Efficient Elicitation Approaches to Estimate Collective Crowd Answers

JOHN JOON YOUNG CHUNG, University of Michigan, USA
JEAN Y. SONG, University of Michigan, USA
SINDHU KUTTY, University of Michigan, USA
SUNGSOO (RAY) HONG, New York University, USA
JUHO KIM, KAIST, Republic of Korea
WALTER S. LASECKI, University of Michigan, USA

When crowdsourcing the creation of machine learning datasets, statistical distributions that capture diverse answers can represent ambiguous data better than a single best answer. Unfortunately, collecting distributions is expensive because a large number of responses need to be collected to form a stable distribution. Despite this, the efficient collection of answer distributions—that is, ways to use less human effort to collect estimates of the eventual distribution that would be formed by a large group of responses—is an under-studied topic. In this paper, we demonstrate that this type of estimation is possible and characterize different elicitation approaches to guide the development of future systems. We investigate eight elicitation approaches along two dimensions: *annotation granularity* and *estimation perspective*. Annotation granularity is varied by annotating i) a single "best" label, ii) all relevant labels, iii) a ranking of all relevant labels, or iv) real-valued weights for all relevant labels. Estimation perspective is varied by prompting workers to either respond with their own answer or an estimate of the answer(s) that they expect other workers would provide. Our study collected ordinal annotations on the emotional valence of facial images from 1,960 crowd workers and found that, surprisingly, the most fine-grained elicitation methods were *not* the most accurate, despite workers spending more time to provide answers. Instead, the most efficient approach was to ask workers to choose all relevant classes that others would have selected. This resulted in a 21.4% reduction in the human time required to reach the same performance as the baseline (i.e., selecting a single answer with their own perspective). By analyzing cases in which finer-grained annotations degraded performance, we contribute to a better understanding of the trade-offs between answer elicitation approaches. Our work makes it more tractable to use answer distributions in large-scale tasks such as ML training, and aims to spark future work on techniques that can efficiently estimate answer distributions.

CCS Concepts: • **Human-centered computing** → *Collaborative interaction.*

Additional Key Words and Phrases: crowdsourcing, annotation, ambiguity, answer distributions

ACM Reference Format:
John Joon Young Chung, Jean Y. Song, Sindhu Kutty, Sungsoo (Ray) Hong, Juho Kim, and Walter S. Lasecki. 2019. Efficient Elicitation Approaches to Estimate Collective Crowd Answers. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 62 (November 2019), 25 pages. https://doi.org/10.1145/3359164

62

Authors' addresses: John Joon Young Chung, jjyc@umich.edu, University of Michigan, Ann Arbor, MI, USA; Jean Y. Song, jyskwon@umich.edu, University of Michigan, Ann Arbor, MI, USA; Sindhu Kutty, skutty@umich.edu, University of Michigan, Ann Arbor, MI, USA; Sungsoo (Ray) Hong, rayhong@nyu.edu, New York University, New York City, NY, USA; Juho Kim, juhokim@kaist.ac.kr, KAIST, Daejeon, Republic of Korea; Walter S. Lasecki, wlasecki@umich.edu, University of Michigan, Ann Arbor, MI, USA.

2573-0142/2019/11-ART62 $15.00
https://doi.org/10.1145/3359164

1 INTRODUCTION

When creating Machine Learning (ML) datasets, it is a common practice to crowdsource a single best answer to represent a data instance (e.g., annotating a data instance with a label). To assure the quality of answers, ML practitioners and crowdsourcing researchers usually collect answers with high levels of agreement [71, 78]. However, in domains where answers are ambiguous or subjective, such as emotion recognition [82] or entity recognition [47], multiple valid interpretations may exist. One key reason for this is *data ambiguity*, where a data instance lacks sufficient contextual information to be annotated with a single label [1, 16, 17, 47, 64]. In these cases, *answer distributions* can more accurately represent people's interpretations than a single best answer would [16]. For example, in emotion annotation, they can indicate whether an emotion is subtly or clearly displayed [82]. As a result, answer distributions have been used as a supervisory signal for training ML models on ambiguous data [3, 17, 24–26, 41, 82]. Unfortunately, crowdsourcing answer distributions will use more human time (and thus, cost) because accurately estimating the proportion of people who would select each label requires more signal than identifying the single majority answer.

In this paper, we show it is possible to estimate the answer distribution of a larger group using fewer workers by eliciting richer responses (e.g., multiple labels with frequency information) from each worker. We investigate eight elicitation approaches along two dimensions: annotation granularity and estimation perspective (Table 1). For annotation granularity, we vary the amount of information that crowd workers are asked to provide, expecting that crowd workers estimate answer distributions more efficiently with finer granularity. We examine four levels of granularity: 1) choosing a single "best" label, 2) choosing all plausible labels, 3) ranking all plausible labels, and 4) annotating the real-valued probability that each label is a plausible answer. For estimation perspective, we ask workers to either respond with their own perspective or estimate from other people's perspective, similar to peer prediction from Bayesian Truth Serum [65, 66].

To evaluate these 4×2 experimental conditions, we measure accuracy and total human time required. Our study collected ordinal annotations from 1,960 crowd workers on the emotional valence of ambiguous facial expressions in images. Our results show that the most accurate and efficient approach was choosing all plausible labels that other workers would have selected. This approach achieved similar performance to the baseline (choosing a single best label) with **40% fewer workers** and **21.4% less human time**. It even outperformed approaches that elicited more fine-grained annotations and took more time. We observed that estimating from other workers' perspective was more effective only when selecting a set of plausible labels (without fine-grained weights). Further analysis showed that for the most fine-grained approach, which involved annotating the probability, workers had a tendency to concentrate probabilities to a smaller number of labels, which explains the trade-offs observed in annotation granularity. Our findings best apply to task domains in which crowd workers generate diverse answers mainly due to data ambiguity. Overall, we make using answer distributions more feasible in tasks that require a large amount of data, such as ML training, and characterize opportunities and challenges in designing elicitation approaches for a more efficient collection of answer distributions.

In this paper, we contribute the following:

- A systematic evaluation of eight elicitation approaches for estimating collective answer distributions, which vary by annotation granularity and estimation perspective.
- Experimental results and analysis on a facial image emotion annotation task, which show fine-grained annotations do not always lead to better estimation, due to heavily skewed estimations in answers from individual workers.
- Guidelines to apply our findings more broadly for efficient and accurate estimation of collective answer distributions in other domains.

2 RELATED WORK

In this section, we review research on 1) causes of answer variation in annotation tasks, 2) benefits of answer distributions in training ML models, 3) elicitation approaches used for ambiguous data annotation, 4) techniques that leverage people's ability to estimate other people's answers, and 5) approaches that leverage diverse answers in other domains.

2.1 Understanding Sources of Answer Variation in Annotation Tasks

Previous research has found that, for ambiguous data, annotators may generate disagreeing and diverse answers that are still valid, which implies a level of inaccuracy in representing such data with a single label. Using the triangle of reference [61], Dumitrache et al. [15] claimed that inter-annotator disagreement comes from three sources, 1) sign: the ambiguity of the data instance itself, 2) interpreter: annotators' different perspectives, and 3) referent: under-specified annotation design. For the aspect of sign, Plank et al. [64] found that even expert annotators disagreed in part-of-speech tagging and claimed that diverse answers can be due to the ambiguity of the data. On the other hand, Sen et al. [68] and Lee et al. [56] reported cases where answer disagreement was due to the different perspectives or expertise of contributors. They found experts and crowd workers generated systematic disagreement in tasks with domain-specific concepts. Kairam et al. [47] found that besides the ambiguity of data and annotators, referent, an unclear annotation design, can also be a source of disagreement in entity extraction tasks. Motivated by previous work that suggested ambiguous data would not be best represented with a single answer, we investigate annotation approaches to collect a distribution of answers more efficiently.

2.2 Benefits of Using Answer Distributions as Annotations

For ML models trained on ambiguous data, answer distributions would be more accurate annotations for data instances than a single answer. With the CrowdTruth metric, which computes the degree of disagreement within an answer distribution, Dumitrache et al. [16] found that answer distributions convey information about the ambiguity of data. Zhang et al. [82] used the answer distribution as the supervisory signal for training an emotion inference model, because it can capture the subtlety in an emotional display. Similarly, Aung et al. [3] used answer distributions when training a machine learning model that infers how much students are engaged in a lecture video from ambiguous facial images. For ambiguous data, using answer distributions as supervisory signals also benefits the performance of machine learning models. Aung et al. [3] and Gao et al. [24] found that using answer distributions gave rise to regularization effects as a model avoided learning from only one answer, but instead learned from multiple plausible answers. In this work, to make such benefits of answer distributions more feasible, we investigate more efficient elicitation approaches for collecting answer distributions.

2.3 Annotation Elicitation Approaches for Ambiguous Data

For ambiguous data instances that cannot be best represented with a single answer, researchers examined various elicitation approaches, but not with the focus on how efficiently and accurately those approaches estimate answer distributions. One approach is allowing annotators to choose multiple labels. In a medical relation extraction task, Dumitrache et al. [17] allowed annotators to choose multiple labels and aggregated responses in a vector in which each dimension is the frequency of crowd workers selecting each label. Dumitrache et al. showed that such an approach improved the performance of the trained algorithm, but did not show how it estimated answer distributions. Cascade and Deluge [5, 8] also allowed workers to select multiple labels to obtain a set of most relevant labels in the classification task. However, these systems focused on obtaining a full range of similar labels to give better context to crowd workers who later choose one final label.

Another approach is for workers to directly generate the answer distribution. Augustin et al. [2] collected annotations in the distributions and measured their accuracy on retrieving objective proportion values, such as the amount of color used in a flag, or getting distributions defined by expert annotators. Jurgens [46] elicited a weighted selection of multiple labels in word sense disambiguation, which is an ambiguous annotation domain, but focused on getting consistent answers with a high inter-annotator agreement. For elicitation approaches motivated by previous work, this work investigates how accurately and efficiently different elicitation approaches can be used in estimating collective answer distributions.

2.4 Estimating Other People's Perspective

The approach of estimating how other people would have answered a question has been used in previous work, but with a different purpose from our use case. Bayesian Truth Serum [65, 66] used distribution estimates on other people's answers to get one correct answer from by comparing the aggregated estimations with aggregated answers from people's own perspectives. For an ideation task, Teevan et al. [77] asked individuals to come up with more diverse ideas by making them assume different expert roles. Unlike these, for ambiguous data annotation tasks, our work evaluates if a low number of workers can estimate eventual answer distributions that many responses would form by assuming perspectives of other annotators.

2.5 Approaches that Leverage Diverse Answers

Our work builds on previous approaches that leveraged multiple answers to achieve a diverse range of goals, from collecting a thorough set of diverse answers to using diverse answers to compute out the best answer.

Data generation tasks for natural language processing, such as natural language elicitation [80], summarization [42] or paraphrase tasks [43] have elicited diverse answers to get thorough datasets that can make machine learning algorithms robust. In order to elicit a more diverse set of responses, previous research has used priming via different instructions or examples to help workers focus on different aspects of the text and hence generate diverse natural language data instances.

Work in crowdsourcing for visual tasks has also leveraged diverse responses to get a single "best" answer. For example, Song et al. [73, 74] elicited and aggregated diverse responses with various error patterns with different tools to get the most accurate results. Gurari et al. [33] expanded this approach to also consider diverse outputs of machine algorithms and more efficiently crowdsourced segmentations. Song et al. [75] aggregated annotations from diverse video frames to more accurately reconstruct 3D scenes from 2D videos.

Crowd ideation is another domain where getting diverse, non-overlapping responses is a crucial goal. To get more diverse responses from crowds, Siangliulue et al. [69] exposed crowds to a more diverse set of examples and showed that the timely exposure of examples can help crowds generate more ideas [70]. Girotto et al. also increased the diversity of ideas with personalized inspiration [27]. Teevan et al. let crowds assume different roles of experts and crowdsourced more diverse ideas [77]. In a civic engagement system for crowd ideation, Grau et al. [30] studied ways to motivate diverse participants, which can lead to diverse answers.

For subjective tasks, systems that aggregate diverse opinions help users to make optimal decisions. Kim et al. [50] built a system that helps citizens deliberate on public policy with an awareness of the opinions of diverse stakeholders. In the context of decision making, Hong et al. [34, 35] built a system that enables users to be aware of the diverse opinions of a small group of users.

In this paper we aim to use different elicitation approaches to more efficiently crowdsource the collection of diverse responses to ambiguous data.

		Annotation Granularity			
		Single	**Multiple**	**Ranking**	**Probability**
Estimation Perspective	***Self***	*Single*	*Multiple*	*Ranking*	*Probability*
	Others	*SingleEsti*	*MultipleEsti*	*RankingEsti*	*ProbabilityEsti*

Table 1. Elicitation approaches examined in this study, and the two dimensions they vary on: 1) the granularity of annotations and 2) whether to annotate based on a worker's own perspective or to estimate what the group as a whole would provide as answers.

(a) Annotation interface for *Single* and *SingleEsti*, in radio buttons. Annotators can only provide a single label.

(b) Annotation interface for *Multiple* and *MultipleEsti*, in checkboxes. Annotator can provide multiple labels.

(c) Annotation interface for *Ranking* and *RankingEsti*, in checkboxes with a ranking function, which ranks labels in the order of selection. Annotators provide multiple labels with ranking information.

(d) Annotation interface for *Probability* and *ProbabilityEsti*, which allows workers to assign tokens to labels such that they sum to 100. Annotators provide multiple labels with probability information.

Fig. 1. Annotation interfaces for our different annotation granularity elicitation approaches.

3 ANSWER ELICITATION APPROACHES

In this work, we investigate answer elicitation approaches to reliably estimate *collective answer distributions* for ambiguous data. A collective answer distribution captures the proportion of each label given by a group sampled from a target population. Since getting reliable collective answer distribution requires more responses (samples) [13], it can be prohibitively expensive for large-scale annotation tasks. In this section, we describe our approaches to elicit richer responses from each crowd worker, to estimate the collective answer distribution with less total crowd worker effort.

We considered three factors that cause answer variation (introduced in Section 2.1): *sign* (input data), *interpreter* (annotator perspective), and *referent* (annotation options). Ambiguous *sign*, or a data instance, serves as the major source of diverse answers that form the distribution and we explain how we included it in the study in Section 4.2. We primarily considered *referent* and *interpreter* when devising the elicitation methods for estimating collective answer distributions. For *referent*, we focused on the *annotation granularity*, which we varied across four levels, from the most coarse-grained single answer to the most fine-grained real-valued probabilities. For *interpreter*, we examined the *estimation perspective*, specifically, whether an annotator is responding from their own perspective or from an indirect perspective by estimating the answers of other workers. Here, we describe eight answer elicitation approaches (Table 1) derived from these two dimensions.

3.1 Varying the Granularity of Annotations

For annotation granularity, we explored four levels. The conventional and simplest annotation approach is *Single*, which only allows a single answer from a worker (Figure 1a). This approach is used as a baseline approach in our study. However, crowd workers would be able to offer more information than a single answer [79] because people can recognize multiple potential interpretations of ambiguous data [58]. Thus, we examine whether receiving more fine-grained annotations than a single answer could improve the efficiency of estimating collective answer distributions. *Multiple* (Figure 1b) allows workers to choose multiple labels that they believe are the most plausibly correct. *Ranking* (Figure 1c) asks workers to provide an ordered set of labels based on which they think will be the most correct. *Probability* (Figure 1d) asks workers to assign a real-valued weight to each label to represent the relative confidence they have in each label.

We expect that workers would estimate answer distributions more accurately with *Probability*, because it is the most precise annotation method which can even represent the exact collective answer distribution. In summary, we varied annotation granularity across four levels by changing the amount of information that workers can convey about their confidence in different labels.

3.2 Varying the Perspective of Workers with Prompts

For estimation perspective, we examined two viewpoints. *Self* asks workers to annotate with their own belief in what the correct label or a set of labels is. However, prior work has shown that asking people to estimate the beliefs of a group can more accurately capture the correct answers [65, 66]. Further, the different backgrounds and perspectives that workers bring to the task [18, 47, 68] may be best elicited if we ask them to estimate what others in a larger group would answer. Thus, *Others* asks workers to estimate the response of the group.

Our final set of eight experimental conditions is composed of all combinations of our four granularities paired with each of our two estimation perspectives (Table 1).

4 EXPERIMENTAL SETUP

To understand how elicitation approaches affect the efficiency of estimating collective answer distributions, we conducted an experiment with the task of annotating the emotional valence of facial expression in images. First, we describe how we measured the efficiency of approaches. Second, we explain why emotion annotation is an adequate domain for the study and which dataset we used. Then, we introduce how we collected gold standard distributions and how we measured the stability of gold standard distributions. After that, we explain how we recruited participants and our experimental procedure and interfaces for the eight elicitation approaches.

4.1 Metrics

We define *efficiency* as the cost required to reach a given level of accuracy. *Accuracy* is defined as the distance between an estimated distribution and a gold standard distribution. We explain how we create gold standard distributions in Section 4.3. To measure the distance, we used Wasserstein distance [63], which is the minimum cost of converting one continuous distribution to another. We did not use the other widely-used metric, KL divergence [52], because it does not reflect the inherently ordered relationship between ordinal labels. For instance, a distribution with weights concentrated on the negative emotion should be measured as more similar to the neutral emotion than to the positive emotion, but KL divergence assigns the same distance (Figure 2).

When calculating the Wasserstein distance, we mapped ordinal labels to continuous values with equal distances between adjacent labels, following the practice in ML [72, 82]. Figure 3 shows an illustration of how a difference between two distributions is expressed in Wasserstein distance. Our

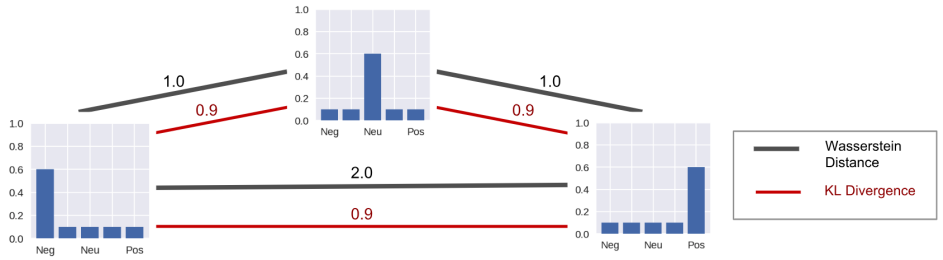


Fig. 2. Distances between distributions measured with two different metrics: Wasserstein distance and KL divergence. Wasserstein distance (depicted with black lines and values) more accurately represents the relationship between ordinal labels, measuring the distance between negative emotion and positive emotion as the farthest. However, KL divergence (depicted with red lines and values) cannot capture this relationship.

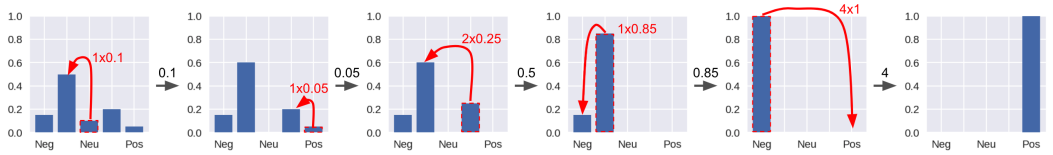


Fig. 3. An illustration of how changes in distribution affect the Wasserstein distance. The red box and the red arrow indicate how the weight moves, and the expression $A \times B$ indicates that the weight of B moved the distance A in the ordinal labels. The next distribution on the right indicates the distribution after the change. The number above the black arrow is the Wasserstein distance between the two distributions.

cost measure is defined in terms of the total human time used to complete the estimation. Because it is common to fix the hourly wage for a worker, the total human time can represent the total cost.

4.2 Dataset

Emotional expressions can have ambiguous characteristics and they are not always interpreted uniformly across different individuals [4]. Prior work has shown that representing emotional expressions with answer distributions can convey richer information, such as how subtly the emotion is expressed [82], but an efficient estimation of collective answer distribution has been under-explored. In this study, for facial expression images, we collected the answer distribution of the positiveness/negativeness of emotional valence, represented in five ordinal labels [82]. For the facial emotion image dataset, we used the FACES dataset [19]. The FACES dataset contains 2,052 facial expression images of 179 faces with diverse ages between 19 and 80. In the FACES dataset, each image in the dataset is intended to display one of six emotions: happiness, neutrality, anger, fear, sadness, or disgust. The main reason we used the FACES dataset is that we could select an ambiguous subset of images with provided validation annotations, which are ratings of how many annotators perceived the emotion of a person in an image as what the image was intended for. We used validation annotations to choose the top 40 most ambiguous images.

4.3 Gold Standard Distributions

For the gold standard of answer distribution estimation, we used collective answer distributions that consist of 50 *Single* annotations and measured their stability via bootstrapping.

4.3.1 Collection of gold standard distributions. For each data instance, we used the distribution of 50 *Single* annotations for gold standard distribution. We used *Single* annotations because it is the most widely-used annotation collection method. To minimize the effects of malicious workers and noisy task results, we used two methods. First, we used gold standard questions [36, 55, 62] with clear displays of positive or negative emotions. For the gold standard questions, we chose four images in which validation annotations in the FACES dataset fully agreed with the intended emotion. When a worker annotated these images incorrectly in the direction of emotional valence, we considered the worker as low-quality or malicious and omitted their data. Second, we collected a worker's reasoning for their annotations to filter out low-quality task results [60]. Specifically, we filtered out annotations in which a crowd worker's reasoning contradicted with their annotation. For instance, we filtered out an annotation where a worker answered with positive emotion and reasoning saying "This person seems very suspicious about something and it is upsetting him". Two authors inspected annotations and reasoning independently and had substantial agreements (Cohen's $\kappa = 0.80$). Disagreements between authors were resolved by discussions. With these filtering methods, we collected gold standard annotations until we reached 50 annotations for each data instance. Each crowd worker annotated five images including one gold standard question, and for the 40 ambiguous images, we recruited 500 workers from Amazon Mechanical Turk (MTurk) usign LegionTools [54]. We recruited crowd workers only in the US with an acceptance rate higher than 97%. We paid each worker \$0.90, which was an hourly wage of \$13.74.

4.3.2 Stability of gold standard distributions. Because we sampled 50 *Single* answer responses for each gold standard distribution, the distribution can be different if we sample answers again. To see how much gold standard distributions vary with resampling, we analyzed their stability with bootstrapping [20]. Specifically, we randomly resampled distributions from each gold standard distribution with replacement $10,000$ times and calculated the Wasserstein distance between the original distribution and the resampled one. Using this method, we estimate how variable the distributions were. The mean distance between gold standard distributions and resampled distributions was 0.11 ($\sigma = 0.07$). This mean is shown as gray dotted lines in Figure 5 and Figure 7.

4.4 Participants

For the 40 images, we collected 30 annotations for each image-condition pair. With eight elicitation approaches, we collected $9,600$ annotations in total. While crowd workers annotated five images, some annotations were lost due to technical issues, and therefore we continued recruiting workers until we collected $9,600$ annotations. Using LegionTools [54], we recruited a total of $1,960$ workers from MTurk, who were in the US and had an acceptance rate higher than 97%. We did not recruit workers who participated in the collection of the gold standard distribution. We paid workers \$1.20 for all annotation approaches, which yielded an average hourly wage of \$8.66.

4.5 Data Collection Procedure and Interface

To collect data, we conducted a between-subject study, where a worker only annotated with one elicitation approach. The study consisted of two parts: instructions and tasks.

When crowd workers entered the experiment, they were randomly assigned to one of the eight elicitation conditions. Workers were first given instructions explaining the annotation approach that they would use. To determine if workers understood the task, we added a quiz at the end of the instructions. For workers who did not pass this quiz, we excluded their data from the analysis.

After the instruction phase, crowd workers started annotating five emotional facial expression images. For *Single* and *SingleEsti*, workers were given radio buttons, and for *Multiple* and *MultipleEsti*, workers were given checkboxes. For *Ranking* and *RankingEsti*, workers were given checkboxes

	In the task page
Single	Annotate the person's emotion as you think. Choose the best value from the options below.
Multiple	Annotate the person's emotion as you think. Choose all possible values from the options below (You must choose at least one option).
Ranking	Annotate the person's emotion as you think. Choose all possible values from the options below and rank them. The most plausible answer should go first and the least plausible answer should go last. (You must choose at least one option, but may not need to choose all options).
Probability	Annotate the person's emotion as you think. Decide how probably each value can represent the person's emotion (You are given 100 tokens to express your subjective probability).
SingleEsti	Imagine 100 crowd workers were asked to choose one value to annotate the data. Estimate the value that most of the other crowd workers are expected to select.
MultipleEsti	Imagine 100 crowd workers were asked to choose one value to annotate the data. Estimate all possible values that the other crowd workers are expected to select (You must choose at least one).
RankingEsti	Imagine 100 crowd workers were asked to choose one value to annotate the data. Estimate all possible values that the other crowd workers are expected to select and rank them. From your estimation, the most popular answer should go first and the least popular answer should go last (You must choose at least one option, but may not need to choose all options).
ProbabilityEsti	Imagine 100 crowd workers were asked to choose one value to annotate the data. Estimate how probable each value is to be selected by other crowd workers (You are given 100 tokens to express worker distribution).

Table 2. Instructions in the task interface

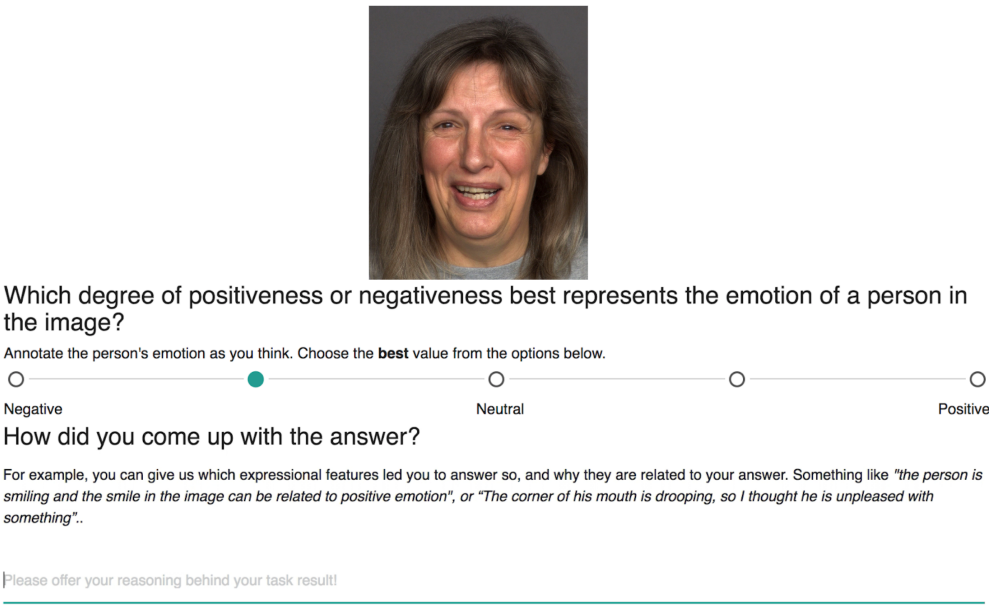


Fig. 4. The task interface screenshot in *Single*. Workers are shown an image to annotate, complete the annotation task with the elicitation interface, and give a reason for their selection.

with an additional function of recording the ranking of each selection with the order that the label was selected with. We decided on this design to minimize the task time, with the manual interaction not being far different from ordinary checkboxes. Other possible options like dragging items and aligning them in a ranking order would have resulted in much more task time compared to the simpler interaction.

For *Probability* and *ProbabilityEsti*, workers were allowed to assign tokens to ordinal labels, with the constraint that they are summed to 100. This approach of using tokens has been shown to be effective for laypeople to understand probabilities [12, 28]. However, unlike previous work which used a graphical representation of tokens, we did not add the graphical representation because

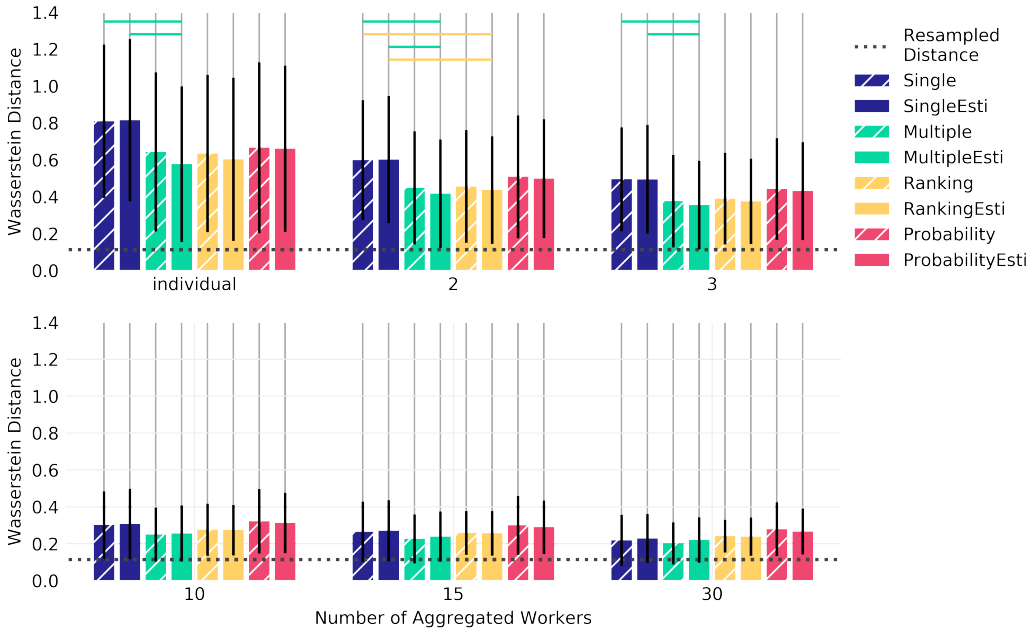


Fig. 5. The performance of elicitation approaches with varying numbers of annotations aggregated (individual, two, three, 10, 15, and 30), measured in Wasserstein distance. Horizontal lines (color-coded according to the approach in the right end of the line) above bar charts indicate whether the performance difference between approaches on the left end of the line and the right end of the line is significant (p-value below $.05/28 = .0018$, after Bonferroni correction) and the effect size is above 0.5 (measured with Cohen's d). When fewer than four annotations were aggregated, only the *MultipleEsti* condition consistently outperformed the baseline (*Single*) condition with the effect size larger than 0.5. Error bars indicate standard deviations. The *esampled distance* refers to the average distance between the gold standard distribution and resampled distributions.

it would result in substantial changes in the interface compared to other approaches, which may influence our experimental results.

Following previous work that crowdsourced collective answer distributions in a similar domain [82], we set the number of ordinal labels to five. Having five labels also can result in less confusion in understanding ordinal labels [10, 11, 44]. All images given to each worker were randomly sampled without replacement. We also asked for the worker's reasoning for their annotations, to better understand their intent [51] and making the interface similar to that of the gold standard collection (Figure 4).

5 RESULTS

In this section, we present results from the comparison of elicitation approaches. First, we compare the accuracy of individual annotations. Then, we evaluate the accuracy when the same number of annotations are aggregated. We also examine the accuracy when we set the total time spent to be the same for each elicitation approach. For elicitation approaches that turned out to be more efficient than the baseline, we evaluate to what extent these approaches reduce the cost. Lastly, we examine how the benefits of fine-grained annotations vary with different data instances.

5.1 For Individual Annotations, *MultipleEsti* Outperforms Other Approaches

5.1.1 Analysis Method. We compared the accuracy of individual annotations between different elicitation approaches with the Wasserstein distance to the gold standard distributions. To evaluate the elicitation approaches along two dimensions, we used the non-parametric Scheirer-Ray-Hare test because the Wasserstein distance was skewed (non-normal). For the same reason, for pairwise comparisons, we used the Mann-Whitney U test. For the eight elicitation approaches, we conducted $\binom{8}{2} = 28$ comparisons, and with Bonferroni correction, we considered the comparison result significant if the p-value was below $.05/28 = .0018$. To show which approach outperformed the other approach with a large difference, we calculated the effect size in Cohen's d and reported the effect size when it was above 0.5, which is medium effect size. For the weights of *Ranking* and *RankingEsti*, we used linearly decaying weights by setting the weight of the first ranked label as 5, and decreasing the weight by 1 as the number of ranking increases to second, third, and so on. We used this method because it performed better than other approaches, such as exponential weights.

5.1.2 Results. The top-left plot in Figure 5 summarizes how the distance between individual annotations and gold standard distributions differs between elicitation approaches. The results were significantly affected by both the granularity of annotations (Scheirer-Ray-Hare test, $df = 3$, $SS = 3.81e+09$, $H = 496.35$, $p < .001$) and the estimation perspective (Scheirer-Ray-Hare test, $df = 1$, $SS = 1.20e+08$, $H = 15.66$, $p < .001$). There was also a significant interaction between two dimensions (Scheirer-Ray-Hare test, $df = 3$, $SS = 1.34e+08$, $H = 17.48$, $p < .001$).

Comparing approaches to each other, *Single* and *SingleEsti* were outperformed by all other approaches ($p < .0001$ for all other approaches). *MultipleEsti* showed the biggest performance difference to *Single* and *SingleEsti*, with the effect size of 0.55 in Cohen's d for both approaches. For approaches in the *Self* perspective (Table 1), no significant differences between conditions were found other than comparisons with *Single* ($p > .0018$ for all approaches). Within approaches of the *Others* perspective, *MultipleEsti* and *RankingEsti* were significantly more accurate than *ProbabilityEsti* ($U = 624445.5$, $n_1 = n_2 = 1200$, $p < .0001$ for *MultipleEsti-ProbabilityEsti* and $U = 653951.5$, $n_1 = n_2 = 1200$, $p < .0001$ for *RankingEsti-ProbabilityEsti*). When comparing approaches with the dimension of the estimation perspective, the *Others* perspective only showed benefits between *MultipleEsti* and *Multiple* ($U = 636014.5$, $n_1 = n_2 = 1200$, $p < .0001$), but not for other granularities of annotations ($p > .0018$ for all other approaches). Overall, *MultipleEsti* was the most accurate elicitation approach for individual annotations.

5.2 For Fewer Aggregated Annotations, *MultipleEsti* Outperforms Other Approaches

5.2.1 Analysis Method. To evaluate the accuracy of aggregated annotations, we randomly sampled teams from the annotation pool consisting of 30 annotations for each elicitation approach. We tested five team sizes, aggregating 2, 3, 10, 15, or 30 annotations. To calculate the average and standard deviation, for each team size, we sampled at maximum 1000 teams with replacement. However, for statistical tests, we used unique teams that do not have overlapping annotations, as sampling with replacement can violate the underlying independence assumption of our statistical tests. For example, total 10 teams with the team size of three annotations can be made out of the pool of total 30 annotations to avoid selecting overlapping annotation across different teams. To test significance, as in Section 5.1, we conducted Scheirer-Ray-Hare test and Mann-Whitney U test ($p < .05/28 = .0018$ considered significant with Bonferroni correction).

5.2.2 Results. Figure 5 shows the performance of elicitation approaches with aggregation. From the Scheirer-Ray-Hare tests on five team sizes (Table 3), we found that annotation granularity affected the performance for all team sizes. The estimation perspective only affected the results

	2 Annotations				3 Annotations				10 Annotations			
	df	*SS*	*H*		*df*	*SS*	*H*		*df*	*SS*	*H*	
Annotation granularity	3	4.82e+8	251	p<.001	3	1.23e+8	144	p<.001	3	1.63e+6	21.25	p<.001
Estimation Perspective	1	1.08e+7	5.64	p<.05	1	9.29e+4	0.11	p>.5	1	603	0.0008	p>.5
Granularity:Perspective	3	4.91e+6	2.55	p>.1	3	2.34e+6	2.74	p>.1	3	22051	0.29	p>.5
	15 Annotations				**30 Annotations**							
	df	*SS*	*H*		*df*	*SS*	*H*					
Annotation granularity	3	5.64e+5	16.49	p<.001	3	1.28e+5	14.99	p<.005				
Estimation Perspective	1	129	0.004	p>.5	1	815	0.10	p>.5				
Granularity:Perspective	3	32922	0.96	p>.5	3	6572	0.77	p>.5				

Table 3. Scheirer-Ray-Hare test results on the eight elicitation approaches with a varying number of annotations being aggregated. Tests were conducted along two dimensions, annotation granularity and the estimation perspective. The interaction between the two dimensions was also tested (Granularity:Perspective).

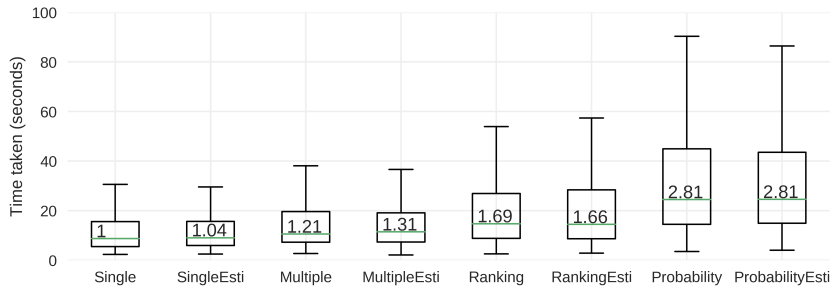


Fig. 6. Task times for different approaches. The ratio of median task times compared to the baseline (*Single*) are shown for each approach.

with two annotations aggregated (Scheirer-Ray-Hare test, $df = 1$, $SS = 1.08e+07$, $H = 5.64$, $p < .05$). The interaction between the two dimensions was not significant for any team size.

When comparing approaches in pairs, with two and three annotations aggregated, we found that *Single* and *SingleEsti* were outperformed by other approaches ($p < .001$ for all other approaches). With two annotations, *MultipleEsti* and *RankingEsti* outperformed *Single* and *SingleEsti* with the effect size larger than 0.5 (p<.0001 for all comparisons), and with three annotations, *MultipleEsti* outperformed them with the effect size above 0.5 (p<.0001 for all comparisons). Overall, with a low number of annotations, *MultipleEsti* was the most accurate in estimating answer distributions.

When more annotations were aggregated, the performance difference was insignificant in most cases ($p > .0018$). Only *Multiple* and *MultipleEsti* outperformed *Probability* and *ProbabilityEsti* ($p < .001$ for all comparisons).

5.3 For Similar Task Times, Only *Multiple* and *MultipleEsti* Outperform *Single*

Our goal is to investigate which approach efficiently estimates collective answer distributions with low human-time cost. To compare the performance of elicitation approaches fairly, we introduce our method for holding the total human time constant across elicitation approaches.

5.3.1 Analysis Method. First, we examined if task time varied significantly across different elicitation approaches. We found that task times for approaches within the same estimation perspective dimension were significantly different (for all pairs with different granularities, $p < .0001$ with the

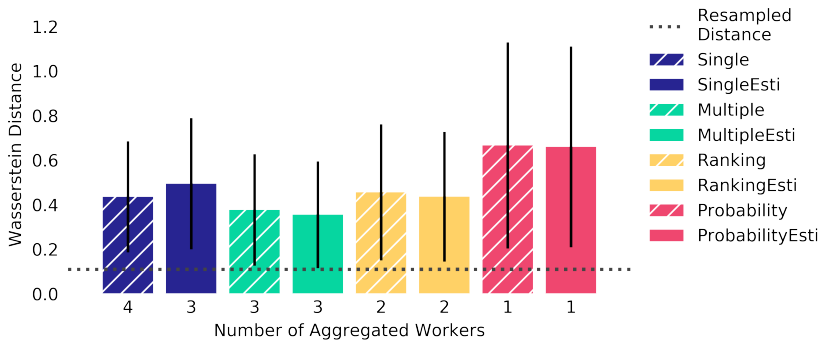


Fig. 7. The performance of different elicitation approaches when the total human time available is set equal to the time required for obtaining four *Single* annotations. Only *Multiple* and *MultipleEsti* outperformed *Single*. Error bars indicate standard deviations. Resampled distance refers to the average distance between the gold standard distribution and resampled distributions.

Mann-Whitney U test, which is still significant with Bonferroni correction over $\binom{8}{2} = 28$ comparisons). Figure 6 shows task times for each elicitation approach in boxplots with the ratio of median task time to that of *Single*.

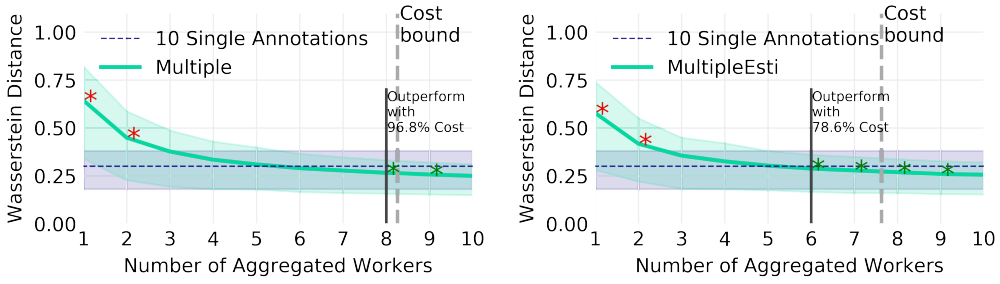
To set the total task time constant across all elicitation approaches, we first defined the time budget as the total human time spent on obtaining a certain number of baseline annotations. Then, we found the maximum number of annotations for each elicitation approach such that the sum of task time would not exceed the time budget. We used the median task time of each approach when comparing. To investigate the performance with a lower budget, we set the number of annotations in *Single* to four, and varied the number of annotations in each answer elicitation approach (the x-axis in Figure 7). To compare the performance across different answer elicitation approaches, we used the same analysis method for the aggregation and statistical tests as in Section 5.2.

5.3.2 Results. Figure 7 shows that, given similar total human time, annotation granularity significantly affected the performances of elicitation approaches (Scheirer-Ray-Hare test, $df = 3$, $SS = 1.11e+09$, $H = 518$, $p < .001$). However, the estimation perspective did not significantly impact the performance (Scheirer-Ray-Hare test, $df = 1$, $SS = 8.47e+04$, $H = 0.04$, $p > .5$). The interaction between the two dimensions was significant (Scheirer-Ray-Hare test, $df = 3$, $SS = 1.74e+07$, $H = 8.08$, $p < .05$). Comparing elicitation approaches in pairs, we found that only *Multiple* and *MultipleEsti* outperformed *Single* and *SingleEsti* ($p < .0001$ for all comparisons). *Multiple* and *MultipleEsti* also outperformed the rest of the approaches within *Self* and *Others*, respectively ($p < .0001$ for all comparisons). When comparing elicitation approaches within the dimension of the estimation perspective, we found that there was no significant difference between the approaches.

5.4 *MultipleEsti* Requires 21.4% Less Human Time than the Baseline, *Single*

For approaches that outperformed *Single* in Section 5.3.2, we measured comparative efficiency, how much the total human time can be reduced to achieve a similar performance as the baseline.

5.4.1 Analysis Method. First, we set the number of aggregated annotations for *Single* as 10 because with more than 10 annotations, the performance did not increase much with the addition of an annotation (improvement < 2.9%). Then, we varied the number of annotations for *MultipleEsti* and *Multiple* from one to 10 and examined when the two approaches started to exceed *Single* (the



(a) *Multiple* (b) *MultipleEsti*

Fig. 8. The amount of total human time (cost) *Multiple* and *MultipleEsti* require to significantly outperforms 10 baseline annotations, *Single*. A significance test was conducted with Scheirer-Ray-Hare test, to consider the effects of each data instance in comparing approaches. If 10 *Single* annotations significantly outperformed *Multiple* or *MultipleEsti*, red star is marked. If *Multiple* or *MultipleEsti* significantly outperformed the baseline, green star is marked. The numbers of annotations with which task times for Multiple and MultipleEsti corresponds to that of 10 *Single* annotations are visualized as Cost bound with a dashed gray vertical line. The numbers of annotations required to outperform *Single* are visualized with thick black vertical lines. Shaded regions indicate interquartile ranges of Wasserstein distance.

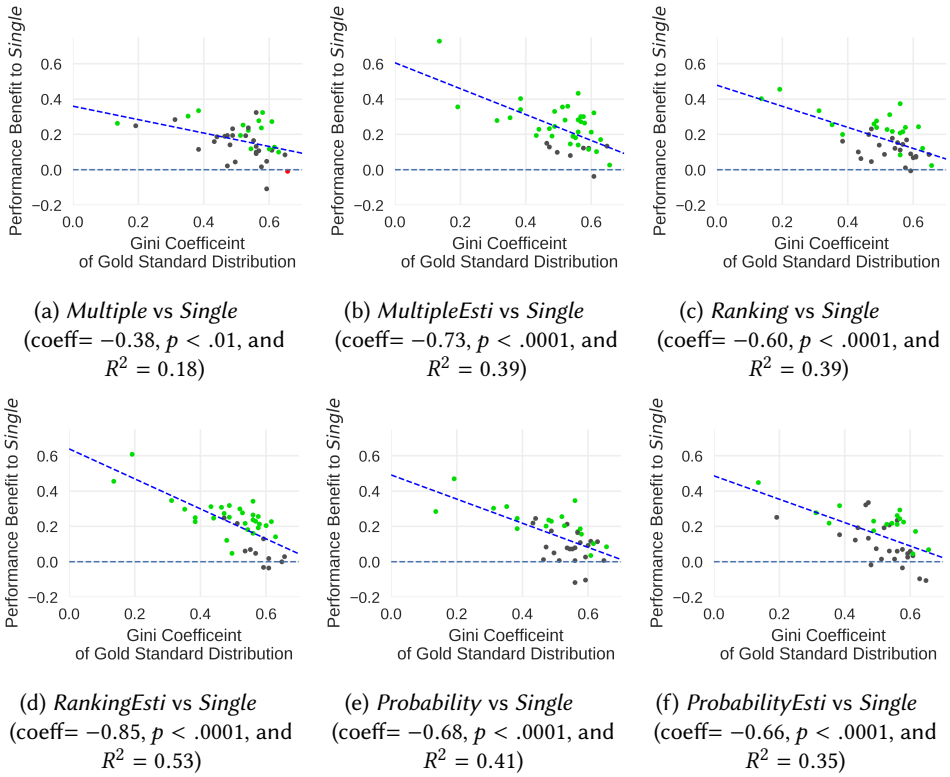
baseline) in performance. In this analysis, for each team size of *Multiple* or *MultipleEsti* annotations, we conducted a Scheirer-Ray-Hare test with respect to two dimensions: 1) the data instances (total 40 facial images) and 2) annotation elicitation approaches (*Single* versus *Multiple* or *MultipleEsti*). Then, we examined if the results of *Multiple* or *MultipleEsti* were significantly different from *Single* across data instances by examining the effects of elicitation approaches. We adopted this particular analysis method because we wanted to consider the mean difference between data instances and the Mann-Whitney U test cannot account for this.

5.4.2 Results. The results (Figure 8) showed that *Multiple* significantly outperformed 10 *Single* annotations with eight annotations (20% fewer workers), while using 3.2% less human time. *MultipleEsti* showed more cost benefit, significantly outperforming 10 *Single* annotations with only six annotations (40% fewer workers), resulting in the use of 21.4% less human time.

5.5 Fine-grained Annotations Are More Beneficial For More Ambiguous Data

The benefits of answer elicitation methods can vary across different data instances. For example, crowd workers might estimate the distribution of a less ambiguous data instance efficiently, even by annotating a single answer. For fine-grained annotations, we examine if there is any correlation between the ambiguity of the data and the amount of the performance benefit that each approach offers compared to the baseline, in Wasserstein distance.

5.5.1 Analysis Method. We measured the level of ambiguity of a data instance with the Gini coefficient of the gold standard distribution. The Gini coefficient is a measure of how dispersed weights are across labels. A higher Gini coefficient indicates that weights are more skewed, and a lower Gini coefficient indicates that weights are more evenly dispersed across labels. For example, in our study, a Gini coefficient of 0.8 indicates that all weights are skewed to a single label, and 0 indicates that all labels have equal weights. To measure the performance benefit for each data instance, we measured the performance difference between *Single* and approaches that receive more fine-grained annotations. For each data instance-elicitation approach pair, because *Single* annotations are compared to other approaches six times, we considered the performance difference



(a) *Multiple* vs *Single* (coeff= -0.38, $p < .01$, and $R^2 = 0.18$)

(b) *MultipleEsti* vs *Single* (coeff= -0.73, $p < .0001$, and $R^2 = 0.39$)

(c) *Ranking* vs *Single* (coeff= -0.60, $p < .0001$, and $R^2 = 0.39$)

(d) *RankingEsti* vs *Single* (coeff= -0.85, $p < .0001$, and $R^2 = 0.53$)

(e) *Probability* vs *Single* (coeff= -0.68, $p < .0001$, and $R^2 = 0.41$)

(f) *ProbabilityEsti* vs *Single* (coeff= -0.66, $p < .0001$, and $R^2 = 0.35$)

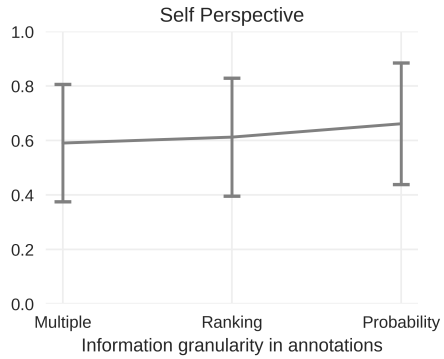
Fig. 9. The performance benefit of approaches that receive more fine-grained annotations than a single answer, compared to the baseline (*Single*), for each data instance. The performance benefit is measured by calculating the difference in Wasserstein distance to the gold standard distribution. Data instances in which the fine-grained annotations significantly outperformed the baseline annotations are visualized in green and those in which the baseline annotations significantly outperformed others are visualized in red (considered significant with $p < .05/6 = .0083$). Those without a significant difference are visualized in gray. For each approach, we ran a linear regression between Gini coefficients of data instances and performance benefits, whose results are presented in the caption. The performance benefit of fine-grained annotations got larger for more ambiguous data – when the Gini coefficients in gold standard distributions lower.

between the elicitation approach and the baseline approach significant when the p-value was lower than $.05/6 = .0083$ with Bonferroni correction. For the Gini coefficients and performance benefits, we conducted linear regression to examine the correlation between them.

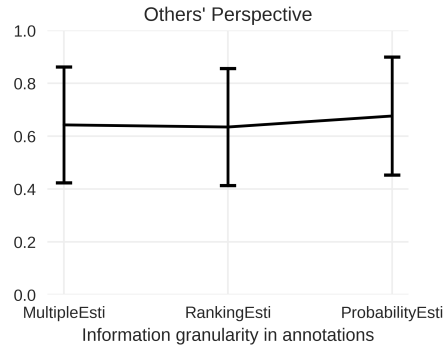
5.5.2 Results. Performance benefits had a negative correlation with the Gini coefficient of the gold standard distribution (Figure 9, $p < .05$ for all approaches), but our results suggest that performance benefits get larger for more ambiguous data. For all data instances and all elicitation approaches, we only found one case where *Single* outperformed the other approach (Figure 9a).

5.6 Discussion

Overall, *Multiple* and *MultipleEsti* showed similar or better performance than more fine-grained elicitation approaches. This result is surprising because more efforts were put into fine-grained elicitation approaches, in terms of task time. We further analyze why more fine-grained elicitation approaches performed worse in the next section.



(a) The selection-level accuracy for *Self*. Approaches with different annotation granularity was significantly different each other.



(b) The selection-level accuracy for *Others*. Except *MultipleEsti* and *RankingEsti*, other pairs of approaches were significantly different .

Fig. 10. For approaches that allow more fine-grained annotations, the selection-level accuracy was higher or similar compared to approaches with lower granularity annotations. Significance is decided with $p < .05/3 = .017$. Error bars indicate the standard deviation.

For the estimation perspective, estimating with the *Others* perspective was only effective with the annotation granularity of selecting multiple labels, but not with higher granularity. The result suggests that people cannot fully estimate answers with others' perspective. Previous work also suggests people's limited capability in estimating other people's perspectives. Bayesian Truth Serum [65, 66], which inspired the estimation of other people's answers, was also more reliable with a higher number of people. Similarly, previous work on perspective taking also showed that people cannot fully take other's perspective [23], relying on their own perspectives to some extent.

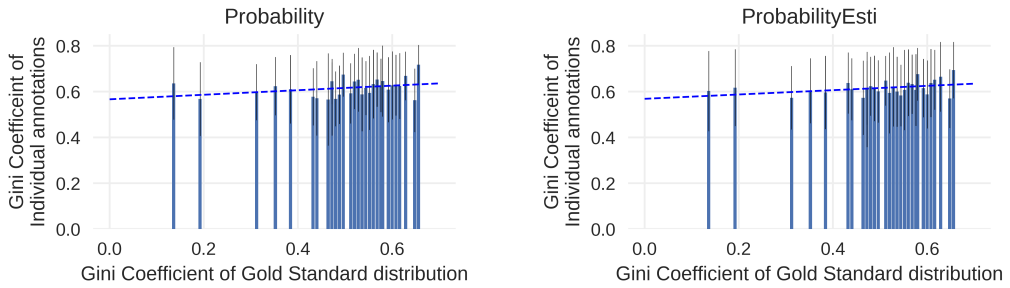
6 EXPLORING WHY FINE-GRAINED ANNOTATIONS ARE NOT ACCURATE

We found that eliciting fine-grained annotations like rankings or probabilities did not have a comparative advantage over more coarse-grained annotations that collect only the selection of relevant labels. To investigate the cause of such results, we conducted additional analysis. The first factor we considered is if crowd workers selected the right labels, and the second is if crowd workers assigned incorrect rankings to the selected labels. The last is what weight patterns are assigned to selected labels for approaches that collect real-valued weights.

6.1 Selection-level Accuracy Did Not Decrease With Finer Granularity

First, we examined selection-level accuracy, which measures how accurate workers were in selecting labels, ignoring the weight they assigned. A label is considered accurate if either a selected label appears in the gold standard, or an unselected label does not appear in the gold standard. For each annotation with five ordinal labels, we computed the selection-level accuracy as the ratio of accurate labels among all labels.

6.1.1 Results. Figure 10 shows the average selection-level accuracy of all annotations from different elicitation approaches. Within the *Self* perspective, the increase in the granularity of annotation led to a slight increase in the selection-level accuracy, with all comparisons between approaches in the *Self* perspective being all significant ($p < .01$ for all comparisons). However, the effect size was small (0.05 for *Multiple-Ranking* and 0.10 for *Ranking-Probability*). For the *Others* perspective, the selection-level accuracy of *MultipleEsti* and *RankingEsti* was not significantly different ($U = 706066$,



(a) *Probability*. With the linear regression on the mean of Gini coefficients, the linear model was $y = 0.10x + 0.57$, with $p < .05$, and $R^2 = 0.11$

(b) *ProbabilityEsti*. With the linear regression on the mean of Gini Coefficients, the linear model was $y = 0.09x + 0.57$, with $p < .05$, and $R^2 = 0.13$

Fig. 11. Correlation between Gini coefficients for annotations that receive probability (y-axis) and gold standard distributions (x-axis). Gini coefficients of probability annotations were constantly high regardless of the Gini coefficients of gold standard distributions. It suggests that people skew weights towards a small number of labels regardless of the ambiguity of data when annotating probability information. Black error bars indicate standard deviations.

$n_1 = n_1 = 1200$, $p > .1$) and other pairs of elicitation approaches were significantly different but with very small effect sizes (Cohen's $d = 0.09$ for *RankingEsti-ProbabilityEsti*, Cohen's $d = 0.07$ for *MultipleEsti-ProbabilityEsti*, and $p < .05/3 = .017$ for both).

6.2 Ranking Level Agreement Did Not Change with Finer Granularity

For approaches that receive rankings or probabilities, we measured the ranking-level agreement to the gold standard distribution and how each elicitation approach was accurate in retrieving the ranking of labels in the gold standard distribution. Then, we investigated how annotation granularity affects the ranking-level agreement. We used Spearman's ρ to compute the degree of agreement in ranking between each annotation and the gold standard distribution.

6.2.1 Results. Within the same dimension of the estimation perspective, in the ranking level agreement, approaches that received probabilities were not significantly different from those that received rankings. For approaches within the *Self* perspective, the mean correlation coefficient was 0.60 for *Ranking*, while it was 0.58 for *Probability* ($U = 714384.5$, $n_1 = n_2 = 1200$, $p > .1$). For those in the *Others* perspective, the mean correlation coefficient for *RankingEsti* was 0.62, and for *ProbabilityEsti*, it was 0.60 ($U = 713106$, $n_1 = n_2 = 1200$, $p > .1$).

6.3 Real-valued Probability Weights Tend to be Skewed Towards Fewer Labels

We analyzed the pattern of how crowd workers assigned probabilities by measuring the Gini coefficients (explained in Section 5.5.1) for individual annotations and the gold standard distributions. By examining the relationship between them, we can understand patterns like if individual workers are distributing weights uniformly across labels or skewing weights compared to the gold standard distribution. If crowd workers estimate answer distributions perfectly, the Gini coefficient of individual annotations would be the same as that of the gold standard distributions.

6.3.1 Results. In Figure 11, the mean of the Gini coefficients for individual probability annotations remained relatively high regardless of the Gini coefficient of a gold standard distribution. The mean of Gini coefficients for all annotations was 0.62 for both of *Probability* and *ProbabilityEsti*. From the

linear regression on Gini coefficients of gold standard distributions and means of Gini coefficients for annotations from *Probability* and *ProbabilityEsti*, we could find that the mean Gini coefficients of approaches that receive probabilities did not change much with the Gini coefficients of gold standard distributions. Our results indicate that the patterns of assigning weights did not change much with data instances, with workers assigning probabilities toward a small subset of labels.

6.4 Discussion

For elicitation approaches except for *Single* and *SingleEsti*, we found that the increase in the granularity of annotations resulted in a slightly higher or similar selection-level accuracy. These results suggest that ranking and probability information might have been the major source of performance degradation for more fine-grained annotation approaches. We also found that approaches that receive rankings and probabilities were similar in the ranking-level agreement to the gold standard distribution. As approaches that receive rankings performed better or similarly compared to those that receive probabilities, it suggests that the real-valued weight information did not add meaningful information for estimating collective answer distributions.

For *Probability* and *ProbabilityEsti*, we found that crowd workers skewed weights to a few labels regardless of the ambiguity of the data instance (Figure 11). One possible explanation for the skewed weights is a bias from the sequential interpretation of multiple possible answers. When perceiving multiple possible interpretations, people tend to detect each interpretation sequentially with time gaps [58]. This sequential interpretation could have caused a confirmation bias [45], resulting in the assignment of higher weights to labels perceived earlier. On the other hand, *Multiple* and *MultipleEsti* might have been less influenced by such a bias, because these approaches force uniform weights across the selected labels. It might be the reason why they performed better than more fine-grained elicitation approaches. However, the current data is not sufficient to fully support this explanation (e.g., workers' selection sequence was not recorded), and more investigation is necessary in future work.

7 DISCUSSION

Our study and analysis showed that, while it is possible to reduce required human efforts in estimating answer distributions, the most fine-grained approach is not the most accurate elicitation approach. In this section, we discuss 1) the scope of the task domains where our findings apply, 2) task interface design, 3) payment design, 4) the dataset we used, and 5) guidelines to applying elicitation approaches.

7.1 Scope and Limitations

To apply the elicitation approaches from our study, diverse answers should result primarily from the ambiguity of the data, not from annotators' personal bias or lack of domain knowledge. In tasks like evaluating whether a political speech supports the liberal or conservative ideology [37], annotators can have a personal bias. In such tasks, it might not be effective for annotators to estimate how other people would have answered, because a strong bias would limit people's ability to consider perspectives of other people [22, 23]. For tasks that require domain knowledge, such as annotating legal decisions [29], the varying answer can be due to the lack of knowledge. Estimating answers of others would also be more challenging as workers without domain knowledge would not know how those with the knowledge would answer. In this type of domains, carefully choosing workers who make estimates would be crucial to efficiently estimating answer distributions.

7.2 Task Interface Design

Given the variation in estimation perspective, the ranking and probability annotations represent different information. For the *Self* approaches, weights indicate the intensity of an annotator's own preference or opinion for each label. For the *Others* approaches, on the other hand, weights indicate the expected popularity of each label among a group of annotators. As future research explores the granularity or the estimation perspective of annotation approaches, the relationship between these two dimensions needs to be considered because they can affect the design of the task, such as how the instructions should be written.

For interfaces that receive ranking and probability annotations, we discuss how specific designs can impact the estimation results. For ranking annotations, we designed interfaces to minimize the task time by receiving ranking information with the order of selections. However, this can affect the results due to confirmation bias from the sequential interpretation [45] (also explained in Section 6.4) and workers might not change ranking inputs when they make premature and inaccurate decisions. Alternative designs would be interfaces that detach selection of items and ranking, so that workers can be less affected by sequential selection when they are ranking. However, if the interaction becomes more complex, the task time would increase as a trade-off.

For probability annotations, the total number of tokens that workers were allowed to assign can affect annotation quality. For instance, in previous work, with a lower number of tokens, people were more accurate at understanding the probability [48]. It needs to be further studied how a varying number of tokens would affect the estimation of the answer distribution.

Task time can be a practical factor, as the effects of the increased cost can be significant for large-scale tasks like annotating ML training datasets [76]. Thus, when designing more complex elicitation approaches that take more task time, it is crucial to consider time-accuracy trade-offs.

7.3 Payment Design

Regardless of the total task time required by different annotation approaches, we paid workers the same amount, $1.20 per task, following our institution's IRB (Institutional Review Board) policies. This resulted in slightly different hourly wages, but this did not vary much between conditions as the task not only included annotating five facial images, but also reading instructions, solving a quiz for attention check, and providing reasoning for each annotation. The resulting average hourly wages for all conditions were over the 2019 minimum wage in the U.S., ranging from $7.71 to $9.89. Prior work has shown that higher payment did not lead to a significant change in response quality given fair payment [59]. Instead, payment primarily affected task adoption and completetion rate, which does not impact our findings.

7.4 Dataset Considerations

We used a popular dataset in facial emotion recognition that showed sufficient ambiguity to result in answer distributions [19]. This variation in interpretation between annotators was also observed in the original research that yielded this dataset [19]. However, it contains only Caucasian faces, meaning that we could not observe the answer distribution that would have been created for different demographics by annotators. Previous work in emotion perception found that interpreting emotion is a universal human ability [21], therefore, for different annotator-image race pairs, we expect estimation-based approaches to still work. However, as recognition ability may be skewed by different pairings of annotator's race and the race of the person in the image, the resulting distributions may differ [21]. Exploring the subtle effects of demographics in estimation-based annotations could be a compelling question for the CSCW community to pursue.

7.5 Guidelines for Estimating Collective Answer Distributions

From our findings, we suggest guidelines for designing crowdsourcing tasks to efficiently estimate collective answer distributions. We believe these guidelines will be helpful to practitioners who want to collect answer distributions for ambiguous data for purposes such as training ML models.

Annotation Granularity Should Not be Too Coarse-grained or Too Fine-grained. Eliciting only one answer from a worker can be expensive, as each response contains minimal information to estimate collective answer distributions from a target group. However, asking workers to provide annotations that are too fine-grained can also be inefficient. This is because workers tend to concentrate their confidence estimations on fewer labels than actually occurs when eliciting responses from a broader group. As a result, it can be more inaccurate than approaches that do not elicit label proportion information at all. Additionally, providing more fine-grained answers requires more time from each worker, which consequently decreases efficiency.

Ask Workers to Estimate How Other Workers Would Have Answered. Ask workers to estimate how other workers would have answered, as prior work [65, 66] and our results suggest that this improves people's ability to estimate the eventual responses of the group compared to asking them about their own beliefs regarding the correct labels. When combined with the correct annotation granularity, we observed that this significantly increased estimation accuracy.

8 FUTURE WORK

We identify two potential directions for future work, 1) improving the estimation capability of workers and 2) maximizing the benefits of elicitation approaches with adaptable task UI.

To improve crowd workers' estimation ability, we can show them other example data and corresponding gold standard collective answer distributions. The worker would learn from the examples and expand their knowledge on how other workers would annotate the data [14, 31]. This idea can be combined with active learning, where the the model being trained fetches potentially related examples that can be most helpful in estimating answer distributions [9]. A key factor would be whether the worker generalizes and applies the learned knowledge to newly observed data instances. However, this direction would increase the cost of setting up the task because collective answer distributions need to be collected for example data instances.

Another direction of improving the estimation capability is making workers interact during the task. For instance, the interface can show intermediate results from other workers [53], allowing the worker to realize how others annotate. A more interactive form would be allowing workers to communicate more closely, similar to having a discussion [6, 7, 67]. For these directions, it would be crucial to avoid groupthink [38] and facilitate workers to estimate with a wider perspective.

Directions for improving the estimation capability can be extended to solve more difficult problems, where a worker's perspective is hard to be changed. For these problems, the aforementioned approaches of showing workers to examples or other perspectives would not work. For instance, with political data, people experienced belief polarization [49], reinforcing one's own opinion even after observing contradicting information. For these challenging problems, more systematic interactions would be required, such as those based on Bayesian models for belief polarization [39, 40].

It may also be possible to improve the efficacy of our approach by adaptively switching the annotation interface based on the ambiguity of the data, or how much disagreement is expected [57]. From the results, we could observe that the benefits of fine-grained annotations were amplified when a data instance was more ambiguous. Because crowd workers [67] and machines [32, 81] are capable of estimating the level of disagreement for certain annotation tasks by observing data, it would be possible to adaptively show the task interface that is expected to maximize efficiency.

9 CONCLUSION

This paper demonstrates that, when crowdsourcing the annotation of ambiguous data, we can reduce the cost of collecting reliable collective answer distributions by eliciting richer answers from each worker. We investigated answer elicitation approaches that vary along two dimensions: 1) annotation granularity and 2) the estimation perspective. Our results show that the choice of elicitation interface matters, and that the best (most efficient) solution is neither the fastest, nor the most potentially-accurate. Instead, we found that the best efficiency can be achieved using an intermediate-granularity approach that asks workers to select multiple labels that they estimate a group of other workers would choose. By analyzing when more fine-grained annotations are less accurate, we found that workers showed a tendency to estimate more heavily-skewed distributions when annotating weights than were actually observed in our ground truth data. While approaches that elicit fine-grained annotations did not result in the best performance when estimating collective answer distributions in our experiments, our work suggests that finding better ways to guide workers to estimate answer distributions by leveraging more fine-grained annotations is a promising future research direction.

ACKNOWLEDGMENTS

We thank crowd workers who participated in our study. We also thank Stephanie D. O'Keefe, other members of CROMA Lab and KIXLAB, and Jane Im for constructive feedback on this work. This research was supported in part by USDOT Center for Connected and Automated Transportation (CCAT) at the University of Michigan, Toyota Research Institute (TRI), the IBM AI Horizons Network, a DARPA Young Faculty Award, and the Next-Generation Information Computing Development Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science and ICT (NRF2017M3C4A7065960).

REFERENCES

[1] 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. https://doi.org/10.1609/aimag.v36i1.2564

[2] Alexandry Augustin, Matteo Venanzi, Alex Rogers, and Nicholas R. Jennings. 2017. Bayesian Aggregation of Categorical Distributions with Applications in Crowdsourcing. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI'17)*. AAAI Press, Menlo Park, CA, USA, 1411–1417. https://doi.org/10.24963/ijcai.2017/195

[3] Arkar Min Aung and Jacob Whitehill. 2018. Harnessing Label Uncertainty to Improve Modeling: An Application to Student Engagement Recognition. In *IEEE International Conference on Automatic Face Gesture Recognition (FG'18)*. IEEE, Piscataway, New Jersey, US, 166–170. https://doi.org/10.1109/FG.2018.00033

[4] Kirsten Boehner, Rogério DePaula, Paul Dourish, and Phoebe Sengers. 2007. How emotion is made and measured. *International Journal of Human-Computer Studies* 65, 4 (2007), 275 – 291. https://doi.org/10.1016/j.ijhcs.2006.11.016

[5] Jonathan Bragg, Daniel S Weld, et al. 2013. Crowdsourcing multi-label classification for taxonomy creation. In *First AAAI conference on human computation and crowdsourcing*. AAAI, Palo Alto, CA, USA.

[6] Joseph Chee Chang, Saleema Amershi, and Ece Kamar. 2017. Revolt: Collaborative Crowdsourcing for Labeling Machine Learning Datasets. In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI'17)*. ACM, New York, NY, USA, 2334–2346. https://doi.org/10.1145/3025453.3026044

[7] Quanze Chen, Jonathan Bragg, Lydia B. Chilton, and Dan S. Weld. 2019. Cicero: Multi-Turn, Contextual Argumentation for Accurate Crowdsourcing. In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI'19)*. ACM, New York, NY, USA, Paper No. 531. https://doi.org/10.1145/3290605.3300761

[8] Lydia B Chilton, Greg Little, Darren Edge, Daniel S Weld, and James A Landay. 2013. Cascade: Crowdsourcing taxonomy creation. In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI'13)*. ACM, New York, NY, USA, 1999–2008. https://doi.org/10.1145/2470654.2466265

[9] 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 ACM Conference on Human Factors in Computing Systems (CHI'19)*. ACM, New York, NY, USA, Paper No. 230. https://doi.org/10.1145/3290605.3300460

[10] Andrew M. Colman, Claire E. Norris, and Carolyn C. Preston. 1997. Comparing Rating Scales of Different Lengths: Equivalence of Scores from 5-Point and 7-Point Scales. *Psychological Reports* 80, 2 (1997), 355–362. https://doi.org/10.2466/pr0.1997.80.2.355

[11] John Dawes. 2008. Do Data Characteristics Change According to the Number of Scale Points Used? An Experiment Using 5-Point, 7-Point and 10-Point Scales. *International Journal of Market Research* 50, 1 (2008), 61–104. https://doi.org/10.1177/147078530805000106

[12] Adeline Delavande and Susann Rohwedder. 2008. Eliciting subjective probabilities in Internet surveys. *Public Opinion Quarterly* 72, 5 (2008), 866–891. https://doi.org/10.1093/poq/nfn062

[13] Don A Dillman, Jolene D Smyth, and Leah Melani Christian. 2014. *Internet, phone, mail, and mixed-mode surveys: the tailored design method.* John Wiley & Sons.

[14] Shayan Doroudi, Ece Kamar, Emma Brunskill, and Eric Horvitz. 2016. Toward a Learning Science for Complex Crowdsourcing Tasks. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*. ACM, New York, NY, USA, 2623–2634. https://doi.org/10.1145/2858036.2858268

[15] Anca Dumitrache. 2015. Crowdsourcing disagreement for collecting semantic annotation. In *European Semantic Web Conference*. Springer, 701–710. https://doi.org/10.1007/978-3-319-18818-8_43

[16] Anca Dumitrache, Lora Aroyo, and Chris Welty. 2018. Capturing Ambiguity in Crowdsourcing Frame Disambiguation. In *Proceedings of the Sixth AAAI Conference on Human Computation and Crowdsourcing, HCOMP 2018, Zürich, Switzerland, July 5-8, 2018*. AAAI, Palo Alto, CA, USA, 12–20.

[17] Anca Dumitrache, Lora Aroyo, and Chris Welty. 2018. Crowdsourcing Ground Truth for Medical Relation Extraction. *ACM Trans. Interact. Intell. Syst.* 8, 2 (2018), 11:1–11:20. https://doi.org/10.1145/3152889

[18] Anca Dumitrache, Oana Inel, Lora Aroyo, Benjamin Timmermans, and Chris Welty. 2018. CrowdTruth 2.0: Quality Metrics for Crowdsourcing with Disagreement. *arXiv preprint arXiv:1808.06080* (2018).

[19] Natalie C Ebner, Michaela Riediger, and Ulman Lindenberger. 2010. FACES—A database of facial expressions in young, middle-aged, and older women and men: Development and validation. *Behavior research methods* 42, 1 (2010), 351–362. https://doi.org/10.3758/BRM.42.1.351

[20] Bradley Efron and Robert J Tibshirani. 1994. *An introduction to the bootstrap*. CRC press.

[21] Hillary Anger Elfenbein and Nalini Ambady. 2002. On the universality and cultural specificity of emotion recognition: a meta-analysis. *Psychological bulletin* 128, 2 (2002), 203.

[22] Nicholas Epley and Eugene M Caruso. 2008. Perspective Taking: Misstepping Into Others' Shoes. *Handbook of imagination and mental simulation* (2008), 295.

[23] Nicholas Epley, Boaz Keysar, Leaf Van Boven, and Thomas Gilovich. 2004. Perspective taking as egocentric anchoring and adjustment. *Journal of personality and social psychology* 87, 3 (2004), 327. https://doi.org/10.1037/0022-3514.87.3.327

[24] 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. https://doi.org/10.1109/TIP.2017.2689998

[25] Xin Geng. 2016. Label Distribution Learning. *IEEE Transactions on Knowledge and Data Engineering* 28, 7 (2016), 1734–1748. https://doi.org/10.1109/TKDE.2016.2545658

[26] Xin Geng, Chao C. Yin, and Zhi-Hua Z. Zhou. 2013. Facial Age Estimation by Learning from Label Distributions. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35, 10 (2013), 2401–2412. https://doi.org/10.1109/TPAMI.2013.51

[27] Victor Girotto, Erin Walker, and Winslow Burleson. 2019. CrowdMuse: Supporting Crowd Idea Generation through User Modeling and Adaptation. In *Proceedings of the Conference on Creativity and Cognition (C&C'19)*. ACM, New York, NY, USA, 95–106. https://doi.org/10.1145/3325480.3325497

[28] Daniel G Goldstein and David Rothschild. 2014. Lay understanding of probability distributions. *Judgment & Decision Making* 9, 1 (2014).

[29] D. G. Gordon and T. D. Breaux. 2014. The role of legal expertise in interpretation of legal requirements and definitions. In *2014 IEEE 22nd International Requirements Engineering Conference (RE)*. IEEE, Piscataway, New Jersey, US, 273–282. https://doi.org/10.1109/RE.2014.6912269

[30] Paul Grau, Babak Naderi, and Juho Kim. 2018. Personalized Motivation-supportive Messages for Increasing Participation in Crowd-civic Systems. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 60. https://doi.org/10.1145/3274329

[31] Thomas L. Griffiths, Charles Kemp, and Joshua B. Tenenbaum. 2008. *Bayesian Models of Cognition*. Cambridge University Press, 59–100. https://doi.org/10.1017/CBO9780511816772.006

[32] Danna Gurari and Kristen Grauman. 2017. CrowdVerge: Predicting If People Will Agree on the Answer to a Visual Question. In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI'17)*. ACM, New York, NY, USA, 3511–3522. https://doi.org/10.1145/3025453.3025781

[33] Danna Gurari, Yinan Zhao, Suyog Dutt Jain, Margrit Betke, and Kristen Grauman. 2019. Predicting How to Distribute Work Between Algorithms and Humans to Segment an Image Batch. *International Journal of Computer Vision* (2019),

1–19.

[34] Sungsoo Ray Hong, Minhyang Mia 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 ACM Conference on Human Factors in Computing Systems (CHI'18)*. ACM, New York, NY, USA, Paper No. 66. https://doi.org/10.1145/3173574.3173640

[35] Sungsoo Ray Hong, Minhyang Mia Suh, Tae Soo Kim, Irian 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 (2019), Article 106. https://doi.org/10.1145/3359208

[36] Shih-Wen Huang and Wai-Tat Fu. 2013. Enhancing Reliability Using Peer Consistency Evaluation in Human Computation. In *Proceedings of the ACM Conference on Computer-Supported Cooperative Work & Social Computing (CSCW'13)*. ACM, New York, NY, USA, 639–648. https://doi.org/10.1145/2441776.2441847

[37] Mohit Iyyer, Peter Enns, Jordan Boyd-Graber, and Philip Resnik. 2014. Political Ideology Detection Using Recursive Neural Networks. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Baltimore, Maryland, 1113–1122. https://doi.org/10.3115/v1/P14-1105

[38] Irving L Janis. 1971. Groupthink. *Psychology today* 5, 6 (1971), 43–46.

[39] Alan Jern, Kai-Min K Chang, and Charles Kemp. 2014. Belief polarization is not always irrational. *Psychological review* 121, 2 (2014), 206. https://doi.org/10.1037/a0035941

[40] Alan Jern, Kai min Chang, and Charles Kemp. 2009. Bayesian belief polarization. In *Advances in Neural Information Processing Systems 22*, Y. Bengio, D. Schuurmans, J. D. Lafferty, C. K. I. Williams, and A. Culotta (Eds.). Curran Associates, Inc., 853–861. http://papers.nips.cc/paper/3725-bayesian-belief-polarization.pdf

[41] Xiuyi Jia, Weiwei Li, Junyu Liu, and Yu Zhang. 2018. Label Distribution Learning by Exploiting Label Correlations. In *Thirty-Second AAAI Conference on Artificial Intelligence*. AAAI, Palo Alto, CA, USA.

[42] 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)*. Association for Computational Linguistics, New Orleans, Louisiana, 628–633. https://doi.org/httts://doi.org/10.18653/v1/N18-2099

[43] 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

[44] Rob Johns. 2010. Likert items and scales. *Survey Question Bank: Methods Fact Sheet* 1 (2010), 1–11.

[45] Eva Jonas, Stefan Schulz-Hardt, Dieter Frey, and Norman Thelen. 2001. Confirmation bias in sequential information search after preliminary decisions: an expansion of dissonance theoretical research on selective exposure to information. *Journal of personality and social psychology* 80, 4 (2001), 557.

[46] 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*. Association for Computational Linguistics, Atlanta, Georgia, USA, 556–562.

[47] Sanjay Kairam and Jeffrey Heer. 2016. Parting Crowds: Characterizing Divergent Interpretations in Crowdsourced Annotation Tasks. In *Proceedings of the ACM Conference on Computer-Supported Cooperative Work & Social Computing (CSCW'16)*. ACM, New York, NY, USA, 1637–1648. https://doi.org/10.1145/2818048.2820016

[48] Matthew Kay, Tara Kola, Jessica R. Hullman, and Sean A. Munson. 2016. When (Ish) is My Bus?: User-centered Visualizations of Uncertainty in Everyday, Mobile Predictive Systems. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*. ACM, New York, NY, USA, 5092–5103. https://doi.org/10.1145/2858036.2858558

[49] Thomas Kelly. 2008. Disagreement, dogmatism, and belief polarization. *The Journal of Philosophy* 105, 10 (2008), 611–633.

[50] Hyunwoo Kim, Eun-Young Ko, Donghoon Han, Sung-Chul Lee, Simon T Perrault, Jihee Kim, and Juho Kim. 2019. Crowdsourcing Perspectives on Public Policy from Stakeholders. In *Extended Abstracts of the ACM Conference on Human Factors in Computing Systems (CHI'19)*. ACM, New York, NY, USA, LBW1220. https://doi.org/10.1145/3290607.3312769

[51] Aniket Kittur, Ed H. Chi, and Bongwon Suh. 2008. Crowdsourcing User Studies with Mechanical Turk. In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI'08)*. ACM, New York, NY, USA, 453–456. https://doi.org/10.1145/1357054.1357127

[52] Solomon Kullback and Richard A Leibler. 1951. On information and sufficiency. *The annals of mathematical statistics* 22, 1 (1951), 79–86. http://www.jstor.org/stable/2236703

[53] Walter Lasecki and Jeffrey Bigham. 2012. Self-correcting crowds. In *Extended Abstracts of the ACM conference on Human Factors in Computing Systems (CHI'12)*. ACM, New York, NY, USA, 2555–2560. https://doi.org/10.1145/2212776.2223835

[54] Walter S. Lasecki, Mitchell Gordon, Danai Koutra, Malte F. Jung, Steven P. Dow, and Jeffrey P. Bigham. 2014. Glance: Rapidly Coding Behavioral Video with the Crowd. In *Proceedings of the 27th Annual ACM Symposium on User Interface Software and Technology (UIST '14)*. ACM, New York, NY, USA, 551–562. https://doi.org/10.1145/2642918.2647367

[55] John Le, Andy Edmonds, Vaughn Hester, and Lukas Biewald. 2010. Ensuring quality in crowdsourced search relevance evaluation: The effects of training question distribution. In *SIGIR 2010 workshop on crowdsourcing for search evaluation*, Vol. 2126. ACM, New York, NY, USA.

[56] 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. In *Proceedings of the Conference on Computer-Supported Cooperative Work & Social Computing (CSCW'18)*. ACM, New York, NY, USA. https://doi.org/10.1145/3274373

[57] Christopher H Lin, Mausam Mausam, and Daniel S Weld. 2012. Dynamically Switching between Synergistic Workflows for Crowdsourcing. In *Twenty-Sixth AAAI Conference on Artificial Intelligence*.

[58] Donald G. Mackay and Thomas G. Bever. 1967. In search of ambiguity. *Perception & Psychophysics* 2, 5 (01 May 1967), 193–200. https://doi.org/10.3758/BF03213049

[59] Winter Mason and Duncan J Watts. 2009. Financial incentives and the performance of crowds. In *Proceedings of the ACM SIGKDD workshop on human computation*. ACM, New York, NY, USA, 77–85. https://doi.org/10.1145/1809400.1809422

[60] Tyler McDonnell, Matthew Lease, Mucahid Kutlu, and Tamer Elsayed. 2016. Why is that relevant? Collecting annotator rationales for relevance judgments. In *Fourth AAAI Conference on Human Computation and Crowdsourcing*. AAAI, Palo Alto, CA, USA.

[61] 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.

[62] David Oleson, Alexander Sorokin, Greg P Laughlin, Vaughn Hester, John Le, and Lukas Biewald. 2011. Programmatic Gold: Targeted and Scalable Quality Assurance in Crowdsourcing. *Human computation* 11, 11 (2011).

[63] Ingram Olkin and Friedrich Pukelsheim. 1982. The distance between two random vectors with given dispersion matrices. *Linear Algebra Appl.* 48 (1982), 257 – 263. https://doi.org/10.1016/0024-3795(82)90112-4

[64] Barbara Plank, Dirk Hovy, and Anders Søgaard. 2014. Linguistically debatable or just plain wrong?. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (volume 2: Short Papers)*, Vol. 2. ACL, Baltimore, MD, USA, 507–511. https://doi.org/10.3115/v1/P14-2083

[65] Dražen Prelec. 2004. A Bayesian Truth Serum for Subjective Data. *Science* 306, 5695 (2004), 462–466. https://doi.org/10.1126/science.1102081

[66] Dražen Prelec, H Sebastian Seung, and John McCoy. 2017. A solution to the single-question crowd wisdom problem. *Nature* 541, 7638 (2017), 532.

[67] Mike Schaekermann, Joslin Goh, Kate Larson, and Edith Law. 2018. Resolvable vs. Irresolvable Disagreement: A Study on Worker Deliberation in Crowd Work. *Proc. ACM Hum.-Comput. Interact.* 2, CSCW, Article 154 (Nov. 2018), 19 pages. https://doi.org/10.1145/3274423

[68] Shilad Sen, Margaret E. Giesel, Rebecca Gold, Benjamin Hillmann, Matt Lesicko, Samuel Naden, Jesse Russell, Zixiao (Ken) Wang, and Brent Hecht. 2015. Turkers, Scholars, "Arafat" and "Peace": Cultural Communities and Algorithmic Gold Standards. In *Proceedings of the ACM Conference on Computer-Supported Cooperative Work & Social Computing (CSCW'15)*. ACM, New York, NY, USA, 826–838. https://doi.org/10.1145/2675133.2675285

[69] Pao Siangliulue, Kenneth C Arnold, Krzysztof Z Gajos, and Steven P Dow. 2015. Toward collaborative ideation at scale: Leveraging ideas from others to generate more creative and diverse ideas. In *Proceedings of the ACM Conference on Computer-Supported Cooperative Work & Social Computing (CSCW'15)*. ACM, New York, NY, USA, 937–945. https://doi.org/10.1145/2675133.2675239

[70] Pao Siangliulue, Joel Chan, Krzysztof Z Gajos, and Steven P Dow. 2015. Providing timely examples improves the quantity and quality of generated ideas. In *Proceedings of the Conference on Creativity and Cognition (C&C'15)*. ACM, New York, NY, USA, 83–92. https://doi.org/10.1145/2757226.2757230

[71] 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*. ACL, Baltimore, MD, USA, 254–263.

[72] Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Seattle, Washington, USA, 1631–1642.

[73] 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 Transactions on Interactive Intelligent Systems (TiiS)* 19, 1 (2019),

Article No.3. https://doi.org/10.1145/3237188

[74] Jean Y Song, Raymond Fok, Alan Lundgard, Fan Yang, Juho Kim, and Walter S Lasecki. 2018. Two tools are better than one: Tool diversity as a means of improving aggregate crowd performance. In *23rd International Conference on Intelligent User Interfaces (IUI'18)*. ACM, New York, NY, USA, 559–570. https://doi.org/10.1145/3172944.3172948

[75] Jean Y Song, Stephan J Lemmer, Michael Xieyang Liu, Shiyan Yan, Juho Kim, Jason J Corso, and Walter S Lasecki. 2019. Popup: reconstructing 3D video using particle filtering to aggregate crowd responses. In *Proceedings of the 24th International Conference on Intelligent User Interfaces (IUI'19)*. ACM, New York, NY, USA, 558–569. https://doi.org/10.1145/3301275.3302305

[76] Chen Sun, Abhinav Shrivastava, Saurabh Singh, and Abhinav Gupta. 2017. Revisiting unreasonable effectiveness of data in deep learning era. In *Proceedings of the IEEE international conference on computer vision*. IEEE, Piscataway, New Jersey, US, 843–852. https://doi.org/10.1109/ICCV.2017.97

[77] Jaime Teevan and Lisa Yu. 2017. Bringing the wisdom of the crowd to an individual by having the individual assume different roles. In *Proceedings of the Conference on Creativity and Cognition (C&C'17)*. ACM, New York, NY, USA, 131–135. https://doi.org/10.1145/3059454.3059467

[78] Petros Venetis, Hector Garcia-Molina, Kerui Huang, and Neoklis Polyzotis. 2012. Max algorithms in crowdsourcing environments. In *Proceedings of the 21st international conference on World Wide Web*. ACM, New York, NY, USA, 989–998. https://doi.org/10.1145/2187836.2187969

[79] Edward Vul and Harold Pashler. 2008. Measuring the Crowd Within: Probabilistic Representations Within Individuals. *Psychological Science* 19, 7 (2008), 645–647. https://doi.org/10.1111/j.1467-9280.2008.02136.x

[80] William Yang Wang, Dan Bohus, Ece Kamar, and Eric Horvitz. 2012. Crowdsourcing the acquisition of natural language corpora: Methods and observations. In *2012 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, Piscataway, New Jersey, US, 73–78. https://doi.org/10.1109/SLT.2012.6424200

[81] Chun-Ju Yang, Kristen Grauman, and Danna Gurari. 2018. Visual Question Answer Diversity. In *Sixth AAAI Conference on Human Computation and Crowdsourcing*. AAAI, Palo Alto, CA, USA.

[82] 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 (ICMI '17)*. ACM, New York, NY, USA, 51–59. https://doi.org/10.1145/3136755.3136792

Received April 2019; revised June 2019; accepted August 2019