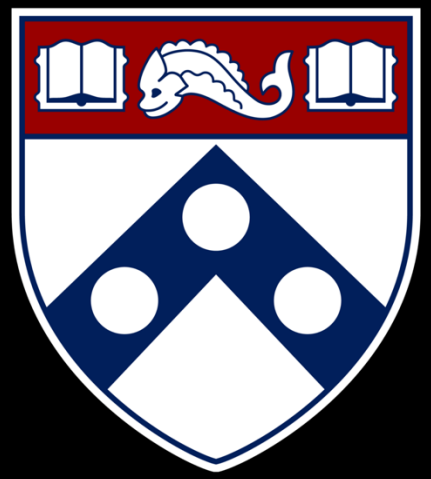
Empirical Security & Privacy, for Humans

UPenn CIS 7000-010

11/04/2025



What Are the Chances? Explaining the Epsilon Parameter in Differential Privacy

Reading

What Are the Chances?
Explaining the Epsilon Parameter in Differential Privacy

Priyanka Nanayakkara[*†]
Northwestern University

Mary Anne Smart[*]
University of California San Diego

Rachel Cummings[†‡]
Columbia University

Gabriel Kaptchuk[‡]
Boston University

Elissa M. Redmiles[‡]
Max Planck Institute for Software Systems

Abstract

Differential privacy (DP) is a mathematical privacy notion increasingly deployed across government and industry. With DP, privacy protections are probabilistic: they are bounded by the privacy loss budget parameter, ε. Prior work in health and computational science finds that people struggle to reason about probabilistic risks. Yet, communicating the implications of ε to people contributing their data is vital to avoiding privacy theater—presenting meaningless privacy protection as meaningful—and empowering more informed data-sharing decisions. Drawing on best practices in risk communication and usability, we develop three methods to convey probabilistic DP guarantees to end users: two that communicate odds and one offering concrete examples of DP outputs.

We quantitatively evaluate these explanation methods in a vignette survey study ($n = 963$) via three metrics: objective risk comprehension, subjective privacy understanding of DP guarantees, and self-efficacy. We find that odds-based explanation methods are more effective than (1) output-based methods and (2) state-of-the-art approaches that gloss over information about ε. Further, when offered information about ε, respondents are more willing to share their data than when presented with a state-of-the-art DP explanation; this willingness to share is sensitive to ε values: as privacy protections weaken, respondents are less likely to share data.

1 Introduction

Differential privacy (DP) [20] is a formal definition of privacy that has been integrated into several high-profile data analysis pipelines, including the 2020 U.S. Census data products [1] and internal metric measurement tools at, e.g., Google [23], Apple [4], Microsoft [18], and Uber [81].

As DP is increasingly applied to protecting people's privacy, it is vital that organizations deploying DP effectively communicate the privacy implications of implementation details that govern the *strength* of systems' privacy protections. Without such transparency, organizations risk engaging in "privacy theater," [19, 76, 77] which may result in people falsely believing they are well-protected [14, 80].

While DP offers a precise framework for measuring worst-case privacy loss, research has found that non-experts struggle to form accurate assessments of the real-world privacy protections DP affords [14, 85]. One source of confusion is the probabilistic (i.e., non-absolute) nature of DP's privacy protection. In particular, DP bounds privacy loss as a function of the unitless privacy loss budget parameter ε. Differentially-private algorithms inject a calibrated amount of statistical noise inversely proportional to ε into either the data or analysis outputs (depending on the DP model), meaning higher values of ε correspond to weaker privacy protections.

Explaining probabilistic systems to end users (i.e., people contributing their data) is a challenging task, as observed by prior social science research on health risk communication [43, 75]. Explaining probabilistic privacy risk, such as that created by DP, is a similarly—or perhaps an even more—challenging problem, given that the probabilistic nature of the system arises from the use of a complex, explicitly mathematical process, rather than variation in population-level behaviors. Moreover, the privacy protections offered by differentially-private mechanisms lack context, i.e., they are agnostic to the social context of a dataset or analysis. Privacy scholars, however, have theorized that people understand privacy contextually [61].

Despite the critical importance of ε, many deployed DP systems only describe ε in technical documentation, while information about the privacy protection accessible to the general public glosses over the implications of the chosen ε altogether [14, 19]. This is particularly problematic, as the values of ε used in practice, and thus the real-world privacy protections afforded by DP systems, vary wildly [11, 16].

Prior research on explaining DP has either sidestepped the

[*]The author conducted part of this work while visiting Columbia University.
[†]The author conducted part of this work while visiting the Simons Institute for the Theory of Computing at UC Berkeley.
[‡]The author contributed equally to advising this work.

What is differential privacy?

- Aims to increase individual's privacy within a collective data analysis
- Many high profile uses
  - 2020 U.S. Census data products
  - Internal metric measurement tools at Google, Apple, Microsoft, and Uber

Calibrating Noise to Sensitivity in Private Data Analysis

Cynthia Dwork[1], Frank McSherry[1], Kobbi Nissim[2], and Adam Smith[3]*

[1] Microsoft Research, Silicon Valley. {dwork,mcsherry}@microsoft.com
[2] Ben-Gurion University. kobbi@cs.bgu.ac.il
[3] Weizmann Institute of Science. adam.smith@weizmann.ac.il

Abstract. We continue a line of research initiated in [10, 11] on privacy-preserving statistical databases. Consider a trusted server that holds a database of sensitive information. Given a query function f mapping databases to reals, the so-called *true answer* is the result of applying f to the database. To protect privacy, the true answer is perturbed by the addition of random noise generated according to a carefully chosen distribution, and this response, the true answer plus noise, is returned to the user.

Previous work focused on the case of noisy sums, in which $f = \sum_i g(x_i)$, where x_i denotes the ith row of the database and g maps database rows to $[0, 1]$. We extend the study to general functions f, proving that privacy can be preserved by calibrating the standard deviation of the noise according to the *sensitivity* of the function f. Roughly speaking, this is the amount that any single argument to f can change its output. The new analysis shows that for several particular applications substantially less noise is needed than was previously understood to be the case.

The first step is a very clean characterization of privacy in terms of indistinguishability of transcripts. Additionally, we obtain separation results showing the increased value of interactive sanitization mechanisms over non-interactive.

1 Introduction

We continue a line of research initiated in [10, 11] on privacy in *statistical* databases. A statistic is a quantity computed from a sample. Intuitively, if the database is a representative sample of an underlying population, the goal of a privacy-preserving statistical database is to enable the user to learn properties of the population as a whole while protecting the privacy of the individual contributors.

We assume the database is held by a trusted server. On input a query function f mapping databases to reals, the so-called *true answer* is the result of applying f to the database. To protect privacy, the true answer is perturbed by the addition

* Supported by the Louis L. and Anita M. Perlman Postdoctoral Fellowship.

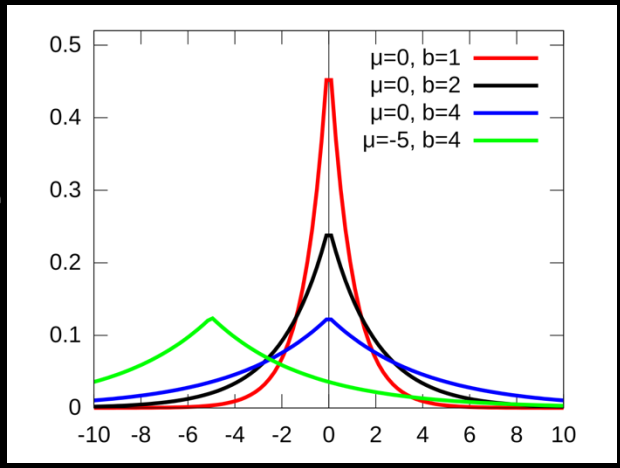
What is differential privacy? Definition

A randomized algorithm $A : \mathrm{D} \rightarrow \mathrm{R}$ is *ε-differentially private* if

for every pair of databases $\mathrm{D}, \mathrm{D}' \in \mathrm{D}$
that differ in at most one entry and
for every subset $S \subseteq \mathrm{R}$,

$$\Pr[A(\mathrm{D}) \in S] \leq e^{\varepsilon} \cdot \Pr[A(\mathrm{D}') \in S]$$

What is differential privacy? Example



$$A\,(\mathrm{D}) = \text{count-ones}(\mathrm{D}) + Lap(0, {}^{1}/_{\varepsilon})$$

Suppose D = [0,0,0,0]

ε = 0.1	ε = 0.5	ε = 2	ε = 4
4.0	0.8	0.2	0.1
-12.2	-2.4	-0.6	-0.3
2.4	0.5	0.1	0.1
23.7	4.7	1.2	0.6
-13.5	-2.7	-0.7	-0.3

More accuracy

Less privacy

Suppose D = [0,0,0,0,**1**]

ε = 0.1	ε = 0.5	ε = 2	ε = 4
7.1	2.2	1.3	1.2
3.4	1.5	1.1	1.1
-14.8	-2.2	0.2	0.6
-23.0	-3.8	-0.2	0.4
0.0	0.8	1.0	1.0

More accuracy

Less privacy

Challenge: Communicating DP protections

- Difficult to explain probabilistic guarantees to most people
  - DP bounds privacy loss as a function of the unitless parameter ε
- Paper offers three explanation methods
  - *Odds-Text*
  - *Odds-Vis*
  - *Sample Reports*
- Methods apply to an individual choosing whether to participate in a DP-protected data analysis, with ε accounted for

Scenario proposition

- Imagine you work on a **team with four other people**. All five of you report to the same manager. The company is requiring each of you to **participate in a survey**. The survey asks the following yes/no question:

  Do you feel adequately supported by your manager?

- You have had negative experiences with your manager and **want to answer no**. However, you don't want your manager to find out you responded no. Your **manager may retaliate** if they believe you responded no. For example, they might give you a negative performance review, assign you extra work, or try to get you fired.
- You must decide how to respond to the survey question.
- You can participate and **respond 'no' truthfully**, or say you would **prefer not to participate**.

Scenario, further context

- Based on lunchtime conversations, it is obvious to your manager that **all your other teammates will respond yes**, indicating they feel adequately supported by your manager.
- On the other hand, your **manager has no idea how you will respond**.
- Your **manager will receive a report** on the results of the survey. This report will **say the total number of people who responded no**.
- Even though the report will **keep your names anonymous**, once your manager gets the report, they can **still use it to guess your response**. For example, imagine the report shows that there was one no response. Your manager will believe it was you because they believe your other teammates all responded yes.

Ends here, when using deterministic control

Baseline [when privacy method used]

- Your company will not report exactly how many employees on your team responded NO.
- Instead, they will generate many potential reports by using a statistical method to modify the total number of NO responses.
- So, each potential report may show a number somewhat lower or higher than the actual number of NO responses. Only ONE report will be randomly sent to your manager.

Alternative [when paper's expl. meth. used]

- However, your company will **use a privacy protection method** to help prevent your manager from correctly guessing anyone's response.
- Your company will **not report exactly how many employees** on your team responded no. Instead, they will **generate many potential reports** by using a **statistical method to modify the total number of no responses**. So, each potential report may show a number somewhat lower or higher than the actual number of no responses.
- Only **ONE report will be randomly sent to your manager**. Note that due to the privacy method, it is possible that

the specific report your manager receives will lead them to believe that you responded no **regardless of what you respond on the survey**.

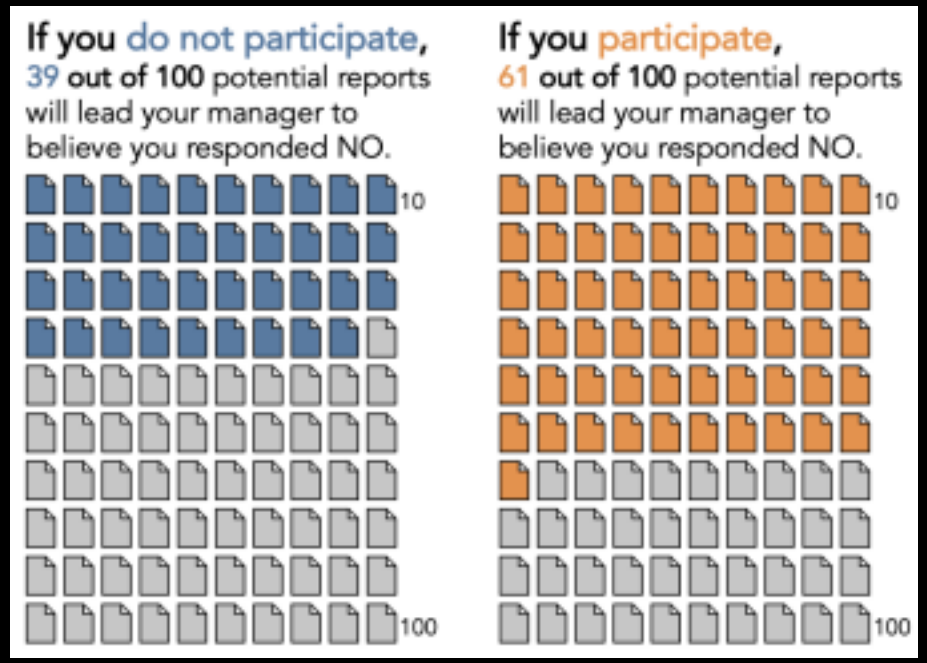
Odds-Text

$\varepsilon = 0.1$	$\varepsilon = 0.5$	$\varepsilon = 2$	$\varepsilon = 4$
$x = 48$ $y = 52$	$x = 39$ $y = 61$	$x = 18$ $y = 82$	$x = 7$ $y = 93$

If you **do not participate**,
39 out of 100 potential reports will lead your manager to believe you responded NO.

If you **participate**,
61 out of 100 potential reports will lead your manager to believe you responded NO.

Odds-Vis


If you do not participate,
39 out of 100 potential reports
will lead your manager to
believe you responded NO.
10
100
If you participate,
61 out of 100 potential reports
will lead your manager to
believe you responded NO.
10
100

Sample reports

If you **do not participate**, below are examples of potential reports your manager might receive. The total number of NO responses may be fractional or negative due to the privacy method. The total number of NO responses is:

Potential Report	0.8
Potential Report	-2.4
Potential Report	0.5
Potential Report	4.7
Potential Report	-2.7

If you **participate**, below are examples of potential reports your manager might receive. The total number of NO responses may be fractional or negative due to the privacy method. The total number of NO responses is:

Potential Report	2.2
Potential Report	1.5
Potential Report	-2.2
Potential Report	-3.8
Potential Report	0.8

Study

Explanation method | Epsilon value | optional/mandatory

- 3x4x2 between-subjects **vignette survey study** (n = 963)
  - Controls: No privacy, Det. explanation
- Participants recruited from Prolific

Demographic Attribute	Count
Gender	
Man	478
Non-binary	24
Woman	458
Prefer to self-describe	2
Prefer not to answer	6
Age	
18-29	346
30-39	286
40-49	142
50+	173
Prefer not to answer	16
Race/Ethnicity	
Hispanic or Latino	104
Black or African American	114
White	718
American Indian or Alaska Native	19
Asian, Native Hawaiian, or Pacific Islander	84
Mixed, Multiracial, or Biracial	5
Unique free-text responses	3
Prefer not to answer	9

Demographic Attribute	Count
Education	
High school or less	140
Some college	319
Bachelor's or above	501
Prefer not to answer	3
Education/work in computer science/IT	
Yes	192
No	735
Prefer not to answer	36
Income	
Less than $10,000	60
$10,000 to under $20,000	64
$20,000 to under $30,000	89
$30,000 to under $40,000	95
$40,000 to under $50,000	79
$50,000 to under $65,000	131
$65,000 to under $80,000	103
$80,000 to under $100,000	93
$100,000 to under $125,000	77
$125,000 to under $150,000	49
$150,000 to under $200,000	38
$200,000 or more	42
Prefer not to answer	43

Willingness to Share

Based on the scenario and description of privacy protection, would you respond no (i.e., respond truthfully) to the survey question? (Yes/No/I prefer not to answer this question.)

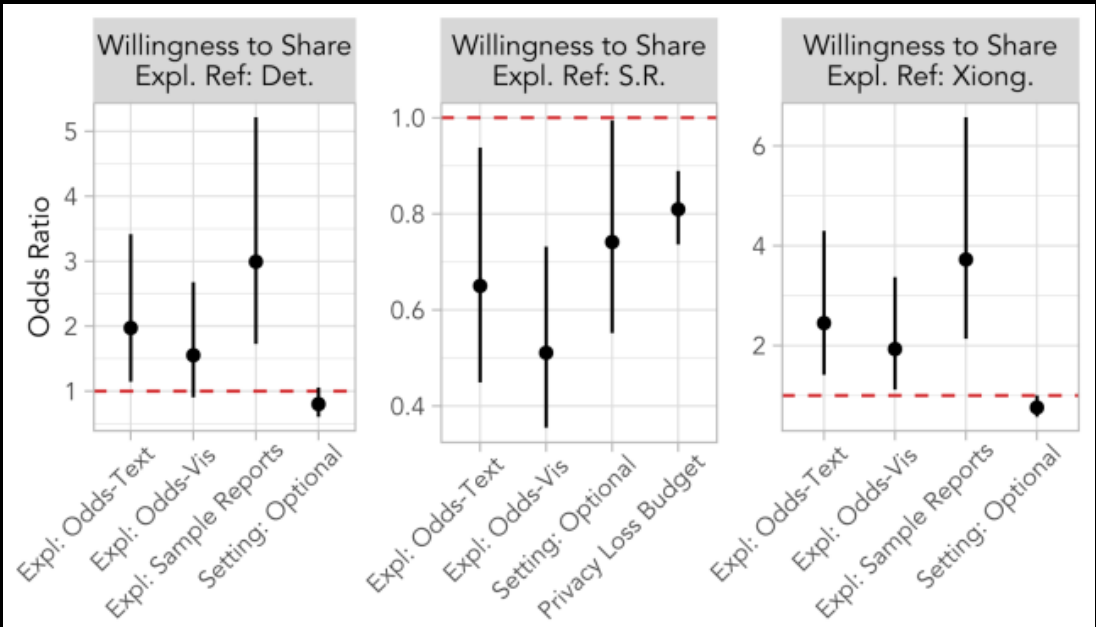


Figure 5: Logistic regression models examining relationships between willingness to share data and our IVs. Det. = Deterministic; S.R. = SAMPLE REPORTS; Xiong. = Xiong et al. See Figure 2 for interpretation.

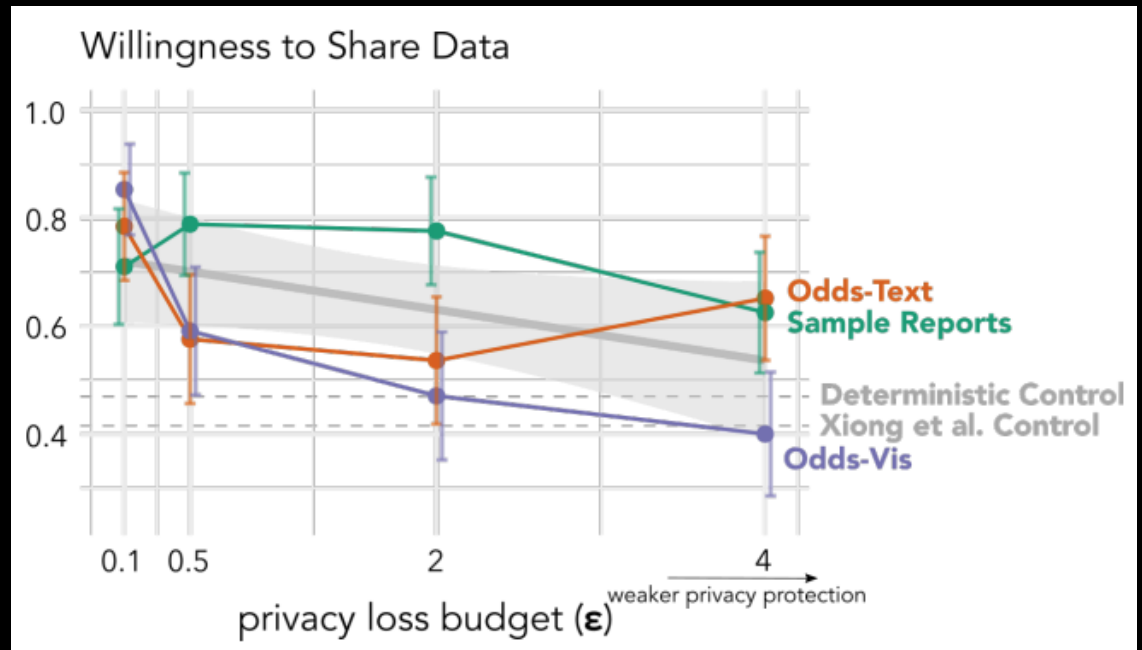
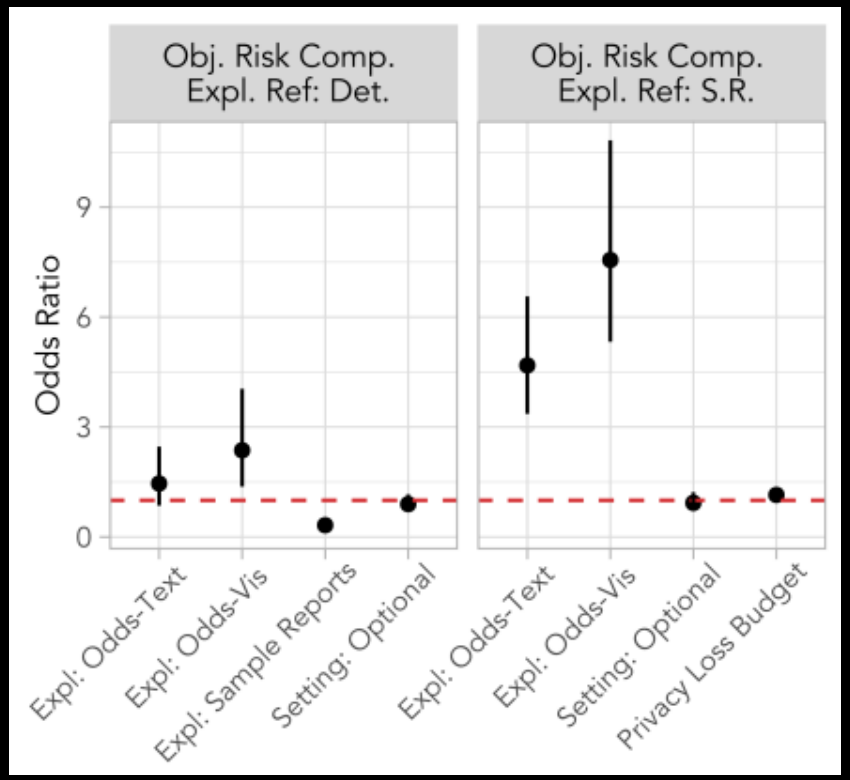
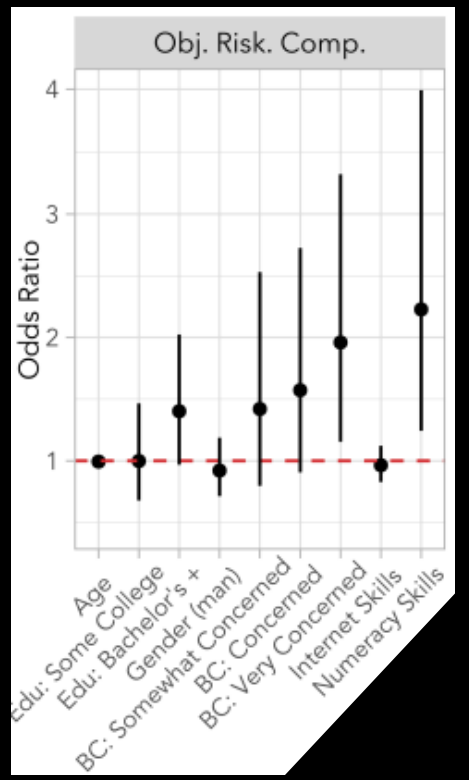


Figure 6: Proportion of respondents willing to share data across explanation methods and ε, shown with 95% binomial CIs. We plot a regression line (solid gray) between proportion of data sharing across our methods and ε.

Objective Risk Comprehension

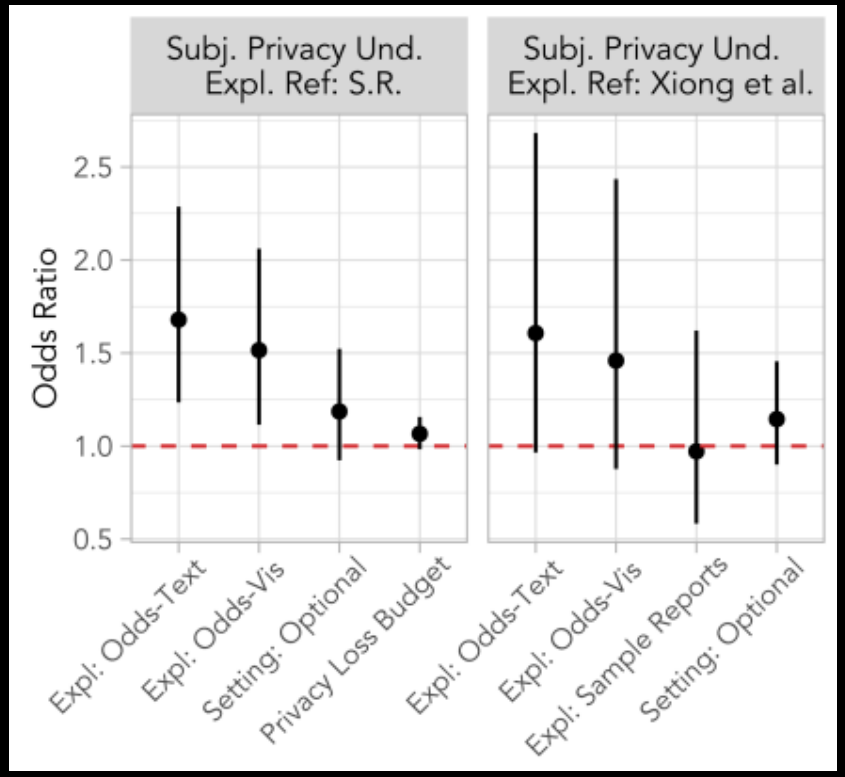
My manager is more **[than twice as]** likely to believe I responded no if I respond no (i.e., respond truthfully) to the survey question than if I respond yes (i.e., respond untruthfully) to the survey question. (True/False/I don't know/I prefer not to answer this question.)

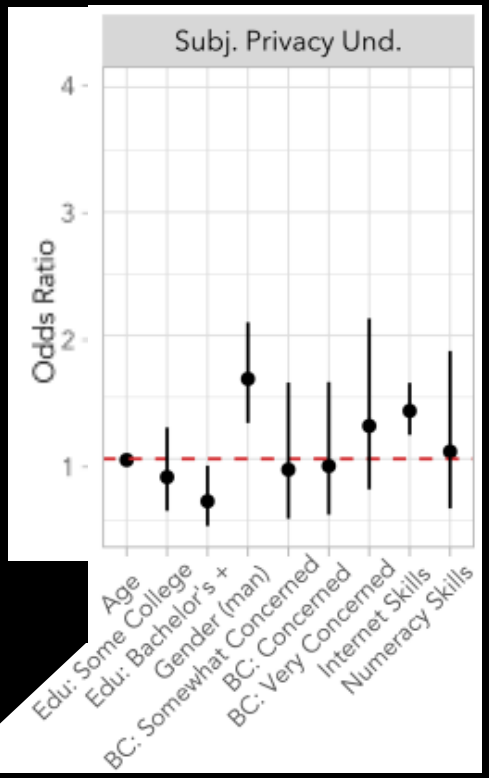




Subjective Privacy Understanding

Based on the scenario and description of the privacy protection, how confident are you that you understand the privacy protection applied to the survey results? (Not at all confident/ Somewhat confident/Confident/Very confident/I prefer not to answer this question.)

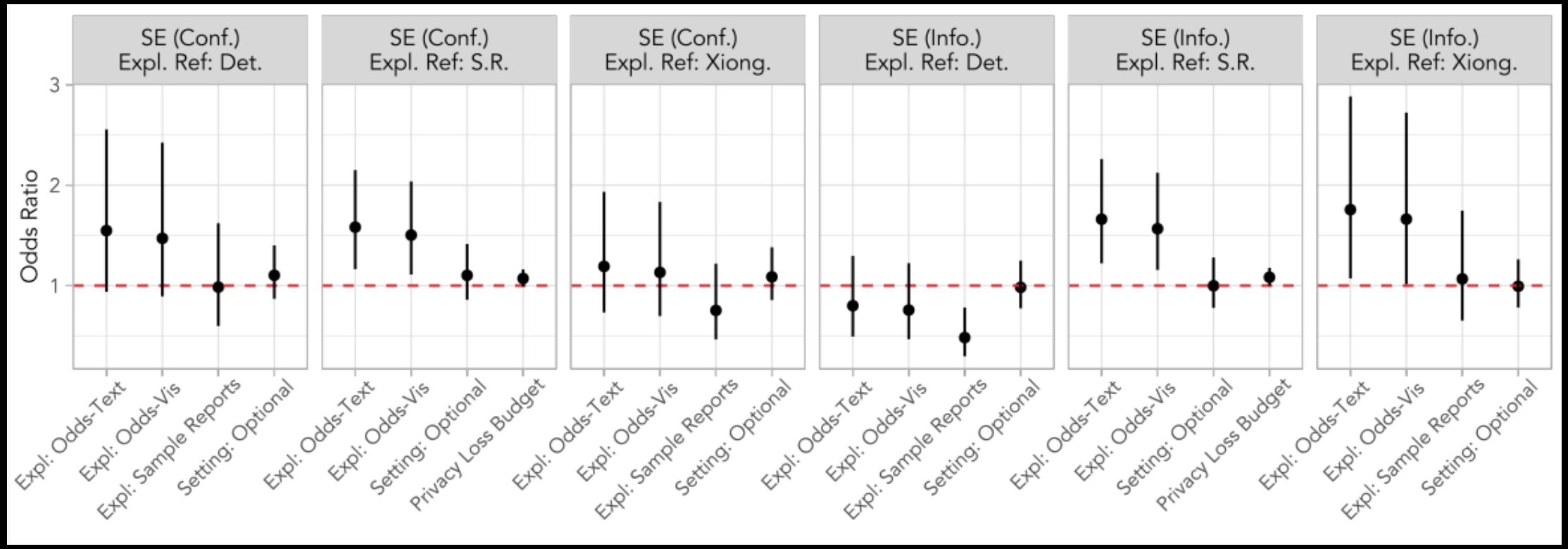




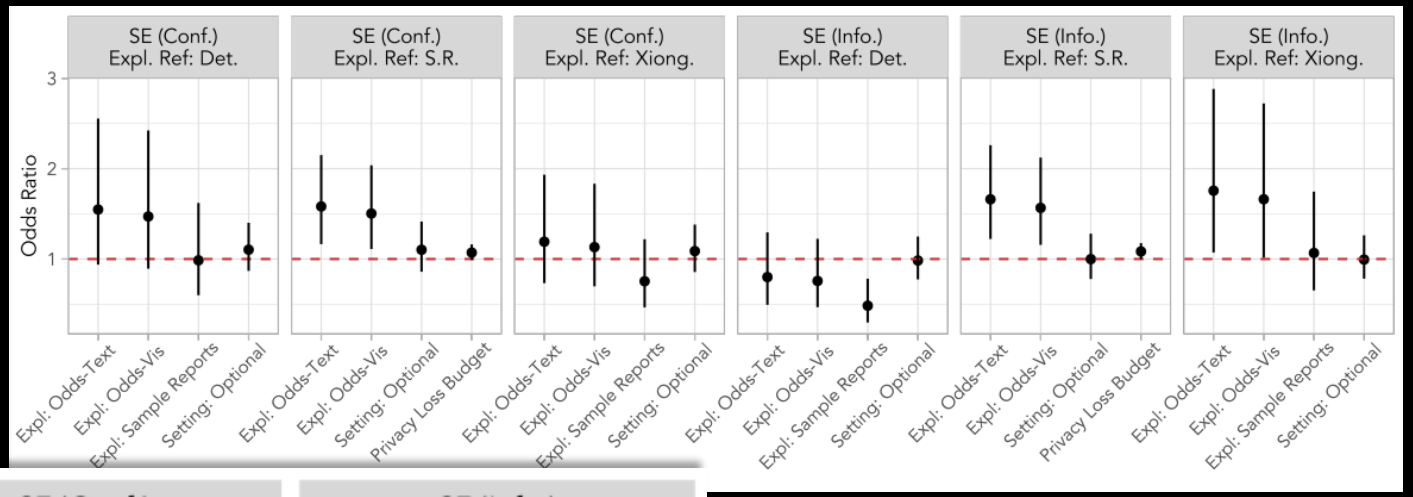
Self Efficacy

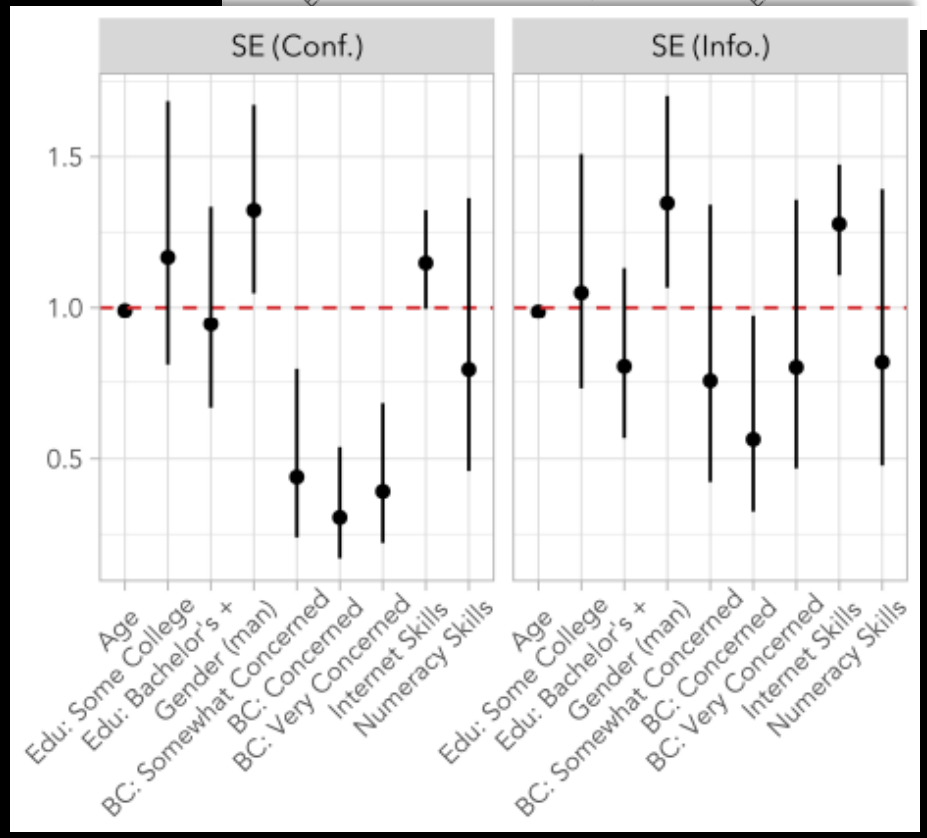
- Based on the scenario and description of the privacy protection, how confident are you that you have enough information to decide whether to respond no (i.e., respond truthfully) to the survey question? (Not at all confident/Somewhat confident/Confident/Very confident/I prefer not to answer this question.)
- What further information, if any, would you like to have to help you decide whether you would respond no (i.e., respond truthfully) to the survey question? (open-text response)
- Based on the scenario and the description of the privacy protection, how confident are you in deciding whether to respond no (i.e., respond truthfully) to the survey question? (Not at all confident/Somewhat confident/Confident/Very confident/I prefer not to answer this question.)

Self Efficacy



Self Efficacy





Summary of results

Compared to Sample Reports, Odds-Based Text and Odds-Based Visual improved:

- Objective risk comprehension (O.R. = 4.7; 7.6)
- Subjective privacy understanding (O.R. = 1.7; 1.5)
- Self-efficacy (enough info) (O.R. = 1.7; 1.6)

Takeaways

- Odds-based methods are promising for explaining ε to end users
- Explanations should include ε information, since it supports self-efficacy
- People's willingness to share data is sensitive to changes in ε
- Explanation methods can support auditing & public deliberation over differential privacy deployments

Discussion

- What do you think about the generality of this approach?
  - When could it broadly apply to other scenarios?
- This method targets users where privacy may be the key concern. What about utility, which may be a concern of adopters?
- Can you think of other explanatory methods that might be better?
  - E.g., show analysis with/without DP entirely, or with varying epsilon