, your innovations should not be exaggerated, and no limitation of your work should be swept under the rug. Graduate students are taught that they need to sell their work and its contribution to the field – and this is an important skill – but good marketing should be about isolating the value of your contribution and presenting it clearly.

Exaggerations, undocumented limitations, or other issues that lead reviewers to suspect they are being mislead will cause them to start reading your paper more suspiciously. This takes their focus away from appreciating the contribution of your work.

Alas, even if you are honest in how you convey your research, it is exceedingly hard to determine if you have con-

veyed the information necessary for someone other than yourself to understand it. The best way to determine if your research will be comprehensible is to ask colleagues who were not involved in the research to read an early draft of your paper. If you are not a native-level speaker of the language in which the work is written, find a native-level speaker to point out and help remove any problems with language, spelling, or idioms.

3. EXPERIMENTAL DESIGN

Experiments are designed to test hypotheses. It is important that you state the hypothesis or hypotheses you are testing precisely.

Your experimental design should be documented in sufficient detail to allow another researcher to replicate your experiment without consulting you for missing details. While many papers fail to reach this standard and still are accepted into top publications, you will be well served by working to ensure that your experiment is documented well enough to allow it to be replicated.

3.1 What to include

One way to collect the details you'll need to present about your experiment is to imagine the chronological progression of your experiment from the perspective of your participants (noting the variations between treatment groups), your perspective as a researcher, and the perspective of anyone else involved in the conduct of the experiment (including recruiters). Often, the description of the experiment in your paper will also follow a chronological time line. Questions that you'll want to answer in your description of the experiment should include:

- How were participants recruited?
- What incentive was provided to participate?
- Where did the participants go to participate?
- What were participants asked to do before, as part of, and following the experiment?
- What information did participants learn along the way and how might this have influenced behaviors later on in the experiment?
- If the study was a between-subjects study, how did the experience (treatment) vary between the groups?
- Did the order of any tasks change for different participants?

Detail the recruiting process and the resulting demographic makeup of the participant pool. If the participant pool does not precisely reflect the population of interest you'll want to discuss the differences.

If the study involved deceiving participants, you'll want to explain the deception and document if and when the deception was revealed to the participants.

Finally, describe how participants were observed and how observations translated into data points used for later analysis. A surprising number of submissions fail to explain how behaviors are observed, scored, and then fed into statistical tests. A statistical test is of little value if the reviewer doesn't know what data are being fed into it and how they were obtained.

Along the way, you should highlight the decision points you came across in designing the experiment and explain the reasoning behind the choices you made. Own up to mistakes, especially if you have suggestions for how the methodology might have been improved. I've never run a study and not wished I'd designed their experiment at least slightly differently when the time came to write up the results.

3.2 Ethics and participant safety

Many institutions (including all U.S. universities) require human subjects experiments to be approved by an Institutional Review Board (IRB). Explain whether you needed to run the study through a review board (IRB) and any concerns that your design addressed.

Regardless of whether your experiment was reviewed by an IRB, you will want to describe any potential ethical or safety risks and how you addressed them. If your study used deception, did you later disclose the deception to participants? If so, how did participants react to the deception? If you collected information from participants that could cause harm if it fell into the wrong hands, how did you work to protect that data?

3.3 Supplementary documentation

No matter how hard you try, you may not be able to fit every detail of your experiment – such as the precise wording of every question posed to participants – in the body of the paper. To assist readers who may have questions that you could not be expected to anticipate or that fall outside the stated contribution of your work, consider attaching your study materials into an appendix at the end of your paper. Appendices are not a substitute for carefully detailing your methodology in the paper, as reviewers are neither required nor expected to read them. You must still ensure that you have explained all of the details essential the the validity of your experimental goals, methodology, and conclusions.

4. SPECIAL CONSIDERATIONS: EXPERIMENTS INVOLVING SIMULATED ATTACKS

Experiments involving behavior in response to privacy and security threats warrant extra consideration in areas that may be less important when writing up other HCI experiments.

4.1 Have a clear threat model

If testing security behavior in response to an attack, clearly explain the assumptions made about the information, capabilities, and resources available to an attacker. These assumptions are your *threat model*. A common failing in papers is to fail to document or justify the assumptions that make up your threat model. Document how any attack you may simulate is similar to, and different from, a real attack.

4.2 Avoid bias in favor of your own system

Beware of the potential or appearance of bias when designing a threat model and testing security behavior as it relates to a system you have designed or built. An attacker will have more incentive to break your system than you do. Reviewers may look at an attack you tested against and

believe, rightly or wrongly, that they could have designed more effective attacks. One way to remove bias is to identify third parties with both the talent and incentive to develop the best possible attack against the system you have built. We've even considered holding contests to identify the most effective attack against which to test our systems.

4.3 Address ecological validity

An experiment is ecological valid if, and only if, participants in the study behave as they would in the real-life situation that the experiment is trying to emulate.

Ecological validity is especially challenging when designing experiments of security and privacy behavior because, for most of the interesting scenarios to study, security or privacy will not be the primary goal of the individual. Furthermore, not completing the primary task may also have risks and consequences. Simulating the forces motivating a user's drive to complete a primary task and response to potential risks is challenging, and so extra attention to ecological validity is warranted as you design and document your experiment.

Questions you'll want to address include whether participants knew the study was about security or that the researchers study security. If they did, would this have led them to pay more attention to security? Did participants believe that they would be negatively impacted if they exhibited an insecure behavior in the same way that they would in real life? Did participants believe they had something to lose if they failed to complete a task because they could not do so securely or did not know how to do so securely? Have you considered the potential for users' behaviors to become habituated in a manner that could negatively impact security or usability?

5. DISCLOSE LIMITATIONS

Be up front and disclose all the limitations of your study design, participant demographics, statistical tests, and other methodological issues.

Part of a reviewer's job is to keep papers with misleading or fundamentally flawed results from entering the peer-reviewed research literature, where their publication may lead others to cite their conclusions as fact. Reviewers always have an eye out for methodological limitations, overstated results, or other statements that might mislead a reader about your research findings. This is your job as well, and the better you can show that you've done it the better reviewers will feel about your work. Reviewers will be less likely to critique your study design if they see that you are aware of the limitations of your work and have fully disclosed them. By showing that you've already looked at your own experiment with a critical eye, you allow your reviewers to focus more on evaluating your contribution and less on searching for flaws that they suspect may be hiding under the covers.

Your discussion of limitations may be placed into its own subsection of your methodology section, results, or discussion sections. You may even want to promote limitations to an independent section of the paper.

6. PRESENTING RESULTS

For each test, remind the reader of the hypothesis to be tested, present the data used to test the hypothesis, explain your choice of statistical test, and then present the result of that test.

6.1 Tables and figures

Clearly label the rows and columns of tables and the axes of figures. Make sure the title or caption precisely describes what the contents of the table or figure represent. You'd be surprised how many papers we receive in which axes are not labeled or for which it is unclear what a graph is representing.

6.2 Statistical validity

There are two common statistical errors that we on the SOUPS committee see every year. First, authors often describe a failure to disprove the null hypothesis as an indication that the null hypothesis is true. Rather, failure to disprove the null hypothesis simply means that there was not enough evidence, given the size of the sample and the random outcomes, to prove the null hypothesis to be false. Failure to prove something is false does not mean it is true.

Similarly, Another common gaffe is to observe that a statistical test failed to show a significant difference between two groups and to conclude that no such difference exists. The only way one could ever prove that no difference existed between two populations would be to test, and obtain a perfect measurement, of every member of both populations, rendering statistical tests unnecessary. Otherwise, the best you can do is to measure the statistical certainty that the difference between two populations is within a certain bound.

The second common mistake is to use more than one data point per participant in a statistical test that assumes data points are independent. The very fact that two data points come from a single participant means they are not independent.

An example of this mistake is to ask 10 participants 10 questions, and then feed the 100 responses into a statistical test. The statistical test will produce a p value as it would if there were actually 1 question asked of 100 participants. If it isn't already clear to you why this is a problem, imagine that one were to ask 50 questions about statistics of one man and one woman. The statistical test has 50 samples for men and 50 samples for women. Let's say the man has no knowledge of statistics, and gets all 50 questions wrong. The woman gets them all right. Misled to believe that it had 50 independent trials from both men and women, the statistical test would indicate that women are better at statistics than men with a p value far below 0.01–a significant result! It is hopefully intuitively obvious that one cannot make such a strong conclusion about two populations by sampling only one member from each.

There are a number of ways to run statistical tests when you have multiple data points from the same participant. One simple one is to take a summary statistic for each participant and run the statistical test on the summary statistic. A student t-test is, after all, a test for comparing students' scores on exams that have many questions. It is designed to be used for a summary statistic, their test score, over a large enough number of questions that the score fits a normal distribution.

Speaking of t-tests, and other statistical tests that rely on scores to be drawn from a normal distribution, you will

want to show that your scores indeed appear to resemble a normal distribution if you are using these tests. At the very least, explain why you believe the scores should fall into a normal distribution. Better yet, use a non-parametric test when there is any doubt that the distribution is normal. If you have any question about the right test to use or how to use it, don't be afraid to ask for help. If you don't have a knowledgeable colleague handy, a number of helpful online guides can be found by searching on the phrase "choosing the right statistical test".

As the number of statistical tests used to test one or more hypotheses grows, so does the chance that one will reach your significance threshold due to random chance. Be sure to correct for multiple comparisons. See, for example, the Wikipedia entry on Multiple Comparisons. This is easy to do but, again, *a surprising number of papers have numerous statistical tests and no correction for multiple comparisons.*

After completing both the experiments and analyses required to test a hypothesis, you'll want to discuss your results. In doing so, be careful not to jump to conclusions beyond those supported by your hypothesis and tests. Speculation about possible implications that could be tested with future work should be presented as such.

Finally, remember that the discovery of a statistical correlation does not prove causation. Consider alternate hypotheses that might explain your results.

7. CITING RELATED WORK

To find related work perform web searches on key terms, scour the HCISEC bibliography and the proceedings of SOUPS, CHI, the IEEE Symposium on Security and Privacy (Oakland), USENIX Security, ACM CCS, and NDSS. As you look at related work, note key terms that may be useful for searching for other related work. You may also want to consult with other researchers who have written work in an area closely related to yours.

7.1 Where does related work go?

The HCI community mostly adheres to a convention of presenting related work before experiments to ensure that the reader has all necessary background information before reading about the experimental methodology. This convention exists, in part, because may experiments build on the methodology of prior experiments and so much of the related work is germane to the experimental design.

The security community mostly adheres to a convention of presenting related work after experiments and results are presented. This convention makes sense when much of the related work you may want to cite is not needed to motivate or provide background on your experiment. When this is the case, the related work may bog down a reader who is interested in getting to the details of your experiment and would prefer to understand the broader context later. If you follow this convention, you may need to cite some papers twice: first in an introductory section to motivate or provide essential background and later, after your experiment and results have been presented, to put your work in broader context.

Either convention is accepted at SOUPS and so you should choose the convention that works best for your paper.

7.2 Citation etiquette

Wherever you cite related work, make sure to use citation numbers as an essential supplement to more descriptive text that describes a work. Do not use citation numbers alone to describe a work. In other words, do not say "[42] presents an experimental methodology for testing the obsessiveness of paper reviewers" but instead say "Zaphod Beeblebrox *et al.* developed one of the first experimental methods for testing the obsessiveness of peer reviewers [42]." By so doing, you'll help familiarize your reader with the names of those working in the field, your paper will read more smoothly, and those familiar with the literature won't have to flip pages forward and backward to identify which work you are citing.

Providing citation context will also help you avoid making the mistake of citing multiple works with one long string of citation numbers, such as "[59,71,72,78,84]". Such bulk citations provide inadequate clues to the reader about what each paper is about, its contribution to the field, and its relation to your work. If work is related enough to cite, it's usually related enough to warrant some explanation.

Citing tenuously related work to increase the reference count will not earn points with reviewers or excuse the absence of key related work that has been overlooked. While there is no prescribed number of references, expect warning bells to go off in reviewers' minds if you have fewer than 10 citations or if more than a quarter of citations are to your own work.

8. ACKNOWLEDGEMENTS

Assuming you received help along the way, your paper should have an acknowledgements section, though this should not be submitted during review if your paper is anonymized.

I am grateful to Lorrie Cranor and Andrew Patrick for encouraging me to write this article, as well as their comments, corrections, and suggestions along the way.