Leveraging Crowd for Collecting and Maintaining Educational Resources for Privacy Learning

Wendy Hui WANG\textsuperscript{1}, Yu TAO\textsuperscript{1,*}, Kai WANG\textsuperscript{2}, Dominik JEDRUSZCZAK\textsuperscript{1} and Benjamin KNUTSON\textsuperscript{1}

\textsuperscript{1}Stevens Institute of Technology, United States
\textsuperscript{2}Kean University, United States
*Corresponding author

Keywords: Crowdsourcing, Educational game, Online privacy.

Abstract. As the Internet grows in importance, it is vital to develop methods and techniques for educating end-users to improve their awareness of online privacy. Web-based education tools have been proven effective in many domains and have been increasingly adopted by many online professional and educational services. However, the design and development of Web-based education tools for online privacy is still in its early stage. The traditional solutions always involve privacy experts who have sophisticated expertise. However, it is not clear how inspiring and effective these education tools are to general users of varying backgrounds. Furthermore, such involvement can make the tool development costly. In this paper, we discuss our experiences of designing a quiz-based privacy learning tool by leveraging the wisdom of the crowd on Amazon Mechanical Turk for two purposes: (1) designing the quiz questions for the tool; and (2) evaluating and assessing the effectiveness of the tool. Empirical study demonstrates that the crowd can provide high-quality educational materials for privacy learning.

Introduction

Online social networking communities have undergone an explosion in recent years, as both the kinds and numbers of sites have grown and their memberships increased. This has led to the proliferation of personal data revealed by users of online communities, which presents a variety of privacy risks. Internet users still know little about online privacy [4], even though their awareness of online privacy has increased [15]. However, the existing privacy education tools are either too technical or impractical [25]. Therefore, there is a need for effective privacy education tools for Internet users of all ages and backgrounds.

In this paper, we consider the privacy education tool that takes the format of online quizzes. By answering questions and studying the correct solutions, quiz-based learning provides effective, interactive experiences that motivate users to actively engage in the learning process. One of the challenges of developing an effective privacy learning quiz is the design of quiz questions. Quiz questions that are either too easy or too difficult can make the users lose interests easily. Traditionally the quiz questions are designed by the privacy experts who have sophisticated expertise and broad experience. However, it is not clear how motivating, inspiring, and/or effective these quiz questions are to general users of varying backgrounds, especially to novice users who have rarely dealt with privacy issues before. Furthermore, the involvement of professionals can make the quiz development costly. Facilitated by advances in web-related technologies, crowdsourcing has become a new technique for practicing open innovation (e.g., T-shirt design [10] and software idea competition [22]). It has shown that crowdsourcing indeed can perform some creative tasks with reduced production costs [23, 32]. In this paper, rather than relying on the privacy experts, we leverage the crowd on Amazon Mechanical Turk (AMT) [1] to design quiz questions for privacy education. AMT is an online labor market where employees (called workers) are recruited by employers (called requesters) for the
execution of tasks (called HITs, acronym for Human Intelligence Tasks) in exchange for a wage (called a reward).

There are several challenges of utilizing AMT for the design of quiz questions for privacy education. First, AMT typically supports microtasks that are simple, fast-paced, and require the least amount of training. It is not clear whether the crowd on AMT can perform the task such as quiz question design that is complex, time-consuming, and requires some background knowledge in online privacy. Second, the AMT workers have various backgrounds and knowledge of online privacy. They may create quiz questions that are dramatically different in terms of both content and format. It is challenging to aggregate these ideas of significant individual heterogeneity. Third, it is unclear how to evaluate the quality of the quiz questions, as well as the effectiveness and educational impact of these quiz questions, with quantitative measurements.

To address these challenges, we develop a prototype of CERPA for crowdsourcing educational resources for privacy learning. The design of CERPA consists of four phases:

- **Idea collection phase**: CERPA collects a set of quiz questions from a crowd of non-experts on AMT;
- **Quiz development phase**: the privacy education quiz is developed based on the questions obtained by the Idea collection phase; and
- **Quiz evaluation phase**: the developed quiz is released to both AMT and privacy experts for effectiveness evaluation. The evaluation feedback information is then used to further improve the quiz.

To our best knowledge, we are the first who exploit crowd innovation for the design of privacy education tools. Our contributions include the following. First, we design novel metrics to measure the quality score of the workers’ ideas in terms of both correctness and novelty. Correctness measures whether the quiz questions designed by the workers are indeed suitable for the given privacy setting in the quiz, while novelty measures whether the ideas are dissimilar to the others. A novel quiz question should be dissimilar to a significant portion of the other questions. Similar quiz questions are likely make the users lose the interests of the quiz. Therefore, intuitively, a correct and novel idea receives a high quality score.

Second, we request AMT workers to rank the workers’ quiz questions in terms of their educational impact. Since the ranking is subjective, it is expected that there exists inconsistency among the workers’ preferences (e.g., worker A prefers idea 1 to idea 2, while worker B prefers idea 2 to idea 1). We design an efficient algorithm to aggregate the workers’ rankings that are potentially inconsistent. We combine the ranking by educational impact and the ranking by the idea’s quality score into a single ranking, and choose a number of ideas (i.e., quiz questions) from the top of the ranking.

Third, based on the chosen top ideas, we implement a prototype of the privacy education quiz, and leverage both privacy experts and a large crowd of non-experts on AMT to evaluate the quality of the quiz in terms of its usability and educational impact. The evaluation results show that the quiz questions designed by crowdsourcing are indeed high-quality and creative. This convinces that the crowd can be leveraged for collecting and maintaining educational resources for privacy learning.

The paper is organized as following. First, we give a brief overview of the work related to using crowd for creative tasks, in the next section. Second, we define the design tasks for the crowd in the next section. Then we explain the details of the idea collection and quiz evaluation phases respectively. After that, we present our experience of using Amazon Mechanical Turk to design and evaluate CERPA, a prototype of privacy education quiz. Finally, we conclude the paper by summarizing our main findings and future work.

**Related Work**

Using the Internet and related technologies, crowdsourcing is able to access the brainpower of many people and achieve a wide range of outcomes, from the simple and mundane task of collecting
street addresses of companies, to more sophisticated achievement, such as Wikipedia, innovation
competitions, and helping solve cutting-edge scientific problems [14, 19]. One type of tasks that are
popular in the field of computer science is to combine human-generated results with computer
algorithms to achieve synergistic outcomes [19]. For example, crowds’ input has been used to
improve automatic extraction of concepts in texts [13] and to improve search results from search
engines [9]. This type of crowdsourcing builds the foundation for more advanced techniques for
large scale human-computer collaboration.

Another popular type of crowdsourcing is to give creative tasks to a crowd. In this type, instead
of having people generate some close-ended answers, people are explicitly told to generate novel
outcomes. Yu and Nickerson [36] conducted studies in which a crowd developed chair design
sketches and successive crowds combined the designs of previous crowds, resembling genetic
algorithm in computer science. The authors showed that crowd members tended to integrate novel
and practical features of design, which helped improve creativity [37]. In another study, crowd-
based genetic algorithm was used to generate graphical advertisements of an online game. The
results showed that having crowds modifying previous generation of ads generated better ads than
having crowds combining previous ads [30].

In addition to making design sketches, textual idea generation is also a popular creative task in
crowdsourcing research. An important question is whether crowds are able to generate ideas of
similar quality as professionals. Poetz and Schreier [28] show that product ideas from an idea
contest with customers have higher novelty and customer benefit than professionals’ ideas, although
the idea feasibility is somewhat lower. Some top ideas that the executives like are from customers,
or crowds. Therefore, it is possible for crowds to generate useful ideas, especially if they are the end
users of the products. Some specific technique in crowdsourcing idea generation show their
effectiveness in improving idea creativity, such as deliberately finding source of analogies from
other web-sites [35] and decomposing the initial creative task into sub-problems [24].

In addition to idea generation, idea evaluation is also an important topic in crowdsourcing
creative tasks. It is found that using multiple scales (e.g., novelty, value, feasibility) to measure idea
quality is beneficial if ideas are not long [31]. If ideas are long, a single scale measurement (only
idea quality) may lead to more accurate evaluation. The study also finds that having 20 ratings for
an idea generates accurate evaluation. In another study of crowd-generated ideas, it is found that
prediction voting (predicting whether an idea can win the competition) is more appropriate when
many poor ideas need to be filtered out, while Likert scale ratings are more appropriate when more
refined distinctions need to be made for ideas that are of reasonable quality [5].

**Preliminaries**

**Privacy Domain**

Internet privacy is a broad term that refers to a variety of factors, techniques and technologies used
to protect sensitive and private data. An education quiz that covers such a broad concept is expected
to be complicated and cumbersome. Therefore, in this paper, we narrow down the privacy domain
to a specific issue, privacy issues of social networking sites. We pick this issue due to the following
two reasons. First, a key requirement of a good educational quiz is to create an effective learning
environment that users are familiar with [29]. According to a recent survey, 74% of online adults
use social networking sites [2]. Therefore, a quiz that focuses on privacy issues of social networking
sites should enable general users to enjoy the quiz by leveraging their real-life social networking
experience. This also should enable general users to get highly engaged in the learning process and
can easily transfer what they learned from the quiz to their real-life. Second, the design of quiz
questions requires the designers to be equipped with knowledge and/or related experience. Normally,
common AMT workers are not experts in privacy. However, given the popularity of
social networks, we expect that non-expert crowd workers still can contribute high-quality quiz
questions by leveraging their real-life experiences of using social networking sites.
Privacy Attributes

Privacy can be viewed from many different perspectives, including political policies, the rights of citizens, and protection for consumers [6]. Informally, any fact about an individual can be treated as a personal *private attribute* [11]. Many personal private attributes, for example, age, gender, sexual orientation, ethnicity, religious and political views are considered highly sensitive. In this paper, we are in particular interested in protecting privacy attributes when sharing information on social networking sites.

Idea Collection

In the *Idea collection* phase, we generated Human Intelligence Task (HIT) on Amazon Mechanical Turk (AMT), asking workers to contribute the quiz questions that aim at online privacy education for social networking. In particular, the workers are asked to provide at least three example postings on social networking sites (e.g., Facebook) that expose private and sensitive aspects of one’s personal life. The workers are also asked to provide some suggestions of how to fix these postings to remove the privacy information. These example postings and the suggested solutions will be implemented as the questions in the privacy education quiz. Figure 1 (a) shows the HIT for the idea collection task. To help workers understand the task, we also included an example of social network posting that reveals some private information (the phone number in this example), as well as two suggestions of how to fix this posting to remove the privacy information. As a part of HIT description, we also collected the workers’ demographic information, including their locations, gender, and education level. We also surveyed their background knowledge in online privacy, including: (1) how often do they use Facebook; (2) whether they have adequately protected their own private information on Facebook; and (3) how much do they know about online privacy. We collect these information aiming to find out the correlation between the quality of ideas and the workers’ background. More details of our correlation study can be found in the Results section.

Quiz Development

We performed the idea collection phase on AMT. Then we aggregated the collected ideas, ranked them, and picked sixteen ideas. We implemented a prototype of Web-based privacy education quiz CERPA, with the sixteen picked ideas implemented as Facebook-style posting examples in the quiz. CERPA was implemented by using MySQL, Javascript, and HTML5.
To ensure that the quiz can be accepted by the general public of a wide age range, we designed four different characters in the quiz. Each character is associated with a set of privacy topics. The characters and their associated privacy attributes are:

- **Sophia** (high school student, age \( \in [16, 18] \)): ID, credit cards, driver license, and academic behavior;
- **Aiden** (college student, age \( \in [19, 24] \)): home address, academic behavior, medical information, and personal emotions;
- **Olivia** (young adult, age \( \in [25, 30] \)): professional issues;
- **Lucas** (older adult, age > 30): information of family members, personal ideology, and medical information.

We design the following character-based quiz rules. The players can pick any of the four aforementioned characters to start the quiz. Each character has five or six online postings, each posting revealing some private and sensitive aspects of one’s personal life. Each posting is also associated with three possible fix solutions. Most of the example postings and the fix solutions in the quiz are generated from the crowds’ ideas (collected by *Idea Collection* Phase). The players have to: (1) decide whether these postings have privacy issues, and what these privacy issues are, if there is any; and (2) what is the best fix solution in terms of balanced privacy and data availability. Users’ different decision of where the privacy problem occurs is assigned with a different performance score. An example of the postings is shown in Figure 3 (a). Besides showing the scores, the quiz gives the players the feedback of their choices. The feedback advises the players why their choices are (in)correct, and what are the real privacy problems of the postings. An example of the feedback is shown in Figure 3 (b). The feedback was designed by the privacy experts. After the players finishing judging the privacy problems of the postings, the quiz shows three to four fix solutions. An example of the fix solutions is shown in Figure 3 (c). The best solution is the one that best addresses the trade-off between privacy and data availability. The players pick one solution. Each solution has a score, depending on how good privacy and data availability is balanced. For instance, consider the four solutions shown in Figure 3 (c). Removing the image alone (Option A) or only the specific medical condition in the text (Option B) cannot protect the privacy entirely. Therefore, the only correct solution is to remove the post. The quiz also
gives the feedback to the players of their picked fix solutions. The feedback states clearly why the picked solution can(not) remove the privacy leak sufficiently.

**Quiz Evaluation**

We generate Human Intelligence Task (HIT) on Amazon Mechanical Turk (AMT) to recruit workers to evaluate the developed quiz. The HIT takes the format of a survey. It evaluates two types of effectiveness of **CERPA** for privacy education and learning:

- **Tool usability**, which measures whether the tool (i.e., the quiz) is user friendly; and
- **Educational impact**, which measures the impacts of the tool on delivery and learning of online privacy knowledge.

There are twenty questions in the survey. Among these questions, fifteen questions are single-choice questions, and five questions are free-text format. The single-choice questions ask the users to select the number that best matches their experience. The numbers specifies a scale of 1 to 5, with 1 being strongly disagree, 2 being disagree, 3 being neutral, 4 being agree, and 5 being strongly agree. The free text-format questions ask the workers to enter their feedback and suggestions. Next, we discuss how we designed the questions in the survey.

**Evaluation of Tool Usability**

We used the usability testing method to evaluate tool usability. Usability tests, with a user-centered approach, have been successfully used in the development of other products such as interactive multimedia software and web-based learning tools [21, 34, 27]. The usability test model by [21] suggests to test: (1) Learnability (the ease of learning to use the system, e.g., clearly labeled components, easy navigation); (2) Performance effectiveness (the ease of using the system in terms of speed and error rate); (3) Flexibility (the level of freedom to use multiple commands to achieve the same goal); (4) Error tolerance and system integrity (the ability to recover from errors and prevent data corruption or loss); and (5) User satisfaction (users’ opinions and perception of the training tool). For each testing component, we designed corresponding questions in the HIT, asking for the crowd’s feedback.

**Evaluation of Educational Impact**

Kirkpatrick proposed a training evaluation model [18] that can objectively analyze the effectiveness and impact of education. We apply Kirkpatrick’s evaluation model to our evaluation phase. Kirkpatrick’s evaluation model measures the following four items.

**Level-1 Reaction Measurement**

In the survey, we ask the questions related to what the workers thought and felt about the quiz. For example, did they feel that the quiz was worth their time? What were the biggest strengths and weaknesses of the quiz? Did they like the topic, the materials, and the format? By measuring reaction, we will understand how well the quiz-based learning was received by the workers.

**Level-2 Learning Measurement**

To measure how much the workers’ knowledge of online privacy has increased as a result of the education quiz, we asked the workers in the survey whether their knowledge of online privacy has increased as a result of the playing the quiz.

**Level-3 Behavior Measurement**

We ask a series of questions regarding how the workers would apply the knowledge learned from the quiz. For example, will the workers put any of their learning of online privacy techniques to their practical use? Are they able to teach their new knowledge and skills to other people? The main focus is the behavioral changes following the quiz playing.
Level-4 Results Measurement

We analyze the effect of the quiz on the improved performance of the workers. Kirkpatrick suggested several key performance indicators, such as volumes, percentages, and other quantifiable aspects of performance, such as whether the number of privacy accidents has been reduced [18]. It is hard to collect such observable performance indicators in the survey. Therefore we did not ask such questions. Instead we asked the workers whether they would recommend this quiz to others for the learning on online privacy. There are twenty questions in the survey. We categorize all the questions into three components:

- Evaluation of quiz interface, including seven questions that cover the evaluation of tool usability;
- Evaluation of quiz content, including seven questions that cover the evaluation of both tool usability and educational impact; and
- Overall evaluation, including six questions that cover the evaluation of educational impact.

Results

In this section, we discuss the main findings of our experience of developing and evaluating CERPA.

Idea Aggregation

Sixteen workers participated in the idea collection phase. We collected the demographic information of these sixteen workers. All users are from US, with the mean of ages as 39. Regarding the education level, ten workers did not finish high school. Four finished high school, and two had college education. Ten workers are females and six were males. Regarding the background knowledge and experiences on social networking privacy, all workers use Facebook at least every month. Most of them considered themselves having protected their privacy on Facebook sufficiently. A majority of these workers considered themselves having intermediate level of privacy knowledge. To summarize, the participated workers are of various demographic backgrounds. They also have different knowledge and experiences in online privacy. We collected forty-eight examples from the workers. After aggregation, it turns out that these examples cover thirteen distinct privacy attributes. Figure 4 (a) shows these privacy attributes and their popularity (i.e., the number of examples that each privacy attribute is associated with). We observe that activity, home address, and personal emotions are the three privacy attributes that are of the highest popularity, while the well-known privacy attributes such as name, age, and birth date has fewer examples. This is surprising as the privacy concerns of revealing activities and personal emotions on social networking sites are paid much less attention as name and birth date [16]. This shows that the workers indeed intended to contribute creative ideas of privacy examples, as required in the HIT instructions. Second, the examples that belong to the same privacy attribute group indeed target at different Web user populations. For instance, in group of the privacy attribute location, one example described an adult’s posting showing his/her exercise routine at a specific gym, while another example mentioned the location of a kid’s favorite playground. We categorize the workers’ examples into four age populations: (1) kids and high school students (age < 18); (2) college students (age in [18, 24]); (3) young professionals (age in [25, 30]); and (4) older adults (age > 30). These four age populations lead to the design of four characters in our quiz prototype. We assign the examples to these four population groups, and count the number of examples in each group. The grouping of examples is not exclusive; some examples can be applied to multiple groups. Figure 4 (b) shows the results. It turned out that the older adult group received the highest number of privacy attributes. This may due to the fact that a majority of workers are in the same age range as the older adult group; they design the examples from their real-life experiences.
Idea Quality Evaluation

In this set of experiments, we measure: (1) both the correctness and novelty of the ideas collected by the Collection phase; and (2) the correlation between the quality of workers’ ideas and their privacy background knowledge and experiences.

Correctness & Novelty

First, we measure the correctness of the collected 48 posting examples and their suggested fix solutions. We define the correctness ratio as the portion of examples (out of 48 examples) whose correctness score is higher than a given threshold (in our setting is 0.75). Seven examples were evaluated as incorrect. All suggested fix solutions (of the correct examples) were correct. The correctness ratio is 0.84, which is sufficiently high. We analyzed the background of those workers who submitted incorrect examples. Most of them use Facebook often (40%, 40%, and 20% use Facebook every day, every week, and every month). The average self-evaluation of privacy scale of those workers who submitted incorrect examples is 2.8 (out of 5). This shows that the workers’ background knowledge indeed impact the correctness of their ideas.

Second, we measure the novelty of the ideas. We tried two different settings:

- **Open-ended** setting: the workers are asked to contribute the examples of any possible privacy attribute;
- **Closed-ended** setting: the workers are asked to contribute the examples of a given set of privacy attributes. In our setting we we narrow down to four specific privacy attributes: (1) medical information, (2) income, (3) work history, and (4) student records.

![Number of examples of each privacy attribute](image1)

![Number of examples for each population group](image2)

Figure 4. Analysis of Crowd-generated Ideas.

We choose different similarity thresholds (we use the same threshold value of δ and θ). Figure 5 shows the results for both the open-ended and closed-ended settings. All reported results are
computed as the average of novelty scores of all example postings. We have the following observations. First, the novelty reduces when the similarity threshold decreases. This is not surprising as higher similarity thresholds lead to fewer similar ideas, and thus higher novelty. Second, for both settings, the novelty is sufficiently high (at least 0.88). This shows that the crowd respected the creativity requirements in the HIT description, and generated high-novelty ideas. Third, the novelty of ideas generated in the closed-setting is higher than in the open-setting. This is not surprisingly as in the open-end setting, the workers incline to generate ideas on the popular privacy attributes. These ideas have high probability to be similar. But in the closed-end setting, the picked privacy attributes are those that the general public rarely pay attention to (as proven by the analysis of the open-end setting results in Section ). Ideas generated on those topics are less likely to collude and more likely to be dis-similar. We also asked the privacy experts to review the ideas from the crowd regarding the novelty. The privacy experts agreed that some privacy examples with high novelty scores were indeed new and urging. Given both high correctness ratio and high novelty, we believe that in terms of quiz-based privacy learning, the non-expert crowd indeed can generate high-quality quiz design ideas from their experiences, as they may be able to generate ideas in new ways and may have access to solutions that the experts do not.

![Figure 5. Novelty of Crowd-generated Ideas.](image)

<table>
<thead>
<tr>
<th>Correlation between idea quality and Workers’ background.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Correctness</strong></td>
</tr>
<tr>
<td>Novelty</td>
</tr>
</tbody>
</table>

**Correlations between Workers’ Background and Idea Quality**

To investigate whether there exists a relationship between the background of an individual and the quality of his generated ideas, we measure the Pearson correlation between the correctness & novelty of ideas and the workers’ privacy background and experiences. We consider the workers’ self-evaluation of expertise level as their privacy expertise, and their self-evaluation of whether having provided sufficient privacy protection on social networking sites as their privacy experience. The correlation results are shown in Table 1. We have the following findings. First, the idea correctness is positively correlated to the workers’ privacy expertise. This is not surprising as with more expertise the workers understand the problem better and thus produce more suitable answer. Second, the correlation between the idea correctness and the workers’ online privacy experience is very weak. This is somewhat surprising but still can be explained: Getting more experiences may not be able to help the workers to understand privacy better. Sometimes the users repeat the same mistakes if they were not aware. Regarding the novelty of ideas, the correlation between the novelty and the expertise, as well as the correlation between the novelty and experience, were negative. This implies that indeed the workers’ background does not help to generate novel ideas. We must admit
that the correlation results may be biased, as the workers may over-estimate their expertise level. They may also be over-confident of their online privacy experience.

**Quiz Evaluation**

The quiz evaluation involves both AMT workers and three privacy experts. In total, 120 workers participated in the evaluation task. All of them are from U.S. with an average age of 33. Regarding gender, 62 workers are females, and 58 are males. Regarding the educational level, 34 workers did not finish high school, 74 workers finished high school, and 12 workers received college education. Regarding their online privacy experience, 11 workers considered themselves not protecting privacy sufficiently, 85 workers considered them with sufficient protection of online privacy, and 24 workers were not sure. Regarding their privacy expertise, one worker knew nothing about online privacy; 5, 45, and 69 workers labeled themselves as level 2, 3, and 4 (out of 5) respectively. None of the workers considered themselves as an expert on online privacy. The average of self-evaluation expertise level is 3.69 (out of 5). To summarize, the workers who evaluated the quiz came from various demographic backgrounds, with different knowledge and experiences in online privacy.

**Detecting Cheating Workers**

We recorded the time that each participant took to finish the survey (including quiz playing). It turns out that the length of the working time is dramatically diverse. Figure 6 shows the distribution of the time length. The length of the working time varies from 2.78 minutes to 33.9 minutes. The average of working time length is 8.3 minutes. As our privacy experts spent two minutes on average to finish playing one quiz character, we expected that the survey should take at least four minutes to finish (including quiz playing). We consider those workers who finished the survey in less than four minutes as cheating. There are 10 (out of 120) cheating workers based on the time analysis. Therefore, we consider our collected evaluation results acceptable.

![Figure 6. Time to Finish Survey.](image)

**Tool Usability**

The tool usability evaluation is performed in the format of survey. The survey includes the questions that ask users to score on both the quiz interface and the content. The evaluation scores are in a scale of 1 to 5, with 1 being strongly disagree and 5 being strongly agree. We compute the average of evaluation scores per evaluation component per character. Figure 7 (a) and (b) show the results of tool usability evaluation. Most of the components received the evaluation score no lower than 3.03. In particular, the crowd are satisfied with the quiz interface, with the evaluation score no lower than 4, for all four characters (Figure 7 (a)). The crowd also enjoyed the examples and the suggested fix solutions (scoring at least 3.36) for all four characters (Figure 7 (b)). We computed
the average score of the quiz interface and the content. The Sophia character received the highest average score, while the Olivia character received the lowest score.

**Educational Impact**

The evaluation of educational impact includes the grading of the four learning components, namely, reaction measurement, learning measurement, behavior measurement, and results measurement. The score scale is the same as those questions for tool usability evaluation. The evaluation results are shown in Figure 7 (c). We computed the average score of the four components. It turns out that the average education score is in the range of [3.23, 3.83]. In other words, the crowd considers the quiz educational in general. From the scores of each individual component, we observe that the crowd agrees that the quiz works the best on the reaction measurement but the worst on the behavior measurement. Regarding the educational impact of different quiz characters, the Aiden character received the highest average score for education, while the Olivia character received the lowest score.

(a) Usability evaluation: quiz interface

(b) Usability evaluation: quiz content
Crowd vs. Experts

We asked the privacy experts to play the quiz and finish the same survey as the AMT workers. We calculated the average scores of both usability and educational impact by the experts, and computed the difference between the average scores between the experts and the crowd. It turns out the experts’ opinions are close to the crowd in terms of tool usability (average score difference < 0.3). The experts also agree that the quiz is educational in general, but with scores 15% lower than the crowd. Surprisingly, the experts picked the character Lucas as the most educational, but Aiden the worst. We interviewed the experts and found that the disagreement comes from some questions that the crowd labeled as inspiring but considered as trivial by the experts. This proves that indeed it is possible that the privacy experts may have different opinions from the general public regarding the effectiveness of the education quiz.

Conclusion

In this paper, we presented our effort of leveraging crowd-sourcing techniques for collecting and maintaining educational resources for privacy learning. We utilized the crowd for both idea collection for quiz content design and quiz evaluation. Our experiments show that the non-expert crowd can provide useful inputs for both quiz design and evaluation.

For future work, we plan to improve CERPA based on the received feedback from the quiz evaluation phase. We also plan to explore if under certain assumptions the “wisdom of the crowd” is able to outperform a smaller group of “experts”. In the longer term, we would like to investigate the potential of recruiting non-expert workers in the context of computer supported collective action, i.e., support participatory, end-to-end collective action in which a crowd can identify opportunities, formulate goals, brainstorm ideas and develop plans, mobilize, and take action [33].

Acknowledgement

This material is based in part upon work supported by the National Science Foundation under Grant Number 1464800. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References


