PSI UI/UX Experiments

REU 2017 Final Paper

Kathryn Taylor

August 4, 2017

I Introduction

The Harvard Privacy Tools Project is a collaboration between computer scientists, social scientists, lawyers, and others within the Harvard community to develop strategies and tools to enhance data privacy. The Private data Sharing Interface (PSI) is one such tool. With funding from the National Science Foundation Secure and Trustworthy Cyberspace Project, PSI has been developed with the goal of allowing researchers depositing sensitive datasets through the Harvard Dataverse platform to publicly release differentially private statistics about their data.

One of the most critical elements of the PSI tool is the Budget interface, where data depositors set parameters that are translated into a privacy "budget" to be distributed across the differentially private statistics that they choose to release. In order to achieve the goal of bringing differential privacy into practice with the PSI tool‘, the Budget feature must be useful and intuitive for data depositors. Much work has been done by Jack Murtagh, James Honaker, and other Harvard researchers to ensure that PSI correctly translates the mathematical principles of differential privacy into interactive features. The next step is to test the usability of those features in the Budget interface.

Because of my background in social science, computer science, and graphic design, the PSI team invited me to work with them this summer to help design and conduct a usability study for PSI's Budget tool. We set out to conduct a pilot test of five subjects, followed by a formal study with 20 subjects aimed at illustrating PSI's readiness for real-world use. After successfully completing a pilot round with five participants, we decided to pause and reassess the interface, as well as the procedures and purpose of the test. Our changes were based on feedback from study participants, members of the Privacy Tools Project, and user interface experts. With the newly improved interface and test protocols, we began our formal usability study at the midpoint of the summer. With 10 completed tests, we have made good progress and have gained valuable insights into the usability and overall potential of the PSI tool. Our primary constraint in terms of completing the formal study has been the challenge of recruiting appropriate test subjects.

In this paper, I describe the original plans for this summer's project, the initial steps we took, the changes we made to our strategies, protocols, and the the PSI tool. With regards to the formal usability study, which remains in progress, I present our methodologies, our current results, and our future plans.

II Original Vision of the Usability Study

At the most simple level, our goal for the summer was to complete a usability study for the PSI prototype. Our only requirements for test subjects were that they had worked with some kind of sensitive or individual-level data before, and that they were above the age of 18. We expected to be able to do productive tests with anyone fitting this description, and to be able to show from these tests that PSI has great utility and that it is ready for widespread use.

I distinguish the "original vision" for the study from where we currently stand in order to highlight how certain elements of our plan have evolved, a process that I describe in detail in subsequent sections of this paper. However, at the core, we have stayed true to and successfully fulfilled the project's original goals, which were, according to Jack Murtagh, Georgios Kellaris, and James Honaker, to:

- Develop appropriate metrics, survey questions, and interface tasks for evaluating the usability of the tool,

- Recruit study participants,
- Run participants through the study procedures: informed consent, questionnaires, and user interface tasks,
- Organize and synthesize the collected data, including translating audio and video recordings into concrete quantitative measures, and
- Analyze the collected data to produce a list of recommended improvements to the tool and common mistakes participants made.

Though the final two goals regarding organizing the analyzing the collected data cannot be fully realized until the formal usability study is complete, I have achieved these goals with regards to the tests we have conducted so far by organizing and quantifying all results in a spreadsheet (found in the PSI Google Drive), and by producing a list of recommendations for improvements to the tool based on our tests up to this point (found later in this document).

III Informal Testing

Before sending out emails to recruit subjects for the study, we decided to conduct a couple of informal user tests. The purpose of this step was to allow us to gain familiarity and comfort with the flow of the test procedures, and to identify and fix any glaring flaws in the protocols or the interface itself before opening ourselves up to making mistakes with real test subjects that we did not know personally.

Two of the Privacy Tools interns graciously agreed to act as test subjects for these informal tests, which proved to be extremely valuable for our usability study overall. I was able to observe my mentors as they conducted the tests, which prepared me to be able to run subsequent tests myself. I made note of the fact that the subjects appeared to be feeling pressured and stressed by our test, and began thinking of ways to minimize that effect through my interactions with future subjects. These tests also helped us to begin to identify the main issues with the PSI interface, including the lack of clarity and excessive length of the explanatory text in the "Introduction" and "Help" pages and the lack of intuitiveness of the editable fields such as the privacy loss parameters, the error estimates, the hold button, and the reserve budget slider.

Going into this informal testing phase, our team had a cache of test materials stored in our Google Drive, including a consent form, a background questionnaire, a list of comprehension questions about differential privacy, the text for a general scenario for users to complete, a dataset to be used in the scenario, a list of specific tasks for users to complete, and a System Usability Scale form. The general flow of presenting these materials to test subjects had been decided, but it became clear during the informal tests that it was necessary to hammer out in greater detail the list of steps we needed to follow during the study. Based on this realization, I compiled a document titled "PSI UI/UX Experiments: Detailed Steps" containing the order and details of every action we needed to perform, all notes related to those actions, as well as the text of the accompanying materials to be presented to subjects during each step.

I compiled the protocols document following the first informal test. We were able to do a practice run of these newly organized procedures during the second informal test, which went well in terms of our testing methods and allowed us to feel comfortable and prepared moving into the pilot study.

IV Pilot Study

IV.I Recruiting

After refining and streamlining our testing strategies, we sent out a recruiting email to faculty and students of the Harvard Institute for Quantitative Social Science (IQSS), a group we felt would have relevant statistical knowledge and experience working with individual-level data. Formal outreach to this group was approved by the IRB. Additionally, we conducted word-of-mouth recruiting, identifying and inquiring with friends and colleagues who appeared to be a good fit for the study.

Through these channels, we were able to recruit five test subjects of varying educational and professional backgrounds. They ranged from having completed some college to having completed a Ph.D, and came from fields including health policy and technology, psychology, and computer science. In keeping with our criteria, all of these subjects were either "Unfamiliar" or "Somewhat Familiar" with differential privacy.

IV.II Testing

For each of these tests, we followed the same testing procedures. We held the tests in closed conference rooms in the Maxwell-Dworkin Building on Harvard's campus. Participants were informed beforehand that the test would take approximately one hour, and that they would be compensated with a $20 Amazon gift card.

The documents referenced in the test protocols can all be found in the PSI Google Drive. The steps are enumerated below:

1. Hand the participant a hard copy of the consent form for them to sign.
2. Hand the participant a hard copy of the background questionnaire.
3. Open a laptop with the PSI website loaded in front of the participant and ask them to read the introduction screen.
4. Show the participant the dataset to be used during their interaction with the PSI tool. Tell them that they should become familiar with it and behave during the simulation as if it is their dataset, containing sensitive information, about which they would like to be able to release some statistics while maintaining privacy.
5. After the participant has explored the dataset, ask them to read the next introductory page in PSI, which describes the concept of privacy loss parameters and secrecy of the sample.
6. Give the participant a hard copy of a list of comprehension questions about differential privacy. If the participant does not answer the questions correctly, explain the correct answers to them before proceeding.
7. Inform the participant that we are turning on audio and video recording, and ask them to begin expressing their thought processes out loud.
8. Give the participant a hard copy of the text of a general scenario for them to read and carry out as they interact with PSI. The scenario tells the participant that they are a social scientist who has collected a sensitive dataset (the one previously shown to them), which they think contains information that would be helpful to other researchers interested in studying the relationship between race and income in Millennials. Next, the text asks them, "What differentially private statistics would you want to release to help these social scientists decide if your dataset contains a rich enough variety of populations to answer their research questions?" Participants are asked to write their answers to this open-ended question on the sheet of paper. Next, the text asks participants to use PSI to release differentially private statistics about their dataset. At this point, we ask the participant to enter the interface and begin making choices, such as the selection of privacy loss parameters and the selection of variables and statistics.
9. Ask the participant specific questions about the decisions they made while using the tool.
10. Give the participant a hard copy of a list of specific tasks to complete within the tool.
11. Give the participant a hard copy of the System Usability Scale for them to fill out about their overall experience with PSI.

IV.III Interventions

Regarding Step 6, which involves participants answering comprehension questions, we decided to go over the participants' responses during the test and explain to them any questions that they answered incorrectly. We understood that this might be unorthodox in terms of usability tests, which generally try to minimize intervention. Our thinking, however, was that people who were entirely unfamiliar with differential privacy would not be able to use PSI, and that if, even after reading the introductory text, participants did not have a basic understanding of the key concepts, they would not be able to proceed with the test. We decided that, even though the test would be somewhat contaminated, participants who answered the comprehension questions incorrectly could still give us valuable usability insights after being corrected. We did not want to simply send people home after they answered incorrectly, or to continue with a useless test where the participant did not understand what was going on.

IV.IV Results

Subjects were generally able to complete the entire test in approximately one hour. One subject did not answer any comprehension questions correctly, three subjects answered only one comprehension question correctly, and one subject answered all comprehension questions correctly. For the participants who had at least some incorrect answers, we discussed their answers with them until they seemed to have a more correct understanding of the concepts.

After completing their tasks in the interface, subjects awarded PSI an average System Usability Score of 59.5 out of 100, with the lowest answer being a 45 and the highest being a 72.5. Information from usability.gov about scoring the System Usability Scale notes that the score should not be interpreted as one would a grade on an academic test, where anything below 70 is considered to be a "failure." An "average" score is considered to be a 68/100. Therefore, during the pilot test, PSI was not rated as an excellent system to use as of yet, but a score of 59.5 suggests that it is not horrible, either. There is reasonable (not discouraging) room for improvement.

Though we only tested five subjects during this pilot period, we gained many qualitative insights to apply to PSI and to our testing protocols. With regards to our test protocols, the main takeaway from the pilot study was that our test was very difficult because it asked subjects to understand and perform challenging, open-ended tasks involving technical terms while providing minimal straightforward guidance. During one of our tests, Derek Murphy from Harvard IQSS, who has experience with user interface testing, observed our session and provided us with notes. His main points were as follows:

- Do not allow tests to run overtime by much. Don't make participants feel obligated to stay. Cut off the time if necessary.
- Have one member of our team primarily interact with subjects to make them feel comfortable and to make the interaction less confusing and intimidating.
- Assure participants that we are not testing them, and that we understand that some of the tasks may not be easy. Try to decrease stress.
- Some of our steps were too open-ended.
- Simplify the language used throughout the tool to make it more digestible.
- Record how long each task takes.
- When participants ask questions about features or concepts that we do not wish to answer for the purposes of the integrity of the test, we can deflect by asking questions back to the participant.
- Consider the option of not correcting subjects on the comprehension questions.
- While participants are completing tasks within the tool and expressing their thoughts out loud, respond with verbal and physical cues to show that we are listening, thereby encouraging them to continue speaking.

With regards to the PSI tool itself, pilot study participants highlighted a some common issues, listed below:

- The explanatory text in the introduction and the help files is too long and complicated, especially for non-native English speakers.
- The question mark buttons leading to the help files are not large or obvious enough.
- Technical concepts such as boolean variables bin names, and confidence levels are difficult to understand within the tool if subjects did not already know their meaning.
- It is unclear how to use different moving parts such as the reserve budget slider, the hold buttons, the editable errors, and the parameter entries.

With these results in mind, we moved into a period of adjustment.

V Adjustment Period

Equipped with this concrete and significant feedback from the pilot study, we decided to pause our testing to solicit additional feedback from the Privacy Tools group through the means of my midterm presentation, to make edits to the PSI tool and the test protocols, and to more explicitly and realistically outline our vision for the formal usability study.

V.I Changes to PSI

Our pilot study participants commonly highlighted that the text used in the PSI interface was too long and difficult to read. I edited all of the interface's text to make it more straightforward. Additionally, in response to subjects' comments that the help files were not easy to find, we enlarged and darkened the question mark buttons that are located at different places around the tool.

Another central theme of the pilot study was that subjects did not seem to be fully equipped, even after reading the introductory materials, to use all of PSI's moving parts when they entered the tool and attempted to complete their tasks. To better orient users to the different features of the interface, Jack developed a tutorial to help users to deliberately explore and understand each of the parts.

V.II Changes to the Test Protocols

V.II.1 Clarity

Following the pilot study, it was clear to us that our test was too difficult and open-ended. We set out to handle this problem by eliminating, adding, reordering, and revising some steps. We eliminated the question asking participants what kinds of statistics they would want to calculate in general. This question, asked before participants interacted with PSI, was intended to inform us of the differences between their outside expectations and their actual experience with the PSI tool. However, in practice, the question only served to confuse and pressure people, so we decided to remove it and instead ask participants *after* completing their tasks whether there were any features they would have wanted or expected from a tool like PSI that were not present. This change allowed us to continue comparing expectations and reality while also reducing stress for participants.

V.II.2 Correcting Comprehension

In response to advice from Derek, we decided not to correct participants on their answers to the comprehension questions. Between the pilot and formal tests, we edited the introductory text to make it more understandable, which should help with subjects' comrprehension. Additionally, as I will discuss more in the next section, we expected the formal study participants' ability to use the tool and to work with differential privacy to be higher because we would be increasing our selectivity. Overall, we felt that, to best simulate a real world interaction with PSI, we should not intervene to explain correct information.

V.II.3 Tutorial

Instead of throwing subjects into the deep end by asking them to use the tool however they would like to release some differentially private statistics, as we did during the pilot study, we decided to help subjects become comfortable with the tool by having them go through the tutorial. This would reduce stress and enhance performance in a realistic way, because we plan to have some sort of tutorial feature in the real version of PSI.

V.II.4 Tasks

Next, instead of the long and complicated tasks from the pilot study, we decided to develop a longer list of smaller, more understandable tasks that would lead subjects to use specific features of the tool. I identified all of the parts of PSI that we wanted to test: adding and deleting statistics to be calculated, editing the confidence level, adjusting the privacy loss parameters, entering a population size for secrecy of the sample, reserving budget for future users, interpreting the error estimates, editing and holding the error, and submitting the selections. After formulating this list, in consultation with Jack, I wrote a list of 10 tasks to prompt subjects to use each of these features without directly telling them to do so. Instead of saying, "Edit the privacy loss parameters," we say, "You are thinking about your dataset, and you realize that it contains some information that makes it more sensitive than you originally thought. Use the tool to make the changes necessary to reflect this shift." This kind of questioning allows us to observe how clear the function of each of PSI's features is, because we see whether test subjects know where to go within the tool to accomplish somewhat open-ended goals.

Below is the list of tasks I wrote corresponding to each of the features we wished to test:

1. **(Adding Statistics)** You decide that the income and race variables are also important for future researchers, so you decide to release statistics for these. Add a mean and a quantile for income, as well as a histogram for race.
2. **(Deleting Statistics)** You no longer wish to include a quantile for income. Delete this statistic.
3. **(Editing Confidence Level)** You decide that you want to be very confident in your error estimates. Use the tool to set a 98 percent confidence level.
4. **(Editing Privacy Loss Parameters)** You are thinking about your dataset, and you realize that it contains some information that makes it more sensitive than you originally thought. Use the tool to make the changes necessary to reflect this shift.
5. **(Entering Population Size)** You have just been informed by a colleague that your dataset was actually randomly sampled from a population of size 1,200,000. Use the tool to make changes to reflect this. Does this make your statistics more or less accurate?
6. **(Reserving Budget)** You decide that it would be useful to allow other researchers who do not have access to your raw data to make some of their own selections for statistics to calculate from your dataset. Use the tool to make changes to reflect this.
7. **(Interpreting Error Estimates)** How much error is there for the mean age statistic? What does this number mean?
8. **(Editing Error)** Make it so that the released mean age is off from its true mean by at most one year. Is this more or less accurate than what you had before?
9. **(Holding and Editing Error)** Make it so that each count in the released race histogram is off from the true count by at most 10 people *without* changing the error you just set for mean age.
10. **(Submitting Selections)** You are satisfied with your statistics and your error estimates. Finalize your selections.

V.II.5 Ordering

For the sake of clarity, we decided to change the order of steps so that subjects would read all introductory text (instead of just the first page) and answer comprehension questions before being given a dataset and a specific scenario. Before, subjects would read the first window, then look at their dataset, then read the next page, then do comprehension questions, then give them the scenario. After conducting the pilot tests with this ordering, we realized that it was not intuitive.

V.II.6 Measurement

A major addition to the test protocols after the pilot was notes about collecting quantitative data. During the pilot, we took extensive notes during the tests. For the formal study, we decided that we needed more consistent metrics, so we decided to collect the following data for each subject:

- Scores:
  - Correct/incorrect answers for comprehension questions
  - Critical and non-critical errors
  - Answers and overall score for System Usability Scale
- Times:
  - How long it takes to read the introductory pages
  - How long it takes to fill in the privacy loss parameters
  - How long it takes to complete the tutorial
  - How long it takes to complete each of the 10 specific tasks

Here, a "critical error" is a mistake made by a participant that leads to them being unable to complete a given task correctly. A "non-critical error" is when a participant does something incorrectly, but it does not lead them to fail the task at hand. This can mean that they made a mistake that did not prevent them from achieving correctness, or that they made a critical mistake but were able to correct themselves.

V.II.7 Updated Protocols

After making these edits, the test protocols (with notes about when and what to measure at each step) are as follows:

1. Hand the participant a hard copy of the consent form for them to sign.
2. Hand the participant a hard copy of the background questionnaire.
3. Open a laptop with the PSI website loaded in front of the participant and ask them to read the two introductory screens, stopping before entering the privacy loss parameters.
4. Give the participant a hard copy of a list of comprehension questions about differential privacy.
5. Present the participant with the text of the scenario, as well as the dataset to be used in the simulation. Tell them that they should become familiar with the dataset and behave during the scenario as if this is their dataset, containing sensitive information, about which they would like to be able to release some statistics while maintaining privacy.
6. Inform the participant that you are turning on audio and video recording before they begin using the tool. Turn on audio and video recording and ask the participant to express their thought process verbally.
7. Ask the participant to enter the tool and go through the tutorial, releasing a mean age for the example.
8. Give the participant a hard copy of the list of specific tasks, asking them questions about their choices after each. Read each task out loud.
9. Ask the participant if they have any additional questions or suggestions. What else would they want or expect from a tool like this?
10. Give the participant a hard copy of the System Usability Scale for them to fill out about their overall experience with PSI.

V.III Revisiting the Goals of the Usability Study

I conducted some brief research into the existing literature on user interface studies to assess how many test subjects are usually used. This exercise resulted in the confirmation of the necessity of testing approximately 20 subjects. This is generally accepted as a number that lends itself to extracting quantitative results.

Before the pilot study, our criteria for test subjects was simply that they are over 18 and have some experience working with data. After observing our pilot subjects, however, we realized that only the subjects with some statistical background were able to be really successful with the tool. We decided to start selecting subjects with statistical knowledge, and also to open up recruiting to people who were already familiar with differential privacy. This is because we realized that at least some of our real-world users will be familiar with differential privacy, so it would be unrealistic to only test people with no outside knowledge.

In terms of our vision for the study as a whole, we decided that we would be successful in the formal study if we completed approximately 20 tests with reasonable subjects, collected qualitative and quantitative insights, and were able to, as a result, prepare and move forward with PSI's first phase of implementation.

VI Formal Study

VI.I Test Subjects

To recruit subjects for the formal testing, we reached out again to our approved mailing lists and engaged in more concerted word-of-mouth efforts. Through these means, we were able to bring in 10 subjects over a two-and-a-half-week period. We realized, however, that subject recruiting is a more difficult challenge than we had anticipated, and that it should be one of the key considerations when approaching a usability study. If we had gotten approval to reach out to a larger number of groups, or if we had been prepared to conduct virtual testing with subjects in different places, we might have been able to complete more tests before the end of the summer.

Figure 1: Education Level of Test Subjects

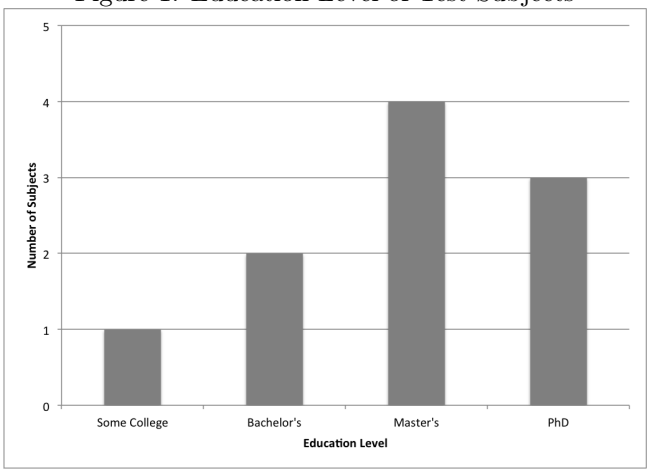
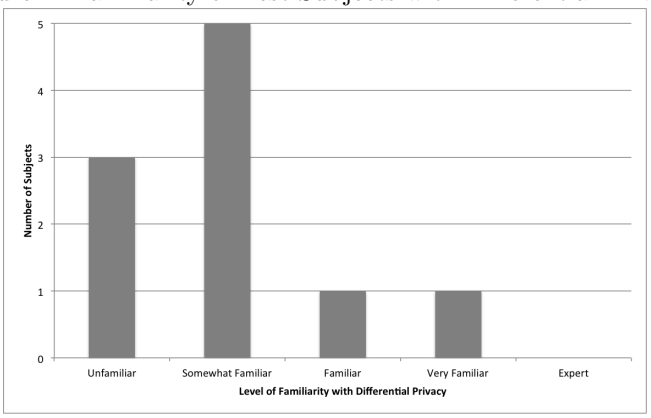


Figure 2: Familiarity of Test Subjects with Differential Privacy



Overall, our formal test subjects had a higher level of general education, statistical knowledge, and familiarity with differential privacy than the pilot test subjects. We hypothesized that this would contribute to more successful testing. Figure 1 shows how many test subjects had achieved each level of education. The most common educational background was a Master's degree, with four test subjects having been educated to that point. An additional three test subjects had achieved a Ph.D.

Figure 2 shows subjects' prior familiarity with differential privacy. Only three of our 10 test subjects were entirely unfamiliar with differential privacy. The rest were at least somewhat familiar.

VI.II Results and Insights So Far

Because the formal testing period is not yet complete, these results are not final. However, with 10 tests done, it is worthwhile to see what we can learn at this point.

One positive result of the formal tests thus far is that the negative feedback we've received has predictably aligned with issues already brought to our attention by the pilot study. We have not been taken by surprise by a host of new problems, but rather have been able to collect more details about clear, exiting areas for improvement.

Figure 3: Task Accuracy of 10 Formal Test Subjects

Tasks	1			2			3			4			5			6			7			8			9			10		
Time	CEs	NCEs	Time	CEs	NCEs	Time	CEs	NCEs	Time	CEs	NCEs	Time	CEs	NCEs	Time	CEs	NCEs	Time	CEs	NCEs	Time	CEs	NCEs	Time	CEs	NCEs	Time	CEs	NCEs
3:45	2	1	0:05	1	0	0:04	0	0	0:30	0	1	0:10	0	0	0:22	0	1	0:02	0	0	0:41	2	0	0:02	0	0	0:05	0	0
4:25	0	0	0:24	0	0	0:04	0	0	0:57	0	0	0:30	0	0	6:25	0	0	0:44	0	0	0:27	0	1	3:54	1	0	0:57	0	0
2:27	1	0	0:01	0	0	1:07	0	0	0:47	0	0	1:30	0	0	0:47	0	0	0:32	1	0	4:06	2	0	0:09	0	0	0:04	0	0
3:11	0	0	0:15	0	0	0:56	0	0	0:54	0	0	0:09	0	0	1:07	0	0	0:34	0	0	0:05	0	0	0:16	0	0	0:05	0	0
6:12	0	0	0:06	0	0	0:03	0	0	0:28	0	1	0:17	0	0	0:07	0	0	0:27	0	0	0:05	0	0	1:01	0	0	0:11	0	0
3:19	0	0	0:01	0	0	0:06	0	0	0:32	0	0	0:11	0	0	0:30	0	0	0:54	0	0	0:11	0	0	0:40	0	1	0:09	0	0
5:33	1	0	0:05	0	0	2:35	0	0	0:44	1	0	0:17	0	0	0:12	0	0	2:40	1	0	0:32	1	0	0:01	1	0	0:34	0	0
4:06	1	1	0:07	0	0	0:06	0	0	1:48	1	0	0:28	0	0	0:56	0	0	0:19	0	0	0:05	0	0	0:07	0	0	0:06	0	0
3:51	1	0	0:03	0	0	1:04	0	0	0:40	0	0	0:06	0	0	0:10	0	0	0:29	0	0	0:15	0	0	0:05	0	0	0:05	0	0
2:44	0	1	0:03	0	0	0:25	1	0	0:54	2	0	0:20	0	0	1:34	0	0	0:57	0	0	0:11	0	0	0:09	0	0	0:07	0	0

Figure 3 contains the details of each of the 10 test subjects' performance on each of the 10 usability tasks. Each

row corresponds to one subject. Cells are highlighted green when the subject performed the task correctly. Yellow is for when the subject committed a non-critical error, and red is for when they committed a critical error, meaning they did not successfully complete the task.

The tasks that we expected to be straightforward, almost all subjects completed correctly. Most or all subjects correctly performed tasks 2, 3, 5, 6, 7, and 10, which correspond to deleting statistics, editing the confidence level, entering population size, using the reserve budget slider, and interpreting the error, respectively. Subjects did still express some confusion with some of the features involved in these tasks, which will be addressed in our upcoming edits to PSI, but they were still able to use these features successfully.

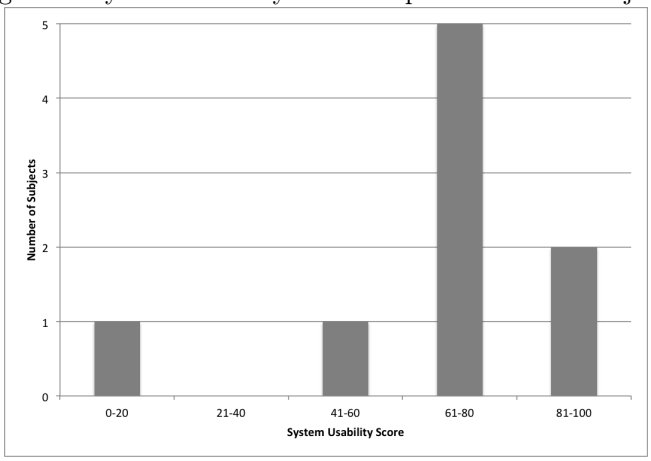
The more difficult tasks understandably presented problems for some subjects. Not only did observing these issues inform us of problems with the tool, but it also gave us ideas for how to fix them. As illustrated by Figure 3, subjects commonly experienced difficulty with tasks 1, 4, 8, and 9, which correspond to selecting and entering metadata for statistics, editing the privacy loss parameters, and editing and holding the error, respectively.

Below is a list of changes to the PSI tool that I compiled as a result of subjects' experience with the tasks:

- Add some sort of "submit" or "OK" button for selecting statistics in the middle area of the tool.
- Change the order in which the variables populate the middle area. Currently, after adding one statistic, the next statistic chosen shows up at the top of the list. Several participants noted that they expected newer additions to appear at the bottom of the list.
- Simplify the language in the introductory and tutorial text, which can be difficult to understand.
- Clarify the concept of a "budget" to help people understand the relationship between the different numbers in the interface, such as reserving the budget and distributing the error.
- Rephrase the "Confidence Level (alpha)" description so that it is either "Confidence Level" and can be filled out with 0.95 for a 95 percent confidence level, or "Alpha" and can be filled out with 0.05 for a 95 percent confidence level.
- Make sure that the help file appears consistently and in the right section across the tool.
- Make sure that all pop-up information appears in modals, not alert windows, which make people think that they have done something wrong.
- Ensure consistency of the appearance of the mouse when hovering over different parts of the tool. Currently, sometimes there is a hand, sometimes a question mark, and sometimes just a cursor (when it should be clickable).
- Change the use of the color red around the tool. Several subjects expressed discomfort with it, thinking that it meant something "bad."

Figure 4 shows the distribution of the scores for the System Usability Scale given by the formal test subjects.

Figure 4: System Usability Scale Responses of Test Subjects



Subjects gave PSI an average score of 65 out of 100. The majority of scores were in the upper two bins, which is an improvement from the pilot study. As discussed previously, scores in the 60s and 70s for the System Usability Scale are not associated with failure, but rather a normal level of usability.

VI.III Notes and Questions

VI.III.1 The Budget Concept

One of the main action items resulting from the formal study is the process of clarifying the concept of a privacy "budget" and how the other calculations relate to it. This remains an open question for us, as we have tried several different iterations of explanatory text to inform the user about the budget concept and we continue to experience problems and confusions. More attention must be paid to developing a clear and relatively simple way to convey this idea to subjects of different backgrounds. This may take the form of a video, a more specific tutorial or scenario, or better introductory text.

VI.III.2 Length of Text

Related to this idea of explaining concepts to the user, we are also still working to find a way to cut down the amount of text people must read before using the tool while maintaining a high level of accuracy and detail in our discussion of technical concepts. Most test subjects have expressed some frustration or hesitation with the high amount of reading required by PSI. During our tests, we encouraged subjects to read all available text in order to use the tool. In practice, however, we suspect that many users would see a large chunk of text and simply skip it. Dealing with this issue is a difficult challenge that may not be able to be solved cleanly. We might simply need to decide between detail and brevity. Choosing detail would mean maintaining a relatively high volume of text for the sake of clarity, while taking the risk that users might skim, skip, or become frustrated with the text. Choosing brevity would make the tool initially more approachable, but might cause problems late on when subjects lack specific information about the tool's features.

VI.III.3 Defining Standards for Time

A question related to our metrics is how to extract insights from the time records for the user tests. We have not yet defined what a "good" time for each task would be, and as such do not have a way to understand how successful our users have been. This is something to think about as we complete the testing and analysis phases, and certainly should be defined more clearly in future studies.

VI.III.4 Recruiting

One of our earliest and greatest challenges throughout our usability studies was the issue of recruiting test subjects. Once we exhausted our outreach by word-of-mouth and through the mailing lists we had approved by the IRB, we were unable to quickly and easily get subjects for testing. This is an issue that must be considered in depth at the very beginning of conceiving a usability study, because as many different options as possible should be IRB-approved and ready to be explored. Additionally, if the idea of conducting remote tests via video chat had been more developed by the time of the start of our study, we could have accessed and tested more subjects from different places, which would also have enhanced the robustness of our findings.

VII Conclusion

VII.I Future Plans

The most obvious next step is to conduct more user tests to complete the formal usability study. After this, we will write up the complete results for an external paper on PSI's usability study.

Completing this testing phase does not mean that no further testing should be done in relation to PSI and differential privacy. Further similar tests could be done on the new version of PSI that will exist after implementing our upcoming changes. Additionally, tests could be done using datasets supplied by the test participants themselves, rather than the pre-populated one that we used for all subjects. More advanced features such as regression will need to be tested, and specific tests for different kinds of research (a psychology-specific test versus an economics test) could be useful as well. More thought should be given to these kinds of future plans.

VII.II Acknowledgments

I would like to thank my mentor, Jack Murtagh, who worked with me closely throughout the summer, for giving me helpful guidance when necessary while respecting and acting upon my ideas for the project. I am grateful to Dr. Salil Vadhan, Dr. James Honaker, Dr. Georgios Kellaris, and the rest of the faculty of the Harvard Privacy Tools Project for their encouragement and insights. Special thanks to Derek Murphy from Harvard IQSS for observing one of our user tests and providing constructive feedback that allowed us to greatly improve our protocols.

Thank you to all of our test subjects for giving their time, energy, and expertise to help us improve PSI. Thank you to my fellow Privacy Tools REU students for making my experience this summer worthwhile through spirited foosball games, culinary outings, and their genuine enthusiasm for data privacy.