Eliciting Confidence for Improving Crowdsourced Audio Annotations

ANA ELISA MÉNDEZ MÉNDEZ, New York University, USA
MARK CARTWRIGHT, New Jersey Institute of Technology, USA and New York University, USA
JUAN PABLO BELLO and ODED NOV, New York University, USA

In this work we explore confidence elicitation methods for crowdsourcing "soft" labels, e.g., probability estimates, to reduce the annotation costs for domains with ambiguous data. Machine learning research has shown that such "soft" labels are more informative and can reduce the data requirements when training supervised machine learning models. By reducing the number of required labels, we can reduce the costs of slow annotation processes such as audio annotation. In our experiments we evaluated three confidence elicitation methods: 1) "No Confidence" elicitation, 2) "Simple Confidence" elicitation, and 3) "Betting" mechanism for confidence elicitation, at both *individual* (i.e., per participant) and *aggregate* (i.e., crowd) levels. In addition, we evaluated the interaction between confidence elicitation methods, annotation types (*binary*, *probability*, and *z-score derived probability*), and "soft" versus "hard" (i.e., binarized) aggregate labels. Our results show that both confidence elicitation mechanisms result in higher annotation quality than the "No Confidence" mechanism for binary annotations at both participant and recording levels. In addition, when aggregating labels at the recording level, results indicate that we can achieve comparable results to those with 10-participant aggregate annotations using fewer annotators if we aggregate "soft" labels instead of "hard" labels. These results suggest that for binary audio annotation using a confidence elicitation mechanism and aggregating continuous labels we can obtain higher annotation quality, more informative labels, with quality differences more pronounced with fewer participants. Finally, we propose a way of integrating these confidence elicitation methods into a two-stage, multi-label annotation pipeline.

CCS Concepts: • **Applied computing** → *Sound and music computing*; • **Human-centered computing** → *Empirical studies in collaborative and social computing*; • **Information systems** → **Collaborative and social computing systems and tools**.

Additional Key Words and Phrases: crowdsourcing, machine learning, audio annotation

ACM Reference Format:
Ana Elisa Méndez Méndez, Mark Cartwright, Juan Pablo Bello, and Oded Nov. 2022. Eliciting Confidence for Improving Crowdsourced Audio Annotations. *Proc. ACM Hum.-Comput. Interact.* 6, CSCW1, Article 88 (April 2022), 25 pages. https://doi.org/10.1145/3512935

1 INTRODUCTION

Crowdsourcing has made it possible to annotate large amounts of data quickly, making it revolutionary in domains where the annotation process is quick (e.g., image annotation) and machine learning models require large training datasets. However, the benefits of scale in crowdsourcing

Authors' addresses: Ana Elisa Méndez Méndez, anaelisa.mendez@nyu.edu, New York University, 370 Jay St, Brooklyn, NY, USA; Mark Cartwright, mark.cartwright@njit.edu, New Jersey Institute of Technology, GITC, Room 3902E, University Heights, Newark, NJ, USA, New York University, 370 Jay St, Brooklyn, NY, USA; Juan Pablo Bello, jpbello@nyu.edu; Oded Nov, onov@nyu.edu, New York University, 370 Jay St, Brooklyn, NY, USA.

2573-0142/2022/4-ART88 $15.00
https://doi.org/10.1145/3512935

are not easily replicated in domains where the annotation process is slower and thus more costly, e.g., subjective tasks such as sentiment annotation on text, and perceptual tasks such as video annotation for visual object detection and tracking, and audio annotation for sound event detection in complex environments. For example, manually tagging 10 s audio recordings (a common example length in audio datasets) is a slower process than it is for images due to the temporal nature of audio (i.e., audio annotation tasks require devoting time to listening to examples). For this reason, crowdsourcing for audio annotation typically does not scale as well as image annotation, i.e., when compared to image annotation each label takes longer to get and thus is more expensive to collect. Moreover, real-world audio data can contain multiple simultaneously occurring sounds, making it hard to disambiguate one sound from the other. Ambiguity in the data leaves room for annotator interpretation, thus increasing chances of disagreement and creating the need to quantify uncertainty in data labels [2].

Machine learning research has shown that training with more informative labels, representing the uncertainty in the data labels, can reduce the amount of required data. More specifically, "soft" labels provide richer training information than "hard" labels, and consequently, models can be trained with much less data [32]. "Soft" labels are continuous labels based on probability estimates, i.e., labels from 0 to 1, where a label of 0.4 means there is a 40% probability that the example corresponds to the target class. Whereas, "hard" labels are discrete labels, i.e., 0 or 1, where 0 indicates the non-presence of the target class and 1 indicates the presence of the target class. Studies have shown that training with "soft" labels of ambiguous data increases model performance when compared to training with "hard" labels [6, 25, 26, 77].

Crowdsourcing researchers have also adopted the idea of "soft" labels and have investigated collecting "soft" labels for ambiguous data to increase annotation quality [5, 17, 25, 39]. However, most previous research did not evaluate the elicitation method but only the aggregation mechanisms used to convert several individual "hard" labels to aggregate "soft" labels [25], or several individual "soft" labels into aggregate "soft" labels [5, 39]. The exception is recent work by Chung et al. [17], which evaluated elicitation methods for collecting "soft" multi-class[1] labels of facial expressions. For their task, the authors found that "fine-grained" annotation methods are more beneficial for more ambiguous data and that collecting aggregate "soft" labels through *multi-label*[2] annotations (for a *multi-class* problem) achieves the highest accuracy while reducing human annotation effort. They found this method resulted in higher quality labels than asking annotators to directly provide "soft" labels. This finding corresponds with studies that show that people tend to misjudge their probability and confidence assessments [69]. In addition, annotation tasks for real-world audio recordings are typically multi-label rather than multi-class, and thus their findings for multi-class annotation may not transfer to this other family of tasks.

We seek to address these limitations by proposing two alternative methods for collecting "soft" labels from individual annotators. Our goal is to obtain "fine-grained" probability estimates from an individual participant rather than a population of participants. With a *binary* (0 or 1) setting or a *multi-label* setting you would need more people to get the same granularity that you would in a *multi-class* setting using the *multi-label* mechanism proposed by [17]. With our proposed methods, we aim to obtain "fine-grained" annotations with fewer annotators than the method proposed in [17]. In our analysis, we first focus on the confidence elicitation method rather than aggregation as a way to address ambiguous tasks where repeated annotations are expensive. Subsequently, we proceed to study the interaction between the elicitation method, the annotation

[1]A multi-class problem is a task where each data point corresponds to one class among multiple possible classes (e.g., classifying dog species).

[2]A multi-label problem is a task where each data point can correspond to multiple possible classes (e.g., tagging attributes of a dog like fur color, fur type, size, among others).

type, and the aggregation method for when repeated annotations are within budget. To make this approach applicable to real-world audio annotation, we initially formulate our approaches for binary annotation but later discuss how to integrate them into a multi-label annotation pipeline. We study this problem in the context of machine listening of urban sound which has numerous applications where the data is ambiguous and the quality of the annotations highly impacts results. 1) Urban noise pollution monitoring, which can help in developing plans for mitigating noise pollution, a "quality of life issues for urban residents in the U.S. with proven effects on health, education, the economy, and the environment" [8]. 2) Urban bioacoustic monitoring, which is key in understanding bird migration patterns, how species communicate in cities, and how they are affected by city living, especially in areas with collision hazards such as buildings, planes, communications towers, and wind turbines [58]. 3) Audio-based traffic monitoring, which is a promising privacy-preserving alternative to vision-based methods. 4) Sound-awareness tools for the hard of hearing, which can increase awareness of critical sound events for deaf and hard of hearing people. 5) Audio forensics, upon on which decisions based could have a big impact in the fate of a person. And 6) machine perception for autonomous robots/vehicles, where incorrect predictions about the environment can potentially be fatal. All of these problems require annotation of real-world environments, which could be highly ambiguous, making it important to collect confidence information to get higher quality labels, and thus, higher quality models. Finally, to the best of our knowledge this study is the first of its kind in the audio domain, although we believe findings are transferable to other domains where the annotation process is inherently slow and labels are often ambiguous.

We propose two methods for collecting "soft" labels from individual annotators: 1) a direct confidence elicitation method, where participants not only provide "hard" labels, but also directly state how confident they are in each label provided — these confidence estimations are then converted to probabilities and used as "soft" labels — and 2) a betting mechanism based on de Finetti's probability theory [21]. De Finetti defined that "the degree of probability attributed by an individual to a given event is revealed by the conditions under which he would be disposed to bet on that event" [21]. This means that if a person is willing to place money on an event, it is because they think the probability of the event occurring exists. We investigate this approach since studies have shown that the value of the monetary incentives in betting tasks improves the accuracy of participants' responses [44, 55, 66]. In this method, participants indirectly select their confidence by placing bets either on their "hard" label or on a lottery. As in the first method, we then convert these confidences to probabilities for use as "soft" labels. We evaluate these two methods in an experiment in which we compare them to a baseline method that collects only "hard" labels. In addition, we explore three annotation types: 1) binary annotations, 2) probabilities transformed directly from the confidence annotations, and 3) probabilities transformed from z-score confidence annotations. Finally, we compare "hard" versus "soft" aggregate labels. In our experiment, we aim to answer two questions:

(1) Does confidence elicitation for collecting "soft" labels affect *individual* (i.e. per participant) annotation quality?
(2) Do the confidence elicitation method and aggregate label type affect *aggregate* (i.e. crowd) annotation quality and cost?

The first question addresses the annotation quality per participant. However, we can also aggregate "hard" labels from multiple annotators into "soft" labels. Thus, when addressing the second question, we investigate which annotations to use, how to aggregate them, and from how many annotators.

2 RELATED WORK

2.1 Ambiguity in crowdsourced labels for machine learning

Crowdsourcing is often used to collect annotations from the crowd that can be used as "ground truth" annotations for solving machine learning problems. The key of crowdsourcing lies in the fact that such tasks quickly reach large amounts of people, hence providing scale. However, there are challenges to crowdsourcing data annotation. One of them is the uncertainty in the quality of the labels it produces [42]. Prior work discusses best practices for achieving expert or "ground truth" performance using novices in domains like image annotation and natural language [52, 63, 64]. Results show that annotation quality improves as redundancy is added.

While these studies have shown that aggregating data can resolve disagreements resulting in higher label quality, they have not taken into consideration the fact that there might be more than one possible "ground truth" label. Ambiguity in data allows for different interpretations and thus creates the necessity of collecting different types of annotations. Moreover, not all instances are created equal. Some examples are easier to identify, while others are more confusing, and in consequence, disagreement is higher [2]. The "ground truth" label, the annotators, and the instances to be labeled can be placed as vertices in the triangle of reference [53] discussed by Aroyo & Welty and Dumitrache et al. [2, 24]. The instances to be labeled correspond to the *sign*, the "ground truth" label corresponds to the *referent*, and the annotator corresponds to the *interpreter*. Some examples of ambiguous tasks are: 1) emotion recognition [50, 57, 76], where the annotation depends on the interpretation of the annotator, 2) text sentiment analysis [22, 54, 75], where words can have different meanings depending on the context, and 3) audio annotation [11, 18, 31], where multiple sources can overlap making it hard to disambiguate them. One solution for solving disagreement between annotators is by aggregating the collected labels.

Furthermore, machine learning has been shown to benefit from representing ambiguity in data [6, 25, 26, 77]. However, until recently, crowdsourcing work has focused on creating "soft" labels by collecting annotations from multiple participants [25], but this solution is expensive. Although directly eliciting confidence from annotators is challenging, recent work has addressed collecting individual "soft" labels [17]. This approach has its challenges and it is not clear how it could be incorporated into multi-label annotations scenarios like real-world audio annotation.

2.2 Improving crowdsourced annotation quality through redundancy

In the recent past, crowdsourcing research seeked to improve the quality of the annotations obtained through crowdsourcing by finding the most efficient way of aggregating such labels. Research in machine learning in different domains have used majority voting [1, 19] for aggregating labels collected through crowdsourcing for training models. However, more recently, work has focused on finding aggregating techniques that can result in higher accuracy labels, and in consequence, higher accuracy models by modeling the correct answer (the *referent*), the reliability of the annotators (the *interpreter*), and the difficulty of the instances to label (the *sign*) [7, 20, 30, 41, 46–48, 56, 62, 67, 68, 72, 73, 78, 79]. Such works do not take into consideration the process of collecting these labels.

Consequently, with the goal of collecting higher quality labels, some studies have focused on the annotation task and tool design as a way of improving aggregate accuracy [15, 16, 37, 38, 45, 65]. Other studies have allowed interaction between participants as a way of solving inter-annotator agreement and improving aggregation quality [13, 14, 33, 34].

2.3 Machine learning with "soft" labels

The previously mentioned studies focused on collecting "hard" (binary) annotations, about the presence of different target classes, leaving unstudied the fact that there might be more than one

true label as discussed by Aroyo & Welty [2]. A way of collecting information for those kind of instances is to collect "soft" labels. Such "soft" labels allow small machine learning models to achieve similar performance to larger models trained using "hard" labels [27, 32], as models are allowed to learn from multiple answers instead of learning from only one answer [6, 29]. "Soft" labels or label distributions are continuous labels based on probability estimates. Dumitrache et al. [25, 26], Zhang et al. [77], and Aung et al. [6] show that, for ambiguous data, collecting "soft" labels and using these labels for training machine learning models, results in better performance when compared to training with "hard" labels. In the audio domain, researchers have shown that "soft" labels train better acoustic models for speech enhancement than "hard" labels [23, 71].

2.4 Crowdsourcing aggregate "soft labels" for representing ambiguous data

To collect "soft" labels for training machine learning models, researchers have investigated collecting multiple "hard" labels through crowdsourcing and aggregating them into "soft" labels for describing ambiguous data. Dumitrache et al. [25] demonstrated that aggregating labels with high disagreement into "hard" values is inappropriate for training a semantic language model. This study collected multi-label annotations from participants that are later aggregated and converted into "soft" labels. However, they do not show how they calculated the probabilities of the labels to model the "ground truth".

2.5 Crowdsourcing individual "soft labels" for representing ambiguous data

Other work focuses on directly collecting probabilities with the goal of solving annotator disagreement [5]. In a study by Augustin et al., participants assign values between 0 and 100 to each of 6 sentences in a sentiment analysis task. To evaluate their results, the authors model spam data and compare their probability aggregates with spammers to approaches without spammers. When 60% of the workers are spammers, they show comparable results to state-of-the-art approaches without spammers. Jurgens [39] created a task for eliciting annotations for word sense disambiguation with the goal of obtaining high inter-annotator agreement.

In response to previous studies, Chung et al. [17] investigated collecting "soft" labels through four different elicitation methods: *multi-class* annotations, *multi-label* annotations, *ranked multi-label* annotations, and *probability multi-label* annotations (for a *multi-class* problem). Their goal was to achieve the highest annotation accuracy while reducing human efforts. Their results suggest that the *multi-label* approach resulted in higher quality labels than directly requesting "soft" labels. At the same time, they also found that as the number of annotators increased, the differences between elicitation methods became insignificant. One possible downside is that if more granularity is needed, using the recommended *multi-label* method would require more annotators to gain probability resolution. In addition, results from this study may not translate well to tasks that are typically multi-label rather than multi-class, which is the case of real-world audio recordings. Using their recommended method, the probability of the classes needs to sum up to 1, while for multi-label annotations each label has its separate probability distribution. If we divide the multi-label problem into multiple binary problems, using the method proposed in [17] would require more annotators to gain probability resolution. For example, if one participant selects "yes" and "no", the probability would be 0.5. The possible probabilities from this one participant are 0, 0.5 or 1. From two participants, the maximum probability resolution would be 0, 0.25, 0.5, 0.75 and 1. As we want to keep the number of annotators per sample to a minimum, we want to increase the resolution per participant. Collecting audio annotations through crowdsourcing remains understudied and is evaluated in this work.

2.6 The challenge of eliciting confidence

To better understand how to collect "soft" labels and the way humans perform confidence annotations, we need to understand human's capabilities to evaluate their own performance. Some studies have focused on understanding how monetary incentives influence confidence accuracy by varying the value of the incentives, allowing the participants to place bets on their answers [44, 55, 66]. Their results show that although participants fail to maximize their winnings when placing bets based on the correctness of their answers, the value of the monetary incentives improves the accuracy of their responses. In addition, when faced with a betting task, participants make decisions faster when the chances of winning are high [66].

However, a downside of collecting annotations using confidence scales has been shown in studies in metacognition where they explain that "while two participants may have a comparable clarity in their experience of a stimulus, they might use different criteria to decide themselves confident" [61].

Although the use of such mechanisms has been growing in crowdsourcing, it is still an open research area, as crowdsourcing often focuses on collecting as many labels as possible in the shortest amount of time. However, inherently slower processes like audio and video annotation could benefit from such mechanisms. In this work, we propose an alternative betting mechanism for collecting annotations for ambiguous data. This method has the potential of being more efficient at getting higher label resolution of the "soft" labels by decreasing the step size for the probabilities when needed, while still keeping the number of annotators low. In addition, this method is appropriate for cases where the multiple choice method, such as those proposed by Chung et al. [17], is inefficient because of the lack of multiple classes. We will study the interaction between the confidence elicitation method, the annotation type, and the aggregation method. Finally, we will demonstrate the use of these methods in an audio annotation task, although we believe our findings are replicable across domains.

2.7 Crowdsourcing audio annotations

Prior work in the sound domain shows that as the number of annotators is increased, on average, aggregate annotations are closer to "ground-truth" annotations [12]. Datasets like OpenMic-2018 [36], Audio Set [31], FSD-50k [28], and SONYC-UST-V2 [10] are some of the most recent works for collecting crowdsourced audio annotations. These four datasets collected redundant annotations, and aggregated them to form the "ground truth" labels. However, the temporal characteristics of such mediums, require a certain amount of time invested to be annotated regardless of the annotation time, making it inefficient to collect as many redundant annotations as usually collected through crowdsourcing. At the same time, sound recordings are often ambiguous, making "soft" labels collected from fewer annotators potentially more appropriate for such instances.

3 EXPERIMENTAL SETUP

We performed a between-subjects study where each participant is randomly assigned to one of the three elicitation mechanism tasks. For the first research question, the independent variables are the confidence elicitation method and annotation type, while the dependent variable is the *individual* annotation quality. For the second research question, the three independent variables are the confidence elicitation method, the annotation type, and the aggregation method. The dependent variables are the *aggregate* annotation quality and the cost (number of annotators).

We collected "hard" and "soft" labels using three different confidence/no-confidence elicitation mechanisms and compared which type of label and elicitation mechanism combination achieves higher-quality labels.

(1) **No confidence** elicitation: The first method consists of a binary task where participants were asked to identify whether there is or not a jackhammer present in the audio recordings. This method does not elicit participant confidence in any way and is used as the baseline for comparison.
(2) **Simple confidence** elicitation: The second method is a two-step mechanism. The first step is identical to the first method. In the second step participants were asked how confident they are in their previous response. During step 2 participants were allowed to change their answer to the question in step 1.
(3) **Betting**-based confidence elicitation: The third method consists of a betting mechanism based on de Finetti's probability theory [1]. As defined in section 1, "the degree of probability attributed by an individual to a given event is revealed by the conditions under which he would be disposed to bet on that event." To collect this information, we asked participants to answer the question about the presence of a jackhammer, in the same way as the first and second methods. After this, we asked them, in a series of questions, to choose between an X% chance lottery and their answer to the first question for the chance to win a bonus, where in each question X was varied from 50% to 90%. Choosing their answer would mean that they are more confident in winning the bonus with their answer than winning a lottery with X% chances of winning. On the other hand, choosing the lottery, would mean they are more confident in the lottery than in their answer.

Section 3.2 will elaborate on these steps in more detail.

3.1 Audio Generation Process

We used the Scaper [60] soundscape generator with audio from the UrbanSound8K [59] dataset to create a controlled dataset for experimentation. UrbanSound8k is a dataset consisting of 8,732 audio excerpts taken from field recordings from Freesound[3] divided into 10 folds for cross-validation. Each of the files contains one of 10 classes: *siren, air conditioner, children playing, gunshot, engine idling, drilling, jackhammer, car horn, dog barking and street music*. Based on the cross-validation folds, we created 10 groups of recordings using Scaper [60]. With Scaper, a user provides a collection of source audio files as input, and then the system samples, sequences, and mixes the source audio into polyphonic soundscapes. This tool allows you to select the length of the final recordings, the minimum and maximum length of each of the input sounds, the minimum and maximum number of classes per recording, among other parameters. For the purposes of this work, we wanted to have a dataset that is ambiguous, in a way that participants find it difficult to identify the target class. This added difficulty allows us to get more useful information from the confidence elicitation method for both the model and the quality of the labels. To accomplish the desired ambiguity level, we selected a minimum of three events per recording and a maximum of nine, with a minimum duration of two seconds and a maximum of six seconds per event,increasing the probability of overlapping sounds.

The target class in this work is *jackhammer*, which can be easily confused with two other categories from the UrbanSound8k dataset: *drilling* and *engine idling*. The chosen categories are the only ones in the UrbanSound8k dataset that follow similar spectro-temporal patterns to those of the *jackhammer*, thus creating higher ambiguity than when overlapped with other sounds. When making the selection of files to present to participants, we considered the difficulty of the recordings as one important aspect. To increase the ambiguity of examples even further, positive examples contain the jackhammer plus at least one of the confusing categories, i.e., *jackhammer + engine idling*, *jackhammer + drilling*, or *jackhammer + engine idling + drilling*. Negative examples do not

[3]www.freesound.org

contain the *jackhammer* class but could contain the other two (*drilling* and *engine idling*). All of the generated examples are 10 seconds long.

For the annotation tasks, we generated 2,000 examples for each of the folds and applied the selection criteria explained earlier in this section for choosing positive examples on each of the folds. The fold with the least number of positive examples had 290 recordings. Using this number, we randomly selected 290 positive recordings and 290 negative recordings for each fold, ending up with a total of 580 recordings per fold. Although using this data to train a machine learning model is beyond the scope of this study, we plan to use the results from this study in future work to test the practical applications of "soft labels", based on the different confidence elicitation methods, for machine learning. For this reason, we determined the number of recordings needed to achieve a performance of 70% accuracy. To do so, we randomly selected groups of 15, 20, 30 and 40 recordings per fold, balanced to have the same number of positive and negative recordings, and trained models using cross-validation based on said folds. In total, the number of examples for training for each variation was 135, 180, 270, and 360, respectively. Accuracy did not increase significantly when selecting more than 20 recordings per fold (180 total training examples for 9 combined folds), thus, we selected 20 recordings per fold to be annotated. The same set of audio recordings was used for all three annotation methods.

3.2 Annotation Tasks

To collect labels through crowdsourcing, we created a web-based annotation application, which had interfaces to support three annotation tasks, one for each confidence elicitation method.

The annotation tasks consisted of listening to 20 sound recordings (all of the recordings in a single fold) and identifying whether there was a jackhammer present or not in the recordings. Depending on the task shown, participants were also asked to select their confidence in the presence/non-presence of the jackhammer. More specifically, we recruited 300 participants who were randomly assigned to one of the three tasks and one of the 10 folds, i.e., 10 participants per fold per task. The three tasks differed in the way the confidence was collected: task 1 did not request confidence responses; task 2 directly requested confidence labels; and task 3 used a betting method for indirectly eliciting confidence.

Participants were recruited on Amazon Mechanical Turk, applying a selection criteria of 95% or greater HIT approval rate on all their tasks, and US location. This study received New York University IRB approval (IRB-FY2019-2872). We paid all participants a $1 participation fee and a bonus calculated based on their performance. All interfaces reminded participants at every step that their payment would depend on their performance. The bonus fee per recording depended on the type of task and was based on the probability of winning this bonus to account for the different probabilities of winning with each task, i.e., to make the expected payment the same. For the "No Confidence" and "Simple Confidence" elicitation tasks, the probability of winning was based on the probability of correctly answering the question about the presence of the *jackhammer*. Since it was a "yes"/"no" question and half of the recordings contained the target class and half did not, the probability of getting a correct answer was 0.5, and thus, the bonus per recording was $0.2 for an average estimated payment of $2.50 per task. For the "Betting" mechanism task, the probability for winning the bonus depended on both the round selected for payment and on whether the participant decided to play the lottery or chose to keep their answer for testing (more details about this process are explained in Section 3.2.3). For an average estimated payment of $2.50 per task, we estimated the bonus per recording to be $0.24.

Once participants accepted the task on Mechanical Turk and the study consent, they were shown the task instructions, which were accessible for the duration of the task. Before proceeding to the

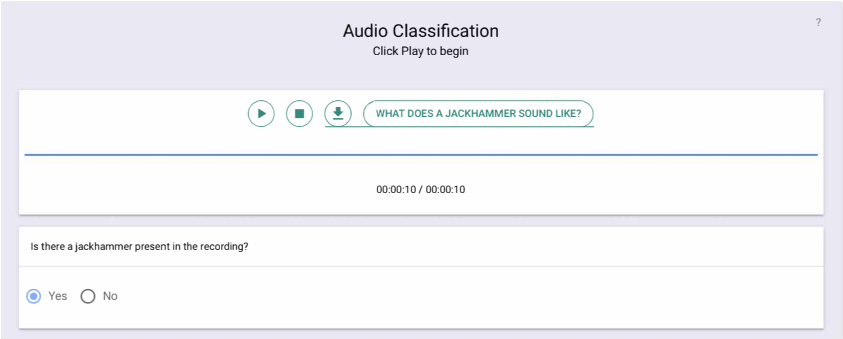


Fig. 1. "No Confidence" elicitation interface. Participants are asked to answer the question: "Is there a jackhammer present in the recording?"

primary annotation task, participants had to annotate at least two example audio clips for practice to familiarize themselves with the task.

3.2.1 Task 1: "No Confidence" elicitation. In the first task we asked participants to listen to an audio recording and then answer the question "Is there a jackhammer present in the recording?", to which they had to reply "yes" or "no". Once participants answered the question, they were directed to the next recording. Figure 1 shows the interface for this task. The step-by-step process is described below:

(1) The first step showed one audio recording with a play button. Participants had to click "play" and listen to the entire recording before proceeding to the next step.
(2) They then were asked to answer the question: Is there a jackhammer present in the recording? To which they had to reply "yes" or "no".
(3) After answering the first question, they were directed to the next recording by clicking on the "next" button, back to Step 1 until the 20 recordings were labeled.
(4) Once all recordings had been annotated, participants received a summary table with a review of their performance and their performance-based monetary compensation.

The annotations from this task consisted of binary annotations where "0" represented the "no" answer and "1" represented the "yes" answer.

3.2.2 Task 2: "Simple Confidence" elicitation. In the second task, we requested probability estimates directly from the participants on how confident they were about the presence of a jackhammer in the recordings. The interface for this task is shown in Figure 2 and the step-by-step process is described below:

(1) The first step showed one audio recording with a play button. Participants had to click "play" and listen to the entire recording before proceeding to the next step.
(2) They were then asked to answer the question: Is there a jackhammer present in the recording? To which they had to reply "yes" or "no".
(3) After answering the first question, they were asked a second question: "How confident are you that there is (is not) a jackhammer present in the recording?", for which they had to select a confidence estimate of the presence of the class on a scale from 50 to 100 in increments of 10. It is worth mentioning that during this step participants were allowed to change their answer to the first question in Step 2 if they considered they made a mistake.

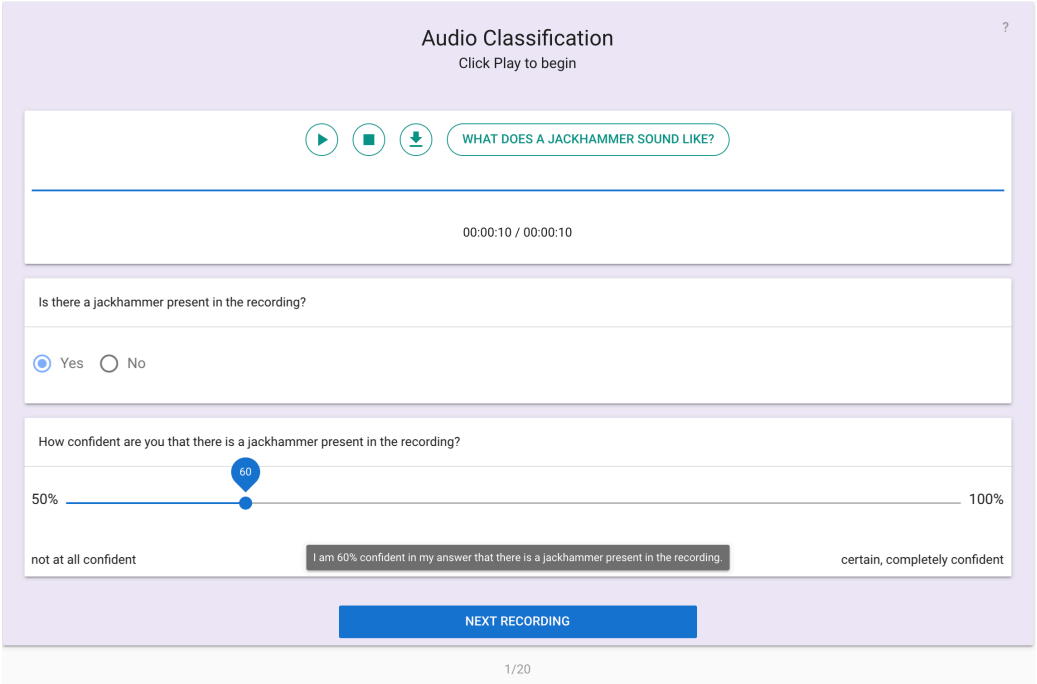


Fig. 2. "Simple Confidence" elicitation interface. Participants get to directly select how confident they are in their answer to the question: "Is there a jackhammer present in the recording?". In this case, the participant selected 60% confident that there is a jackhammer.

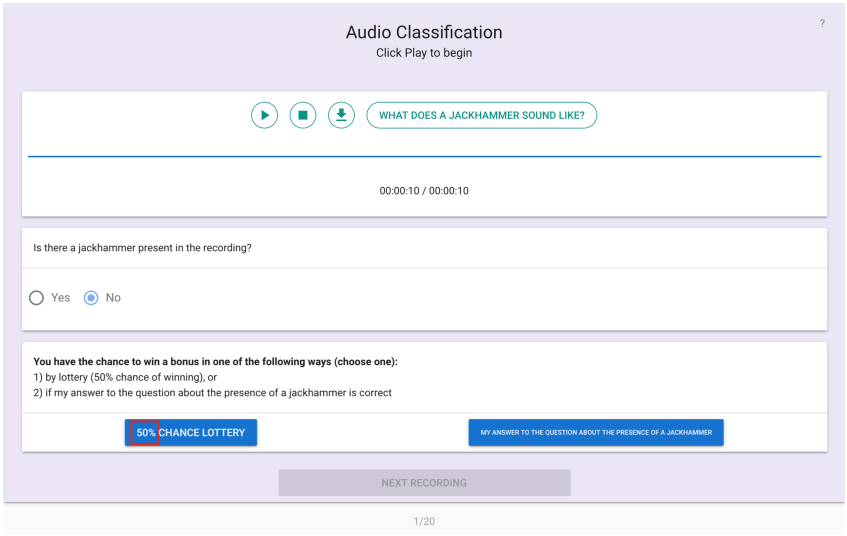
(4) Then, participants clicked on the "next" button and went to the next recording, back to Step 1 until the 20 recordings were done.
(5) Once all recordings had been annotated, participants received a summary table with a review of their performance and their performance-based monetary compensation.

The annotations from this task consist of a pair of a binary annotation, in the same way as in the previous task, and a confidence in that annotation ranging from 50% to 100%.

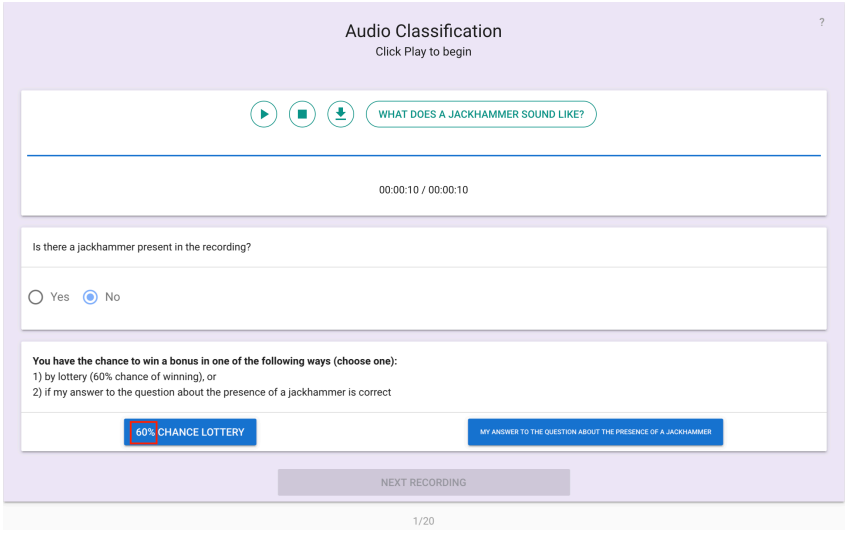
3.2.3 Task 3: Betting mechanism. In Task 3 we elicited confidence through a betting mechanism based on de Finetti's [21] probability theory. Figure 2 shows the interface for this task, and the step-by-step process is described below:

Steps (1) and (2) are the same as in Task 2.

(3) After answering the question, they were presented with a follow-up step: "You have the chance to win a bonus in one of the following ways (choose one): 1) by lottery (X% chance of winning), or 2) by correctly answering the question."
(4) Step 3 was repeated 5 times per recording, progressively increasing the chances of winning the lottery from 50% to 90%. Each of these options happened only once. After the fifth round was played, the round for payment was randomly chosen (e.g., if the participant selected their answer for rounds 1, 2 and 3, and the lottery for 80% and 90%, the payment round selected uniformly at random could either test if their answer was correct if rounds 1 through 3 were selected or play an 80 or 90% chance lottery.) It is worth mentioning that during this



(a) 50% chance lottery selected



(b) 60% chance lottery selected

Fig. 3. "Betting" mechanism interface. Participants play a game where they can selected between a lottery with increasing chances of winning from 50% to 90% and their answer to the question: "Is there a jackhammer present in this recording?", for a chance to win a bonus. In 3a, the participant selected to play a 50% chance lottery and in the following round (shown in 3b, the participant selected to play a 60% chance lottery)

step participants were allowed to change their answer to the first question in Step 2 if they considered they made a mistake.

(5) Once participants were done with Step 4, they had to select the "next" button, where they were shown a popup window letting them know which of the 5 rounds had been selected for payment, and go to the next recording, back to Step 1 until the 20 recordings were done.
(6) Finally, participants received a review of their performance and their monetary compensation based on their performance or lottery results, depending on the round selected for payment in Step 4. For each round the lottery was selected, participants were presented with a button with a running timer, up to millisecond precision, based on their computer clock. To play the lottery, they had to click on this button. If, when stopped, the centiseconds were less than the lottery chances, then the participant won the bonus for that round, otherwise, the participant lost. The use of the clock mechanism is meant to give the participant trust in the lottery process.

The annotations from this task consist of a pair of: 1) a binary annotation, as in the previous task, and 2) a list of binary responses representing whether the participant selected to play the lottery or their answer. For example, a list consisting of [1, 1, 0, 0, 0], means that the participant selected their answer on rounds 1 and 2, and the lotteries for 70%, 80% and 90%.

3.3 Processing of Collected Annotations

To use the confidence labels collected in both of the confidence elicitation tasks for training machine learning models, we first converted them into probability estimates. For the direct elicitation task, the confidence labels collected range from 50% to 100% in increments of 10%. For those to which the answer to the question about the presence of the jackhammer was "yes", the probability estimates remain the same, for those to which the answer was "no", the probability estimates are calculated by subtracting the confidence from 1, e.g., if the answer was "no" and the confidence was 60%, the probability estimate is 0.4 (40%).

For the "Betting" task, the confidence estimates were first calculated based on the five responses to the second question. For example, if one participant selected their answer to the first question instead of the lottery, a "1" was assigned to that response, if they selected the lottery, a "0" was assigned to that response. After all responses were collected, a list with the responses is saved (e.g., [1, 1, 0, 0, 0], which would mean the participant selected their answer the first two times and then the lottery for 70, 80 and 90%). The switch from their answer to the lottery would indicate the confidence they have in their answer. In the previous example, it would mean that the participant is between 60 and 70% confident that there is (or is not) a jackhammer present in the recording. In this case, we assigned a 65% confidence estimate. After we did this process with all responses for the betting task, we calculated the probability estimates using the same procedure as for the direct elicitation task.

For both of these methods, we also explored a variation of the probability calculation in which we added an additional per-participant standardization step. In this variation, we computed z-scores based directly on their confidence responses, min-max scaled over all participants to convert them back to confidences, and then converted to probabilities in the same manner as described earlier. Z-scores are calculated because the distribution of the collected labels depends on the personal interpretation of the scale presented [61]. In our analysis, we refer to these labels as "z-score derived probability" labels.

3.4 Metrics

We measured annotation quality based on two types of metrics: metrics that are dependent on the decision threshold for binarizing annotations and those that are not. The threshold-dependent metrics we used were accuracy, precision, recall and F-measure. The metric we used that was

not dependent on a fixed decision threshold for binarizing annotations was the area under the precision-recall curve (AUPRC). The AUPRC curve shows the trade-off between precision and recall across different decision thresholds. Because it requires probability estimates, this metric was not used to evaluate the baseline "no confidence" annotation method.

3.5 Estimating annotation time

Due to a data collection mishap, our web application did not collect completion time information for the original 200 participants assigned to the "Betting" and "Simple Confidence" annotation tasks, but the application did collect this information for the original 100 participants assigned to the baseline annotation task. To remedy this situation, we collected data from an additional 20 participants in the "Betting" annotation task and additional 20 participants in the "Simple Confidence" annotation task. This data was only used to estimate the average annotation time for these tasks and was not used in the annotation quality analysis.

4 RESULTS

4.1 Summary

We collected 10 annotations per recording for each of the three tasks, for a total of 300 unique participants. The average task completion times are in Table 1 and show that while the baseline "No confidence" task had the lowest median completion time (6.53 mins), the "Betting" task had the next lowest (7.19 mins), with "Simple Confidence" having the highest (7.65 mins). Based on these completion times, we estimate we paid participants a median of $37.45 per hour for the "Simple Confidence" elicitation tasks, $35.43 per hour for the "Betting" mechanism task, and $35.73 per hour for the "No Confidence" task. Since there was a possibility of annotation conflicts for the "Betting" task (e.g., [1, 0, 1, 0, 1]), which could affect the quality of the results, we checked the annotations and found that conflicts were rare (44 out of 2000, i.e., 2.2%). In these situations, we selected the confidence value as that corresponding to the first transition from 1 to 0.

In our analysis, we aim to answer two research questions:

(1) Does confidence elicitation for collecting "soft" labels affect *individual* (i.e. per participant) annotation quality?
(2) Do the confidence elicitation method and aggregate label type affect *aggregate* (i.e. crowd) annotation quality and cost?

We can answer the first research question by saying that the confidence elicitation method for collecting "soft" labels affects individual annotation quality. More specifically, when comparing confidence elicitation methods on the quality of the binary labels per participant, both "Simple Confidence" and "Betting" methods are statistically significantly better than the baseline "No Confidence" elicitation method. When measuring confidence elicitation method and confidence annotation type at the individual level, the "Simple Confidence" method is borderline significantly better than the "Betting" method, with no statistical differences between the annotation types.

The second research question (do the confidence elicitation method and aggregate label type affect *aggregate* (i.e. crowd) annotation quality and cost?) is answered by the fact that confidence elicitation method and annotation type affect aggregate annotation quality and cost. More specifically, when using "hard aggregate" labels to compare performance against the ground-truth binary labels, the "Betting" method performs better than the "Simple Confidence" method for AUPRC, with differences more pronounces at higher number of participants. Moreover, when comparing "hard aggregate" labels to "soft aggregate" labels based on annotator agreement, we can see that "soft aggregate" labels perform better than "hard aggregate" labels at any number of participants, which means we can reduce the cost of the task using the "soft aggregate" labels.

Table 1. Completion time for annotation tasks in seconds

Task	Median	Mean	Standard deviation	N
No confidence	6.53	8.41	4.78	100
Simple confidence	7.65	8.36	3.49	20
Betting	7.19	6.58	3.56	20

We elaborate more on the results in the sections below.

4.2 Does confidence elicitation for collecting "soft" labels affect *individual* (i.e. per participant) annotation quality?

In order to answer our first research question we investigate: 1) the effect of confidence elicitation method on the quality of the binary labels per participant, and 2) the effect of both confidence elicitation method and annotation type (*binary* labels, continuous *probability* labels and *z-score derived probability* labels) on the quality of the annotations per participant.

We first explain the different annotation types: 1) the *binary* labels are the direct answers to the question about the presence of the jackhammer, 2) the continuous *probability* labels are the probabilities calculated directly from the confidence annotations as described in Section 3.3, and 3) the *z-score derived probability* labels are probabilities computed from the per-participant normalized confidence scores as described in Section 3.3.

While we are primarily interested in "soft" labels, analyzing the "hard" (i.e., binary) labels from the tasks gives us some insight into whether label quality differences in confidence elicitation tasks are due to changing behavior of the annotator or the richer information of the "soft" labels. Thus we analyze "hard" labels and "soft" labels separately.

4.2.1 Effect of confidence elicitation method on the quality of the "hard" labels per participant. To measure the effects of the elicitation method on the quality of the *binary* annotations per participant, we calculated a one-way ANOVA for each of the quality metrics (with 0.5 threshold) using the cross-validation fold as blocking variable. After testing the residuals, the normality assumption does not hold for all of the metrics, so we proceeded with the Aligned Ranks Transformation ANOVA (ART anova) [74] for all the quality metrics.

The results for accuracy are shown in Figure 4. These results suggest that both the "Simple Confidence" and the "Betting" mechanisms have better performance than the "No Confidence" method. We proceeded to perform statistical analysis to understand if these differences are significant. The ART anova ($F(2, 270) = 5.403$, $p = 0.005$) shows that we reject the null hypothesis that the means of all confidence elicitation methods are equal. Similar results were obtained for precision, shown in Figure 4 as well ($F(2, 270) = 4.271$, $p = 0.0149$). Given these results, we performed a post hoc Tukey HSD test to understand which differences are statistically significant. Table 2 shows the results for the Tukey HSD tests for accuracy and precision, the two metrics that showed statistical significance in the ART anova. For accuracy there is a significant difference, at $\alpha = 0.05$, between the "No Confidence" elicitation method and the "Simple Confidence" method, and between the "No confidence" and "Betting" method, and no significant difference between the "Simple Confidence" and "Betting" methods. For precision, there is significant difference, at $\alpha = 0.05$, between the "Simple Confidence" method and the "No Confidence" method, and between the "Betting" method and the "No Confidence" method. This result is important, as it shows us that eliciting confidence results in an improvement of the accuracy and precision per participant. No significant differences were found between mechanisms for either recall nor F-measure.

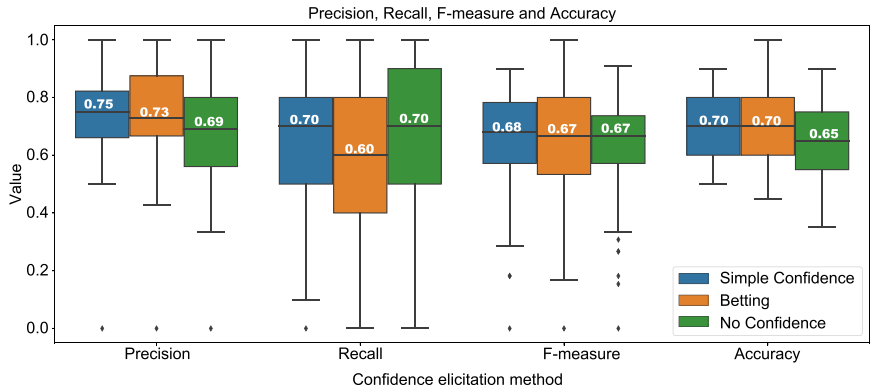


Fig. 4. *Individual* Precision, recall, F-measure and accuracy boxplots, averaged over cross-validation folds, for each of the confidence elicitation methods for the binary annotation type. The lines in the boxplots correspond to the medians, the boxes are the inner quartiles, and the whiskers extend to 1.5 the IQR.

Table 2. Tukey HSD Results for Accuracy and Precision when Comparing Confidence Elicitation Methods

Metric	Method	Estimate	p-value
Accuracy	Simple confidence - No confidence	37.85	0.008
	Betting - No confidence	33.61	0.022
	Simple confidence - Betting	4.24	0.94
Precision	Simple confidence - No confidence	31.46	0.036
	Betting - No confidence	32.59	0.028
	Simple confidence - Betting	-1.13	0.996

Table 3. Tukey HSD Results for AUPRC when Comparing Confidence Elicitation Methods and Annotation Types

Metric	Method	Estimate	p-value
AUPRC	Simple confidence - Betting	23.1	0.0565

4.2.2 Effect of both confidence elicitation method and confidence annotation type on the quality of "soft" labels per participant. To measure the effects of the elicitation method, annotation type, and their interaction on the quality of the "soft" labels per participant, we calculated a two-way ART anova for each of the quality metrics using the cross-validation fold as blocking variable. We focused on the AUPRC metric that does not have a set threshold for binarizing the annotations. Figure 5 shows the AUPRC for both confidence elicitation methods and confidence annotation types. We found that the "Simple Confidence" elicitation method performs borderline statistically significantly higher than the "Betting" mechanism at $\alpha = 0.1$ (AUPRC: $F(1, 360) = 3.66$, $p = 0.0565$). However, no significant effects were found for the continuous annotation types (probability versus z-score derived probability).

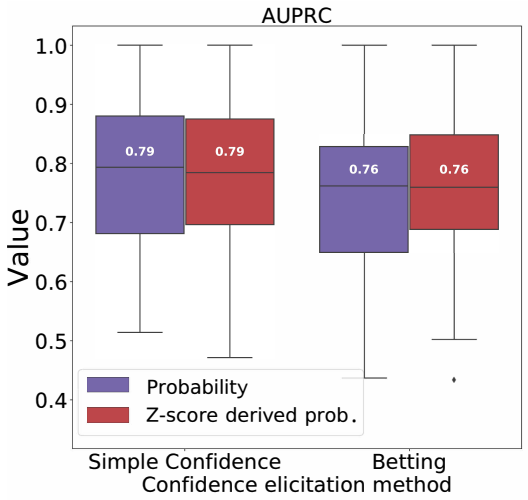


Fig. 5. *Individual* AUPRC boxplots averaged over cross-validation folds, for each of the confidence elicitation methods and annotation type. The lines in the boxplots correspond to the medians, the boxes are the inner quartiles, and the whiskers extend to 1.5 the IQR.

4.3 Do the confidence elicitation method and aggregate label type affect *aggregate* (i.e. crowd) annotation quality and cost?

We also want to understand the effect of elicitation and per-recording aggregation methods on aggregated annotation quality.

For each elicitation method, we:

(1) Averaged the binary labels per recording.
(2) Averaged the probability labels per recording.
(3) Averaged the z-score derived probability labels per recording.

Finally, the aggregate labels were directly binarized using a threshold of 0.5 to form what we call "hard aggregate" labels. The non-binarized aggregates are called "soft aggregate" labels.

4.3.1 Effect of confidence elicitation method and annotation type on aggregate *labels.* Similarly to Section 4.2, we measured performance with AUPRC, as it does not depend on choosing a threshold for binarizing the labels. Each curve in Figure 6 represents the average over 500 iterations when calculating AUPRC (Figure 6), where in each iteration we randomly ordered the participants per condition, i.e., adding one randomly selected participant per cross-validation fold at a time for each confidence elicitation mechanism and annotation type.

For averaged binary labels, both confidence elicitation mechanisms produce aggregate labels of higher quality than the baseline "No Confidence" method at all levels of participant averaging. Within each of the two confidence elicitation methods, we see that "soft" labels have higher quality than "hard" labels for all levels of participant averaging. In addition, much like the per-participant results presented in Section 4.2, when averaged over few participants, there is little difference between the two elicitation mechanisms. However, as we increase the number of participants, the quality of labels from the "Betting" mechanism achieve higher performance than those for the "Simple Confidence" mechanism.

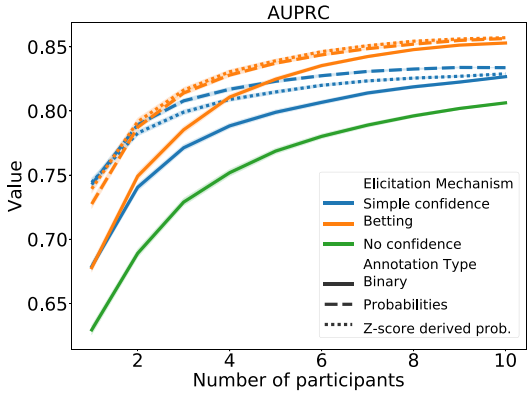


Fig. 6. *Aggregate* AUPRC quality metric while varying the the number of annotators aggregated per example shown for each elicitation mechanism and annotation type

*4.3.2 Effect of confidence elicitation method and annotation type on aggregate label type (*hard *or* soft*) based on annotator agreement.* While the previous analysis investigates the effect of the annotation mechanisms on the binary annotations, we cannot perform a similar per-participant annotation quality analysis for the "soft annotations" because we do not have ground-truth "soft annotations" to compare to. Thus, to measure the quality of the "soft annotations", we instead measured annotator agreement, using the Krippendorff's α coefficient, as we increased the number of annotators and compared it to the aggregated annotations for all 10 participants. By measuring the "soft annotation" quality relative to the 10-participant aggregate annotations we can establish when the annotator agreement gain, as a result of adding one participant, plateaus for "soft annotation". At the same time, we can directly compare the performance of "hard" versus "soft" aggregate label types. Figure 7 shows how the elicitation mechanism, annotation type and number of participants affect annotator agreement for "soft" and "hard" aggregate label type. When comparing "soft" and "hard" aggregate label types for each confidence elicitation method and annotation type as in Figure 7, we can see that agreement for "soft aggregate" labels is higher than agreement for "hard aggregate" labels at any number of participants. Similarly to findings by Chung et al. [17], these results suggest that using "soft aggregate" labels, we can collect annotations from fewer participants than if we use "hard aggregate" labels.

5 DISCUSSION

5.1 Findings

In this study we have evaluated two confidence elicitation methods: 1) "Simple confidence" and 2) "Betting-based" confidence, evaluated three annotation types: 1) binary responses, 2) probabilities directly transformed from confidence submitted by participants, and 3) probabilities transformed from z-scores of confidence submitted by participants, and compared two aggregate label types: 1) "hard" and 2) "soft".

In our analysis, we first studied if the confidence elicitation method affects the *individual* annotation quality. The results show that individual binary annotations collected using either of the confidence elicitation methods have higher quality than those collected using the baseline "No Confidence" task. This finding suggests that giving participants an additional step to provide their confidence increases annotator performance. This may be explained by participants' increased attention to the task driven by the need to re-evaluate their performance of it. Such an explanation

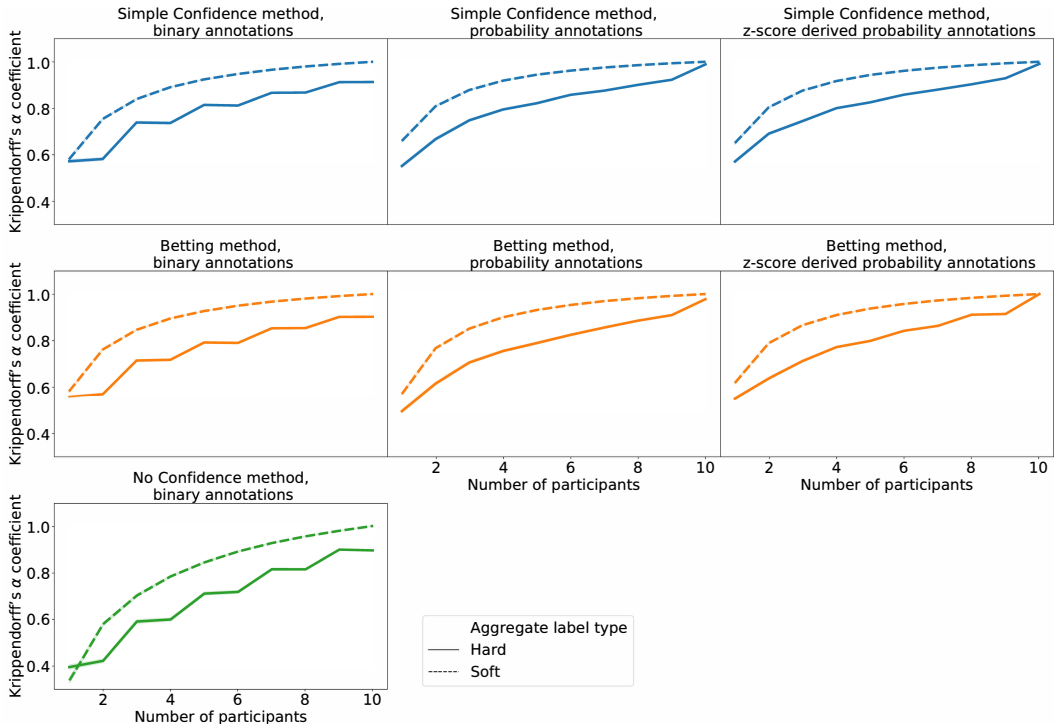


Fig. 7. Annotator agreement measured by the Krippendorff's α coefficient as the number of annotators is increased, when compared to the 10-participant aggregation, for each confidence elicitation method and annotation type.

is supported by studies that have found that participants who were are given the opportunity to change their answers achieve higher quality answers than those who are not allowed [3, 4, 49, 51, 70]. However, we did not collect participant interaction logs and thus more evidence is needed to support this claim. Moreover, these results suggest that eliciting confidence is beneficial even for tasks where the model is trained with only binary targets, e.g. binary classification with a support vector machine or decision tree.

We evaluated the quality of "soft" labels by calculating AUPRC, a metric that does not depend on a set decision threshold for binarizing the labels. For the *individual* annotations, the "Simple Confidence" method achieves higher quality than the "Betting" mechanism but with only borderline statistical significance. On the *aggregate* labels, the "Betting" and "Simple Confidence" mechanisms achieve similar performance at low participant aggregation levels, but as the number of participants averaged in the aggregates increases, the quality of the labels from "Betting" mechanism increases to a higher level than those from the "Simple Confidence" mechanism. In addition, for all participant levels, the "soft aggregate" labels yield higher quality labels than the "hard aggregate" labels. Moreover, when analyzing our data sample containing completion times, the "Betting" task took less time to complete than the "Simple Confidence" task (7.19 vs 7.65 minutes). Although directly eliciting confidence seems more straightforward, providing a slider interface seems to make the overall task slower than the buttons in the "Betting" method. Additionally, one way to speed up more the "Betting" method would be to stop asking participants for their choice once they select the lottery and assign the lottery to the remaining rounds when randomly selecting the round for

payment. We did not do this in the current study because we wanted to make sure that participants were selecting their true bets instead of randomly going from betting to their answer and providing conflicting annotations. However, given that we did not find conflicting annotations to be an issue, this is an easy change that may reduce task completion times.

Thus, similarly to previous work [5, 17, 26, 39], we find that when using confidence elicitation methods, the resulting annotations have higher accuracy and precision when compared to those of the baseline ("No Confidence" elicitation), giving further evidence that collecting more fine-grained annotations is beneficial when dealing with ambiguous data. With the methods evaluated in this paper, our aim was to be able to collect continuous labels not just in the aggregate from multiple annotators but from individuals. Because some mediums take longer to annotate, asking a bit more from individual annotators is preferable to having multiple annotators per data item. In addition, the methods evaluated in this paper fall between the *multi-label* and *probability multi-label* approaches from Chung et al. [17], and while the methods from Chung et al. [17] are better suited to multi-class problems, as explained in the introduction, we believe the methods presented in this paper are more suitable for binary and multi-label tasks. We propose the "Betting" method as a way of collecting probabilities while avoiding directly asking participants for their confidence, since it has been shown that people misjudge their own confidence [69].

The proposed methods can be used in both binary annotation tasks and expanded to multi-label tasks, as discussed in Section 5.2. We evaluate our methods in an audio annotation task which has applications to many important real-world problems, e.g., noise pollution monitoring, urban bioacoustic monitoring, audio-based traffic monitoring, sound-awareness tools for the hard of hearing, audio forensics, and machine perception for autonomous robots/vehicles. Moreover, the annotation methods are not limited to an audio use case. The confidence elicitation methods evaluated could be applied to other annotation scenarios where data can be ambiguous, binary or multi-label, e.g., tasks such as video annotation for visual object detection and tracking, or multiple evidence-based diagnosis.

The results of this study imply the following recommendations. If your machine learning method supports "soft" labels as targets, then we recommend eliciting confidence labels for ambiguous data, but which elicitation method depends on your label quality requirements and your annotation budget. For lower annotation budgets in which only one annotator per instance is collected, the differences between the two evaluated mechanisms are small. However for larger annotation budgets which can support increased quality through additional participants, the results show that the "Betting" mechanism yields higher quality labels. In fact, even if your machine learning method only supports "hard" labels as targets, we still recommend using the confidence elicitation mechanisms for higher quality labels on ambiguous data if the annotation budget is there. For example, for only a 10% increase in annotation cost (due to the increase in annotation time in the elicitation tasks), the increase in quality of the binary labels is equivalent to adding an additional annotator per example in the "No confidence" task (100% increase in annotation time).

5.2 Incorporating our approach into a multi-label annotation pipeline

To collect annotations for real-world audio data we would need to make some adjustments. First, we would need to some "ground-truth" or "gold standard" data in order to pay participants. Our approach depends on "ground-truth" annotations for verifying performance and determining payment. If we mix in recordings for which we know this information, we would base the payment on those recordings. This is a standard approach in crowdsourcing for verifying the quality of the annotations and has been used and discussed in multiple works [17, 35, 40, 43, 64].

Given that real-world audio recordings are typically multi-label, it would also be important to incorporate multi-label annotation into the annotation pipeline. In previous work, researchers

found that people tend to under annotate in multi-label settings, finding a lot of disagreement in some classes [11]. Moreover, fine-grained annotations benefit more ambiguous data [17], but in non-ambiguous data, this type of annotations are not necessary. This is especially important when the number of classes present is too large for setting them up in a one-stage pipeline, which is a common case in real-world audio recordings. For future work, we propose a two-stage annotation setup where we first collect multi-label annotations with fewer participants. On those classes where there is disagreement, we setup a second stage, in which participants provide probability estimates using the most appropriate confidence elicitation for the task, incorporating this work into a multi-label annotation pipeline. This approach would allow us to focus specifically on classes with disagreements — which are common in many real-world settings such as healthcare — and which require the collection of useful "soft" labels. In future work, we would like to explore how to aggregate the *binary* annotations from the first stage with the "soft" annotations from the second stage efficiently.

5.3 Limitations

The data that we created for this experiments was generated using the UrbanSound8k dataset [59], a dataset containing 10 urban sound classes. We chose this source dataset because its classes are similar to the scenario of our real-world use case. While there are many possible class combinations that could result in ambiguous acoustic scenes, exploring all such combinations is infeasible in a controlled human subjects experiment. Thus, we limited target classes and distraction class combinations to a feasible set for evaluation and ones likely to result in ambiguous acoustic scenes. To generate our positive *jackhammer* examples, we incorporated *engine idling* and *drilling* as distraction classes. These two sources follow similar spectro-temporal patterns to those of the *jackhammer* making it more difficult to disambiguate them [9]. While this set is limited, we believe the results of our experiment should generalize to other class combinations with overlapping spectro-temporal patterns.

While we believe these results are applicable to other domains in which data is ambiguous, the "Betting" elicitation mechanism only increased the task completion time by an average of 1.98 s per item ($60 * (7.19 - 6.53)/20 = 1.98\ s$). While this increase in cost is a small percentage of the cost for audio annotation, this increase may be too high for domains in which annotation is fast. Therefore, the results of this paper are most applicable slower annotation tasks (e.g., audio, video, long-text annotation, etc.).

As a workaround for a data collection mishap in which we initially only collected timing information for the "No Confidence" tasks, we collected an additional sample of data for the "Betting" and "Simple Confidence" tasks (20 participants each) at a later point in time. Because of this, we don't have paired timing and quality data for our entire sample of data. It is possible that the distribution of Mechanical Turk workers changed in the time between the two samples, however our selection criteria was the same for both samples. In addition, while the task completion times currently show that participants took less time with the "Betting" task, if a task requires an increase in confidence granularity (i.e., increasing the number of confidence levels per task), it is unclear how such an increase will affect each of the task completion times. While one may predict that additional questions will slow down the granularity in the "Betting" task, it may also be that additional slider levels may increase the cognitive load and thus completion time for the "Simple Confidence" task.

Finally, to draw conclusions about the listening behavior from the elicitation tasks, it is necessary to collect data about the interaction with the interface, e.g., interaction logs. With this type of data, we could understand if participants listened to the recordings more than once and if they switched their binary responses after listening to the recordings, or after doing some bets. Unfortunately,

this interaction logging data was not collected, and thus such an investigation is left for a future study.

5.4 Future Work

One of our main motivations is to collect annotations that can help achieve more accurate machine learning models. For future work, we will explore the use of the collected "soft" labels for training machine learning models, as we expect that the use of the "soft" labels will help us achieve higher model performance than models trained with "hard" labels, while using fewer participants.

Moreover, we will test the generalizability of our results to other tasks and domains with ambiguous sources. Also, we will evaluate how different sources of ambiguity affect the quality of the annotations. Additionally, we will test different step sizes for increasing granularity of the probabilities and its effect on the annotation quality and the cost (number of annotators). Finally, we will expand our experiments to real-world audio data in a multi-label, two-stage annotation approach.

6 CONCLUSION

In this study we investigated three methods for collecting crowdsourced audio annotations based on confidence elicitation: 1) one in which we don't elicit confidence, 2) one in which we elicit confidence with a simple confidence slider, and 3) one in which we elicit confidence through a betting mechanism. In addition, we also investigated whether discretizing the continuous confidence-based labels to binary affects annotation quality. We analyzed the resulting annotations both individually and aggregated over all annotators.

We show that confidence elicitation results in *individual* binary annotations that are of higher quality at the cost of only a minimal increase in annotation time. This implies that we should collect confidence information when dealing with ambiguous data even when machine learning models have only binary targets, e.g., if a classification model only accepts "yes"/"no" labels for training. As an example, a noise monitoring system that detects whether a recording contains sound sources such as "car horns", "sirens", "construction", and "dogs barking", would benefit from collecting confidence information even if the model only accepts binary labels for each class.

When comparing confidence elicitation methods, we found that eliciting confidence with a simple slider mechanism resulted in *individual* confidence annotations of marginally higher quality than the betting mechanism. However, the betting mechanism resulted in higher quality *aggregate* annotations, with differences being more pronounced as more annotators were added. Thus we recommend using the betting mechanism for confidence elicitation when the annotation budget allows it, and when in need of higher quality data. Finally, we show that aggregating confidence annotations results in higher quality aggregate labels than when directly aggregating the binary labels, a finding that is consistent with previous literature [17].

In conclusion, using a confidence elicitation mechanism and aggregating continuous labels results in higher annotation quality. We propose using a betting-based mechanism to indirectly elicit confidence. In our experiments, we found it to be a faster mechanism for collecting annotations than the asking annotators to rate confidence with a slider, and we found it results in higher quality *aggregate* annotations, a desired result when dealing with potentially life changing tasks such as sound-awareness tools for the hard of hearing, audio forensics, and machine perception for autonomous robots/vehicles. While we evaluate our methods in a binary classification task, we discuss how we can expand this work into a multi-label approach, i.e., where there is more than one class present at a time, which is the case for most of the tasks discussed above. We investigate these annotation tasks in the context of urban audio classification, but we believe these confidence

elicitation mechanisms may be useful in a variety of subjective and perceptual annotation tasks in which data is ambiguous and annotation is slow.

7 ACKNOWLEDGEMENTS

This work was partially supported by National Science Foundation awards 1544753 (https://www.nsf.gov/awardsearch/showAward?AWD_ID=1544753) and 1928614 (https://www.nsf.gov/awardsearch/showAward?AWD_ID=1928614). This work also would not be possible without the participation and effort of many workers on Amazon's Mechanical Turk platform.

REFERENCES

[1] Vamshi Ambati, Stephan Vogel, and Jaime G Carbonell. 2010. Active Learning and Crowd-Sourcing for Machine Translation.. In *LREC*, Vol. 1. 2.

[2] Lora Aroyo and Chris Welty. 2015. Truth is a lie: Crowd truth and the seven myths of human annotation. *AI Magazine* 36, 1 (2015), 15–24.

[3] Yigal Attali and Don Powers. 2010. Immediate feedback and opportunity to revise answers to open-ended questions. *Educational and Psychological Measurement* 70, 1 (2010), 22–35.

[4] Yigal Attali and Fabienne van der Kleij. 2017. Effects of feedback elaboration and feedback timing during computer-based practice in mathematics problem solving. *Computers & Education* 110 (2017), 154–169.

[5] Alexandry Augustin, Matteo Venanzi, J Hare, A Rogers, and NR Jennings. 2017. Bayesian aggregation of categorical distributions with applications in crowdsourcing. AAAI Press/International Joint Conferences on Artificial Intelligence.

[6] A. M. Aung and J. Whitehill. 2018. Harnessing Label Uncertainty to Improve Modeling: An Application to Student Engagement Recognition. In *2018 13th IEEE International Conference on Automatic Face Gesture Recognition (FG 2018)*. 166–170.

[7] Yoram Bachrach, Tom Minka, John Guiver, and Thore Graepel. 2012. How to Grade a Test without Knowing the Answers: A Bayesian Graphical Model for Adaptive Crowdsourcing and Aptitude Testing *(ICML'12)*. Omnipress, Madison, WI, USA, 819–826.

[8] Juan Pablo Bello, Cláudio T. Silva, Oded Nov, R. Luke DuBois, Anish Arora, Justin Salamon, Charles Mydlarz, and Harish Doraiswamy. 2018. SONYC: A System for the Monitoring, Analysis and Mitigation of Urban Noise Pollution. *CoRR* abs/1805.00889 (2018). arXiv:1805.00889 http://arxiv.org/abs/1805.00889

[9] Albert S Bregman. 1994. *Auditory scene analysis: The perceptual organization of sound*. MIT press.

[10] Mark Cartwright, Jason Cramer, Ana Elisa Mendez Mendez, Yu Wang, Ho-Hsiang Wu, Vincent Lostanlen, Magdalena Fuentes, Graham Dove, Charlie Mydlarz, Justin Salamon, et al. 2020. SONYC-UST-V2: An Urban Sound Tagging Dataset with Spatiotemporal Context. *arXiv preprint arXiv:2009.05188* (2020).

[11] Mark Cartwright, Graham Dove, Ana Elisa Méndez Méndez, Juan P. Bello, and Oded Nov. 2019. Crowdsourcing Multi-Label Audio Annotation Tasks with Citizen Scientists *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–11. https://doi.org/10.1145/3290605.3300522

[12] Mark Cartwright, Ayanna Seals, Justin Salamon, Alex Williams, Stefanie Mikloska, Duncan MacConnell, Edith Law, Juan P Bello, and Oded Nov. 2017. Seeing Sound: Investigating theEffects of Visualizations and Complexity on Crowdsourced Audio Annotations. *Proceedings of the ACM on Human-Computer Interaction* 1, CSCW (2017), 29.

[13] Joseph Chee Chang, Saleema Amershi, and Ece Kamar. 2017. Revolt: Collaborative crowdsourcing for labeling machine learning datasets. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. 2334–2346.

[14] Quanze Chen, Jonathan Bragg, Lydia B. Chilton, and Dan S. Weld. 2019. Cicero: Multi-Turn, Contextual Argumentation for Accurate Crowdsourcing *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3290605.3300761

[15] Lydia B Chilton, Greg Little, Darren Edge, Daniel S Weld, and James A Landay. 2013. Cascade: Crowdsourcing taxonomy creation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 1999–2008.

[16] Minsuk Choi, Cheonbok Park, Soyoung Yang, Yonggyu Kim, Jaegul Choo, and Sungsoo Ray Hong. 2019. Aila: Attentive interactive labeling assistant for document classification through attention-based deep neural networks. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–12.

[17] John Joon Young Chung, Jean Y Song, Sindhu Kutty, Sungsoo Hong, Juho Kim, and Walter S Lasecki. 2019. Efficient elicitation approaches to estimate collective crowd answers. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–25.

[18] Albin Andrew Correya. 2017. *Retrieving Ambiguous Sounds Using Perceptual Timbral Attributes in Audio Production Environments*. Ph.D. Dissertation. Diploma thesis, 2017 10, 30.

[19] Joana Costa, Catarina Silva, Mário Antunes, and Bernardete Ribeiro. 2011. On using crowdsourcing and active learning to improve classification performance. In *Intelligent Systems Design and Applications (ISDA), 2011 11th International Conference on*. 469–474.

[20] Nilesh Dalvi, Anirban Dasgupta, Ravi Kumar, and Vibhor Rastogi. 2013. Aggregating crowdsourced binary ratings. In *Proceedings of the 22nd international conference on World Wide Web*. 285–294.

[21] Bruno de Finetti. 1992. *Foresight: Its Logical Laws, Its Subjective Sources*. Springer New York, New York, NY, 134–174. https://doi.org/10.1007/978-1-4612-0919-5_10

[22] Shuyuan Deng, Atish P Sinha, and Huimin Zhao. 2017. Resolving ambiguity in sentiment classification: The role of dependency features. *ACM Transactions on Management Information Systems (TMIS)* 8, 2-3 (2017), 1–13.

[23] Pranay Dighe, Afsaneh Asaei, and Hervé Bourlard. 2017. Low-Rank and Sparse Soft Targets to Learn Better DNN Acoustics Models. In *Proceedings of the 2017 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. 5265–5269.

[24] Anca Dumitrache. 2015. Crowdsourcing Disagreement for Collecting Semantic Annotation. In *The Semantic Web. Latest Advances and New Domains*, Fabien Gandon, Marta Sabou, Harald Sack, Claudia D'Amato, Philippe Cudré-Mauroux, and Antoine Zimmermann (Eds.). Springer International Publishing, Cham, 701–710.

[25] Anca Dumitrache, Lora Aroyo, and Chris Welty. 2018. Capturing ambiguity in crowdsourcing frame disambiguation. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing (HCOMP) 2018*.

[26] Anca Dumitrache, Lora Aroyo, and Chris Welty. 2018. Crowdsourcing Ground Truth for Medical Relation Extraction. *ACM Trans. Interact. Intell. Syst.* 8, 2, Article 11 (July 2018), 20 pages. https://doi.org/10.1145/3152889

[27] Soufiane El Jelali, Abdelouahid Lyhyaoui, and Aníbal R. Figueiras-Vidal. 2008. Applying emphasized soft targets for Gaussian mixture model based classification. *Proceedings of the International Multiconference on Computer Science and Information Technology, IMCSIT 2008* 3 (2008), 131–136. https://doi.org/10.1109/IMCSIT.2008.4747229

[28] Eduardo Fonseca, Xavier Favory, Jordi Pons, Frederic Font, and Xavier Serra. 2020. FSD50k: an open dataset of human-labeled sound events. *arXiv preprint arXiv:2010.00475* (2020).

[29] Bin-Bin Gao, Chao Xing, Chen-Wei Xie, Jianxin Wu, and Xin Geng. 2017. Deep label distribution learning with label ambiguity. *IEEE Transactions on Image Processing* 26, 6 (2017), 2825–2838.

[30] Chao Gao, Yu Lu, and Dengyong Zhou. 2016. Exact exponent in optimal rates for crowdsourcing. In *International Conference on Machine Learning*. 603–611.

[31] J. F. Gemmeke, D. P. W. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter. 2017. Audio Set: An Ontology and Human-Labeled Dataset for Audio Events. In *Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 776–780.

[32] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the Knowledge in a Neural Network. (2015), 1–9. https://doi.org/10.1063/1.4931082 arXiv:1503.02531

[33] Sungsoo Hong, Minhyang Suh, Nathalie Henry Riche, Jooyoung Lee, Juho Kim, and Mark Zachry. 2018. Collaborative dynamic queries: Supporting distributed small group decision-making. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–12.

[34] Sungsoo Hong, Minhyang Suh, Tae Soo Kim, Irina Smoke, Sangwha Sien, Janet Ng, Mark Zachry, and Juho Kim. 2019. Design for Collaborative Information-Seeking: Understanding User Challenges and Deploying Collaborative Dynamic Queries. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–24.

[35] John J Horton, David G Rand, and Richard J Zeckhauser. 2011. The online laboratory: Conducting experiments in a real labor market. *Experimental economics* 14, 3 (2011), 399–425.

[36] Eric Humphrey, Simon Durand, and Brian McFee. 2018. OpenMIC-2018: An Open Data-set for Multiple Instrument Recognition. In *Proceedings of the 19th International Society for Music Information Retrieval Conference*. ISMIR, Paris, France, 438–444. https://doi.org/10.5281/zenodo.1492445

[37] Youxuan Jiang, Catherine Finegan-Dollak, Jonathan K Kummerfeld, and Walter Lasecki. 2018. Effective crowdsourcing for a new type of summarization task. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*. 628–633.

[38] Youxuan Jiang, Jonathan K. Kummerfeld, and Walter S. Lasecki. 2017. Understanding Task Design Trade-offs in Crowdsourced Paraphrase Collection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics, Vancouver, Canada, 103–109. https://doi.org/10.18653/v1/P17-2017

[39] David Jurgens. 2013. Embracing ambiguity: A comparison of annotation methodologies for crowdsourcing word sense labels. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 556–562.

[40] Aniket Kittur, Ed H Chi, and Bongwon Suh. 2008. Crowdsourcing user studies with Mechanical Turk. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 453–456.

[41] Matthäus Kleindessner and Pranjal Awasthi. 2018. Crowdsourcing with arbitrary adversaries. In *International Conference on Machine Learning*. 2708–2717.
[42] Edith Law and Luis von Ahn. 2011. Human computation. *Synthesis Lectures on Artificial Intelligence and Machine Learning* 5, 3 (2011), 1–121.
[43] Edith Law and Luis Von Ahn. 2011. *Human Computation*. Morgan & Claypool, San Rafael, CA.
[44] Maël Lebreton, Shari Langdon, Matthijs J. Slieker, Jip S. Nooitgedacht, Anna E. Goudriaan, Damiaan Denys, Ruth J. van Holst, and Judy Luigjes. 2018. Two Sides of the Same Coin: Monetary Incentives Concurrently Improve and Bias Confidence Judgments. *Science Advances* 4, 5 (2018). https://doi.org/10.1126/sciadv.aaq0668
[45] Sang Won Lee, Rebecca Krosnick, Sun Young Park, Brandon Keelean, Sach Vaidya, Stephanie D O'Keefe, and Walter S Lasecki. 2018. Exploring real-time collaboration in crowd-powered systems through a ui design tool. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 1–23.
[46] Qiang Liu, Jian Peng, and Alexander T Ihler. 2012. Variational inference for crowdsourcing. In *Advances in neural information processing systems*. 692–700.
[47] Yao Ma, Alexander Olshevsky, Csaba Szepesvari, and Venkatesh Saligrama. 2018. Gradient descent for sparse rank-one matrix completion for crowd-sourced aggregation of sparsely interacting workers. In *International Conference on Machine Learning*. PMLR, 3335–3344.
[48] Edoardo Manino, Long Tran-Thanh, and Nicholas R Jennings. 2016. Efficiency of active learning for the allocation of workers on crowdsourced classification tasks. *arXiv preprint arXiv:1610.06106* (2016).
[49] Jeremy D Merrel, Pier F Cirillo, Pauline M Schwartz, and Jeffrey Webb. 2015. Multiple-Choice Testing Using Immediate Feedback-Assessment Technique (IF AT®) Forms: Second-Chance Guessing vs. Second-Chance Learning? (2015).
[50] Emily Mower, Angeliki Metallinou, Chi-Chun Lee, Abe Kazemzadeh, Carlos Busso, Sungbok Lee, and Shrikanth Narayanan. 2009. Interpreting ambiguous emotional expressions. In *2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops*. IEEE, 1–8.
[51] Susanne Narciss, Sergey Sosnovsky, Lenka Schnaubert, Eric Andrès, Anja Eichelmann, George Goguadze, and Erica Melis. 2014. Exploring feedback and student characteristics relevant for personalizing feedback strategies. *Computers & Education* 71 (2014), 56–76.
[52] Stefanie Nowak and Stefan Rüger. 2010. How Reliable are Annotations via Crowdsourcing. In *Proceedings of the International Conference on Multimedia Information Retrieval*. 557. https://doi.org/10.1145/1743384.1743478
[53] Charles Kay Ogden and Ivor Armstrong Richards. 1923. *The Meaning of Meaning: A Study of the Influence of Language upon Thought and of the Science of Symbolism*. Vol. 29. K. Paul, Trench, Trubner & Company, Limited.
[54] Sylvester Olubolu Orimaye, Saadat M Alhashmi, and Siew Eu-gene. 2012. Sentiment analysis amidst ambiguities in YouTube comments on Yoruba language (nollywood) movies. In *Proceedings of the 21st International Conference on World Wide Web*. 583–584.
[55] Navindra Persaud, Peter McLeod, and Alan Cowey. 2007. Post-Decision Wagering Objectively Measures Awareness. *Nature Neuroscience* 10, 2 (2007), 257–261. https://doi.org/10.1038/nn1840
[56] Vikas C Raykar, Shipeng Yu, Linda H Zhao, Gerardo Hermosillo Valadez, Charles Florin, Luca Bogoni, and Linda Moy. 2010. Learning from crowds. *Journal of Machine Learning Research* 11, 4 (2010).
[57] Gaurav Sahu. 2019. Multimodal speech emotion recognition and ambiguity resolution. *arXiv preprint arXiv:1904.06022* (2019).
[58] Justin Salamon, Juan Pablo Bello, Andrew Farnsworth, Matt Robbins, Sara Keen, Holger Klinck, and Steve Kelling. 2016. Towards the Automatic Classification of Avian Flight Calls for Bioacoustic Monitoring. *PLOS ONE* 11, 11 (2016), 1–26. https://doi.org/10.1371/journal.pone.0166866
[59] Justin Salamon, Christopher Jacoby, and Juan Pablo Bello. 2014. A Dataset and Taxonomy for Urban Sound Research *(MM '14)*. ACM, New York, NY, USA, 1041–1044. https://doi.org/10.1145/2647868.2655045
[60] Justin Salamon, Duncan MacConnell, Mark Cartwright, Peter Li, and Juan Pablo Bello. 2017. Scaper: A library for soundscape synthesis and augmentation. In *Applications of Signal Processing to Audio and Acoustics (WASPAA), 2017 IEEE Workshop on*. IEEE, 344–348.
[61] Kristian Sandberg, Bert Timmermans, Morten Overgaard, and Axel Cleeremans. 2010. Measuring Consciousness: Is One Measure Better Than the Other? *Consciousness and Cognition* 19, 4 (2010), 1069–1078. https://doi.org/10.1016/j.concog.2009.12.013
[62] Nihar B. Shah, Sivaraman Balakrishnan, and Martin J. Wainwright. 2016. A Permutation-based Model for Crowd Labeling: Optimal Estimation and Robustness. *CoRR* abs/1606.09632 (2016). arXiv:1606.09632 http://arxiv.org/abs/1606.09632
[63] Victor S Sheng, Foster Provost, and Panagiotis G Ipeirotis. 2008. Get another label? improving data quality and data mining using multiple, noisy labelers. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 614–622.

[64] Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Y Ng. 2008. Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks. In *Proceedings of the conference on empirical methods in natural language processing*. Association for Computational Linguistics, 254–263.

[65] Jean Y. Song, Raymond Fok, Juho Kim, and Walter S. Lasecki. 2019. FourEyes: Leveraging Tool Diversity as a Means to Improve Aggregate Accuracy in Crowdsourcing. *ACM Trans. Interact. Intell. Syst.* 10, 1, Article 3 (Aug. 2019), 30 pages. https://doi.org/10.1145/3237188

[66] Bettina Studer and Luke Clark. 2011. Place Your Bets: Psychophysiological Correlates of Decision-Making Under Risk. *Cognitive, Affective, & Behavioral Neuroscience* 11, 2 (2011), 144–158. https://doi.org/10.3758/s13415-011-0025-2

[67] Tian Tian and Jun Zhu. 2015. Max-margin majority voting for learning from crowds. In *Advances in neural information processing systems*. 1621–1629.

[68] Matteo Venanzi, John Guiver, Gabriella Kazai, Pushmeet Kohli, and Milad Shokouhi. 2014. Community-based bayesian aggregation models for crowdsourcing. In *Proceedings of the 23rd international conference on World wide web*. 155–164.

[69] Willem A. Wagenaar and Gideon B. Keren. 1986. Does The Expert Know? The Reliability of Predictions and Confidence Ratings of Experts. In *Intelligent Decision Support in Process Environments*, Erik Hollnagel, Giuseppe Mancini, and David D. Woods (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 87–103.

[70] Kittitas Wancham and Kamonwan Tangdhanakanond. 2020. Effects of Feedback Types and Opportunities to Change Answers on Achievement and Ability to Solve Physics Problems. *Research in Science Education* (2020), 1–18.

[71] Shinji Watanabe, Takaaki Hori, Jonathan Le Roux, and John R. Hershey. 2017. Student-Teacher Network Learning with Enhanced Features. In *Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 5275–5279.

[72] Peter Welinder, Steve Branson, Pietro Perona, and Serge J Belongie. 2010. The multidimensional wisdom of crowds. In *Advances in neural information processing systems*. 2424–2432.

[73] Jacob Whitehill, Ting-fan Wu, Jacob Bergsma, Javier R Movellan, and Paul L Ruvolo. 2009. Whose vote should count more: Optimal integration of labels from labelers of unknown expertise. In *Advances in neural information processing systems*. 2035–2043.

[74] Jacob O Wobbrock, Leah Findlater, Darren Gergle, and James J Higgins. 2011. The aligned rank transform for nonparametric factorial analyses using only anova procedures. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 143–146.

[75] Ali Yadollahi, Ameneh Gholipour Shahraki, and Osmar R Zaiane. 2017. Current state of text sentiment analysis from opinion to emotion mining. *ACM Computing Surveys (CSUR)* 50, 2 (2017), 1–33.

[76] Yi-Hsuan Yang, Yu-Ching Lin, Ya-Fan Su, and Homer H Chen. 2008. A regression approach to music emotion recognition. *IEEE Transactions on audio, speech, and language processing* 16, 2 (2008), 448–457.

[77] Biqiao Zhang, Georg Essl, and Emily Mower Provost. 2017. Predicting the distribution of emotion perception: capturing inter-rater variability. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction*. 51–59.

[78] Dengyong Zhou, Sumit Basu, Yi Mao, and John C Platt. 2012. Learning from the wisdom of crowds by minimax entropy. In *Advances in neural information processing systems*. 2195–2203.

[79] Dengyong Zhou, Qiang Liu, John Platt, and Christopher Meek. 2014. Aggregating ordinal labels from crowds by minimax conditional entropy. In *International conference on machine learning*. 262–270.

Received January 2021; revised July 2021; accepted November 2021