Exploring Expandable-Grid Designs to Make iOS App Privacy Labels More Usable

Shikun Zhang and Lily Klucinec, *Carnegie Mellon University;* Kyerra Norton, *Washington University in St. Louis;* Norman Sadeh and Lorrie Faith Cranor, *Carnegie Mellon University*

https://www.usenix.org/conference/soups2024/presentation/zhang

This paper is included in the Proceedings of the Twentieth Symposium on Usable Privacy and Security.

August 12–13, 2024 • Philadelphia, PA, USA

978-1-939133-42-7

Open access to the Proceedings of the Twentieth Symposium on Usable Privacy and Security is sponsored by USENIX.

Exploring Expandable-Grid Designs to Make iOS App Privacy Labels More Usable

Shikun Zhang
Carnegie Mellon University

Lily Klucinec
Carnegie Mellon University

Kyerra Norton
Washington University in St. Louis

Norman Sadeh
Carnegie Mellon University

Lorrie Faith Cranor
Carnegie Mellon University

Abstract

People value their privacy but often lack the time to read privacy policies. This issue is exacerbated in the context of mobile apps, given the variety of data they collect and limited screen space for disclosures. Privacy nutrition labels have been proposed to convey data practices to users succinctly, obviating the need for them to read a full privacy policy. In fall 2020, Apple introduced privacy labels for mobile apps, but research has shown that these labels are ineffective, partly due to their complexity, confusing terminology, and suboptimal information structure. We propose a new design for mobile app privacy labels that addresses information layout challenges by representing data collection and use in a color-coded, expandable grid format. We conducted a between-subjects user study with 200 Prolific participants to compare user performance when viewing our new label against the current iOS label. Our findings suggest that our design significantly improves users' ability to answer key privacy questions and reduces the time required for them to do so.

1 Introduction

Privacy policies have long been criticized for their complexity and lack of usability [32]. In response to these challenges, standardized and concise privacy nutrition labels have emerged as a potential solution to help users better understand the privacy practices of both websites and mobile apps [19–21]. Usable privacy nutrition labels can not only aid lay users' understanding of how their personal data is used, but also serve as valuable tools for privacy advocates and regulators, functioning as clear points of reference for assessing privacy practices and a foundation to enforce transparent and fair privacy regulations. Prior studies have shed light on the challenges and user frustrations associated with the existing iOS and Google privacy labels, particularly when it comes to label terminology and information layout [10, 25–27, 44, 45].

To address the information layout challenges faced by current iOS privacy labels, we built on prior research on privacy labels and access control interface design to develop and iteratively refine a prototype expandable-grid [38] privacy label that represents all iOS data categories and purposes in a color-coded compact format.

To compare our prototype labels with the existing labels, we conducted a between-subjects survey study with 200 Prolific participants. The main goals of this survey were to compare label comprehension between the existing iOS privacy labels (control condition) and our proposed label design (treatment condition), as well as explore what components contribute positively and negatively to the usability of both designs. We asked survey participants to look at the privacy labels for two existing apps with different label content and to answer comprehension questions based on the information presented in the labels. Additionally, we asked participants to provide the reasoning for their answers, which allowed us to qualitatively code their responses for sources of confusion.

Our work explores the following research questions:

RQ1: Does the proposed iOS privacy label design aid in user comprehension of iOS app data practices?

RQ2: Is the proposed iOS privacy label design effective in decreasing the time it takes for users to answer questions about mobile app data practices?

RQ3: Which elements of the existing and proposed iOS labels are most conducive or disruptive to user comprehension?

Our contributions include:

- A proposed design for an expandable-grid-based privacy label to communicate iOS app data practices.

USENIX Symposium on Usable Privacy and Security (SOUPS) 2024.
August 11–13, 2024, Philadelphia, PA, United States.

- An empirical between-subjects study showing that the proposed design improves users' ability to answer key privacy questions and reduces the time taken to do so.
- Identification of key areas for further improvement of privacy label designs.

2 Background and Related Work

The advent of smartphones has significantly expanded the realm of mobile data processing, offering convenience and productivity to billions of users worldwide. With an increasingly diverse set of sensors and constant proximity to users, consumers are growing increasingly concerned about privacy issues associated with their mobile devices [4, 11]. The major mobile app stores have implemented and refined permission interfaces and privacy controls over the years, and have recently introduced mobile app privacy labels. In this section, we review research on mobile app privacy, privacy notices and nutrition labels, the usability of mobile app privacy labels, and tabular and grid interfaces that inspired our prototype label design.

2.1 Mobile App Privacy

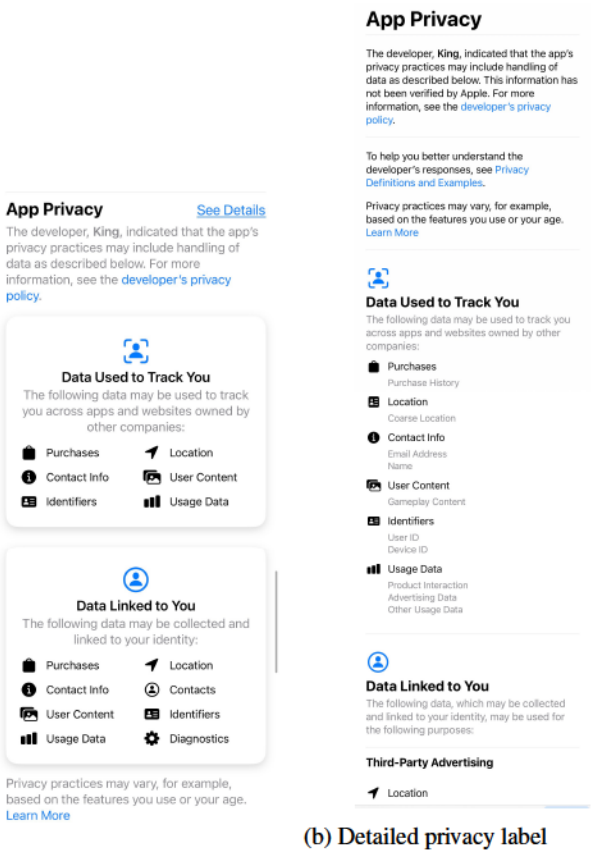
Mobile devices can collect diverse and sensitive data about users, including but not limited to their location, contacts, health data, and photos. When the iPhone was first introduced in 2007, there were no permission settings until three years later [6]. Starting with the location permission, new permission settings were introduced [7]. Currently, the prevalent method of presenting privacy information and seeking consent for app permissions management systems on Android and iOS is the "ask on first use" approach, functioning as both a notice and choice mechanism. In addition, research has demonstrated the considerable influence of privacy nudges on users [1, 3, 17], and iOS added "Do you want to continue allowing this?" nudges, aimed at alerting users about background data collection.

As the number of apps grows and each app potentially requires multiple permissions, managing each and every privacy permission places an overwhelming burden on users. Recent studies highlight these usability challenges and propose the concept of "privacy assistants." These assistants can inform users about sensitive data practices and assist users in configuring privacy settings [9, 12]. Assistants can also leverage machine learning models of individual privacy preferences to further reduce user burden for privacy management [28, 30, 31, 41, 42].

2.2 Privacy Notices and Nutrition Labels

Privacy policies are the de facto standard for informing consumers about data practices, yet research has shown that these policies are prohibitively long and difficult to read [5, 32, 40]. Privacy nutrition labels were first developed by Kelley et al. as a way of addressing these issues by providing consumers with succinct descriptions of key data practices, similar to FDA food and drug labels [16]. Kelley et al. developed website privacy labels and showed they made disclosures easier to understand and reduced the amount of time people need to answer typical privacy questions [19, 20]. Later Kelley and collaborators proposed mobile app privacy labels and reported on a study suggesting that the labels would help smartphone users make better informed privacy decisions when considering apps to install on their devices [21].

In 2020, Apple introduced its own mobile app privacy labels (shown in Figure 1) and started requiring app developers to provide labels for new apps published in the iOS app store. In 2021, Google followed suit with its own variation of mobile app labels for the Google Play store.



(a) Compact privacy label

(b) Detailed privacy label (partially shown)

Figure 1: Existing compact and detailed Apple Privacy Labels as found in the App Store in iOS 16.6 for the Candy Crush app. Users can click on "See Details" in the compact label to see the detailed privacy label.

2.3 Usability of Mobile App Privacy Labels

The privacy labels in the iOS App Store and Google Play Store have been widely criticized. Studies have shown that labels suffer from accuracy problems [23, 25–27], few users are aware of and use the labels, and those who try to use them find them confusing. Zhang et al. reported on a detailed analysis of iOS privacy labels, looking at the extent to which users were aware of their existence and able to use them effectively. This study revealed a number of shortcomings, including confusing label terminology (e.g., unconventional use of terms like "tracking"), confusing information organization, label complexity, and a disconnect between the labels and privacy controls made available to users [44]. This work highlighted the need to better structure label content, which is the focus of the present paper. Android data safety labels are formatted differently than iOS privacy labels and include additional information about app security. However, they suffer from similar problems as iOS labels, including confusing terminology and a complex and confusing structure [10, 29]. The diverse data practices of mobile apps pose a challenge in summarizing relevant information into an easy-to-understand format.

To make matters worse, studies have shown that despite their complexity, existing iOS and Android privacy labels may address only about half of the privacy questions typical mobile app users have [45]. However, despite recent progress towards the development of automated tools to answer users' questions by analyzing the text of privacy policies, privacy Q&A assistants are far from fully accurate [18, 35, 36]. Moreover, effective use of privacy Q&A assistants in their current state presupposes that users can both identify and articulate meaningful privacy questions. Conversely, privacy labels offer users answers to a plethora of likely questions without necessitating users to generate or articulate them independently.

2.4 Tabular and Grid Interfaces

Tables and grids, familiar to most people, present data in a concise and structured manner. Tables can typically be scanned quickly and allow for easy side-by-side comparison. They have been shown to be an effective mechanism for organizing information found in privacy policies. In particular, Kelley et al. compared tabular interfaces for website privacy policies with short- and long-text interfaces and found that people preferred the tabular format. They found their tables, which showed data types in rows and data uses in columns, were easy for study participants to use when scanning for information and comparing policies [20]. Researchers who designed and evaluated standardized financial privacy notices also found that consumers responded positively to a tabular approach [24]. In addition, tabular approaches have been used for IoT security and privacy labels [13].

Grids have been used in the design of access-control interfaces. Reeder et al. deployed a grid interface to compactly represent what access each user and group has for each file and folder in a file system. As users are often members of groups and files are often members of folders, they developed an *expandable-grid* interface that could display a grid of folders and groups, with the ability to expand any folder to show its files or expand any group to show its users. They used green and red colored cells in the grid to indicate that users were allowed or denied access to a particular file. When access permissions were the same for all files in a folder or all users in a group, the green and red colors were used on the folder or group cells. However, when permissions varied for different users in a group or different files in a folder, a yellow cell was used to indicate that expansion was needed to view detailed access permissions. Reeder et al. demonstrated that the expandable-grid approach was more effective than the traditional Windows XP access control system in making users aware of file permissions and allowing them to adjust access control settings. [37, 38]. Reeder et al. also used expandable grids in the design of a privacy label but found less success, largely due to their attempts to represent three dimensions in a two-dimensional space without using color. They offered a number of recommendations for future designers who want to use expandable grids, including representing only one dimension per axis and using short, understandable terms [39]. We leveraged expandable grids in our interface, benefiting from the lessons learned in past work.

3 Designing a New Privacy Label

Our focus in redesigning iOS app privacy labels is to improve their information layout. Both Apple and Google adopt a layered approach, offering a compact version that users can click through to get full details. But in both cases the compact version provides only minimal information, and the full version is difficult to navigate, potentially overwhelming users. Navigating the full iOS label involves extensive scrolling, and users often fail to recognize that it is a linear representation of a matrix of data types and purposes [44]. Google attempted to manage some of the complexity of the full version with an accordion interface, but users who want a full understanding of a policy must individually expand every line of the accordion, with no way to quickly scan to determine whether the app engages in a particular data practice [29].

The core principles guiding our iterative design approach were as follows: maintain a compact format suitable for mobile screens, structure the label in a more intuitive and user-friendly manner, and incorporate interactive elements to enhance user engagement and comprehension. We did not address the confusing terminology in this redesign as it requires a separate and systematic approach to identify more usable privacy terms, which is beyond the scope of this work.

3.1 Adopting an Expandable-Grid Structure

In iOS privacy labels, data practices are described along three dimensions: the data type being collected, the purpose for which that data type is collected, and whether the data being collected is linked to the user or used to track the user. In contrast to prior work by Reeder et al. [39], whose attempts to re-organize three-dimensional privacy policy data along two dimensions produced mixed results, we opted to use color to represent one dimension. We introduced a simple color scheme to represent whether data is linked to the user. We represent purpose and data type using the X and Y axis of the grid, respectively. We observed that whether data is used to track the user is actually a purpose, and therefore fold that into the purpose dimension. With 14 types of data and 33 subcategories present in Apple's privacy labels, accommodating all of them on a small mobile screen is difficult. Leveraging the inherently hierarchical relationship between data categories and subcategories (e.g., "email address" being a subcategory of "contact info"), we opted for an expandable grid format. Initially, users only see the 14 top-level data categories. Upon expanding a row associated with one of these top-level categories, the underlying subcategories of data types are revealed (see Figure 2a for an example of an expanded row). Our current label design does not include column expansion. We use color to indicate linked versus not-linked practices associated with subcategories of data, as further detailed below.

3.2 A Simple Color Scheme

In Apple's privacy labels each category of collected data may be linked ("Data Linked to You") or not linked to the user ("Data Not Linked to You"). This distinction can be captured with two colors. We use red when the collected data is linked to the user (the more privacy-invasive option), and blue, a more calming color, when the collected data is not linked to the user (the less privacy-invasive option). Entries in grey represent data types that are not collected at all.

As part of our design, we wanted to provide a summary of data practices for all the sub-categories beneath a top-level category that had not been expanded in the grid. We opted for a simple design that highlights privacy invasive practices. In this design we have five possible colors for a top-level data category with multiple underlying data types. These colors are explained in a legend accompanying our tabular format (see Figure 3). Grey indicates that no data is collected. Red indicates that a data category and any underlying sub-categories are collected and linked to the user. Dark blue indicates that a data category and any underlying sub-categories are collected but in a manner that is not linked to the user. We also introduced two additional colors to represent situations where sub-categories may be collected and used in heterogeneous ways. Salmon indicates that a subset of the underlying data types are used in a manner that is linked to the user, thereby highlighting the existence of a privacy-invasive practice for at least one of the underlying data types (but not all). Salmon is used independently of whether some of the other underlying data types are blue or grey. The goal is simply to highlight the existence of a privacy invasive practice while also indicating that not all underlying data types are linked to the user. The light blue color is used to indicate that, while only some subcategories of data are collected, none are linked to the user. Our salmon and light blue shades are somewhat similar in meaning to the yellow color used by Reeder et al. to represent user groups or folders in a file system with heterogeneous access permissions [38].

We considered a number of possible options including various shades of blue, red, and purple reflecting the mix of red, blue, and grey cells in underlying sub-categories. We experimented with dynamic colors based on the number of data types present and explored designs with square cells split diagonally to represent linked and unlinked data subcategories. Additionally, we considered numbers inside squares to indicate the number of underlying sub-categories. However, we opted against these options for accessibility and clarity. Our more complex designs still required expansion to understand which sub-categories were present and thus there is limited gain from such added complexity. The light colors in our design serve as a cue for row expansion.

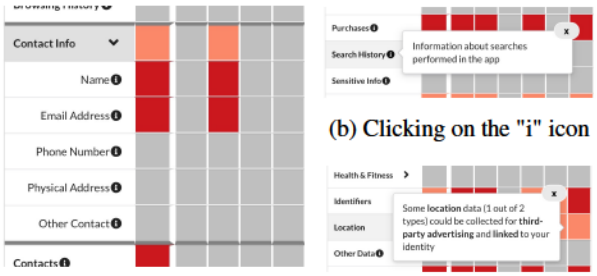
3.3 Adding Interactive Elements

We designed the grid to be expandable so that users could tap on a row label or chevron to expand a row or collapse a row already expanded. In addition, users can tap on individual cells in our table to access more detailed information about the meaning of each cell, the data it corresponds to, and the practices it describes, including how many subcategories of data it represents and the purpose of data collection associated with this particular entry (see Figure 2c for an example).

When Apple first introduced its privacy labels, they were static notices that lacked any interactive features. There was only a "See Details" link at the top right corner of the compact label (Figure 1a), linking to the detailed label (Figure 1b). In prior studies of iOS privacy labels [44], users expressed a desire for more interactive labels. Later, Apple changed its labels so that users who tap on each section within the compact labels are brought to the corresponding section of the detailed privacy label. However, the iOS labels still do not offer a direct link to definitions of terms used (a list of definitions is available only in the detailed view after users tap on "See Details"). To make definitions of terms more accessible, we placed information icons next to relevant terms; tapping one of these icons triggers a pop-up with a definition of the term, as shown in Figure 2b. To make it easier to expand the grid and access the popovers on a small screen, we designed the interface so that a tap anywhere near a row label expands the row and a tap anywhere in a cell triggers the

popover. Tapping outside the popover or on another element closes it. This seemed to work well for our pilot participants.

As our legend does not always fit on the same screen as the label, we incorporated a hyperlink within the table. This hyperlink ("What do the colors and symbols mean?") enables users to readily jump to the legend for details. Figure 4 shows our label on two different screen sizes.



(a) Clicking on a row

(b) Clicking on the "i" icon

(c) Clicking on a cell

Figure 2: Interactive elements in treatment labels

Figure 3: Legend used in treatment labels

3.4 Interview Pilot

We conducted two rounds of small-scale semi-structured interviews to help gain rich insights into the strengths and weaknesses of our prototypes. All pilot participants were assigned to view either the iOS privacy labels or our prototype labels in a round-robin fashion. These interviews were conducted over Zoom on their iPhone and participants shared their screens while interacting with the privacy labels on a mobile website that we created. This enabled us to record what actions they took with the label while answering our questions.

We asked participants about their prior experience with privacy labels and whether privacy ever influenced their decision to stop using an app. Then, we sent participants links in Zoom to open the label on their iPhone. The second section of the interview assessed participants' ability to accurately answer questions based on label information. Afterwards, we asked participants about the definitions of terms used on the labels. Finally, we asked participants to identify helpful or unhelpful aspects of the labels and provide additional feedback.

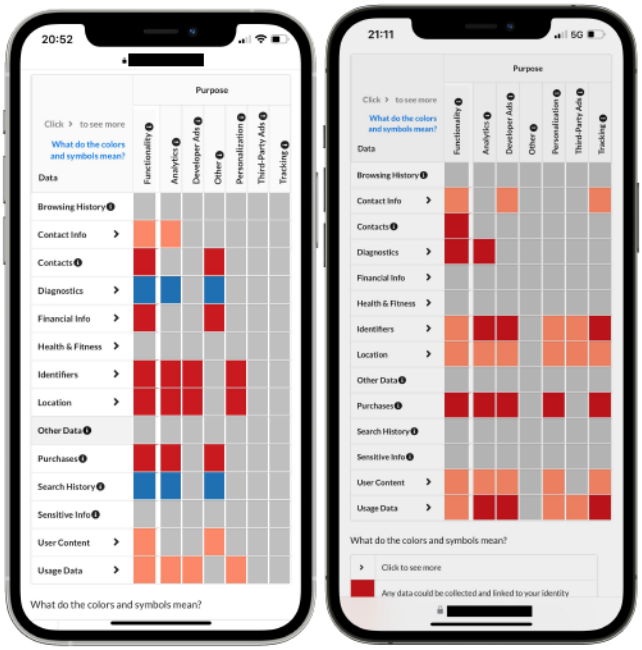


Figure 4: Treatment labels on two different screen sizes: Venmo label on an iPhone 12 Pro (left) and Candy Crush label on an iPhone 12 Pro Max (right)

Insights from the interview phase informed some modifications to the label and the development of our survey protocol. For example, our early label design hid some of the less common data categories under a "see more" row, but we found this confused pilot participants so we showed rows for all categories. In addition, the early version ordered data types and purposes by frequency in the App Store. However, pilot participants did not understand this so we switched to alphabetical order.

4 Methods

In this section, we describe our study design. We describe our recruitment process, survey procedure, survey pilots, thematic analysis, and limitations of this study.

Ethical considerations. Our interview pilot, survey pilots, and main survey were reviewed and approved by the Carnegie Mellon University Institutional Review Board. All study participants completed online consent forms.

4.1 Recruitment

We recruited participants on the Prolific[1] research participant recruitment platform who were iPhone users running iOS 14

[1] https://www.prolific.com/

or a newer version of the operating system, and thus had iOS labels available on their phones. The number of participants was determined after performing a power analysis as detailed in Section 4.3. We recruited participants who were over 18, fluent in English, and residing in the United States. We required participants to take the survey on their computers while viewing the privacy labels on their iPhones. All of these criteria were checked using Prolific's built-in screening capabilities to prevent ineligible participants from accessing the survey. Additionally, we set parameters on Prolific to create a balanced sample in terms of gender. We did not set any screening criteria for other demographic factors and we did not collect any demographic data in our survey beyond what was collected automatically by Prolific. Participants were paid $5 for successfully completing the survey.

4.2 Survey Design

We used a between-subjects survey design where participants were randomly assigned to either view Apple's privacy labels or our prototype labels. The survey consisted of four main parts: general questions about privacy and privacy labels, questions about the information found in the privacy labels for two different apps, questions about terms used in privacy labels, and feedback about the labels they were shown.

4.2.1 Study Apps

We selected two popular apps that represent significantly different types of privacy labels: Candy Crush Saga and Venmo. Candy Crush Saga has a "Data Used to Track You" section, whereas Venmo does not and instead has a "Data Not Linked to You" section. Both apps have "Data Linked to You" sections. Neither app has all three sections since iOS privacy labels with three sections are less common than those with two [2]. See Figure 1a for an example of Candy Crush's compact iOS privacy label. In our label design (shown in Figure 4), the Venmo label has blue squares to represent data that is collected but not linked to identity and red and salmon squares to represent data linked to identity, while the Candy Crush label only has red and salmon squares that represent data collected and linked to identity.

4.2.2 Survey Procedure

The survey began with general questions about privacy and privacy labels. Next, we prompted participants to use their iPhone's camera to scan a dynamically generated QR code, which encoded their Prolific ID and sent them to a specific privacy label based on their condition. We opted to show the labels on participants' phones instead of on computers to ensure that participants interact with the privacy labels in a more ecologically valid setting. Upon scanning the QR code, participants were directed to a webpage simulating the Apple App Store environment for either the Candy Crush or Venmo app. Full webpage representations shown to participants can be found in Appendix C. Participants then responded to six comprehension questions related to the app privacy label they were viewing. They were encouraged to interact with the label on their iPhone while answering the questions. Then, they scanned another QR code to view the second app and answer the same set of six questions for the second app label. The study website used Javascript to record participants' actions, including scrolls, taps, and associated timestamps. In addition, the study website also checked the browser user agent string, confirming that participants were indeed viewing the labels on an iPhone running iOS 14 or above.

4.2.3 Conditions

Half of the participants were randomly assigned to the control condition ($N = 100$) (viewing the current iOS privacy labels) and the other half to the treatment condition (viewing our prototype labels). In both conditions, participants saw the corresponding labels for two apps. Within each condition, we also randomized the order in which participants encountered each app (Venmo first or Candy Crush first). As a result, participants were randomly assigned to one of four possible groups in Qualtrics: `Control Candy`, `Control Venmo`, `Treatment Candy`, and `Treatment Venmo`, each *group* comprising approximately 50 participants. For instance, a participant in the `Control Venmo` group was in the control condition and saw Apple's Venmo label first and then Apple's Candy Crush label. This allowed us to both compare the treatment and control labels as well as see whether the order in which the apps were viewed affected participant performance.

4.2.4 Comprehension Questions

We asked six multiple-choice comprehension questions about each app's privacy label, totaling twelve questions. The questions were designed to elicit all potential types of interactions users could have with the labels in both conditions. These questions represent typical user privacy questions that can be answered using the labels, i.e., questions about types of data collected and purpose. Prior research [45] found that about 30% of contextualized user questions about mobile apps are related to data types and purpose of data collection. They include questions such as whether an app might collect photos and videos, or whether diagnostic data might be linked to a user's identity. Each question had different answers for each of the two apps of interest, preventing participants from using the same answers for both questions. Table 1 shows the questions that were asked, their respective question category, the correct answers for each label, and the actions participants would need to take to find the correct answer for each condition. We also asked participants to provide open text explanations for each multiple-choice question. In Section 5, we denote the

#	Question		Answer		Action	
	Text	Category	Venmo	Candy	Treatment	Control
Q1	Does this app collect data for Analytics purposes and, if so, what data? (Select all that apply)	Any data type for one purpose	Contact info, Diagnostics, Identifiers, Location, Purchases, Usage data	Diagnostics, Identifiers, Location, Purchases, User content, Usage data	Look down a column	See details and find all data types under a purpose
Q2	Does this app collect location data for Third-Party Ads purposes?	One data subtype for one purpose	No	Coarse location	Expand row and look at a cell	See details and find a data type under a purpose
Q3	Does this app collect Photo and Video data and, if so, for what purpose(s)? (Select all that apply)	One data subtype for any purpose	(App) Functionality, Other	(App) Functionality	Expand row and look at a row	See details and find a data subtype under all purposes
Q4	Does this app collect Purchases data and, if so, for what purpose(s)? (Select all that apply)	One data type for any purpose	(App) Functionality, Analytics, Other	(App) Functionality, Analytics, Developer Ads/Advertising, Other, Tracking	Look at a row	Find a data type in compact view, see details, find a data type under all purposes
Q5	Does this app link Diagnostics data to your identity?	One data type is linked or not	No	Yes	Look at a row's color + legend	Find a data type in compact view
Q6	Does this app collect data for Tracking purposes, and if so, what data? (Select all that apply)	All data types for one purpose	No	Contact Info, Identifiers, Location, Purchases, User content, Usage data	Look down a column	Find all data types in compact view

Table 1: Questions used in the survey, corresponding correct answers for each app, and the action needed for participants in the treatment and control conditions to answer each question correctly

questions for Candy Crush and Venmo as "CQ1–CQ6" and "VQ1–VQ6" respectively.

4.3 Survey Pilots

We conducted two rounds of survey pilots to make sure the survey protocol (including the server hosting and recording participant actions and the Qualtrics survey flow) worked as intended and to collect data for use in our power analysis.

We conducted the first pilot survey with 40 participants on Prolific under the same recruitment criteria as our main survey. We conducted an a priori power analysis using G*Power. T-test was chosen as the test family, and the Wilcoxon-Mann-Whitney test of the mean accuracies between two conditions was selected. We chose the usual alpha level of 0.05 and the most common beta value of 0.2 (indicating a power of 0.8) [33] to calculate the minimum sample size necessary for detecting the expected effect as estimated by the pilot sample. The detailed results of the power analysis for each of the 12 questions (Table 4) can be found in the Appendix. This analysis ensured that our study (with 100 participants per condition) was adequately powered to detect significant differences in the accuracy between the control and treatment conditions for Questions Q1, Q2, Q4, and Q6. As the effect sizes were small for Q3 and Q5, the power analysis suggested we would need a much larger sample size to detect significant differences between the control and treatment. We selected our sample size based on the other questions, but we still included Q3 and Q5 to observe the effectiveness of design mechanisms (e.g., color, row expansion) noted in Table 1.

4.4 Thematic Analysis

In our thematic analysis of the open text explanations provided by participants for the comprehension questions, we employed both inductive and deductive coding methods [8, 15]. To ensure internal reliability, three of the authors participated as coders and inductively coded the responses. Each response was coded by two authors. Our coding process included the following steps: three authors read through the responses to develop a set of codes. The first author reviewed all responses. The second author focused on the treatment responses, while the third author focused on the control responses. After developing these initial codes, the authors discussed the definitions and adjusted the codes based on their discussion. As our interest was primarily in the reason for incorrect answers, the three authors independently coded the explanations for a set of 812 incorrect answers and compared their coding. Any disagreements were resolved, which resulted in adjusted definitions. Finally, the authors proceeded to code participant responses using the revised codebook and resolved all conflicts.

4.5 Limitations

We enrolled participants whose iPhones were running iOS 14 or above because the iOS privacy labels were only available for those users. Additionally, our participant pool was limited to Prolific users in the United States who were proficient in English. We focused only on participants from one region, the United States, because App Store interfaces and available apps vary by region. This allows us to ensure the consistency of our simulated presentation of the privacy labels in the App Store, aligning with participants' prior experiences and mitigating the introduction of unaccounted variables. However, our results may not generalize to users in other regions of the world. Subsequent investigations could delve into the potential influence of using various languages in the labels or broaden the scope to encompass additional cultural variables. Moreover, our study focused on just two apps, Candy Crush and Venmo, and their corresponding iOS privacy labels. Users might have different experiences using other app labels or after becoming more acquainted with the labels over time.

Finally, our study focused on use of the labels by participants assigned to use them and may not fully reflect the experience of users who are motivated to review labels of apps they actually use or are considering using.

5 Results

We first present information about our participants, followed by results on accuracy, errors, time answering questions, perceived confidence, learning effect, interaction with treatment labels, and understanding of iOS label section headers.

5.1 Participants

We manually removed 15 participants due to low-quality free-text responses, lacking necessary interactions (e.g., scrolls, visiting both app labels), or answering too quickly. We analyzed the demographic information provided by Prolific. Our sample is balanced with 100 male and 100 female participants, all of whom met our specified criteria: fluent in English, iPhone users, and residing in the United States. Further details regarding the distribution of participant ages and ethnicities can be found in the Appendix. Our participants are experienced Prolific users with an average total approval count [2] of 1262 ± 1168 tasks. Our minimum approval count is 16 with 3 participants having less than 50 approvals. The median completion time was 21.2 minutes.

5.2 Accuracy Analysis

Our survey included six comprehension questions (Table 1) for each of two apps. We assessed the performance of participants in both the control and treatment conditions based on the accuracy of answers they provided.

5.2.1 Significant Differences in Half of the Questions

Figure 5a shows the accuracy percentages (the proportions of correct responses) for each of the 12 questions in each condition. We observed that the treatment group outperformed the control group in 9 out of the 12 questions. In one question (CQ6), both conditions had the same accuracy percentage. However, in two instances (CQ2 and VQ5), the control group outperformed the treatment group. To assess the statistical significance of these differences, we conducted pairwise Fisher's exact tests between the control and treatment groups for all 12 questions; the Holm-Bonferroni corrected p-values for these tests are also marked in Figure 5a. We obtained statistically significant results for half of the questions: for five of these questions the treatment group outperformed the control group and for one question the control group outperformed the treatment group.

[2] The total approval count represents an individual participant's number of approved submissions for tasks on Prolific.

5.2.2 Treatment Outperforms Control When Data Collection Is Absent

In the case of questions VQ2 and VQ6, which require participants to determine that Venmo does not engage in the data practices discussed in these questions, the treatment condition performed significantly better than the control group. For VQ2, this improvement arises because the treatment condition clearly indicates the absence of data collection with a gray-colored square, whereas participants in the control condition need to inspect the relevant sections to discern this absence i.e., participants need to search for "coarse location" within the detailed label under the "third-party advertising" purpose category and recognize that it is not there. This distinction becomes particularly evident when comparing the same question between Candy Crush (CQ2) and Venmo (VQ2) apps in Figure 5a for the control condition. In CQ2, around 90% of the participants were able to answer correctly when the data practice is there; while in VQ2, less than 40% of the control participants answer it correctly.

5.3 Error Analysis

We examined the number and type of errors made by participants in each condition, identifying common error themes across both conditions as well as errors that frequently occurred in just one of the two conditions.

5.3.1 Treatment Significantly Reduced Errors

We calculated the mean number of incorrect answers for each condition. In the treatment group, the mean number of incorrect answers was 2.68 with a standard deviation of 2.19, while the control group had a mean of 5.08 incorrect answers with a standard deviation of 2.38. We employed the Mann-Whitney U test to compare two independent groups (control and treatment) as the data is not normally distributed. The test confirmed the significant difference between the two groups ($U = 2230.0$, $p = 4.53e\text{-}12$) with a large effect of size 0.55. The treatment significantly improved on the control, reducing errors by approximately half.

5.3.2 Common Error Themes Across Conditions

We analyzed 812 explanations for incorrect answers and identified a number of common error themes for both conditions during our qualitative analysis. First, many participants were confused by the terminology used in the labels, such as conflating "identifiers" with "linked to your identity" in Q5, mixing up "contacts" and "contact info" in Q1 and Q6, and struggling to differentiate between "developer advertising" and "third-party advertising" in Q2. This confusion often led them to search in the wrong part of the label for answers. Second, some participants misunderstood the questions or provided responses based on their personal beliefs or prior knowledge

rather than the information provided in the labels. For instance, in the case of asking whether Venmo collects purchase data, one participant answered, "They do keep record of your bank account login information, routing numbers and credit cards linked to your account, but they do not disclose and[sic] information to third party social networking services." Third, some participants made accidental errors or mistakes when answering the multiple-choice questions but quickly realized and explained them in their free-text justifications. Fourth, some participants provided vague or brief justifications, making it difficult for us to pinpoint the reasons behind their error.

5.3.3 Challenges with Color Coding for Treatment

In two of the questions (CQ2 and VQ5), where the treatment condition showed worse performance compared to the control condition, errors were related to the use of color coding within the treatment labels. In the case of CQ2, the correct answer for Candy Crush is indicated by a salmon-colored square in the treatment label. Participants first need to understand that the color signifies certain sub-categories, but not all "Location" sub-categories are collected. Then participants must expand the row to know whether the salmon-colored square signifies "coarse location" or "precise location" being collected. The qualitative analysis revealed that 19 participants (60% of the incorrect participants) could not find the info or provided answers that suggested they did not expand the row. This is also consistent with the recorded taps where 18 (58% of the incorrect participants) did not expand the row. The 30% error rate for this question also aligns with participants' comprehension rate of color cues, as evaluated in the treatment condition later in the survey, where 31% of participants did not seem to understand that a salmon-colored square indicates less data is collected than a red square.

VQ5 pertains to whether diagnostics data is linked or not to user identity. In the treatment condition, participants need to recognize that a blue square signifies that the data is not linked to user identity. From our qualitative analysis of participants' justification, we found that 14 participants (50% of the incorrect answers) misinterpreted the colors, and another 5 participants (18% of the incorrect answers) accidentally selected the wrong answer or immediately realized that they had selected the wrong answer as explained in their free-text justifications. We also assessed participants' ability to correctly interpret the blue color in a later question, with 80% of participants correctly interpreting the meaning of the color.

5.3.4 Incomplete Answers for the Control

We observed that a major reason for incorrect answers in the control was incomplete answers. This pattern is very evident in the case of Q1 (finding all data types for analytics purposes). For VQ1, participants need to find all data types used for analytics across sections "Data Linked to you" and "Data Not Linked to You," which required a lot of scrolling in the control condition. No control participants answered correctly. In CQ1, where all correct answers fell under the "Data Linked to You" section, participants were more accurate, with a 45% error rate. Our qualitative analysis showed that 79% of the errors were due to participants not scrolling enough to find all the information. Another common error was to select all data types as answer choices (20% of the errors).

Furthermore, a significant drop in performance was observed in the control group when comparing CQ4 and VQ4. The sole difference between the two was that participants had to identify 3 purposes for CQ4 and 5 for VQ4. Control participants were more likely to provide incomplete answers when faced with a higher number of purposes. Conversely, treatment participants responded with high accuracy regardless of the number of purposes they had to identify.

Each accuracy question also asked participants to explain how they arrived at their answers through a free-text response. As described in Section 4.4, we thematically coded these responses. Below, we present the primary themes that emerged during this analysis, along with their respective frequencies.

5.4 Time to Answer Comprehension Questions

We computed the time spent on comprehension questions, excluding the time for free-text responses. For the control condition, the mean time was 10m59s, while for the treatment condition, it was 8m28s. Since the time spent does not follow a normal distribution, we conducted the Mann-Whitney U test, which revealed a statistically significant difference between the control and treatment conditions ($U = 2202.0$, $p = 4.78e\text{-}07$) with a large effect size of 0.56.

These findings indicate that participants in the treatment condition spent significantly less time compared to those in the control condition. Figure 5b provides a detailed breakdown of the time spent answering each of the 12 questions. Our timing data includes time for both correct and incorrect answers. When we specifically examined the time for correct answers, we found the same trends. We further conducted pairwise Mann-Whitney U tests between the control and treatment conditions for each of the 12 questions and applied Holm-Bonferroni correction to the p-values. As shown in Figure 5b, 8 out of the 12 questions produced significant results.

In all questions except one (CQ2), the control group took more time than the treatment. In VQ2 and VQ6, where the answer is "no" and thus there is no mention of that type of data collection in the control, the control took significantly more time than the treatment with a large difference.

5.5 Perceived Confidence Analysis

For each question, we also asked participants to rate their confidence in their answers on a Likert scale ranging from 1 (not at all confident) to 5 (extremely confident). The distribution

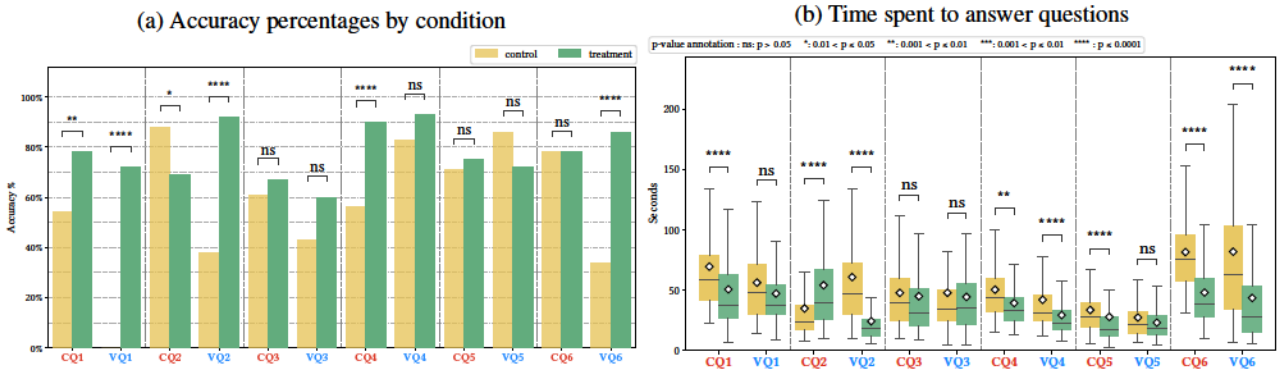


Figure 5: CQ1 denotes Question 1 for the app Candy Crush, and VQ1 denotes Question 1 for the app Venmo. All p-values adjusted by the Holm-Bonferroni method.

	Incorrect Count						Time Spent in Minutes					
	Candy 1st	Candy 2nd	p	Venmo 1st	Venmo 2nd	p	Candy 1st	Candy 2nd	p	Venmo 1st	Venmo 2nd	p
Control	2.39 ± 1.40	1.43 ± 1.38	**	3.39 ± 1.30	2.94 ± 1.41	ns	5.26 ± 1.48	5.87 ± 3.60	ns	6.89 ± 3.70	4.01 ± 1.87	****
Treatment	1.73 ± 1.30	1.12 ± 1.05	*	1.41 ± 1.44	1.10 ± 1.17	ns	5.70 ± 3.64	3.70 ± 3.73	****	4.49 ± 2.75	3.03 ± 1.60	**

Table 2: The difference in the number of incorrect answers (left) and time spent in minutes (right) among participants viewing Candy or Venmo as either their first or second app across both the control and treatment conditions. Eight one-sided Mann-Whitney U tests were conducted in total, and p-values were adjusted using the Holm-Bonferroni method.

of responses is graphed in Appendix A, Figure 6. The control condition exhibited a significant level of uncertainty regarding VQ2 and VQ6, both of which pertain to situations where the data collection mentioned in the question is absent from the label. In 9 out of 12 questions, more participants in the treatment group felt "extremely confident" in their answers compared to the control group. The mean confidence score for the control condition is 4.29 ± 0.49, while for the treatment condition, it is 4.48 ± 0.63. We also calculated the Kendall's Tau-b rank correlation between the average confidence level and the number of incorrectly answered questions across both conditions. The Kendall's Tau-b is a non-parametric measure of association that exists between two ordinal variables [22]. In the control group, Kendall's tau correlation revealed a statistically significant negative and *weak*[3] relationship between the two variables ($\tau = -0.17, p = 0.02$). For the treatment, a statistically significant negative *strong* correlation was observed between the two variables ($\tau = -0.35, p < 0.0001$). This stronger correlation in the treatment group suggests that participants in this condition were more likely to feel confident about their answers when they were indeed correct compared to the control group.

5.6 Learning Effect

To assess the presence of a potential learning effect between the first and second exposure to the labels, we further divided each of the control and treatment conditions into two distinct groups: the "first" group consists of participants encountering the label as the first label they viewed in the study, while the "second" group comprises participants encountering the label as the second label they viewed in the study. We consider the learning effect in two dimensions: accuracy and time.

Even though we observed a decrease in errors from first to second for both apps and both conditions (Table 2), after correcting p-values for multiple tests, only the decreases in errors for Candy Crush are significant across both conditions. Table 2 shows the average time for participants in each condition for both groups. There is a significant decrease in the time needed to answer questions for the treatment group no matter which app they see first. However, the decrease in time only appears for participants answering questions for the Venmo app in the control condition.

5.7 Interaction with Treatment Labels

As noted in Section 4.2.2, we captured participants' interactions with treatment labels, including taps and scrolls.

Our analysis revealed that among 100 participants in the treatment condition, 80% of participants expanded one or more rows during the study, 68% of participants tapped at least one of the information icons to access definitions, 61% of participants tapped at least one cell, and 32% of participants tapped the hyperlink (i.e., "What do the colors and symbols mean?") located inside the table that brings them to the legend. These findings indicate that participants actively engaged with

[3] https://www.spss-tutorials.com/kendalls-tau/#kendalls-tau-formulas

the interactive elements incorporated into the treatment labels.

Analyzing participant interaction with information icons revealed that the purpose category "Tracking" garnered the highest number of taps at 76, followed by "Other" with 36 taps. The data type "Other" received 59 taps, while "Purchases" received 41 taps. Other information icon interactions with more than 10 taps include: "Sensitive info" (23 taps), "Other diagnostic data" (19 taps), "Third-party advertising" (14 taps), "Analytics" (13 taps), "App functionality" (13 taps), "Diagnostics–crash data" (12 taps), and "Diagnostics–performance data" (11 taps).

5.8 Participants' Understanding of iOS Privacy Label Section Headers

To briefly explore participants' understanding of the terminology used in the privacy labels, we asked participants multiple-choice questions regarding the definitions of "Data Linked to You," "Data Not Linked to You," and "Data Used to Track You." 74% of the participants correctly identified the definition of data linked to you and 49% correctly identified the definition of data not linked to you, with no significant differences between the control and the treatment conditions.

For "Data Used to Track You," only 53% of the control and 33% of the treatment were correct, showing a significant difference with a p-value of 0.016 after Bonferroni correction. One potential explanation for the treatment label's poorer performance is treatment participants were exposed to the term "Tracking" rather than "Data used to track you" in the interface, but still asked about "Data used to track you" in the survey. Another potential explanation is that the treatment label displayed "Tracking" as a purpose, alongside other purposes such as "Personalization" and "Third-Party Ads." The correct definition of tracking—"Identifiable data that is shared with third parties to personalize ads" contains words similar to these other purposes. This might have led participants to believe that tracking should be distinct from these listed purposes. To delve deeper, we observed that out of 28 treatment participants who clicked on the information icon for tracking at some point during the study, 18 answered this question correctly. In contrast only 15 out of 72 treatment group participants who did not click on the information icon for tracking were correct. This indicates that the information icon likely contributed to participants selecting the correct definition of tracking.

We also described two data collection scenarios and asked participants whether they consider each to be tracking or not: 1) app sharing your location/email address with third party advertisers, 2) app using location to show you nearby stores. For the first scenario, 81.5% correctly consider that to be tracking, but for the second scenario only 6% correctly identified that it is not tracking under Apple's definition. The responses to these scenario questions were not significantly different between control and treatment conditions.

6 Discussion

Below, we summarize the main findings of our research and discuss future possible avenues for extending this work, including addressing privacy label limitations not addressed by the proposed grid layout (e.g., confusing terminology) and opportunities to offer personalized label presentations.

6.1 What Made the Treatment Effective

Expandable grids have been evaluated in various contexts with mixed results: they were shown to be well suited for windows file permission control but less effective for P3P policies [37–39]. Our study reaffirmed the advantage of displaying the complete policy [38]. In one early design variation, we opted to display only selected rows, requiring users to click "see more" for additional content. However, many pilot interview participants missed accessing the complete content. The expandable grid format enabled us to accommodate the limited screen real estate available on mobile devices and present the entire label in a compact, organized format. In contrast, the lengthy format of the control label resulted in incomplete answers due to the need for extensive scrolling and compiling answers from multiple sections. This improvement was instrumental in helping participants answer questions such as Q4 correctly, which requires them to consider all purposes associated with the collection of a specific data type. Treatment participants could readily answer the question by inspecting a single row in the table, whereas control participants had to scroll through a number of purpose sections spanning multiple screens within the "See Details" view.

Prior research suggests that an effective approach involves developing an expandable grid representing one dimension per axis and incorporating color to represent a third dimension [38, 39]. Reeder et al. also found that juxtaposing two dimensions on a single axis was confusing to users [39]. The current full iOS privacy labels represent the two dimensions of data using a list, which did not work well with users [44]. In our design, we arranged data type and purpose along the X and Y axes, while employing color as the third dimension.

Reeder et al. also noted that despite multiple cues in the P3P Expandable Grid, 14.5% of participants did not seem to notice that they could expand the grid [39], a problem we also encountered with 20% of our participants failing to expand rows. On the other hand, Zhang et al. observed that iOS label users expected interactive privacy labels on mobile screens, and were disappointed when they could not tap on the label to access privacy choices or additional information [44]. 91% of users in our study did engage with the interactive components of our labels. This interaction seems to facilitate user comprehension and enhances usability of the labels.

Treatment participants performed significantly better and more quickly than control participants when the particular data collection practices they were looking for were not

present within the labels. In such scenarios, they could spot the gray-colored squares that effectively signaled the absence of certain data practices. This aids users in swiftly identifying apps not collecting certain types of data at a glance. Additionally, our correlation analysis indicates a notably stronger negative correlation between participants' confidence levels and errors in the treatment condition (namely, treatment participants answered more correctly and more confidently)

6.2 How To Improve the Treatment

Introducing Users to Row Expansion and Legend. Many of the treatment errors were attributed to users not expanding rows, as noted above. We did not provide any training to help users become familiar with the labels in either condition. It would be beneficial for the interface to include a quick integrated tutorial or animated cues to help users understand the legend and the row expansion, which could improve accuracy.

Addressing Accessibility Concerns. Considering that our treatment prototype relies on color coding, there is an accessibility issue for individuals who are color blind. To mitigate this concern, we carefully selected colors that are accessible for people with various color vision conditions except monochromacy. We recognize the limitations of relying solely on color and future research could explore the integration of dot or stripe patterns and other features to further enhance clarity and accessibility. Additionally, the use of a grid may also raise further accessibility concerns for individuals with visual impairments, including those who have low vision. These elements may be difficult to handle for screen readers, which are tools commonly used by visually impaired users. We note that the current version of the label deployed in the app store is also tedious to navigate with a screen reader. While our results suggest that our proposed design could help many users, addressing the needs of the visually impaired community when it comes to benefiting from privacy labels will require more work.

Improving Terminology. It is also worth noting that our treatment labels did not address the issue of confusing terminology, a pain point identified by participants in prior studies [27, 44]. This decision was deliberate, because we believe that rectifying this problem necessitates a systematic and comprehensive approach to identifying more intuitive terminology. Our findings also provide further evidence of the confusion created by some of the terms used in existing privacy labels, especially when it comes to Apple's definition of tracking. Not only did participants fail to answer the questions regarding tracking correctly, their interaction with the information icons echoed the same trend. The interactive information icon for "tracking" received the highest number of taps at 76. In contrast, other purpose terms such as "analytics" and "third-party advertising," which were also included in the comprehension questions, each received under 20 taps.

In addition, our results suggest that the terminology used to refer to some top level categories of data types is also unintuitive, with users struggling to identify the top-level data category for some data types (e.g., "Photos and Videos" falling under "User Content"). The information icons for the two "Other" terms (one for purpose and one for data type), also attracted a great number of clicks from our participants, indicating participants' need for additional information. This aligns with previous research findings [44], indicating that participants expressed confusion when encountering terms in the label associated with other data types or other purpose. Further research will be needed to address these issues.

6.3 Future Directions

Comparing App Labels. Ultimately, we believe that privacy interfaces should be designed to empower users to readily compare the data practices associated with similar apps such as two apps in the same category. We believe that the tabular format presented in this paper will naturally lend itself to a comparison interface that can highlight cells where two apps have diverging data practices, allowing the user to quickly zoom in on key differences. Future work could also explore ways to use the proposed grid layout to highlight practices that are atypical of similar apps in the app store.

Personalized Label Presentations. While the grid format in this study is clearly improving usability, privacy labels remain complex. A further opportunity to enhance usability might involve exploring personalized presentations of privacy labels, as has been prototyped for IoT labels [14], letting users choose which practices interest them and which they don't care about. Such an approach could also benefit from the use of machine learning to assist users in making these selections (e.g., [30, 31, 41, 43]). A tabular format similar to the one evaluated in this study could be adapted to highlight data practices of interest or highlight practices that are likely to deviate from the user's expectations (e.g., [34]).

7 Conclusion

We propose an expandable-grid-based privacy label designed to improve the usability and mobile app privacy communication over current iOS labels. Our between-subjects study with 200 Prolific participants shows significant user improvement in answering privacy questions more accurately and faster. We believe that our redesign contributes to better informing consumers about the privacy implications of their future app downloads. We hope that this research will inform the design of more effective mobile app privacy labels and the development of effective privacy labels in other domains such as websites and IoT devices.

Acknowledgments

This research has been supported in part by grants from the National Science Foundation (grant CNS-1801316, grant CNS-1914486, grant CNS-2207216, and grant CCF-1852260), an unrestricted research grant from Google under its "privacy-related faculty award" program, and a gift from Innovators Network Foundation. We thank the students who contributed to the pilot interviews in their class project. These individuals include Terren Gurule, Oliver Marguleas, Alex Qiu, Ziping Song, and Cameron Wu.

References

[1] Hazim Almuhimedi, Florian Schaub, Norman Sadeh, Idris Adjerid, Alessandro Acquisti, Joshua Gluck, Lorrie Faith Cranor, and Yuvraj Agarwal. Your location has been shared 5,398 times! A field study on mobile app privacy nudging. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15)*, pages 787–796, 2015.

[2] David G. Balash, Mir Masood Ali, Xiaoyuan Wu, Chris Kanich, and Adam J. Aviv. Longitudinal analysis of privacy labels in the apple app store, 2023.

[3] Rebecca Balebako, Jaeyeon Jung, Wei Lu, Lorrie Faith Cranor, and Carolyn Nguyen. "Little brothers watching you": Raising awareness of data leaks on smartphones. In *Proceedings of the Ninth Symposium on Usable Privacy and Security*, SOUPS '13, New York, NY, USA, 2013. Association for Computing Machinery.

[4] Jan Lauren Boyles, Aaron Smith, and Mary Madden. Privacy and data management on mobile devices. *Pew Internet & American Life Project*, 4:1–19, 2012.

[5] Rex Chen, Fei Fang, Thomas Norton, Aleecia M McDonald, and Norman Sadeh. Fighting the fog: Evaluating the clarity of privacy disclosures in the age of CCPA. In *Proceedings of the 20th Workshop on Workshop on Privacy in the Electronic Society*, pages 73–102, 2021.

[6] Jacqui Cheng. Ars reviews iOS 4: What's new, notable, and what needs work. https://arstechnica.com/gadgets/2010/06/ars-reviews-ios-4-whats-new-and-notable/7/, Jun 2010.

[7] Jacqui Cheng. Review: iOS 6 gets the spit and polish treatment. https://arstechnica.com/gadgets/2012/09/review-ios-6-gets-the-spit-and-polish-treatment/, Sep 2012.

[8] Victoria Clarke, Virginia Braun, and Nikki Hayfield. Thematic analysis. In *Qualitative psychology: A practical guide to research methods*, page 248. SAGE, 3rd edition, 2015.

[9] Jessica Colnago, Yuanyuan Feng, Tharangini Palanivel, Sarah Pearman, Megan Ung, Alessandro Acquisti, Lorrie Faith Cranor, and Norman Sadeh. Informing the design of a personalized privacy assistant for the internet of things. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–13, 2020.

[10] Lorrie Faith Cranor. Mobile-app privacy nutrition labels missing key ingredients for success. *Commun. ACM*, 65(11):26–28, oct 2022.

[11] Cybersecurity and Infrastructure Security Agency. Privacy and mobile device apps. https://www.cisa.gov/news-events/news/privacy-and-mobile-device-apps. Accessed: 2024-02-14.

[12] Anupam Das, Martin Degeling, Daniel Smullen, and Norman Sadeh. Personalized privacy assistants for the internet of things: Providing users with notice and choice. *IEEE Pervasive Computing*, 17(3):35–46, 2018.

[13] Pardis Emami-Naeini, Yuvraj Agarwal, Lorrie Faith Cranor, and Hanan Hibshi. Ask the experts: What should be on an IoT privacy and security label? In *2020 IEEE Symposium on Security and Privacy (SP)*, pages 447–464. IEEE, 2020.

[14] Pardis Emami-Naeini, Janarth Dheenadhayalan, Yuvraj Agarwal, and Lorrie Faith Cranor. An informative security and privacy "nutrition" label for internet of things devices. *IEEE Security & Privacy*, 20(2):31–39, 2022.

[15] Jennifer Fereday and Eimear Muir-Cochrane. Demonstrating rigor using thematic analysis: A hybrid approach of inductive and deductive coding and theme development. *International Journal of Qualitative Methods*, 5(1):80–92, 2006.

[16] U.S. Food and Drug Administration (FDA). The nutrition facts label. https://www.fda.gov/food/nutrition-education-resources-materials/nutrition-facts-label. Accessed: 2024-04-21.

[17] Huiqing Fu, Yulong Yang, Nileema Shingte, Janne Lindqvist, and Marco Gruteser. A field study of run-time location access disclosures on android smartphones. In *Symposium on Usable Security and Privacy (USEC) 2023*, Feb 2014.

[18] Hamza Harkous, Kassem Fawaz, Remi Lebret, Florian Schaub, Kang G Shin, and Karl Aberer. Polisis: Automated analysis and presentation of privacy policies using deep learning. *arXiv preprint arXiv:1802.02561*, 2018.

[19] Patrick Gage Kelley, Joanna Bresee, Lorrie Faith Cranor, and Robert W Reeder. A "nutrition label" for privacy. In *Proceedings of the 5th Symposium on Usable Privacy and Security*, pages 1–12, 2009.

[20] Patrick Gage Kelley, Lucian Cesca, Joanna Bresee, and Lorrie Faith Cranor. Standardizing privacy notices: An online study of the nutrition label approach. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '10, page 1573–1582, New York, NY, USA, 2010. Association for Computing Machinery.

[21] Patrick Gage Kelley, Lorrie Faith Cranor, and Norman Sadeh. Privacy as part of the app decision-making process. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 3393–3402, 2013.

[22] MG Kendell. Rank correlation methods. *Charles Griffin and Company: London*, 1955.

[23] Rishabh Khandelwal, Asmit Nayak, Paul Chung, and Kassem Fawaz. Comparing privacy labels of applications in android and iOS. In *Proceedings of the 22nd Workshop on Privacy in the Electronic Society*, WPES '23, page 61–73, New York, NY, USA, 2023. Association for Computing Machinery.

[24] Kleimann Communication Group. Evolution of a Prototype Financial Privacy Notice: A Report on the Form Development Project, September 2010. [Online; posted 13-September-2012].

[25] Simon Koch, Malte Wessels, Benjamin Altpeter, Madita Olvermann, and Martin Johns. Keeping privacy labels honest. *Proc. Priv. Enhancing Technol.*, 2022(4):486–506, 2022.

[26] Konrad Kollnig, Anastasia Shuba, Max Van Kleek, Reuben Binns, and Nigel Shadbolt. Goodbye tracking? Impact of iOS app tracking transparency and privacy labels. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 508–520, New York, NY, USA, 2022. Association for Computing Machinery.

[27] Tianshi Li, Kayla Reiman, Yuvraj Agarwal, Lorrie Faith Cranor, and Jason I. Hong. Understanding challenges for developers to create accurate privacy nutrition labels. In *CHI Conference on Human Factors in Computing Systems*, CHI '22, New York, NY, USA, 2022. Association for Computing Machinery.

[28] Jialiu Lin, Bin Liu, Norman Sadeh, and Jason I. Hong. Modeling users' mobile app privacy preferences: Restoring usability in a sea of permission settings. In *Proceedings of the Tenth Symposium on Usable Privacy and Security (SOUPS '14)*, pages 199–212, 2014.

[29] Yanzi Lin, Jaideep Juneja, Eleanor Birrell, and Lorrie Faith Cranor. Data safety vs. app privacy: Comparing the usability of android and ios privacy labels. In *Proc. Priv. Enhancing Technol.*, pages 182–210, 2024.

[30] Bin Liu, Mads Schaarup Andersen, Florian Schaub, Hazim Almuhimedi, Shikun (Aerin) Zhang, Norman Sadeh, Yuvraj Agarwal, and Alessandro Acquisti. Follow my recommendations: A personalized privacy assistant for mobile app permissions. In *Twelfth Symposium on Usable Privacy and Security (SOUPS '16)*, pages 27–41, 2016.

[31] Bin Liu, Jialiu Lin, and Norman Sadeh. Reconciling mobile app privacy and usability on smartphones: Could user privacy profiles help? In *Proceedings of the 23rd International Conference on World Wide Web (WWW '14)*, pages 201–212, New York, NY, USA, 2014.

[32] Aleecia M McDonald and Lorrie Faith Cranor. The cost of reading privacy policies. *I/S: A Journal of Law and Policy for the Information Society*, 4:543, 2008.

[33] J.H. McDonald and University of Delaware. *Handbook of Biological Statistics*. Sparky House Publishing, 2009.

[34] Ashwini Rao, Florian Schaub, Norman Sadeh, Alessandro Acquisti, and Ruogu Kang. Expecting the unexpected: Understanding mismatched privacy expectations online. In *Twelfth Symposium on Usable Privacy and Security (SOUPS 2016)*, pages 77–96, Denver, CO, June 2016. USENIX Association.

[35] Abhilasha Ravichander, Alan W Black, Thomas Norton, Shomir Wilson, and Norman Sadeh. Breaking down walls of text: How can NLP benefit consumer privacy? In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4125–4140, 2021.

[36] Abhilasha Ravichander, Alan W Black, Shomir Wilson, Thomas Norton, and Norman Sadeh. Question answering for privacy policies: Combining computational and legal perspectives. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference*

on Natural Language Processing (EMNLP-IJCNLP), pages 4949–4959, Hong Kong, China, November 2019. Association for Computational Linguistics.

[37] Robert W. Reeder, Lujo Bauer, Lorrie F. Cranor, Michael K. Reiter, and Kami Vaniea. More than skin deep: Measuring effects of the underlying model on access-control system usability. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '11, page 2065–2074, New York, NY, USA, 2011. Association for Computing Machinery.

[38] Robert W. Reeder, Lujo Bauer, Lorrie Faith Cranor, Michael K. Reiter, Kelli Bacon, Keisha How, and Heather Strong. Expandable grids for visualizing and authoring computer security policies. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '08, page 1473–1482, New York, NY, USA, 2008. Association for Computing Machinery.

[39] Robert W. Reeder, Patrick Gage Kelley, Aleecia M. McDonald, and Lorrie Faith Cranor. A user study of the expandable grid applied to P3P privacy policy visualization. In *Proceedings of the 7th ACM Workshop on Privacy in the Electronic Society*, WPES '08, page 45–54, New York, NY, USA, 2008. Association for Computing Machinery.

[40] Joel R Reidenberg, Travis Breaux, Lorrie Faith Cranor, Brian French, Amanda Grannis, James T Graves, Fei Liu, Aleecia McDonald, Thomas B Norton, and Rohan Ramanath. Disagreeable privacy policies: Mismatches between meaning and users' understanding. *Berkeley Tech. LJ*, 30:39, 2015.

[41] Daniel Smullen, Yuanyuan Feng, Shikun Zhang, and Norman M. Sadeh. The best of both worlds: Mitigating trade-offs between accuracy and user burden in capturing mobile app privacy preferences. *Proc. Priv. Enhancing Technol.*, 2020(1):195–215, 2020.

[42] Primal Wijesekera, Joel Reardon, Irwin Reyes, Lynn Tsai, Jung-Wei Chen, Nathan Good, David Wagner, Konstantin Beznosov, and Serge Egelman. Contextualizing privacy decisions for better prediction (and protection). In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*, pages 1–13, 2018.

[43] Shikun Zhang, Yuanyuan Feng, Anupam Das, Lujo Bauer, Lorrie Cranor, and Norman Sadeh. "Did you know this camera tracks your mood?": Understanding privacy expectations and preferences in the age of video analytics. *Proc. Priv. Enhancing Technol.*, 2021(2):282–304, 2021.

[44] Shikun Zhang, Yuanyuan Feng, Yaxing Yao, Lorrie Faith Cranor, and Norman Sadeh. How usable are iOS app privacy labels. *Proc. Priv. Enhancing Technol.*, 2022(4):204–228, 2022.

[45] Shikun Zhang and Norman Sadeh. Do privacy labels answer users' privacy questions? In *Symposium on Usable Security and Privacy (USEC) 2023*, Feb 2023.

A Supplemental Tables and Figures

Gender		Age		Ethnicity	
Female	50.0%	18–25	21.5%	Asian	11.0%
Male	50.0%	26–35	34.0%	African American	7.0%
		36–45	23.5%	Caucasian	71.5%
		46–55	10.0%	Mixed	6.0%
		56–65	8.0%	Other	3.5%
		66+	3.0%	No data	1.0%

Table 3: Demographics of our study participants $N = 200$

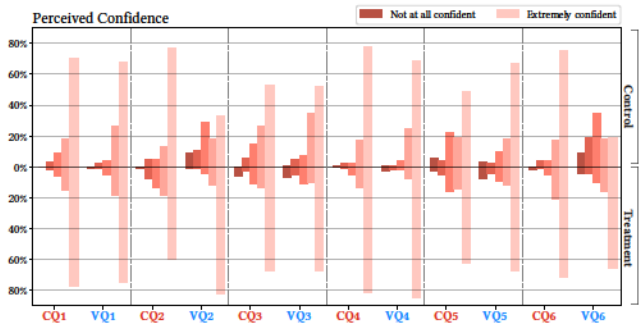


Figure 6: Distribution of participant confidence in their answers to 12 questions across both conditions. The top half represents the control group, while the bottom half shows the treatment group.

Question	Control %	Treatment %	Effect Size	Size per Condition
CQ1	0.35	0.74	0.82	26
VQ1	0	0.79	2.66	4
CQ2	1	0.74	0.82	26
VQ2	0.2	0.89	1.79	7
CQ3	0.65	0.63	0.04	10276
VQ3	0.55	0.68	0.27	227
CQ4	0.25	0.89	1.58	8
VQ4	0.45	0.95	1.26	12
CQ5	0.55	0.68	0.27	227
VQ5	0.7	0.74	0.08	2570
CQ6	1	0.6	1.10	15
VQ6	0.7	0.9	0.49	70

Table 4: A Priori Sample Size for 12 Questions Based on Pilot Results using G*Power

B Survey Text

Consent Form

- I am at least 18 years of age.
- I have read and understand the consent information above.
- I want to participate in the research and continue with the survey.

Introduction This survey is being conducted for research at Carnegie Mellon University. We will ask you to view two websites on your iPhone and answer questions about them. This survey should take about 20 minutes to complete. You will receive your compensation via Prolific upon completion of the study. To participate in this survey, you must use an iPhone with iOS 14 and above and have access to your iPhone throughout the duration of the survey. We recommend that you take this survey on a desktop, laptop, tablet, or other device besides your iPhone. Your participation is voluntary. Please do not reveal any private or personally-identifiable information about yourself or others during the survey.

- What is your Prolific ID? Please note that the text box should auto-fill with the correct Prolific ID.

General Questions about Privacy Label and Apps. Please open the first link by scanning the QR code below with your iPhone's camera. If you are unable to scan the QR code, you cannot participate in this study and you will not get paid.

Please scroll down and view the App Privacy section of this page so that the privacy label is visible on your screen. We are going to ask you a few questions about this section, so please explore the label before continuing the survey.

- Have you seen an iOS app privacy label like this before?
- (Follow-up Yes) How often do you check privacy labels before downloading an app?
- Was privacy ever a reason you decided to not download or stop using an app?

To answer these questions, you will have to interact with the privacy label. Feel free to explore the label for as long as you would like before answering the following questions.

App Comprehension Questions

Q1: Does this app collect data for Analytics purposes and, if so, what data? (Select all that apply)

- Browsing History
- Contact Info
- Contacts
- Diagnostics
- Financial Info
- Health & Fitness
- Identifiers
- Location
- Other Data
- Purchases
- Search History
- Sensitive Info
- User Content
- Usage Data
- This app does not collect data used to track you or for tracking purposes
- I'm not sure

Q2: Does this app collect location data for Third-Party Ads/Advertising purposes?

- It collects precise location for Third-Party Ads purposes
- It collects coarse location for Third-Party Ads purposes
- It collects both precise and coarse location for Third-Party Ads purposes
- It does not collect location for Third-Party Ads purposes
- I'm not sure

Q3: Does this app collect Photo and Video data and, if so, for what purpose(s)? (Select all that apply)

- Analytics
- Developer Ads
- Functionality
- Other
- Personalization
- Third-Party Ads
- Tracking or Data Used to Track You
- This app does not collect [photo and video] data for any purpose
- I'm not sure

Q4: Does this app collect Purchases data and, if so, for what purpose(s)? (Select all that apply) [answers same as Q3]

Q5: Does this app link Diagnostics data to your identity?

- Yes
- No
- I'm not sure

Q6: Does this app collect Data Used to Track You or for Tracking purposes and, if so, what data? (Select all that apply) [answer choices same as Q1]

[For each of the 6 questions above, we asked the following 2 questions]

- What helped you to arrive at this answer? [short response]
- How confident do you feel that the answers you gave about the information on the privacy label are correct? Completely confident (5) to Not at all confident (1).

Second app prompt We will now ask you to complete the same questions for a second app. You can access the second link by scanning the QR code below with your iPhone's camera. If you are unable to scan the QR code, you cannot participate in this study and you will not get paid. Please make sure that you scroll down to the App Privacy section so that the privacy label is visible.

[Repeat Q1 to Q6 for the 2nd app]

Treatment only Questions

QT1: Using the screenshots below, which app collects Diagnostics data and links it to your identity for any purpose? [Candy label only[4]] [Venmo label only[4]]

- App A collects Diagnostics data and links it to your identity
- App B collects Diagnostics data and links it to your identity
- Both apps collect Diagnostics data and link it to your identity
- Neither app collects Diagnostics data and links it to your identity
- I'm not sure

QT2: Using the screenshots below, which app collects more Usage Data for Analytics purposes?

[Candy label only[4]] [Venmo label only[4]]

- App A collects more Usage Data for Analytics purposes
- App B collects more Usage Data for Analytics purposes
- Both apps collect the same amount of Usage Data for Analytics purposes
- Neither app collects Usage Data for Analytics purposes
- I'm not sure

QT3: How useful were the colors in the grid as you answered the questions above?

- Very useful
- Moderately useful
- Somewhat useful
- A little useful
- Not at all useful

QT4: Did you notice the legend? If so, did you use it?

- Yes I noticed it and used it as I answered the questions
- Yes I noticed it but did not use it to answer the questions
- I'm not sure if I saw it
- No I did not notice it
- Other [short response]

Term Definition Questions

QTD1: What does "data linked to you" mean?

- Data that is transferred when you use an app and stored in a database
- Data from your account or device that could be used to identify you
- Data that is used to track you and your activity while using the app
- Information you've given during the sign-up process of an app
- Data that includes your real name, or phone number, or address
- I'm not sure
- Other [short response]

QTD2: What does "data not linked to you" mean?

- Data that is not personal information, but could be used to determine information about you
- Contact information, such as an email address or phone number
- Data not connected to you, even if it is collected by the app
- Data that developers can use to identify you, but is not shared with third parties
- Data that does not include your real name or location
- I'm not sure
- Other [short response]

QTD3: What does "data used to track you" mean?

- Identifiable data that is shared with third parties to personalize ads
- Your location and physical address are collected by the app
- Patterns of using an app, such as frequency or search history

[4] no legend

- Data sent to third parties only for security purposes
- I'm not sure
- Other [short response]

QTD4: If an app shared your location and email address with third party advertisers, do you think that would be considered "tracking"? Yes/No/I'm not sure

QTD5: If an app used your location to show you nearby stores, do you think that would be considered "tracking"? Yes/No/I'm not sure

General Perceptions

QGP1: How helpful did you find the privacy label to be?

QGP2: Generally, how easy or difficult was it to understand the privacy labels? Very Easy (1) to Very Difficult (5)

QGP3: Please rate how easy each element of the privacy label is to understand on a scale from Very Easy (1) to Very Difficult (5). [matrix question]

- Terms used in the label
- Finding definitions of terms used in the label
- Icons [control] / Colors [treatment]
- Label structure

QGP4: Was any part of the label confusing, and if so, please explain. [short answer]

QGP5: Do you think this privacy label provides enough information about how an app collects and uses your data? Yes; No; I'm not sure; Other [Follow up if No] What information would you like to see added to the label, if any? [short answer]

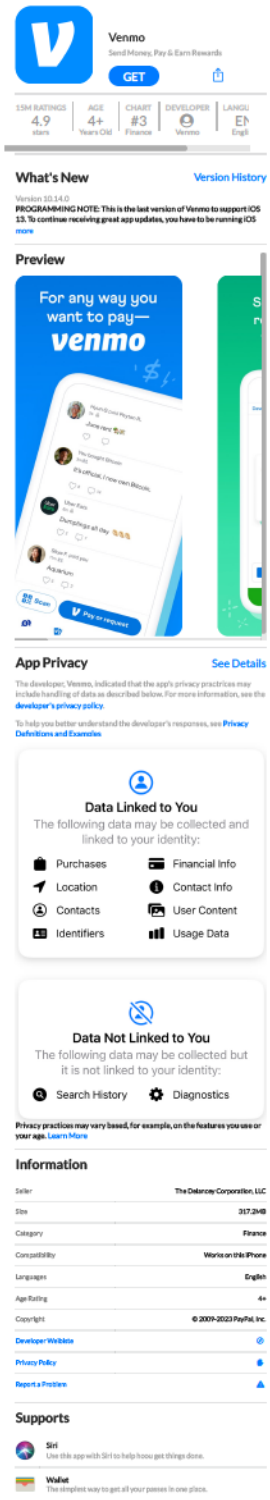
QGP6: If you have any suggestions for improving the privacy label, please provide them below.

QGP7: In the future, do you plan to look at these labels before deciding to download an app?

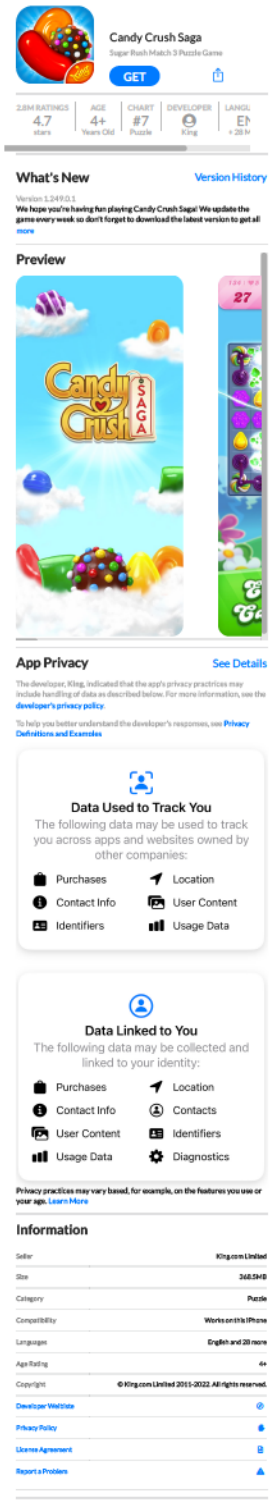
QGP8: Do you have any other comments or feedback regarding the privacy labels or the survey? [short answer]

Wrap-up You will receive payment on Prolific for completing this survey. We thank you for your time spent taking this survey. Your response has been recorded.

C Study Screenshots



(a) Control Venmo



(b) Control Candy

(c) Treatment Venmo

(d) Treatment Candy

Figure 7: Webpages shown to participants