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*

https://www.usenix.org/conference/usenixsecurity23/presentation/nanayakkara

This paper is included in the Proceedings of the 32nd USENIX Security Symposium.

August 9–11, 2023 • Anaheim, CA, USA

978-1-939133-37-3

Open access to the Proceedings of the 32nd USENIX Security Symposium is sponsored by USENIX.

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.

complicated task of explaining ε to end users [14, 24, 45, 85] or focused on addressing the impact of ε for very specific deployments of DP [9, 15, 76], e.g., explaining the randomized response mechanism [9]. As a result, we currently lack explanations of DP that include ε and can be used with all differentially-private mechanisms. Without access to such explanations, organizations deploying DP systems must either write new deployment-specific descriptions of DP that are unlikely to be scientifically evaluated, or risk leaving their users unable to make well-informed decisions.

We fill this gap by developing and evaluating explanation methods for DP that *directly* address the implications of ε. The result of our work is a *framework* for conveying ε to end users that is highly *portable*, in that it can be adapted to many deployment settings. Our explanation methods avoid relying on describing mathematical details of the mechanism and focus on the concrete ramifications of the choices a user might potentially face, e.g., if they are to share their data. This is a conceptual departure from prior work on DP communication, which focuses on the implications of using DP instead of running the equivalent, non-private system [9, 76] or learning attributes using information that is available even if an individual chooses not to share their data [24, 84, 85]. We also evaluate our methods by testing an instantiation of our explanation methods in a scenario with binary count queries. In Section 3.3, we offer direction for how our methods can be ported into other scenarios.

New Explanations for Differentially-Private Systems. We draw on best practices in risk communication and usability [22, 27, 28, 31, 39, 44, 49, 75] to develop explanation methods designed to allow people to quickly and easily reason about probabilistic privacy guarantees under DP.

Specifically, we design three explanation methods for ε. Our first explanation method (ODDS-TEXT) leverages best practice methodology for risk communication to give a textual description of the odds that an information leak might occur if a person decides to share their data; this is a stylized version of the "textbook" understanding of DP [21] which compares the outputs of differentially-private mechanisms applied to neighboring databases, each corresponding to a situation where a person does or does not share their data. Our second explanation method (ODDS-VIS) conveys the same information using a frequency-framed visualization approach which may help people with low numeracy skills more accurately make probability judgments [27]. Our third explanation method (SAMPLE REPORTS) draws on prior work in usable security and privacy (S&P) on improving user comprehension of S&P technologies [31] to provide people with several potential *outputs* of the DP mechanism in an effort to make the implications of their data-sharing choice concrete.

We evaluate the efficacy of these explanation methods using three metrics: (1) objective risk comprehension, (2) subjective privacy understanding, and (3) self-efficacy (personal belief in decision-making capacity [55]). We additionally study the relationship between our explanation methods and (4) willingness to share data.

In summary, we are interested in answering the following research questions (RQs):

RQ1: Which practices in risk communication work best for communicating the probabilistic privacy guarantees offered by DP? Specifically, which practices are effective at increasing people's

(a) *objective risk comprehension* of DP guarantees,

(b) *subjective privacy understanding* of DP guarantees,

(c) *self-efficacy* around making data-sharing decisions?

RQ2: How do the explanation methods we develop influence people's data-sharing decisions?

We answer our RQs via a vignette survey study ($n = 963$) in which we embed our DP explanations into concrete information-sharing scenarios and evaluate them using the aforementioned criteria against each other and multiple control explanations.

Summary of Findings. We find that people have better objective risk comprehension (RQ1a) of DP protections when presented with odds-based explanations (ODDS-TEXT or ODDS-VIS) than with SAMPLE REPORTS, which presents sample outputs from the privacy mechanism.

Despite our findings about objective risk comprehension, none of our explanations meaningfully improve people's *subjective privacy understanding* (RQ1b), i.e., people *feeling* as though they understand the privacy protection.

Further, to assess self-efficacy (RQ1c), we ask respondents if they feel as though they (1) have enough information to make a data-sharing decision and (2) are confident making said decision. The odds-based explanations we test increase people's sense that they have enough information to make a data-sharing decision compared to a state-of-the-art [85] explanation that does not feature information about ε, suggesting that there is merit to not glossing over the probabilistic nature of DP privacy protection. However, our SAMPLE REPORTS method for explaining ε had the opposite effect: it reduced feelings of having enough information to make data-sharing decisions as compared to a very simple and clear description of the scenario without any probabilistic privacy protections, suggesting that this explanation actively confused respondents.

Interestingly, we do not find evidence that any of our explanations meaningfully impact people's confidence in making data-sharing decisions compared to a state-of-the-art explanation. Instead, we find that their overall concern about the ramifications of their data-sharing significantly relates to their confidence.

Last, we study the influence of our explanations on people's willingness to share data (RQ2). Our findings

indicate that people are much more likely to share their data when presented with an explanation of DP that offers information about ε (compared to one that does not), regardless of which explanation method is used. Finally, when offered information about ε, respondents' data-sharing decisions are sensitive to changes in ε, empirically validating theoretical proposals that willingness-to-share depends on the strength of the privacy mechanism [21].

2 Background

Differential Privacy. DP [20] is a mathematical privacy definition which ensures, at a high level, that the results of an analysis should be similar regardless of the inclusion of any given individual's information in that analysis. Differentially-private mechanisms add carefully calibrated random noise at some point in the data analysis process in order to obscure details at the individual level while maintaining accuracy at the aggregate level. If too much noise is added, it will overwhelm the signal in the data, and the analysis results will be useless. If too little noise is added, the privacy protection offered to individuals may not be meaningful. The privacy loss budget parameter, ε, controls this trade-off; a *smaller* privacy loss budget provides a *stronger* privacy guarantee. We state the formal definition of DP below:

Definition: [Dwork et al. [20]] A randomized algorithm $\mathcal{A} : \mathcal{D} \rightarrow \mathcal{R}$ is ε-differentially private if for every pair of databases $D, D' \in \mathcal{D}$ that differ in at most one entry and for every subset $S \subseteq \mathcal{R}$, $Pr[\mathcal{A}(D) \in S] \leq e^{\varepsilon} \cdot Pr[\mathcal{A}(D') \in S]$.

Implementations of DP typically adopt either the *local* or *central* model.[1] In the central model, a trusted curator stores the collected data and adds noise as necessary when releasing statistics, charts, or other aggregate insights about the data. In the local model, noise is added to each individual's data before it is sent to the curator. While prior work has already explored the task of explaining the privacy loss budget to end users in the local model [9, 15, 76], our study addresses the more challenging task of explaining the privacy loss budget in the context of the central model.

One of the simplest mechanisms for achieving DP is the *Laplace Mechanism* [20, 21]. In this mechanism, a data collector releases results of a simple counting query by adding noise sampled from a Laplace distribution centered at zero with scale parameter $\frac{1}{\varepsilon}$. The resulting outcome is distributed according to $Lap(\mu, b = \frac{1}{\varepsilon})$, where μ is the true value of the counting query before noise is added.[2]

Communicating DP to End Users. Prior work has begun to study the task of communicating DP to the general population [9, 14, 15, 24, 45, 76, 85, 86]. For the local model, Smart et al. [76] and Bullek et al. [9] have explained the strength of privacy protections in terms of the probability of bits being "flipped." However, this style of explanation does not work for the central model since noise is added at the aggregate level instead of at the individual level. Metaphors provide a different approach for explaining DP. For example, Karegar et al. [45] use the metaphor of blurring images as an analogy for adding statistical noise to collected data. Such metaphors can help people understand that there exists a privacy-accuracy trade-off, such that increasing the injected noise strengthens the privacy guarantee but harms accuracy. However, these metaphors do not explain the implications of particular settings of ε, a challenge we address in our work.

Xiong et al. [86] studied how to communicate the implications of the privacy loss budget for both privacy and accuracy of location data. The authors develop illustrations for the local, central, and shuffle models (see [5]) that show how the amount of added noise affects privacy and accuracy. They use heatmaps to compare the accuracy of collected location data before and after noise is added. They inject positive rather than unbiased noise to avoid the problem of negative counts that may confuse people who are unfamiliar with DP. In our study, we instead choose to embrace the sometimes unintuitive results produced by adding unbiased noise and investigate end users' perceptions of them.

Franzen et al. [24] borrowed from the literature on quantitative risk communication to explain the protections offered by DP. Although quantitative risk communication formats can aid comprehension, individuals with low numeracy skills struggle to understand these explanations. We include a measure of numeracy skill in our survey to determine whether our explanations similarly disadvantage individuals with low numeracy skills. An important difference between Franzen et al. [24] and our work is the comparison probability (to the probability of a negative outcome given that an individual shares their data) we each present: Franzen et al. present the probability of a negative outcome given no data collection takes place, while we show the probability of a negative outcome given the individual does not share their data, but all other factors remain the same. Because people rarely have the power to immediately stop an entire data collection process, we suggest that it is important to explore this separate decision context in an effort to closely align with real-world decisions people make.

Supporting Decision-Making Around ε. Research on communicating implications of ε has tended to focus on data curators or analysts who are setting privacy loss budgets. For example, there have been several interfaces for DP (DPComp [37], Overlook [82], PSI (Ψ) [25], Bittner et al. [6], DPP [41], ViP [59]) that portray accuracy and/or risk implications of ε to support more informed privacy loss budget setting. Although these tools are aimed at data analysts and curators, they are also relevant to communicating DP to non-experts because these analysts/curators typically

[1] Although, other models also exist [17].

[2] Note that the standard presentation of the Laplace Mechanism focuses on the distribution of the *noise* rather than the *output*. Our presentation is mathematically equivalent and is consistent with our use of the Laplace Mechanism in Section 3.2.2.

are not assumed to have DP background or expertise. As such, these tools must express relevant DP concepts well enough to support decision-making about privacy loss budgets. At the same time, we note that end users are usually not tasked with setting or allocating privacy loss budget, but rather must make individual data-sharing decisions.

Hsu et al. [40] propose an economic framework for people considering sharing their data, e.g., as part of a scientific study, to weigh monetary costs of sharing versus not sharing their data. Wood et al. [84], in explanations of ε in a primer on DP, similarly frame data-sharing decisions in terms of worst-case monetary losses (e.g., in terms of increases to insurance premiums) people could incur if they share their data under DP. Heffetz and Ligett [38] describe ε to economists in the context of calculating a mean salary value, focusing primarily on accuracy outcomes. Finally, Lee and Clifton [48] model disclosure risk by considering a potential attacker who conducts a Bayesian update on their beliefs of whose information is included in an analysis based on seeing a release from a differentially-private mechanism. Our odds-based explanation methods similarly model an attacker's updated beliefs given a DP output.

Probabilistic Risk Communication. Many studies have identified best practices for effective probabilistic risk communication, especially in the medical context [22,49]. Prior work has found benefits of framing probabilities as frequencies [28,39]. One challenge in probabilistic risk communication is that people often misinterpret probabilities expressed as ratios—for example, people may mistakenly interpret an event with a probability of 1 out of 10 as less likely than an event with a probability of 10 out of 100, simply because the former ratio is expressed with smaller numbers [3,88]. Thus, it is best practice to use a consistent denominator when presenting ratios for comparison [49]. Frequency-framed visualizations, such as icon arrays, can also complement numeric risk communication. Compared to purely numeric presentations of risk, icon arrays may improve understanding particularly among people with low-numeracy skills [27]. We incorporate these findings into our explanations of DP by framing probabilities as frequencies and employing icon arrays.

3 Explanation Methods for ε

We introduce three methods to explain ε to end users. These methods work for two common data-sharing settings: one where providing data is *optional*, so people must decide whether to participate (or opt-out), and one where providing data is *mandatory*, so people must decide whether to respond truthfully (or respond untruthfully).

Drawing on best practices from the literature in health risk communication [22,39,49], we develop two explanation methods (ODDS-TEXT and ODDS-VIS) that focus on explaining the *odds* of a negative event occurring by contextualizing privacy guarantees in terms of outcomes that could occur based on decisions users can make. We develop a third method (SAMPLE REPORTS) that provides concrete examples of the protected data, based on findings that indicate concrete examples help people comprehend S&P topics [31]. Examples of each explanation method instantiated in our survey scenario (see Section 3.2.1) are in Figure 1.

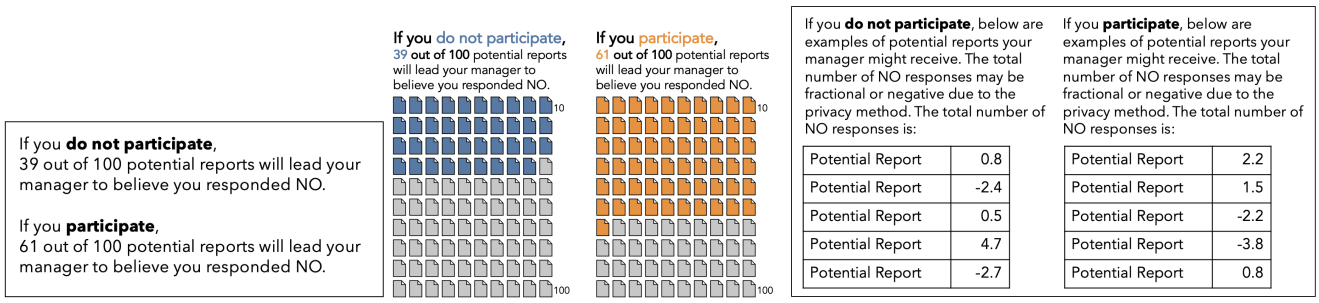
3.1 Description Approaches

ODDS-TEXT. In line with research [75] finding that people reason more effectively about the odds of a risk when framed as frequencies versus percentages, including in the context of privacy decisions [44] and DP specifically [24], we present all probabilities in the ODDS-TEXT explanation as frequencies (Figure 1a). Specifically, probabilities are shown in the form of "z out of 100 potential DP outputs" where z is a natural number and "DP outputs" can be customized to specific scenarios (e.g., if the DP output is published in a report, the explanation may instead say "potential reports"). Specifically, this explanation method comprises of two probabilities corresponding to the chances that an adversary $\mathcal{A}$ believes, based on prior knowledge combined with a DP output, that a data sharer provided information corresponding to the actual value d_{true} if the data sharer participates vs. does not participate OR responds truthfully vs. untruthfully. Data sharers have immediate agency over these actions, and hence showing probabilities aligning with these actions is directly relevant to their data-sharing decision-making process. Specifically, ODDS-TEXT explanations take the following form:

> *If you [do not participate/respond d_{false}], x out of* 100 *potential [DP outputs] will lead $\mathcal{A}$ to believe you responded d_{true}.*
> *If you [participate/respond d_{true}], y out of* 100 *potential [DP outputs] will lead $\mathcal{A}$ to believe you responded d_{true}.*

ODDS-VIS. Research has found that icon arrays—a frequency-framed visualization approach (Figure 1b)—sometimes help people with low-numeracy skills more accurately estimate risk reduction [27]. Thus, we add an icon array to the text-based description of the risk ratios in the ODDS-TEXT condition. We use icon arrays to help people concretely visualize that there are many potential DP outputs, but $\mathcal{A}$ will only receive one. The shape of the icon can be adapted to suit the scenario. We fill in icons top to bottom, as this arrangement has been shown to be optimal for supporting accurate probability judgments [87]. Icon colors follow from the Tableau 10 color palette [78], which was designed keeping in mind common forms of color-vision deficiencies.

SAMPLE REPORTS. Drawing on prior work in S&P showing that concrete examples improve user comprehension of privacy enhancing technologies and secure behavior [31], the SAMPLE REPORTS method shows five potential DP outputs



(a) ODDS-TEXT (b) ODDS-VIS (c) SAMPLE REPORTS

Figure 1: Examples of our explanation methods applied to the manager's performance review scenario under the optional setting and $\varepsilon = 0.5$.

from an analysis if the data sharer were to participate/respond truthfully (i.e., share d_{true}) and five potential DP outputs if they were to not participate/respond untruthfully (i.e., provide no data or share d_{false}) (Figure 1c). Presenting both sets of potential outputs allows the data sharer to make comparisons between how these values would differ based on their decision, and the extent to which their survey response sways the DP output.

3.2 Contextualizing Explanations

In order to leverage the description approaches outlined above, we require the following: (1) a concrete scenario in which the data sharer should think about the explanation and (2) the values for parameters described above (e.g., probabilities for both odds-based methods). In this section, we introduce a hypothetical data-sharing scenario and outline how to appropriately compute values for explanations for the given scenario. We use this hypothetical scenario in surveys with respondents (more details in Section 4). As detailed below and in Section 6, our methods and calculation techniques for each method can be extended to numerous other DP applications.

3.2.1 Workplace Evaluation Scenario

Imagine that an employee (the data sharer) is asked to share an evaluation of their manager ($\mathcal{A}$), but fears their manager will retaliate against them if they believe the employee reviewed them negatively. Everyone reporting to the manager is asked to share an evaluation, which specifically asks the following YES/NO question: *Do you feel adequately supported by your manager?* This scenario describes data collection and analysis occurring under the central DP model: the company will collect un-noised answers and create a report with the total number of NO responses calculated using the Laplace Mechanism with a particular ε. The report will *not* include names of team members and how they responded, however. For our demonstrative scenario, we focus on the Laplace Mechanism because it is both canonical and commonly-used in real-world deployments [2, 65, 72]. Note that our explanation methods can be used to convey privacy strength of various other DP mechanisms (e.g., Gaussian) simply by plugging in other noise distributions in the analysis that follows. We focus on binary count queries because they are an often-used SQL query and have been studied extensively in the DP literature [42, 58].

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

Table 1: Values for odds-based explanation methods.

Since DP guarantees are often framed through a worst-case lens, we further design the scenario to represent a worst-case situation where the manager has prior knowledge on how all the other teammates will respond. For concreteness, we suppose that they will all respond YES, while the employee wants to respond NO (d_{true}), thus putting the employee at risk of being singled out in summary statistics. The other teammates' responses can be modified to suit other contexts, resulting in modifications to μ in the following analysis. We also imagine that the manager's prior belief that the employee will respond NO is 50%. See Appendix A for how the analysis specifically for ODDS-TEXT and ODDS-VIS can be updated to accommodate non-uniform priors.

3.2.2 Calculating Values for Explaining ε

We describe how to calculate values for explanations in the workplace data-sharing context described above. Values computed for our explanation methods are the same regardless of whether the employee is in the mandatory or optional data-sharing setting.[3]

ODDS-TEXT and ODDS-VIS. We calculate values for the ODDS-TEXT and ODDS-VIS explanations by modeling the manager's guessing process. We assume that the manager correctly believes all other teammates will respond YES, consistent with the scenario details described in Section 3.2.1. Thus, the true count of NO responses is either 0 or 1, depend-

[3]These values are the same in our case because we consider count queries, and hence the sensitivities do not change.

ing on whether the employee participates/responds truthfully. Values consistent with this scenario for multiple privacy loss budgets are shown in Table 1.

We model the manager's inference process as estimating the employee's most likely survey response, given the observed DP output. Using distributional knowledge of the DP mechanism, the manager compares posterior probabilities of the employee's response given the DP output to determine a maximum likelihood estimate (MLE). Let the output of the DP mechanism be r. The manager compares $\Pr[E = r \mid E \sim Lap(\mu = 1, b = \frac{1}{\varepsilon})]$ with $\Pr[E = r \mid E \sim Lap(\mu = 0, b = \frac{1}{\varepsilon})]$, respectively corresponding to when the employee: reports truthfully/participates and lies/do not participate, and then guesses the action corresponding to a higher posterior probability. We compute a threshold value $r_{\text{threshold}}$ where $\Pr[E = r_{\text{threshold}} \mid E \sim Lap(\mu = 1, b = \frac{1}{\varepsilon})] = \Pr[E = r_{\text{threshold}} \mid E \sim Lap(\mu = 0, b = \frac{1}{\varepsilon})]$ and report $\Pr[r < r_{\text{threshold}}]$ (i.e., $\frac{x}{100}$ in Table 1) and $\Pr[r > r_{\text{threshold}}]$ (i.e., $\frac{y}{100}$ in Table 1).

SAMPLE REPORTS. In our scenario, the manager will receive a report with the total number of NO responses. To obtain potential outputs for SAMPLE REPORTS under a given ε, we make five random draws from each of $Lap(\mu = 1, b = \frac{1}{\varepsilon})$, corresponding to potential DP outputs when d_{true} is shared, and $Lap(\mu = 0, b = \frac{1}{\varepsilon})$, corresponding to potential DP outputs when there is no participation/d_{false} is shared. We do not post-process sampled values, meaning that they can be fractional or negative. As such, we include a statement preceding these values explaining: *"The total number of* NO *responses may be fractional or negative due to the privacy method."*

3.3 Realizing Portability

Our methods for explaining ε can be used across a range of scenarios. For example, they can be applied to the Gaussian Mechanism and therefore (ε, δ)-DP—a relaxation of ε-DP [21]—using the same probability calculation as for the Laplace Mechanism, except the Laplace CDF is replaced with the Gaussian CDF. Furthermore, our framework is easily adapted to scenarios where the the adversary has a non-uniform prior over beliefs of the data sharer's action. We provide details in Appendix A. In Section 6 we offer further directions to extend our methods to other queries, such as mean queries.

Our odds-based explanation methods can also be extended to other common models of DP, like local DP [46]. When using local randomizers, one can apply, for instance, the Laplace Mechanism on a database of size one and use the same explanation methods. Note that most local DP deployments in practice are on binary/categorical data, meaning that our specific calculations for binary count queries in the central model map nicely. Note, however, that local DP algorithms for binary cases have an even simpler "explanation" for ε, which is the probability of a coin flip (as explored by Bullek et al. [9]). However, that probability does not provide information about the probability of a bad outcome; rather, it explains the randomness of the mechanism.

4 Evaluation

We evaluate our explanation methods by conducting online vignette surveys ($n = 963$), which are designed to mimic real-world decision-making behavior [30], where we present respondents with the survey scenario described in Section 3.2.1 and an explanation of ε created using one of our three methods, as detailed in Section 3.1. Columbia University's IRB approved this research.

4.1 Survey Scenario

Survey respondents are told to imagine themselves as the data sharer (employee) described in Section 3.2.1, either in the optional or mandatory setting. Respondents who see a scenario in the optional setting are told that while participation in their company's survey is optional, participating means they will necessarily respond truthfully (they will respond NO), i.e., if they participate they cannot lie. On the other hand, respondents who see the mandatory setting are explicitly told they can either respond truthfully (NO) or untruthfully (YES). Across settings, scenario details are the same except for slight differences in wording expressing whether the company's survey is mandatory or optional.

Following best practices for vignette studies [71], we designed this scenario to mimic common performance evaluations conducted across workplace and educational settings, thus increasing the chances that respondents would find the scenario believable and relatable. To ensure that all respondents read our explanations of privacy guarantees with similar assumptions about why they would want their data protected in the hypothetical scenario, we clearly define the negative consequences that could transpire if the manager were to correctly guess a negative survey response (*"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."*). We then ask the respondent to imagine that the team's performance reviews will be protected using a "privacy protection method" and do not include the term "differential privacy" anywhere in the survey to prevent respondents from searching for external materials.

The impact of the privacy-protection mechanism is then explained using a method described in Section 3, with one of four privacy loss budgets ($\varepsilon \in \{0.1, 0.5, 2, 4\}$), which represent a range of privacy protection strengths. There is no standard for setting ε [19], so we chose these values to represent what is often recommended in the academic DP literature (small ε values, like 0.1) and larger values that are more consistent with real-world deployments [16]. Some real-world applications of central DP use privacy loss budgets much

larger than our largest value (e.g., the U.S. Census Bureau set a total privacy budget of 19.61 for the 2020 Census redistricting data [11]), but these budgets refer to ε accumulated over many queries, whereas our scenario includes just one.

All respondents who receive SAMPLE REPORTS under the same ε are presented with the same set of random draws. That is, to maintain consistency in what respondents see, we make these draws in advance and do not dynamically make new draws for each respondent. All generated values shown in SAMPLE REPORTS explanations are available on OSF.[4] Values shown in both odds-based explanations are in Table 1.

We tested our questionnaire via cognitive interviews ($n = 12$), following best practices [69] and using a think-aloud protocol [83], with potential study respondents to further refine the scenario for clarity and believability. Based on these interviews, we iterated on the introduction to our explanations and further specified potential negative consequences of information disclosure. We additionally tested our questionnaire via expert reviews ($n > 10$) by experts in DP, survey methodology, and visualization.

4.2 Experimental Design

We use a $3 \times 4 \times 2$ between-subjects study design where each respondent sees one explanation, computed using a particular explanation method (ODDS-TEXT, ODDS-VIS, SAMPLE REPORTS) and privacy loss budget ($\varepsilon \in \{.1, .5, 2, 4\}$), in a given scenario type (optional, mandatory). We also have two control explanations to which we compare our experimental explanations: one where there are no privacy protections (*deterministic control*) and another that includes a high-level explanation of DP that does not mention ε, which is adapted from Xiong et al. [85] (*Xiong et al. control*).

Deterministic Control: The worst-case situation for a hypothetical employee in our vignette is that no privacy protection is applied. In this case, the risk of the negative consequence in the scenario is deterministic (i.e., respondents can expect deterministic outcomes in their manager's beliefs): If the respondent participates/answers NO, their manager will believe they responded NO with probability 1 and if they do the opposite, their manager will believe they responded NO with probability 0. Not only is this "explanation" a worst case, but prior work [3, 7, 75, 88] on risk communication also suggests that it will be the simplest for respondents to understand (objective risk comprehension) and may give them the greatest self-efficacy because of its determinism, in contrast to our probabilistic explanations. Hence, we use this deterministic setting as a control: to do so, we include the same scenario text as in the experimental conditions but omit the stylized description of DP and explanation of privacy guarantees.

[4] OSF link: https://osf.io/w59fv/?view_only=c42a3d68bf9d4f35abe488aab831e775
For reproducibility, we seeded the mechanism with the date. Our code for making draws is also on OSF.

Xiong et al. Control: We also compare our experimental explanations to the current state-of-the-art explanation of DP from Xiong et al. [85]. Because this explanation does not offer information about ε, it helps us assess the impact of adding information about ε on people's understanding of DP protections, self-efficacy in data-sharing decisions, and willingness to share data. To present this control, we include the same scenario text as in the experimental conditions but replace our stylized DP description and privacy explanation with a DP description adapted from Xiong et al. [85] (precise wording on OSF[4]). While Xiong et al. propose several descriptions, we adapt their "DP without names" description since it aligns best with our scenario, and because their evaluation indicates that people found it easy-to-understand and that it supported comprehension on a relevant evaluation question about third-party viewers of the data.

4.2.1 Evaluation Metrics and Willingness to Share Data

Respondents answered questions to evaluate our explanations on three metrics: (1) objective risk comprehension, (2) subjective privacy understanding, and (3) self-efficacy. We also study the relationships between our explanations and (4) respondents' willingness to share data.

(1) Objective Risk Comprehension: We included two TRUE/FALSE questions to evaluate whether the explanations help people understand the risk inherent in the scenario: whether or not their manager will think they responded NO. In addition to options for TRUE and FALSE, we also provide an "I don't know" option [69] to minimize random guesses that are correct by chance. Prior work on communicating DP to people [24, 76, 85] has similarly asked objective-risk-comprehension questions.

The first question, in the mandatory setting, reads:

> *My manager is more likely to think I responded* NO *(i.e., respond truthfully) if I respond* NO *on the survey than if I respond* YES *(i.e., respond untruthfully).*

The version shown in the optional setting is nearly identical but asks about the manager's beliefs if the person were to participate/not participate.

Respondents in all experimental conditions and the deterministic control answer this question. Respondents in the Xiong et al. control do not answer this question, as the explanation does not specify ε; thus ground-truth answers do not exist.[5] For our experimental conditions, the correct answer under all privacy loss budgets we test is TRUE. For the deterministic control, the correct answer is TRUE.

[5] Because the Xiong et al. control indicates that DP is used, it can be argued that there is a correct answer (TRUE) to the first objective-risk-comprehension question. However, as ε tends toward zero, the difference in probabilities goes to zero. Hence for large ε the correct answer would clearly be TRUE, but for small ε values very close to 0, there may be little practical distinction between the probabilities and FALSE would be approximately correct.

The second objective-risk-comprehension question is more challenging but takes a similar form to the first question (again we provide options of TRUE, FALSE, and "I don't know"):

> *My manager is more than twice as likely to think I responded* NO *if I respond* NO *(i.e., respond truthfully) on the survey than if I respond* YES *(i.e., respond untruthfully).*

For the same reason as the first objective-risk-comprehension question, only respondents in the experimental conditions and deterministic control see it. For our experimental conditions, the correct answer is TRUE for $\varepsilon \in \{2,4\}$ and FALSE for $\varepsilon \in \{0.1,0.5\}$. For the deterministic control, the correct answer is TRUE. We create a score ranging from 0–2, which is the total number of correctly-answered objective-risk-comprehension questions. "I don't know" responses are considered incorrect.

(2) Subjective Privacy Understanding: We ask respondents to rate their confidence that they understand the privacy protection on a 4-point semantic scale (not at all confident–very confident). Previous studies explaining DP to people have similarly asked questions around subjective privacy understanding of DP [9, 24, 76]. Respondents are also asked to describe the privacy protection in their own words via an open-text response. Only respondents in the experimental conditions and Xiong et al. control see these questions; respondents who receive the deterministic control are not shown this question because this control does not describe privacy protections.

(3) Self-Efficacy: To understand how empowered people feel to make data-sharing decisions based on our explanation types, respondents are asked three questions about confidence in their decision making. First, they are asked to rate on a 4-point semantic scale their confidence that they have enough information to decide which action to take (similar to a question Franzen et al. [24] classify as "Subjective Understanding"). Second, they are asked to describe in an open-text response what further information, if any, they would like to have to help them with their decision. Third, they are asked to rate on a 4-point semantic scale their confidence in deciding which action to take. These questions are shown to respondents in all conditions.

(4) Willingness to Share Data: To assess respondents' willingness to share data, each respondent was asked whether they would participate (in the optional condition) or answer truthfully (in the mandatory condition) on the survey. They were then asked to explain their reasoning in an open-text response. Although respondents could navigate backward through the survey at any time, their answer to this question was locked after advancing. All respondents were shown these questions.

The 4-point semantic scales were randomly reversed for roughly half of respondents in line with best practices [69]. If scales were reversed for a particular respondent, all corresponding scales in their survey were also reversed.

4.2.2 Questionnaire Structure

Survey respondents are first instructed that they will read a fictional scenario and answer follow-up questions. Next, they read the first section of the scenario, which introduces the hypothetical survey about their manager, how they want to respond NO, and the potential repercussions of doing so. We then assess whether respondents are indeed concerned about these consequences by asking them to rate their level of concern on a 4-point semantic scale (not at all concerned–very concerned), which we later refer to as "baseline concern" in Section 5. We also include an easy-to-answer comprehension check question, and filter out respondents who fail to answer this question correctly after two attempts.

Respondents in the deterministic control end the scenario at this point. Respondents in the Xiong et al. control read the adapted explanation of DP. Respondents in the experimental conditions are provided with the following abstraction of a random distribution:

> *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.*

Then, the respondent is shown one computed explanation.

Subsequently, all respondents answer a series of questions on willingness to share data, objective risk comprehension, subjective privacy understanding, and self-efficacy. Finally, respondents answer questions on numeracy skills [50], internet skills [32], and demographics (gender, age, race, education, computer science/IT educational/work background, and income). The final question is an open-text question for bot detection [47]. Respondents may go back to previous pages of the survey to review any information and are also provided with links to PDFs with complete descriptions of the scenario and privacy protections for easy access. The full survey text is available on OSF[4].

4.3 Participant Recruitment

Participants were all recruited on Prolific, based in the U.S., and were at least 18 years old. We performed a power analysis to estimate the appropriate sample size for the survey [12]; 963 respondents completed the online survey and 12 completed cognitive interviews. We paid cognitive interview participants $7.50 for 15 minute interviews ($30/hour). Survey respondents were paid $2.95 in each of the experimental conditions (median completion time: 10.9 minutes; ∼$16/hour) and $2.30 in each control condition (median completion time: 9.2 minutes; ∼$15/hour). A detailed breakdown of survey respondents' demographics can be found on OSF[4].

4.4 Analysis

Our analysis[6] aims to study (1) the effect of our explanation methods on several dependent variables (DVs) and (2) relationships between these outcomes and sociodemographic attributes, which prior work finds may influence privacy-related decisions (see, e.g., [29, 33, 51, 54, 64, 66, 67, 70]). To guide our analysis, we first construct causal directed acyclic graphs (DAGs) informed by prior work on relationships between demographic attributes and privacy concerns (e.g., [62]), internet skills (e.g., [34, 36]), and numeracy skills (e.g., [26]). These DAGs are on OSF[4].

For our primary analysis, we construct a set of regression models studying the effect of our independent variables (IVs)—explanation method (categorical, see details below), scenario setting (binary: optional or not), and privacy loss budget (ε as a numeric)[7]—on our DVs: objective risk comprehension, subjective privacy understanding, self-efficacy, and willingness to share data. We construct logistic regression models for binary DVs (willingness to share data) and ordinal regression models for ordinal DVs (subjective privacy understanding, both self-efficacy measures, and number of correctly-answered objective-risk-comprehension questions).[8]

When constructing regression models, we treat the explanation IV as a categorical variable. Depending on the DV in the model, we use one or both of the control explanations as the reference (baseline) level for comparison. As described in Section 4.2.1 not all DVs are applicable for both control conditions. In cases where both controls are applicable, we construct multiple models, with each control as the reference level for the explanation IV, respectively. In order to study relationships between ε and the IVs, we cannot use either control condition because there exists no state-of-the-art or control for presenting information about ε. To compare the efficacy of our odds-based methods (ODDS-TEXT and ODDS-VIS) with our concrete-example usability-based method (SAMPLE REPORTS), we also construct models for each DV using SAMPLE REPORTS as the explanation reference level.

Next, we conduct a secondary analysis on relationships between our IVs and age (numeric), education (categorical), gender (binary[9]), baseline concern (ordinal), internet skills (numeric), and numeracy skills (numeric), which prior work finds may influence respondents' ability to interpret numerically-related information such as that presented in our explanations [24, 75]. Note that we include baseline concern in this analysis in line with best practice guidance on vignette surveys and on privacy vignettes in particular [56]. We fit models with only demographics (age, education, gender) and models with baseline concern, internet skills, and numeracy skills adjusted for said demographics.

To support results from the statistical analysis, we qualitatively analyze the open-text responses to help contextualize salient quantitative results. Two of the authors reviewed a subset of about 10% of the respondents' open-text responses and together developed a codebook (available on OSF[4]) capturing themes from responses that help illustrate findings from the statistical analysis. One of the authors then coded an additional subset of about 20% of responses to ensure that we had captured a majority of general sentiments in our codebook; no additional codes were identified during this second round of coding. In total, we coded 283 responses. Because the qualitative data are neither quantified via counts nor our primary research focus, we do not report inter-coder reliability [57].

4.5 Limitations

Our study's results are limited by multiple factors which apply to large-scale survey studies. First, our sample may have failed to capture a representative population. Research has found that Prolific has relatively high external validity for questions about beliefs and perceptions related to privacy [79]. However, as is typical for crowdsourced studies, our sample skews toward younger and more educated individuals, and thus does not fully represent the U.S. population. Our study is also conducted on people based in the U.S., which may limit the applicability of our findings to cultural contexts outside the U.S. Second, we aimed to make the survey scenario as realistic and understandable to respondents as possible by refining it through several cognitive interviews while maintaining a reasonable survey length. However, it is possible that in obscuring certain details, the scenario does not reflect important aspects of various workplace environments that may impact how people make decisions about sharing data. Third, although we aimed to elicit respondents' actual responses by following best practices [69] such as providing "I don't know" answer choices where applicable, it is possible that respondents' answers do not always align with their actual feelings/decisions they would make in similar real-life scenarios. Prior work suggests that vignette studies can be powerful tools for understanding real-life behavior, especially when respondents are highly engaged. We suspect that the high level of concern expressed by respondents about the fictional scenario—over 75% said they were "concerned" or "very concerned"—suggests a high level of engagement. Finally, our survey only focused on the data-sharing scenario of workplace surveys; our findings may not generalize directly to other settings. Although we examine our explanations within a single scenario, we note that the explanation methods we develop for DP are scenario-agnostic, so future work could

[6] Data and analysis code available on OSF.[4]

[7] We include ε instead of e^{ε} in models because e^{ε} artificially compresses relevant privacy conditions, especially for small values of ε [15].

[8] We treat number of correctly-answered objective-risk-comprehension questions as an ordinal variable because there are only three values it can take ($\{0,1,2\}$). Thus, we avoid treating it as a continuous variable.

[9] We provided respondents four options (man, woman, non-binary, self-describe) and allowed them to choose multiple [74]. 20 respondents identified as non-binary and one as fluid gender. Our sample sizes are too small to draw meaningful conclusions about how each of these groups interprets our explanations. Thus, for modeling purposes, we code gender as whether someone identifies as a man (regardless of their other gender identities).

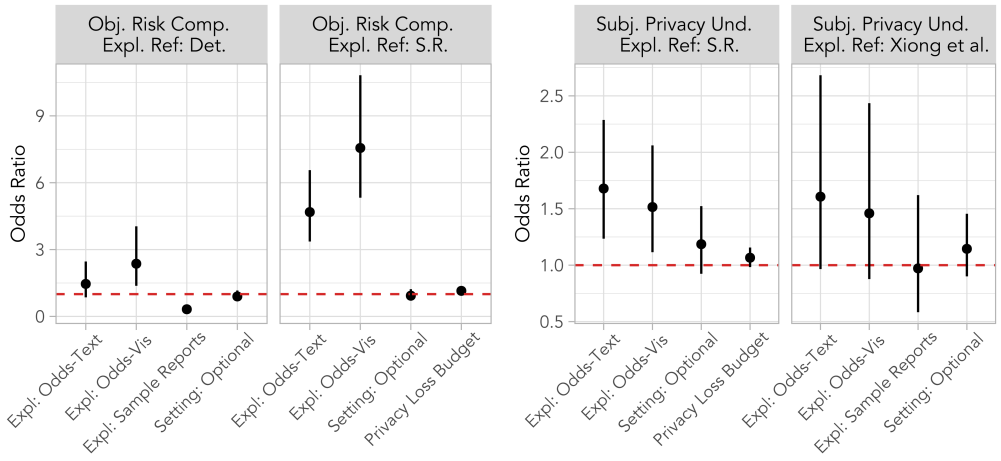


Figure 2: *Left two subplots:* results from ordinal regression models examining relationships between number of correctly-answered objective comprehension questions and experimental IVs. In the first subplot, the reference level for the explanation IV is deterministic control, and in the second the reference is SAMPLE REPORTS. We report odds ratios and corresponding 95% CIs. An OR > 1 indicates an increase in odds, while an OR < 1 indicates a decrease. The dashed red lines depicts OR $= 1$—i.e., no difference in odds. *Right two subplots:* results from ordinal regression models examining relationships between subjective privacy understanding and IVs. Explanation reference levels are SAMPLE REPORTS and the Xiong et al. control, respectively.

port these explanation methods into new scenarios as needed.

5 Results

Based on our survey results, we seek to evaluate the effectiveness of our explanations (**RQ1**) and study how our explanations impact people's data-sharing decisions (**RQ2**).

5.1 Effectiveness of Explanations (RQ1)

We evaluate the effectiveness of our explanations, ODDS-TEXT, ODDS-VIS, and SAMPLE REPORTS, via three metrics: (1) objective risk comprehension, (2) subjective privacy understanding, and (3) self-efficacy.

5.1.1 Objective Risk Comprehension (RQ1a)

We construct an ordinal regression model (Figure 2, left)[10] where the DV is the total number of correctly-answered objective-risk-comprehension questions (per respondent) following the methods in Section 4.4. Compared to the deterministic control, we find that ODDS-VIS explanations have a significant positive effect on objective comprehension of privacy risks (OR $= 2.36$, 95% CI $= [1.38, 4.05]$) and SAMPLE REPORTS explanations have a significant negative effect (OR $= 0.32$, CI $= [0.19, 0.54]$).

We also construct an ordinal regression model with the same DV, but where the explanation reference level is the SAMPLE REPORTS explanation method, which allows us to include ε in the model (as described in Section 4.4). We observe that increased ε has a slight positive effect on objective risk comprehension (OR $= 1.15$, CI $= [1.05, 1.26]$). As ε grows, the disparity between outcomes we provide (i.e., odds or sample DP outputs) also grows, which we hypothesize makes comparison easier. In addition, we find significant positive effects of ODDS-TEXT (OR $= 4.68$, CI $= [3.35, 6.54]$) and ODDS-VIS (OR $= 7.56$, CI $= [5.30, 10.77]$) on objective comprehension compared to SAMPLE REPORTS.

Our secondary analysis (Figure 4) reveals that when adjusted for age, gender, and education (hereafter referred to as "demographics"), higher numeracy skills have a significant positive effect on objective comprehension (OR $= 2.23$, CI $= [1.24, 3.99]$), as does the highest level of baseline concern (OR $= 1.96$, CI $= [1.16, 3.32]$). We posit this could be because highly concerned respondents may give more effort to understanding the presented information [10].

5.1.2 Subjective Privacy Understanding (RQ1b)

Next, we compare people's subjective privacy understanding of DP guarantees when presented with our explanations versus the Xiong et al. control (Figure 2, right), but do not find that any of our explanation methods have significant effects versus the control. We construct a second model where SAMPLE REPORTS is the explanation method reference level. Here we find that our ODDS-TEXT and ODDS-VIS explanation methods are associated with increased perceptions of understanding the privacy protection compared to SAMPLE REPORTS (OR $=$ 1.68, 95% CI $= [1.23, 2.28]$; OR $= 1.52$, CI $= [1.11, 2.06]$).

In our secondary analysis (Figure 4), we find that the highest level of education is associated with lower subjective privacy understanding (OR $= 0.66$, CI $= [0.46, 0.94]$), while identifying as a man is associated with higher subjective understanding (OR $= 1.65$, CI $= [1.29, 2.11]$). There is also a small effect of increased age on lower subjective understanding (OR $= 0.99$, CI $= [0.98, 1.00]$). Finally, when adjusting for demographics, we find that increased internet skills are associated with an increase in subjective

[10] Tables that include p-values are on OSF.[4]

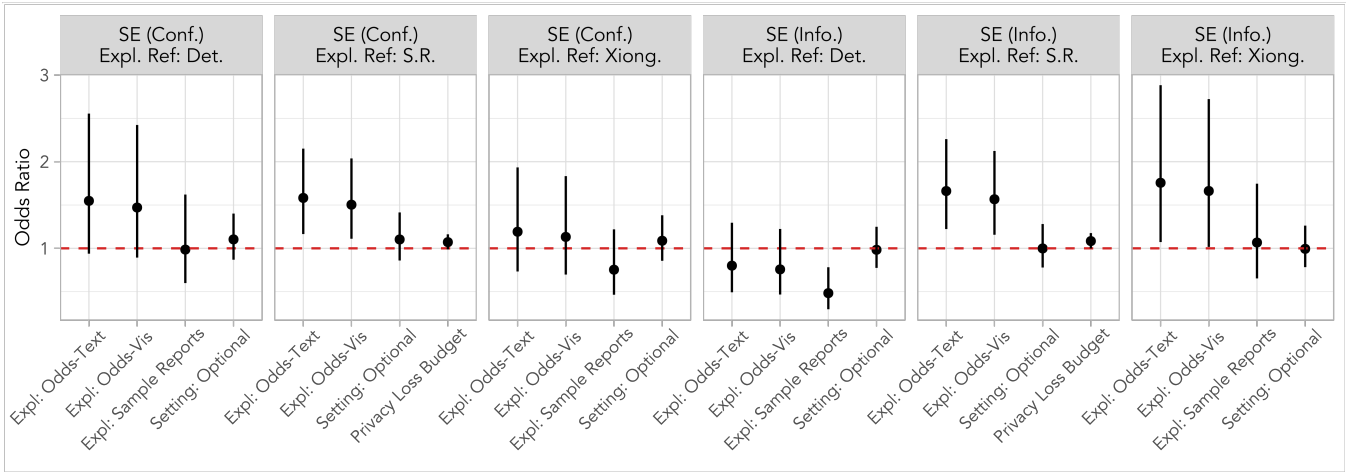


Figure 3: Ordinal regression models examining self efficacy ("enough info" (Info.) and "confidence deciding" (Conf)). SE = Self Efficacy; S.R.=SAMPLE REPORTS; Det. = Deterministic; Xiong. = Xiong et al. See Figure 2 for interpretation.

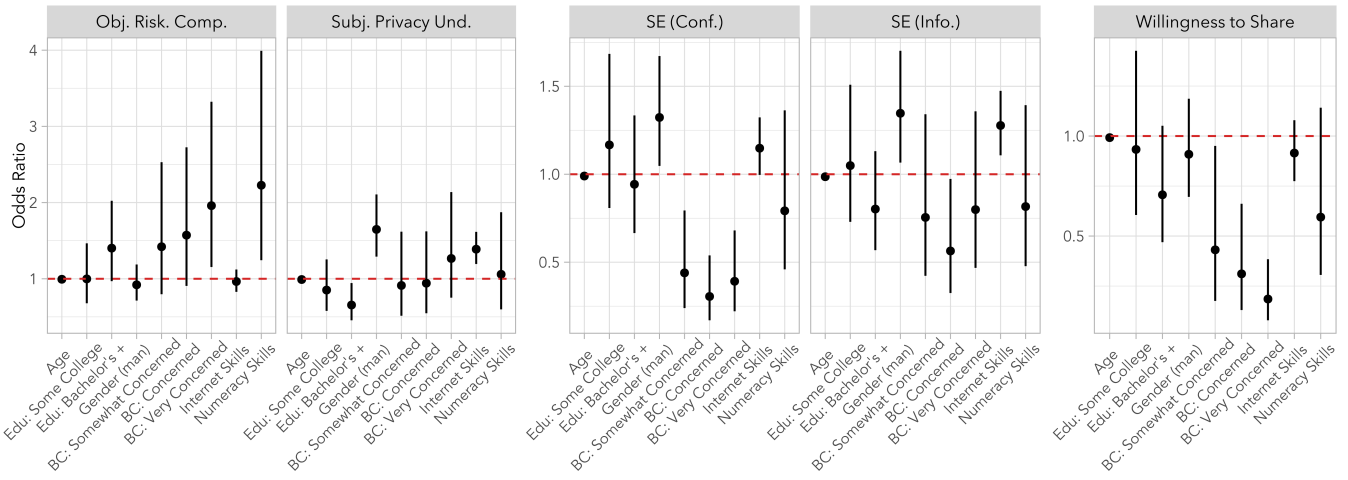


Figure 4: Results from our secondary analysis where we examined relationships between DVs and demographics (age, education, gender (man)) & baseline concern ("BC"), internet skills, and numeracy skills. SE = Self Efficacy. For interpretation of coefficients, see Figure 2.

understanding (OR $= 1.39$, CI $= [1.19, 1.61]$). Consistent with prior work, these findings suggest multiple social factors mediate people's perceptions of their understanding of privacy guarantees. For example, older adults may have lower confidence in their knowledge about S&P topics [54], men may report higher self-confidence across a variety of domains including digital skills and use [13, 35, 53, 63, 73], and those with higher internet skills may perceive themselves as having greater understanding of digital concepts [35].

5.1.3 Self-Efficacy (RQ1c)

We measure self-efficacy in terms of (1) the extent to which respondents feel they have *enough information to make these decisions* and (2) the extent to which they feel *confident in making data-sharing decisions* (Figure 3).

Enough Information to Decide. We find that SAMPLE REPORTS have a significant negative effect on feelings of having enough information to decide when compared to the deterministic control (OR $= 0.48$, CI $= [0.30, 0.78]$). We hypothesize that respondents felt that the information presented in SAMPLE REPORTS misaligned with key pieces of information they needed. For example, one respondent wrote that they "*would like to know the likelihood that the [figure in the] report [the manager] receives is higher or lower*" (than the total number of NO responses), indicating they may have wanted a summary of probabilities like in the ODDS-TEXT or ODDS-VIS explanations. Another respondent wrote: "*I would want to know the chances that my [manager] gets a higher number.*"

Compared to the Xiong et al. control group, respondents who received the ODDS-TEXT explanation were over 75% more likely to report feeling a point higher in our 4-point semantic scale for having enough information to decide

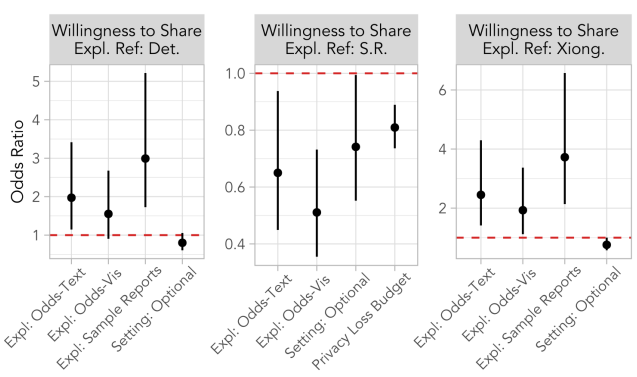


Figure 5: Logistic regression models examining relationships between willingness to share data and our IVs. Det. = Deterministic; S.R. = SAMPLE REPORTS; Xiong. = Xiong et al. See Figure 2 for interpretation.

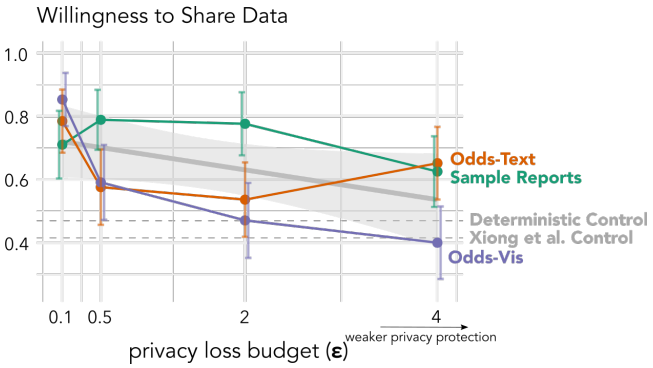


Figure 6: Proportion of respondents willing to share data across explanation methods and ε, shown with 95% binomial CIs. We plot a regression line (solid gray) between proportion of data sharing across our methods and ε.

(OR = 1.76, CI = [1.07, 2.88]); ODDS-VIS respondents were over 65% more likely (OR = 1.66, CI = [1.02, 2.72]). No such association was found for SAMPLE REPORTS.

Compared to the SAMPLE REPORTS respondents, we find that participants who received the ODDS-TEXT and ODDS-VIS explanation methods were over 65% and 55% more likely, respectively, to report feeling a point higher on the scale for having enough information to decide (OR = 1.66, CI = [1.22, 2.26]; OR = 1.57, CI = [1.16, 2.12]). We do not find a significant effect of ε on feelings of having enough information to decide. Finally, our secondary analysis (Figure 4) indicates a small effect of increased age on decreased feelings of enough information to decide (OR = 0.99, CI = [0.98, 0.99]) and that identifying as a man is associated with higher such feelings (OR = 1.35, CI = [1.07, 1.70]). When adjusting for demographics, increased internet skills are also associated with higher such feelings (OR = 1.28, CI = [1.11, 1.47]) and baseline concern of "concerned" is associated with lower such feelings (OR = 0.56, CI = [0.33, 0.98]).

Confidence Deciding. We do not find significant relationships between our experimental explanations and confidence deciding when compared to either the deterministic control or the Xiong et al. control. For the third model where we hold SAMPLE REPORTS as the explanation reference, we do not find that ε has a significant effect on confidence deciding. However, our ODDS-TEXT and ODDS-VIS explanations are associated with an increase in feelings of confidence deciding compared to the SAMPLE REPORTS method (OR = 1.58, CI = [1.16, 2.15]; OR = 1.50, CI = [1.11, 2.04]). Through our secondary analysis (Figure 4), we find that being a man is associated with higher feelings of confidence deciding (OR = 1.32, CI = [1.05, 1.67]). We also find and that all three higher levels of baseline concern (adjusted for demographics) are associated with lower confidence deciding. We posit that those more concerned about negative repercussions may feel more conflicted about which data-sharing decision to make.

5.2 Influence on Data Sharing (RQ2)

To answer RQ2, we investigate the extent to which our experimental explanations influence people's data-sharing decisions (Figure 5). Compared to both the deterministic and Xiong et al. controls, the SAMPLE REPORTS and ODDS-TEXT explanations have significant relationships with willingness to share data. Respondents are about 2 times as likely to share their data when shown the ODDS-TEXT explanation compared to the deterministic control (95% CI = [1.14, 3.42]), and about 3 times as likely to share their data when shown SAMPLE REPORTS compared to the deterministic control (95% CI = [1.73, 5.22]), an interesting finding considering the SAMPLE REPORTS explanation did not seem to well-support respondents in objective risk comprehension (see Section 5.1.1). Respondents were over twice as likely to share their data if shown the ODDS-TEXT explanation (95% CI = [1.41, 4.3]) compared to the Xiong et al. control, nearly two times as likely when shown the ODDS-VIS explanation (95% CI = [1.12, 3.37]), and nearly four times as likely when shown the SAMPLE REPORTS explanation (95% CI = [2.14, 6.57]).

We additionally construct a logistic regression model to examine the relationship between the DV, willingness to share data, and ε. We find that as ε increases, and privacy protections become weaker, respondents are less likely to share their data (OR = 0.81, CI = [0.74, 0.89]). Figure 6 shows the proportion of respondents who said that they would share their data, for each explanation method and ε value. Furthermore, we find that respondents are less likely to share data if given the ODDS-TEXT or ODDS-VIS explanations than the SAMPLE REPORTS explanation (OR = 0.65, CI = [0.45, 0.94] and OR = 0.51, CI = [0.35, 0.73]). We hypothesize that this may be related to differences in feelings of self-efficacy between the explanations (see Section 5.1.3). In the model with SAMPLE REPORTS as reference, we also find that compared to answering truthfully in the mandatory setting, people are less likely to share their data in the optional setting (OR = 0.74, CI = [0.55, 0.99])—in the mandatory setting, people may feel

uncomfortable lying. When asked to explain their decision-making, many participants explicitly expressed a desire to be honest. For example, one respondent wrote that they "*would tell the truth regardless because [they] refuse to lie regardless of the outcome.*" Finally, when adjusting for demographics, the higher three baseline concern levels understandably have a small decreased effect on sharing data (Figure 4).

The open-text responses provide further context on respondents' decision-making. Many respondents explicitly reasoned about how their behavior would change the odds of their manager believing they responded NO. For example, one respondent wrote: "*The chance that the manager will believe I responded no is only slightly higher if I participate than if I don't, so I may as well give my opinion.*" Others expressed less concern about privacy and instead focused on the utility of the collected data. For example, one respondent argued that participating in the survey was "*the right thing to do,*" since it might lead to improved conditions for their coworkers.

5.3 Summary

Below, we summarize salient findings from our results:

- Our ODDS-VIS explanation method improves objective risk comprehension over our SAMPLE REPORTS method. Furthermore, both ODDS-VIS and ODDS-TEXT improve subjective privacy understanding and self-efficacy over SAMPLE REPORTS.
- ODDS-TEXT and ODDS-VIS improve feelings of having enough information to make privacy decisions over the existing state-of-the-art [85].
- All of our explanation methods, which provide information about ε, increase willingness to share data over the existing state-of-the-art [85], which omits ε information. However, note that respondents given SAMPLE REPORTS were more likely to share data compared to those given either odds-based explanation, despite their comparatively poor objective risk comprehension.
- Respondents were less likely to be willing to share data when privacy protections weakened (ε increased).

6 Discussion

We reflect on how methods around communicating DP can be improved by including utility implications, extended to other scenarios, and support accountability around DP deployments. Our results suggest that providing more detail about privacy protection—even probabilistic information with which people may struggle—has the potential to improve people's agency in data sharing decisions.

Communicating Utility Implications. Our work builds on prior work illustrating that odds can effectively communicate privacy risk in comparison with a no-privacy or no-data-collection alternative [24, 44] to show that explanation methods that communicate odds can help people better understand *probabilistic* privacy guarantees. However, our qualitative results indicate that, in at least some contexts, people may also be concerned about utility implications of DP. For example, some respondents felt that it was important for their input to be faithfully communicated, e.g., to improve their workplace environment. Higher amounts of DP noise, while affording stronger privacy protections, reduce the accuracy and "utility" of released statistics. At face value it may seem that accuracy concerns are more in the domain of data curators (e.g., the interfaces described in Section 2), but we suggest that people may similarly require depictions of accuracy to make effective judgments about the downstream utility of their data. This in turn supports people to make informed decisions about data sharing. In line with prior research illustrating that presenting information about both accuracy and privacy improves ability to predict people's data-sharing decisions in medical contexts [44] and prior work aiming to convey accuracy implications of DP to people [86], people may benefit from reasoning not only about privacy, but also about the accuracy-privacy trade-off in the DP context. Thus, we emphasize the need to go "beyond" privacy to consider utility as a key part of respectful data use [68]. Our results yield insight into *how* utility implications may be most effectively communicated to people. Our ODDS-TEXT and ODDS-VIS explanation methods demonstrate how to map ε to outcome probabilities. It may be additionally useful for explanations to provide mappings from ε to utility-related outcomes. For example, if receiving a certain number of NO responses will require the manager to complete additional training, odds-based methods can similarly be used to communicate the probability that that threshold value will be met given a person's response under a given ε. Such information could help them assess utility of their response in terms of leading to a tangible outcome.

Toward More Complex Scenarios. Our work contributes a framework for communicating ε implications to end users; details about our specific scenario can be changed to apply our methods to varied applications of DP. For example, to apply our odds-based explanation methods to other types of queries, such as mean queries, we suggest computing probabilities based on a hypothesis testing framework [52]. While our explanation methods for probabilistic DP guarantees may be easily extended to more complex queries, explaining nuances in more complex algorithmic settings may pose challenges. For example, binary counting queries have sensitivity of 1, which means that every data sharer's answer changes the function's value exactly by the sensitivity. This is not true for many other queries, and may require future work to resolve additional explanation-related challenges. Also as future work, we see promise in creating guidance and tooling that allows people employing DP to generate explanations consistent with specifics of their application settings. Such work would encourage practical use of explanation methods that clarify the impact of ε.

Public Deliberation Around ε. Dwork et al. [19] have pro-

posed a registry of ε values (and other implementation details) used by organizations applying DP. They argue that such a registry could enable comparisons across differentially-private systems and increase accountability around the use of DP, lowering chances of privacy theater. Our explanations can increase the impact of such a registry by building on work that aims to translate the implications of ε to a wider audience, thus helping facilitate public deliberation around privacy loss budgets. More generally, our explanation methods and other methods of translating ε can support external audits by making implementation decisions like ε more widely interpretable and easier to discuss among interested parties with a diverse set of expertise and backgrounds. For example, recent debates around the U.S. Census Bureau's use of DP for the 2020 Census demonstrated challenges of effectively discussing aspects of DP among several groups (computer scientists, policymakers, demographers, non-profit organizations, the public, etc.) [8, 60]. These challenges arose in part because the technical specifics of DP (such as privacy loss budgets) can be abstract or counterintuitive, especially when discussed outside the computer science literature. Explanations of ε effective for a broad population can close gaps in communication and encourage public deliberation over privacy policies.

Acknowledgements

We would like to thank Elizabeth Chase and Sean Kross for guidance on our analysis, Jessica Hullman for reviewing an earlier draft of the paper, Frauke Kreuter for feedback on survey design, and Chase Stokes for visualization feedback. In addition, we thank the attendees of Theory and Practice of Differential Privacy (TPDP) 2022, attendees of the Symposium on Applications of Contextual Integrity 2022, members of the MU Collective at Northwestern University, and members of a Columbia University privacy reading group for valuable discussion and feedback.

All authors were supported by DARPA (contract number W911NF-21-1-0371). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the United States Government or DARPA. In addition to DARPA support, the third author was supported in part by NSF grant CNS-1942772 (CAREER), a Mozilla Research Grant, a JP-Morgan Chase Faculty Research Award, and an Apple Privacy-Preserving Machine Learning Award. The fourth author was also supported by NSF grant #2030859 to the Computing Research Association for the CIFellows Project, and the fifth author was also supported by a Google Research Scholar Award.

References

[1] John M. Abowd. The U.S. Census Bureau adopts Differential Privacy. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '18, page 2867, New York, NY, USA, 2018. Association for Computing Machinery.

[2] Ahmet Aktay, Shailesh Bavadekar, Gwen Cossoul, John Davis, Damien Desfontaines, Alex Fabrikant, Evgeniy Gabrilovich, Krishna Gadepalli, Bryant Gipson, Miguel Guevara, et al. Google covid-19 community mobility reports: anonymization process description (version 1.1). *arXiv preprint arXiv:2004.04145*, 2020.

[3] Diego Alonso and Pablo Fernandez-Berrocal. Irrational decisions: Attending to numbers rather than ratios. *Personality and Individual Differences*, 35(7):1537–1547, 2003.

[4] Apple Differential Privacy Team. Learning with privacy at scale. 2017.

[5] Andrea Bittau, Úlfar Erlingsson, Petros Maniatis, Ilya Mironov, Ananth Raghunathan, David Lie, Mitch Rudominer, Ushasree Kode, Julien Tinnes, and Bernhard Seefeld. Prochlo: Strong privacy for analytics in the crowd. In *Proceedings of the 26th Symposium on Operating Systems Principles*, pages 441–459, 2017.

[6] Daniel M Bittner, Alejandro E Brito, Mohsen Ghassemi, Shantanu Rane, Anand D Sarwate, and Rebecca N Wright. Understanding privacy-utility tradeoffs in differentially private online active learning. *Journal of Privacy and Confidentiality*, 10(2), 2020.

[7] Sidney T Bogardus Jr, Eric Holmboe, and James F Jekel. Perils, pitfalls, and possibilities in talking about medical risk. *Jama*, 281(11):1037–1041, 1999.

[8] boyd, danah and Sarathy, Jayshree. Differential perspectives: Epistemic disconnects surrounding the US Census Bureau's use of differential privacy. *Harvard Data Science Review (Forthcoming)*, 2022.

[9] Brooke Bullek, Stephanie Garboski, Darakhshan J. Mir, and Evan M. Peck. Towards understanding differential privacy: When do people trust randomized response technique? In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, CHI '17, pages 3833–3837, Denver, Colorado, USA, May 2017. Association for Computing Machinery.

[10] Richard L Celsi and Jerry C Olson. The role of involvement in attention and comprehension processes. *Journal of Consumer Research*, 15(2):210–224, 1988.

[11] United States Census Bureau. Census Bureau sets key parameters to protect privacy in 2020 Census results. 2021.

[12] Stephane Champely. *pwr: Basic Functions for Power Analysis*, 2020. R package version 1.3-0.

[13] Regina Ju-chun Chu. How family support and internet self-efficacy influence the effects of e-learning among higher aged adults–Analyses of gender and age differences. *Computers & Education*, 55(1):255–264, 2010.

[14] Rachel Cummings, Gabriel Kaptchuk, and Elissa M. Redmiles. "I need a better description": An investigation into user expectations for differential privacy. In Giovanni Vigna and Elaine Shi, editors, *ACM CCS 2021*, pages 3037–3052. ACM Press, November 2021.

[15] Inbal Dekel, Rachel Cummings, Ori Heffetz, and Katrina Ligett. The privacy elasticity of behavior: Conceptualization and application. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4145498, 2022.

[16] Damien Desfontaines. A list of real-world uses of differential privacy, Oct 2021. Ted is writing things (personal blog).

[17] Damien Desfontaines and Balázs Pejó. Sok: Differential privacies. *Proceedings on Privacy Enhancing Technologies*, 2020(2):288–313, 2020.

[18] Bolin Ding, Janardhan Kulkarni, and Sergey Yekhanin. Collecting telemetry data privately. *Advances in Neural Information Processing Systems*, 30, 2017.

[19] Cynthia Dwork, Nitin Kohli, and Deirdre Mulligan. Differential privacy in practice: Expose your epsilons! *Journal of Privacy and Confidentiality*, 9(2), 2019.

[20] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In Shai Halevi and Tal Rabin, editors, *TCC 2006*, volume 3876 of *LNCS*, pages 265–284. Springer, Heidelberg, March 2006.

[21] Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4):211–407, 2014.

[22] Adrian Edwards, Glyn Elwyn, and Al Mulley. Explaining risks: Turning numerical data into meaningful pictures. *The BMJ*, 324(7341):827–830, 2002.

[23] Úlfar Erlingsson, Vasyl Pihur, and Aleksandra Korolova. RAPPOR: Randomized aggregatable privacy-preserving ordinal response. In Gail-Joon Ahn, Moti Yung, and Ninghui Li, editors, *ACM CCS 2014*, pages 1054–1067. ACM Press, November 2014.

[24] Daniel Franzen, Saskia Nuñez von Voigt, Peter Sörries, Florian Tschorsch, and Claudia Müller-Birn. "Am I private and if so, how many?"–Using risk communication formats for making differential privacy understandable. *arXiv preprint arXiv:2204.04061*, 2022.

[25] Marco Gaboardi, James Honaker, Gary King, Jack Murtagh, Kobbi Nissim, Jonathan Ullman, and Salil Vadhan. Psi (ψ): A private data sharing interface. *arXiv preprint arXiv:1609.04340*, 2016.

[26] Mirta Galesic and Rocio Garcia-Retamero. Statistical numeracy for health: a cross-cultural comparison with probabilistic national samples. *Archives of internal medicine*, 170(5):462–468, 2010.

[27] Mirta Galesic, Rocio Garcia-Retamero, and Gerd Gigerenzer. Using icon arrays to communicate medical risks: Overcoming low numeracy. *Health Psychology*, 28(2):210, 2009.

[28] Gerd Gigerenzer and Ulrich Hoffrage. How to improve Bayesian reasoning without instruction: Frequency formats. *Psychological Review*, 102(4):684, 1995.

[29] Hana Habib, Jessica Colnago, Vidya Gopalakrishnan, Sarah Pearman, Jeremy Thomas, Alessandro Acquisti, Nicolas Christin, and Lorrie Faith Cranor. Away from prying eyes: Analyzing usage and understanding of private browsing. In *Fourteenth Symposium on Usable Privacy and Security (SOUPS 2018)*, pages 159–175, 2018.

[30] Jens Hainmueller, Dominik Hangartner, and Teppei Yamamoto. Validating vignette and conjoint survey experiments against real-world behavior. *Proceedings of the National Academy of Sciences*, 112(8):2395–2400, 2015.

[31] Marian Harbach, Markus Hettig, Susanne Weber, and Matthew Smith. Using personal examples to improve risk communication for security & privacy decisions. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 2647–2656, 2014.

[32] Eszter Hargittai. Digital na(t)ives? Variation in internet skills and uses among members of the "net generation". *Sociological Inquiry*, 80(1):92–113, 2010.

[33] Eszter Hargittai et al. Facebook privacy settings: Who cares? *First Monday*, 2010.

[34] Eszter Hargittai, Anne Marie Piper, and Meredith Ringel Morris. From internet access to internet skills: digital inequality among older adults. *Universal Access in the Information Society*, 18(4):881–890, 2019.

[35] Eszter Hargittai and Steven Shafer. Differences in actual and perceived online skills: The role of gender. *Social Science Quarterly*, 87(2):432–448, 2006.

[36] Eszter Hargittai and Aaron Shaw. Mind the skills gap: the role of internet know-how and gender in differentiated contributions to wikipedia. *Information, communication & society*, 18(4):424–442, 2015.

[37] Michael Hay, Ashwin Machanavajjhala, Gerome Miklau, Yan Chen, Dan Zhang, and George Bissias. Exploring privacy-accuracy tradeoffs using DPComp. In *Proceedings of the 2016 International Conference on Management of Data*, pages 2101–2104, 2016.

[38] Ori Heffetz and Katrina Ligett. Privacy and data-based research. *Journal of Economic Perspectives*, 28(2):75–98, 2014.

[39] Ulrich Hoffrage and Gerd Gigerenzer. Using natural frequencies to improve diagnostic inferences. *Academic Medicine*, 73(5):538–540, 1998.

[40] Justin Hsu, Marco Gaboardi, Andreas Haeberlen, Sanjeev Khanna, Arjun Narayan, Benjamin C Pierce, and Aaron Roth. Differential privacy: An economic method for choosing epsilon. In *2014 IEEE 27th Computer Security Foundations Symposium*, pages 398–410. IEEE, 2014.

[41] Mark F St John, Grit Denker, Peeter Laud, Karsten Martiny, Alisa Pankova, and Dusko Pavlovic. Decision support for sharing data using differential privacy. In *2021 IEEE Symposium on Visualization for Cyber Security (VizSec)*, pages 26–35. IEEE, 2021.

[42] Noah Johnson, Joseph P Near, and Dawn Song. Towards practical differential privacy for sql queries. *Proceedings of the VLDB Endowment*, 11(5):526–539, 2018.

[43] Daniel Kahneman, Stewart Paul Slovic, Paul Slovic, and Amos Tversky. *Judgment under Uncertainty: Heuristics and Biases*. Cambridge University Press, 1982.

[44] Gabriel Kaptchuk, Daniel G Goldstein, Eszter Hargittai, Jake Hofman, and Elissa M Redmiles. How good is good enough for COVID19 apps? The influence of benefits, accuracy, and privacy on willingness to adopt. *arXiv preprint arXiv:2005.04343*, 2020.

[45] Farzaneh Karegar, Ala Sarah Alaqra, and Simone Fischer-Hübner. Exploring user-suitable metaphors for differentially private data analyses. In *Eighteenth Symposium on Usable Privacy and Security (SOUPS 2022)*, pages 175–193, 2022.

[46] Shiva Prasad Kasiviswanathan, Homin K Lee, Kobbi Nissim, Sofya Raskhodnikova, and Adam Smith. What can we learn privately? *SIAM Journal on Computing*, 40(3):793–826, 2011.

[47] Courtney Kennedy, Nick Hatley, Arnold Lau, Andrew Mercer, Scott Keeter, Joshua Ferno, and Dorene Asare-Marfo. Assessing the risks to online polls from bogus respondents. *Pew Research Center*, 2020.

[48] Jaewoo Lee and Chris Clifton. How much is enough? Choosing ε for differential privacy. In *International Conference on Information Security*, pages 325–340. Springer, 2011.

[49] Isaac M Lipkus. Numeric, verbal, and visual formats of conveying health risks: Suggested best practices and future recommendations. *Medical Decision Making*, 27(5):696–713, 2007.

[50] Isaac M Lipkus, Greg Samsa, and Barbara K Rimer. General performance on a numeracy scale among highly educated samples. *Medical Decision Making*, 21(1):37–44, 2001.

[51] Eden Litt. Understanding social network site users' privacy tool use. *Computers in Human Behavior*, 29(4):1649–1656, 2013.

[52] Changchang Liu, Xi He, Thee Chanyaswad, Shiqiang Wang, and Prateek Mittal. Investigating statistical privacy frameworks from the perspective of hypothesis testing. *Proceedings on Privacy Enhancing Technologies*, 2019(3):233–254, 2019.

[53] Mary A Lundeberg, Paul W Fox, and Judith Punćcohaŕ. Highly confident but wrong: Gender differences and similarities in confidence judgments. *Journal of Educational Psychology*, 86(1):114, 1994.

[54] Mary Madden. *Privacy, Security, and Digital Inequality*. Data & Society Research Institute, 2017.

[55] James E Maddux and Jennifer T Gosselin. *Self-efficacy*. The Guilford Press, 2012.

[56] Kirsten Martin and Katie Shilton. Putting mobile application privacy in context: An empirical study of user privacy expectations for mobile devices. *The Information Society*, 32(3):200–216, 2016.

[57] Nora McDonald, Sarita Schoenebeck, and Andrea Forte. Reliability and inter-rater reliability in qualitative research: Norms and guidelines for CSCW and HCI practice. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–23, 2019.

[58] Frank D McSherry. Privacy integrated queries: an extensible platform for privacy-preserving data analysis. In *Proceedings of the 2009 ACM SIGMOD International Conference on Management of data*, pages 19–30, 2009.

[59] Priyanka Nanayakkara, Johes Bater, Xi He, Jessica Hullman, and Jennie Rogers. Visualizing privacy-utility trade-offs in differentially private data releases. *Privacy Enhancing Technologies (PETS)*, 2022.

[60] Priyanka Nanayakkara and Jessica Hullman. What's driving conflicts around differential privacy for the US census. *IEEE Security & Privacy*, (01):2–11, 2022.

[61] Helen Nissenbaum. Privacy in context. In *Privacy in Context*. Stanford University Press, 2009.

[62] Dara O'Neil. Analysis of internet users' level of online privacy concerns. *Social Science Computer Review*, 19(1):17–31, 2001.

[63] Frank Pajares. Gender and perceived self-efficacy in self-regulated learning. *Theory into Practice*, 41(2):116–125, 2002.

[64] Yong Jin Park. Do men and women differ in privacy? Gendered privacy and (in) equality in the internet. *Computers in Human Behavior*, 50:252–258, 2015.

[65] Mayana Pereira, Allen Kim, Joshua Allen, Kevin White, Juan Lavista Ferres, and Rahul Dodhia. Us broadband coverage data set: a differentially private data release. *arXiv preprint arXiv:2103.14035*, 2021.

[66] Lee Rainie, Sara Kiesler, Ruogu Kang, Mary Madden, Maeve Duggan, Stephanie Brown, and Laura Dabbish. Anonymity, privacy, and security online. *Pew Research Center*, 5, 2013.

[67] Elissa Redmiles. Net benefits: Digital inequities in social capital, privacy preservation, and digital parenting practices of US social media users. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 12, 2018.

[68] Elissa M Redmiles. *The need for respectful technologies: Going beyond privacy*, pages 307–316. Springer, 2021.

[69] Elissa M Redmiles, Yasemin Acar, Sascha Fahl, and Michelle L Mazurek. A summary of survey methodology best practices for security and privacy researchers. Technical report, University of Maryland, 2017.

[70] Elissa M Redmiles, Sean Kross, and Michelle L Mazurek. Where is the digital divide? A survey of security, privacy, and socioeconomics. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pages 931–936, 2017.

[71] Amy Henderson Riley, Elizabeth Critchlow, Lyena Birkenstock, MariaLisa Itzoe, Katherine Senter, Nichole M Holmes, and Steven Wesley Buffer. Vignettes as research tools in global health communication: A systematic review of the literature from 2000 to 2020. *Journal of Communication in Healthcare*, 14(4):283–292, 2021.

[72] Ryan Rogers, Adrian Rivera Cardoso, Koray Mancuhan, Akash Kaura, Nikhil Gahlawat, Neha Jain, Paul Ko, and Parvez Ahammad. A members first approach to enabling linkedin's labor market insights at scale. *arXiv preprint arXiv:2010.13981*, 2020.

[73] John A Ross, Garth Scott, and Catherine D Bruce. The gender confidence gap in fractions knowledge: Gender differences in student belief–achievement relationships. *School Science and Mathematics*, 112(5):278–288, 2012.

[74] Morgan Klaus Scheuerman, Katta Spiel, Oliver L Haimson, Foad Hamidi, and Stacy M Branham. HCI guidelines for gender equity and inclusivity, 2020.

[75] Paul Ed Slovic. *The perception of risk.* Earthscan publications, 2000.

[76] Mary Anne Smart, Dhruv Sood, and Kristen Vaccaro. Understanding risks of privacy theater with differential privacy. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW2), Nov 2022.

[77] Christopher Soghoian. An end to privacy theater: Exposing and discouraging corporate disclosure of user data to the government. *Minnesota Journal of Law Science and Technology*, 12:191, 2011.

[78] Stone, Maureen. How we designed the new color palettes in Tableau 10. *Tableau Website*, 2016.

[79] Jenny Tang, Eleanor Birrell, and Ada Lerner. Replication: How well do my results generalize now? The external validity of online privacy and security surveys. In *Eighteenth Symposium on Usable Privacy and Security (SOUPS 2022)*, pages 367–385, 2022.

[80] Jun Tang, Aleksandra Korolova, Xiaolong Bai, Xueqiang Wang, and Xiaofeng Wang. Privacy loss in Apple's implementation of differential privacy on macOS 10.12. *arXiv preprint arXiv:1709.02753*, 2017.

[81] Katie Tezapsidis. Uber releases open source project for differential privacy, 2017.

[82] Pratiksha Thaker, Mihai Budiu, Parikshit Gopalan, Udi Wieder, and Matei Zaharia. Overlook: Differentially private exploratory visualization for big data. *arXiv preprint arXiv:2006.12018*, 2020.

[83] Gordon B Willis. *Cognitive interviewing in practice: Think-aloud, verbal probing and other techniques*, pages 42–63. Sage Publications, London, 2005.

[84] Alexandra Wood, Micah Altman, Aaron Bembenek, Mark Bun, Marco Gaboardi, James Honaker, Kobbi Nissim, David R O'Brien, Thomas Steinke, and Salil Vadhan. Differential privacy: A primer for a non-technical audience. *Vanderbilt Journal of Entertainment and Technology Law*, 21:209, 2018.

[85] Aiping Xiong, Tianhao Wang, Ninghui Li, and Somesh Jha. Towards effective differential privacy communication for users' data sharing decision and comprehension. In *2020 IEEE Symposium on Security and Privacy*, pages 392–410. IEEE Computer Society Press, May 2020.

[86] Aiping Xiong, Chuhao Wu, Tianhao Wang, Robert W Proctor, Jeremiah Blocki, Ninghui Li, and Somesh Jha. Using illustrations to communicate differential privacy trust models: An investigation of users' comprehension, perception, and data sharing decision. *arXiv preprint arXiv:2202.10014*, 2022.

[87] Cindy Xiong, Ali Sarvghad, Daniel G Goldstein, Jake M Hofman, and Çagatay Demiralp. Investigating perceptual biases in icon arrays. In *CHI Conference on Human Factors in Computing Systems*, pages 1–12, 2022.

[88] Kimihiko Yamagishi. When a 12.86% mortality is more dangerous than 24.14%: Implications for risk communication. *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition*, 11(6):495–506, 1997.

A Modeling Adversary's Priors

To make the scenario discussed in this paper more realistic and flexible, we can account for the fact that the manager may have a prior belief about how the respondent will respond even before seeing the privacy-protected report. We can model the manager's thought process by specifying their prior belief and supposing that they perform a Bayesian update to obtain a posterior belief based on the DP output in the report. We show that under this model, changing the manager's prior is equivalent to shifting the threshold ($r_{\text{threshold}}$) on the DP output at which the manager believes that one outcome (i.e., the respondent said NO) is more likely than the other (i.e., the respondent did not say NO).

If the respondent answers NO, the manager will see a sample drawn from a Laplace distribution centered at 1. Otherwise, the manager sees a sample drawn from a Laplace distribution centered at 0. Thus, the manager's task is to guess whether the DP output r was drawn from $Lap(\mu = 1, b = \frac{1}{\varepsilon})$ or $Lap(\mu = 0, b = \frac{1}{\varepsilon})]$.

We use Bayes' Theorem to calculate the probability that the manager finds it more likely that the respondent answered NO after viewing a DP output r (i.e., the manager's posterior). Let f_i denote the probability density function for $Lap(\mu = i, b = \frac{1}{\varepsilon})]$, and let P_{no} denote the manager's prior belief that the respondent answered NO.

The updated probability that the respondent answered NO (the posterior probability) is given by:

$$\frac{f_1(r)P_{no}}{f_1(r)P_{no} + f_0(r)(1-P_{no})} = \frac{e^{-\varepsilon|r-1|}P_{no}}{e^{-\varepsilon|r-1|}P_{no} + e^{-\varepsilon|r|}(1-P_{no})}$$

Similarly, the updated probability that the respondent did not answer NO is given by:

$$\frac{f_0(r)(1-P_{no})}{f_1(r)P_{no} + f_0(r)(1-P_{no})} = \frac{e^{-\varepsilon|r|}(1-P_{no})}{e^{-\varepsilon|r-1|}P_{no} + e^{-\varepsilon|r|}(1-P_{no})}$$

We can find the new threshold, $r_{\text{threshold}}$, by finding the value of r for which the manager finds it equally likely that the respondent answered NO or did not answer NO. In other words, we set the above equations equal to each other and solve for r to obtain:

$$r_{\text{threshold}} = \frac{\ln(1-P_{no}) - \ln(P_{no})}{2\varepsilon} + \frac{1}{2}$$

when $\max\{\frac{1-P_{no}}{P_{no}}, \frac{P_{no}}{1-P_{no}}\} \leq e^{\varepsilon}$ (i.e., as long as the prior is not too extreme).

Once this new threshold is obtained, it is straightforward to apply our explanation methods. We simply need to calculate the probability that the DP output will be greater than the threshold, given each choice the individual can make (e.g., participate or not participate). Note that in our study we use 0.5 as the threshold. This is the value for $r_{\text{threshold}}$ that one obtains when $P_{no} = 0.5$.