Quantifying and Leveraging Uncertain and Imprecise Answers in Multiple Choice Questionnaires for Crowdsourcing
Quantifier et exploiter des réponses incertaines et imprécises dans des questionnaires à choix multiples pour le crowdsourcing

Constance Thierry
name.surename@irisa.fr
Univ. Rennes, CNRS, IRISA, DRUID, France
Lannion, France

Géry Casiez
name.surname@inria.fr
Univ. Lille, Inria Lille - Nord Europe, CRIStAL - UMR 9189, France
Lille, France

Jean-Christophe Dubois
name.surename@irisa.fr
Univ. Rennes, CNRS, IRISA, DRUID, France
Lannion, France

Yolande Le Gall
name.surename@irisa.fr
Univ. Rennes, CNRS, IRISA, DRUID, France
Lannion, France

Sylvain Malacria
name.surname@inria.fr
Univ. Lille, Inria Lille - Nord Europe, CRIStAL - UMR 9189, France
Lille, France

Thomas Pietrzak
name.surname@inria.fr
Univ. Lille, Inria Lille - Nord Europe, CRIStAL - UMR 9189, France
Lille, France

Arnaud Martin
Univ. Rennes, CNRS, IRISA, DRUID, France
Lannion, France

Pierrick Uro
Univ. Lille, Inria Lille - Nord Europe, CRIStAL - UMR 9189, France
Lille, France

ABSTRACT

Questionnaires are efficient for collecting numerous user feedback. However, the reliability of results is a major issue, even with honest participants. Indeed, they face situations of doubt, and usually do not have the option to express their hesitations. We describe a user study in which we provide participants with the possibility to give 1) a certitude rate 2) imprecise answers, 3) both a certitude rate and imprecise answers. Firstly, we observe that contributors express their hesitations consistently: there is a correlation between the task difficulty on the one hand, and the uncertainty and imprecision of the answer, on the other hand. Secondly, our results demonstrate the effectiveness of the decision-making process by using this additional information with the belief functions theory. Indeed, this process helps to reduce the error rate and fewer participants are required to reach a satisfactory correct answers rate.

RÉSUMÉ

Les questionnaires sont utiles pour le recueil de retours utilisateurs. Cependant, la fiabilité des résultats est un problème majeur, même avec des participants honnêtes. Les contributeurs confrontés à des situations d'indécision n'ont généralement pas la possibilité d'exprimer leurs hésitations. Nous décrivons une étude utilisateur dans laquelle nous donnons aux participants la possibilité de donner 1) un degré d'incertitude, 2) des réponses imprécises, 3) à la fois un degré d'incertitude et des réponses imprécises. Nous observons que les contributeurs expriment leurs hésitations de manière cohérente : il existe une corrélation entre la difficulté de la tâche, l'incertitude et l'imprécision de la réponse. Nos résultats démontrent l'efficacité du processus de prise de décision en utilisant cette information supplémentaire avec la théorie des fonctions de croyance. Nous réduisons le taux d'erreur grâce à cette analyse et nous montrons que moins de participants sont nécessaires pour atteindre un taux satisfaisant de réponses correctes.

CCS CONCEPTS

• **Human-centered computing → HCI theory, concepts and models**; *Pointing*; Visualization techniques.

KEYWORDS

Questionnaires, Imprecision, Uncertainty, Crowdsourcing, user expressivity

MOTS CLÉS

Questionnaires, Imprécision, Incertitude, Crowdsourcing, Expressivité de l'utilisateur

ACM Reference Format:
Constance Thierry, Géry Casiez, Jean-Christophe Dubois, Yolande Le Gall, Sylvain Malacria, Thomas Pietrzak, Arnaud Martin, and Pierrick Uro. 2024. Quantifying and Leveraging Uncertain and Imprecise Answers in Multiple Choice Questionnaires for Crowdsourcing. In *Proceedings of the 35th Conference on l'Interaction Humain-Machine (IHM '24), March 25–29, 2024, Paris, France.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3649792.3649806

IHM '24, March 25–29, 2024, Paris, France

1 INTRODUCTION

Questionnaires frequently use closed-ended questions, providing participants with predefined options for selecting a response. Depending on the nature of a closed-ended question, respondents may be required to choose a unique option (typically when using radio buttons), or can select multiple options (typically with checkboxes). Interpreting responses to closed-ended questions is a delicate task as such questions do not allow for the expression of *uncertainty*. Therefore, when analyzing answers, it is generally assumed that respondents were *certain* of their choices. Currently, few interfaces enable users to modulate their answers and to integrate any hesitations they may have [10, 12, 14, 15]. Moreover, existing interfaces are mostly related to educational systems [10, 12, 15] where a correct answer is defined for each question. However, it is not always possible to have a ground truth in questionnaires, as in the case of surveys. For example, when asking a person whether a given book excerpt is relevant or not, the answer depends on the respondent's subjective perception and domain knowledge. As a result, contributors may be prompted to provide additional information about their knowledge on a subject [14], but this differs from measuring the certainty of their answers.

The aim of this work is to study the relevance of an information-gathering interface designed for the general public, capable of taking into account participants' potential hesitations. Therefore, we propose enhancing contributors' expressive capabilities by requesting additional information regarding their uncertainty and imprecision, two inherent imperfections in human contributions, alongside their initial responses. *Imprecision* is defined as participants' indecision among several options, and uncertainty refers to participants' self-assessment of the confidence in their answer. To validate this approach, the article addresses two key questions: Is there a correlation between the difficulty of a task and the imprecision and uncertainty expressed by a contributor? Does considering the imperfections in responses contribute to improve decision-making during data processing?

Our study focuses on multiple-choice questionnaires (MCQs), which involve only one correct answer and no ground truth. The objective is to acquire knowledge through the analysis of responses provided by multiple contributors. Achieving a reliable answer requires a substantial number of participants, leading to the common practice of conducting such studies on crowdsourcing platforms. Crowdsourcing platforms facilitate the efficient gathering of information by compensating a crowd of contributors for answering a predefined set of questions, constituting a task. While this approach is quick and accessible, it can also incur significant costs. Consequently, our goal is to minimize expenses by reducing the number of participants required for these campaigns. In this article, we aim to demonstrate that a more expressive interface has the potential to reduce the need for a large number of contributors by relying on more reliable information. The analysis of responses is based on the theory of belief functions.

This article begins by presenting related work on crowdsourcing interfaces and task creation in the section 2, then, in the section 3, moves on to a user study and its results.

2 RELATED WORK

In this work we focus on leveraging the trade-off between certitude and precision for MCQs crowdsourcing tasks. This section reviews previous work on certainty and on crowdsourcing campaign interfaces. The theory of belief functions, which will be used in the latter part of this article to model imprecise and uncertain MCQ answers, is also introduced.

2.1 Overcoming the rigidity of MCQs

Multiple-choice questionnaires (MCQs) are a specific type of questionnaire that proposes a restricted set of options to a question. . An inherent problem with MCQs is that their restricted format limits respondents' ability to express themselves, as these questionnaires do not allow respondent's to express their ignorance, imprecision and uncertainty. As a result, if they hesitate between two (or more) answers, they will choose one of them randomly. . In a crowdsourcing context, a random response from contributors introduces noise into the collected data. This forces the employer to hire a large number of contributors to reduce this noise, but also increases the cost of the campaign.

2.2 Crowdsourcing interfaces

The literature on crowdsourcing generally deals with issues of answers aggregation [1, 4, 17, 21, 23], contributor profile [27, 28], motivation [25] and question assignment[2, 5, 13]. However, to the best of our knowledge, only a handful of articles deal with crowdsourcing interfaces. Some of these studies focus on different tasks such as image classification and notation [22] or text characterization [14]. Very few, however, relate to MCQs. For example, Thierry *et al.* [26] asked participants to provide a confidence rate of their answers on a 5-point Likert scale, and allow them to select two offering magnitude in the precision of the reported answer. However, this level of imprecision is limited to up to two choices, and the authors did not study the relationship and possible trade-off between precision and certitude. This is of particular interest, given that the way the task is communicated has a real impact on the quality of the contributor's work. Kazai *et al.* [14] argue that task design plays a crucial role, because even well-meaning contributors can generate erroneous data if the task design and interface are of poor quality. This fact is corroborated by the contributors of crowdsourcing platforms interviewed by Kittur *et al.* [16], who claim that the proposed tasks are often poorly designed, leading to misunderstandings between contributors and the employer. The authors propose a task design where the system would explain the importance of the job, offer feedback from peers and experts, and encourage self-assessment.

Feedback is an important element for contributors and can help improve their performance, which is why AMT [3] recommends introducing feedback into the crowdsourcing campaign using, for example, gold data. Dow *et al.* [11] are interested in the possibility for the contributor to resume work after self-correction or external feedback and to do so create an interface to facilitate synchronous feedback. A comparison of three scenarios is carried out, in the first one there is no correction as it is in our crowdsourcing campaign; in the second one an auto-correction is possible; and in the last one, an external correction is carried out by an expert. Open questions are asked to the contributors so that the feedback from the experts can improve the quality of answers. Results show that both self-correction and external correction lead to better quality answers. The external review led to higher performance than the self-correction, but not to better work overall.

The main commonality of the interfaces mentioned above is that most use MCQs. Indeed, they are often used in crowdsourcing platforms because they are easier to aggregate. Marcus and Parameswaran [18] also recommend using closed questions for crowdsourcing when the subject allows it.

2.3 Belief Functions

The theory of belief functions was introduced by Dempster [8] and formalized by Shafer [24]. It is a generalization of fuzzy and probabilistic approaches and allows to model the imprecision and uncertainty of imperfect sources of information. Because of their human nature, contributors to crowdsourcing campaigns are imperfect sources of information and may display uncertainty in their responses. This is why [1, 17, 20, 26] use belief functions to model responses from crowdsourcing campaigns.

The theory considers a set Ω of hypotheses called the frame of discernment. In a crowdsourcing context, the set of answers $\omega_1 ... \omega_n$ proposed to the contributor constitute the discernment framework $\Omega = \{\omega_1 ... \omega_n\}$. Mass functions $m : 2^\Omega \to [0,1]$ model the elementary degree of belief of the source and respect the normalization condition:

$$\sum_{X \in 2^\Omega} m(X) = 1 \tag{1}$$

Thus, if we consider a contributor c answering a question q whose set of possible answers is Ω. If the contributor chooses answer $X \in 2^\Omega$ with $\alpha \in [0,1]$ a confidence in its answer, then we can model it as $m^\Omega_{cq}(X) = \alpha$. The higher the mass $m^\Omega(X)$, the stronger the contributor's belief in his X answer. The X contribution may be imprecise if the contributor selects several $\omega_i \subset \Omega$ answers. Ignorance is modeled in the theory of belief functions by the element Ω such as $m^\Omega(\Omega) > 0$.

There are specific mass functions such as simple support mass functions which reflects an uncertain and possibly imprecise response from the information source. For a contributor c answering a question q whose set of possible answers is Ω we have:

$$\begin{cases} m^\Omega_{cq}(X) = \alpha \text{ with } X \in 2^\Omega \setminus \Omega \\ m^\Omega_{cq}(\Omega) = 1 - \alpha \end{cases} \tag{2}$$

The contributor has an uncertain knowledge because it partially believes in X with a certainty α but not totally since a non-zero mass is present on Ω. Once again, X can be imprecise if it's a subset of Ω.

For information fusion, the K contributors c all report on the same frame of discernment Ω and the same question q. It is possible to average mass functions:

$$m_{Avg}(X) = \frac{1}{K} \sum_{c=1}^{K} m^\Omega_c(X) \tag{3}$$

Many other rules of combination exist in the theory of belief functions, this section lists some conjunctive rules, [19] presents more.

2.4 Summary of related work

MCQs are more common in crowdsourcing platforms as it eases processing the aggregation of answers. However, current platforms do not provide the possibility for contributors to express their doubt in case of hesitation, which can lead them to select a random answer and introduce noise in the collected data. The goal of this paper is to allow contributors to be imprecise in case of hesitation , and to give them the opportunity to express their certainty about their contribution. There are few studies of questionnaires that take into account both uncertainty and imprecision and, to our knowledge, none of them establish a link between these two elements and the task difficulty. Thierry *et al.* [26] do include uncertainty and imprecision in their MCQs. However, they does not focus on the interface itself, but on modeling answers and contributors profiling. Moreover, in the author's work, the imprecision that participants may report is limited and directed, and participants assess their certainty using numbers.

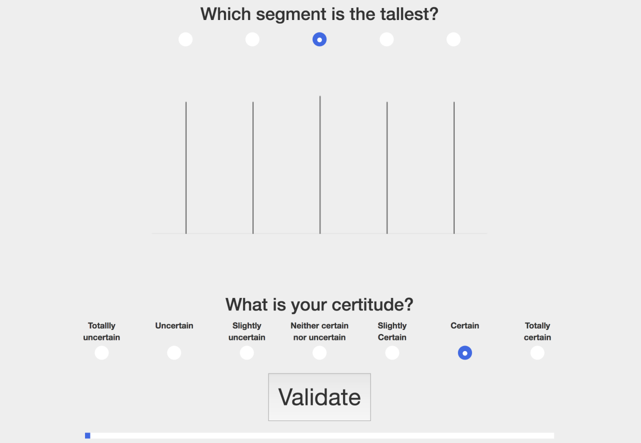
In our work, we offer contributors the option of selecting as many choices as they wish, up to and including all options if necessary. We also require them to estimate the certainty of their answer using a Likert scale. In what follows, we study the impact of this new questionnaire configuration on participants' answers, and propose an aggregation method that leverages the additional information to obtain more accurate final results with fewer participants.

When a contributor is required to formulate a certainty in his answer, two alternatives exist: authors proposing a scale of certainty and those requiring a numerical value. For our part, we have chosen to use a scale of certainty in our interface, as we believe that it is easier for the contributor to choose between different levels of certainty, rather than giving a numerical value, which is a complicated exercise. Moreover, requesting a numerical value of certainty may introduce a bias, as not all contributors associate the same degree of certainty to a given numerical value. For example, a value of 0.43 could mean few certain for one contributor and rather uncertain for another.

The next section introduces the interfaces designed for our experiments on crowdsourcing platform and the results obtained.

3 EXPERIMENT

We performed four tasks, the first one for certainty, the second one for imprecision, the third one for both in view to validate the next hypotheses and the last one as control experience. For the first hypothesis, we assume that the more difficult the task, the less the contributor will be certain of the answer (H1). Second, we estimate that the certainty self-assessment of the contributor depends on the global difficulty of the task (H2). Third, we believe that offering to the contributor the possibility to be imprecise increase the correct answer rate (H3) as the set of answer increase, the probability to give the good answer increase too. In addition, the contributor should be more confident by being imprecise than if he is constrained to select randomly an answer in a set of responses he believes correct. Moreover, thanks to uncertainty and imprecision our last hypothesis is that fewer contributors are required to perform a crowdsourcing task (H4) because these elements are included in the answer modeling for the data aggregation.


Which segment is the tallest?

○ ○ ● ○ ○

What is your certitude?

○ Totally uncertain ○ Uncertain ○ Slightly uncertain ○ Neither certain nor uncertain ○ Slightly Certain ● Certain ○ Totally certain

Figure 1: Interface for Task 1. Participants choose one segment, and provide a self-assessment of their certainty on a 7-point Likert scale.

3.1 Methodology

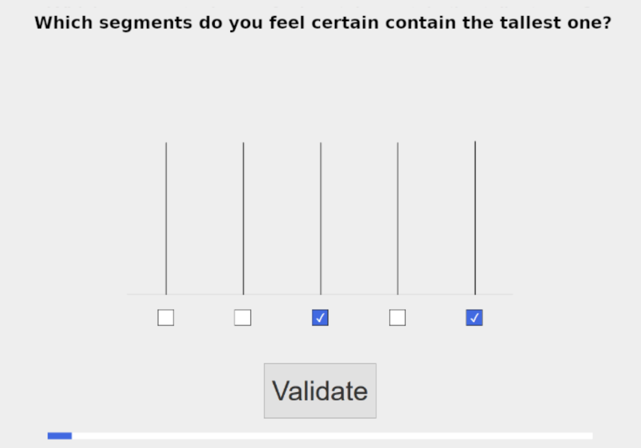
In this work, we are interested in understanding the trade-off between the certainty and precision of croudworkers when answering MCQs. In order to accurately investigate this point, we need questions with ground-truth answers, and we have to be able to control certainty and precision independently. This is why we opted for a psychophysics task in which we show participants five parallel segments of similar length, except (maximum) one. Participants have to answer which of the segments is the tallest. This task allows us to control its difficulty by adjusting the difference in length between the tallest segment and the other. Thus, we can design the four following tasks that respectively enable participants to provide a self-assessment of their certainty in their answer, provide several answers in case they have hesitations, or both.

Task 1: certainty. For this task, participants have to provide a precise answer (only one selection is allowed) and a self-assessment of their certainty in their answer. Figure 1 presents the interface designed for this task, that shows five vertical segments, aligned at the bottom. Participants are asked: "Which segment is the tallest?" and have to answer the question by selecting the segment they perceive as the tallest, thanks to the corresponding radio button. They are also asked to provide a self-assessment of their certainty: "What is your certitude?" by selecting a radio button in a 7-point Likert scale with the following choices: "Totally uncertain", "Uncertain", "Slightly uncertain", "Neither certain nor uncertain", "Slightly certain", "Certain", and "Totally certain".

The interface displays 5 segments: 1 *target* segment and 4 *control* segments. The length of control segments is 40 mm, the length of the target segment is $40 + \Delta$ mm and the horizontal space between all segments is 20 mm. We note Δ the difference between the length of control segments and the length of the target segment. In this task we used 2 sets of Δ:

$$\Delta_0 = \{0\ mm, 0.3\ mm, 0.6\ mm, 0.9\ mm, 1.2\ mm, 1.5\ mm\}$$

$$\Delta_1 = \{0\ mm, 0.3\ mm, 0.6\ mm, 1.2\ mm, 1.8\ mm, 2.4\ mm\}$$


Which segments do you feel certain contain the tallest one?

□ □ ☑ □ ☑

Figure 2: Interface for Task 2. Participants can select as many segments as they like to ensure they get the correct answer.

In the case of $\Delta = 0\ mm$, the target segment and the control segments have the same length, that is 40 mm. Therefore the expected error rate for these trials is the chance level: 20% (one chance out of 5). For these trials, half of the participants use Δ_0, and the other half use Δ_1. Our intent is to verify if participants provide an absolute or relative assessment of their certainty, depending on the difficulty of their trials (H2).

We explain to the contributors that no penalty is applied for uncertainty in their answers.

Task 2: precision. In this task, when hesitating between several segments, participants can provide answers with more or less *precision*, depending on the number of selected segments. The interface for this task (see Figure 2) is similar to that of Task 1, with the following differences: participants are asked: "Which segments do you feel certain contain the tallest one?"; to answer this question, they can select between one and several segments thanks to the corresponding checkboxes. No certainty level can be provided by the contributor.

The length of control segments and spacing between segments are the same as in Task 1. The set of Δ used in this task is the following:

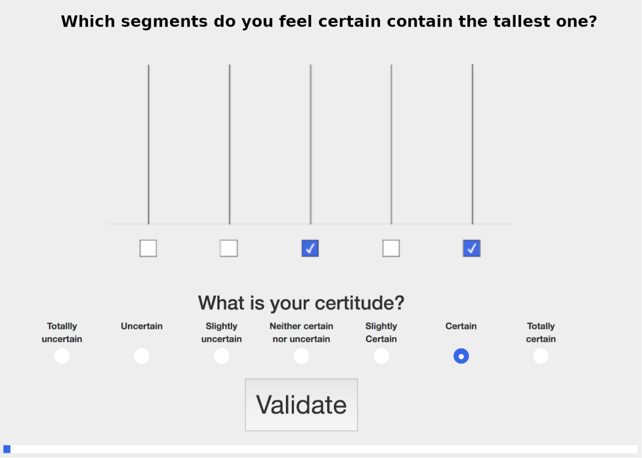
$$\Delta_2 = \{0\ mm, 0.3\ mm, 0.6\ mm, 0.9\ mm, 1.2\ mm, 1.8\ mm, 2.4\ mm\}$$

Given the instructions, we expect participants to select all the segments when $\Delta = 0$, since all segments are of the exact same length. As before, participants are informed that there is no penalty for multiple choices.

Task 3: certainty and precision. This task is a combination of Tasks 1 and 2. Indeed, they can select several options and have to provide a self-assessment of their certainty. The interface for this task is shown on Figure 3.

As for the Tasks 1 and 2, we instruct participants there is no penalty for being uncertain or selecting several options. The set of Δ used in this task is the same one as in Task 2:

$$\Delta_2 = \{0\ mm, 0.3\ mm, 0.6\ mm, 0.9\ mm, 1.2\ mm, 1.8\ mm, 2.4\ mm\}$$



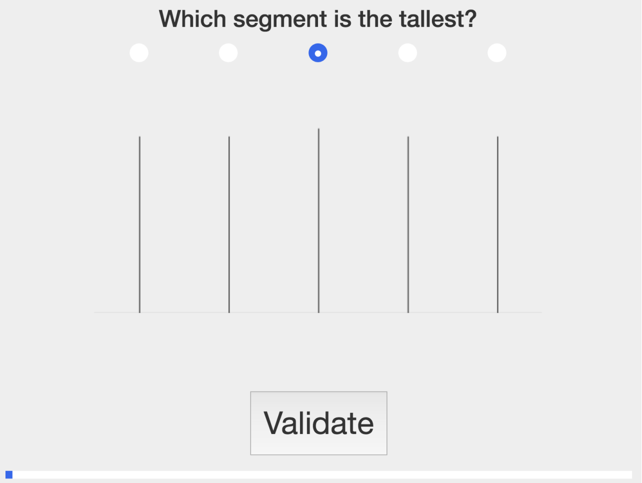
Which segments do you feel certain contain the tallest one?

□ □ ☑ □ ☑

What is your certitude?

Figure 3: Interface for Task 3. Participants can select as many segments as they like to ensure they get the correct answer. They also provide a self-assessment of their certainty on a 7-point Likert scale.

Task 4: control experience. This task has been created to reflect those that can be encountered in crowdsourcing platforms. Participants are asked: "Which segment is the tallest?" and answer the question by selecting the segment they consider to be the tallest using the corresponding radio button. They have to provide a precise answer, knowing that only one segment is allowed. They cannot report a self-assessment of their certainty.



Which segment is the tallest?

○ ○ ● ○ ○

Figure 4: Interface for Task 4. Participants can only choose one segment.

The set of Δ used in this task is the same one as in Task 2 and 3:

$$\Delta_2 = \{0\ mm, 0.3\ mm, 0.6\ mm, 0.9\ mm, 1.2\ mm, 1.8\ mm, 2.4\ mm\}$$

Common considerations. In all four tasks, the vertical position of the segments slightly changes in a random manner from one trial to another to avoid second-guessing due to persistence of vision. The screen is blackened for 1 second between trials to stimulate the eyes to respond to the change, in order to avoid comparisons between trials. A progress bar at the bottom of the page shows participants how far they have come in the task. We introduce one attention question per Block to dissuade participants from rushing the trials.

The experiment followed a mixed design. Task was a between-subject factor. Task 1 used 2 Delta series as a between-subject factor, while Tasks 2, 3 and 4 used only one. Therefore, each participant performed one of the four tasks, with one Delta series. Delta series in Task 1 contained 6 values each, while the Delta series in Tasks 2, 3 and 4 contained 7 values. All tasks had 2 within subject factors: the Position of the tallest segment (5 possibilities), and the number of Blocks (3 in task 1; 2 in tasks 2, 3 and 4). 100 participants have realized tasks 1 to 3, notice that as for task 1 there were two delta series, that means 50 participants have performed the task 1 for Δ_0 and 50 for Δ_1. For task 4, unlike the other tasks, it is mainly the response time that interests us, which is why the crowd performing this experiment is smaller and composed of 25 contributors.

The experiment design is the following:

Task 1 : 2 Delta series × 6 Delta values × 5 Positions × 3 Blocks × 50 Participants = 9000 trials.

Tasks 2 & 3 : 7 Delta values × 5 Positions × 2 Blocks × 100 Participants = 7000 trials.

Task 4 : 7 Delta values × 5 Positions × 3 Blocks × 25 Participants = 2625 trials.

We recruited participants on Crowdpanel[1], and made sure they participated in only one of the Tasks and one of the Delta series. We implemented the experiment as a node.js web server, the crowdpanel platform redirected participants to the web interface. We used Pointingserver from libpointing [6] to get the screen size dimensions in *mm* in order to ensure the same segment length was presented to every participant, independently of the screen they use.

3.2 Results and discussions

In this paper, we wish to show that offering contributors the possibility to be imprecise allow them to be more certain about their answers. We therefore expect the average certainty of the contributors who participated in experiment 3 to be higher than that of the contributors who performed experiment 1 for which they had to give a precise answer. In addition, we propose an answers aggregation using the belief function theory in order to process the information collected via this interface. This method offers better results than the majority vote, traditionally used in crowdsourcing platforms.

We discuss our results in light of our hypotheses. First, we discuss the relation between the task difficulty and the contributors' self-assessment of their certainty in their answers. Then, we discuss the relative against the absolute aspect of participants' certitude self-assessment. Finally, we present aggregation methods that leverage imprecision and incertitude to obtain good answers with fewer participants.

[1] https://crowdpanel.io/

3.2.1 Relation between certainty and difficulty. Figure 5 presents the average participants' self-assessment of certitude in Task 1 for both Δ_0 and Δ_1, and in Task 3. Value 0 represents "Totally uncertain", and 6 represents "Totally certain".

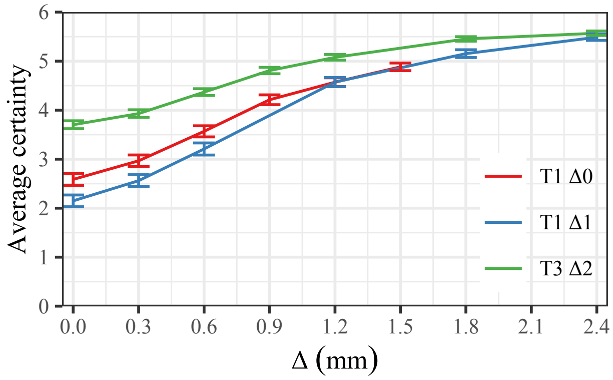


Figure 5: Average certainty on the answer for the different values of the augmented size of the tallest segment. Error bars are 95% confidence intervals.

As the crowd of the first experiment could select only one segment and considering that for $\delta = 0$ the segments were indistinguishable, the contributors should be totally uncertain of the validity of their answer and the average certainty should be around 0. But in Figure 5, the lowest value of the average certainty is about 2 which corresponds to the proposition "Slightly uncertain". In our opinion, the contributor probably imagines that his work will be depreciated if he is not confident enough in his answer. Of course, this is not the case, but it may explain why he does not admit that he is totally uncertain.

The Bravais-Pearson correlation coefficients calculated for the values of the axes (certainty and Δ) are the followings : $r_{\Delta_0} = 0.5, r_{\Delta_1} = 0.7, r_{\Delta_2} = 0.5$. According to the coefficients r the average certainty is positively correlated to the increased size of the segment and negatively correlated to the difficulty of the question which validate the hypothesis H1.

For task 1, Mann–Whitney U tests did not reveal significant differences (p>0.1) between Δ0 and Δ1 for each common $\delta = \{0mm, 0.3mm, 0.6mm, 1.2mm\}$, in spite of Figure 5 suggesting some differences. This suggests that the two groups of contributors had a similar degree of certainty for these values. We observe on Figure 7 that the correct answer rate is close to 100% when $\delta \geq 1.2mm$, which explains participants share the same certitude in both conditions. We therefore partially validate H2, because participants did not have the same scale of certitude self-assessment depending on the δ values they were presented in their trials. This effect is smaller when the task is easy enough. The certainty interval used by the contributor for their self-assessment is restricted by the amplitude of the difficulty interval for Δ_0, which confirms hypothesis H2, the self-assessed certainty used by the contributor is relative to the overall difficulty of the task. Moreover, the average certainty of the crowd in task 3, which had the possibility of being imprecise, is higher than the average certainty of the crowd in task 1, which had to be precise in its answers. Offering contributors the opportunity to be imprecise in order to cope with the difficulty of the task enables them to be more confident as mentioned in H3.

3.2.2 Leveraging imprecision for difficult questions. Figure 6 shows the average imprecision of contributors for the Tasks 2 and 3. We define imprecision as the number of segments that have been selected by participants in a trial. It ranges from 1 to 5 because participants could select 1 to 5 segments.

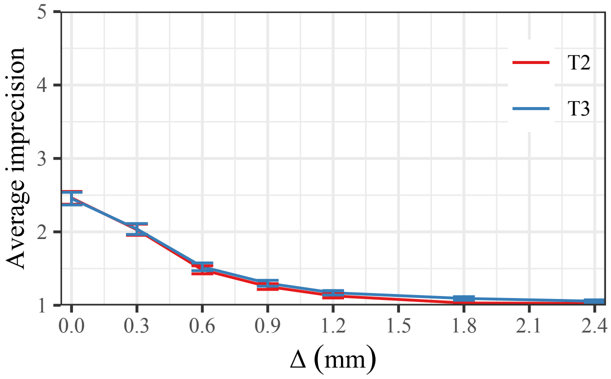


Figure 6: Average imprecision of the answer for the different values of the augmented size of the tallest segment.

Figure 7 presents the correct answer rate for the four tasks. Answers from Tasks 2 and 3 are considered as correct when the tallest segment is included in the set of selected segments.

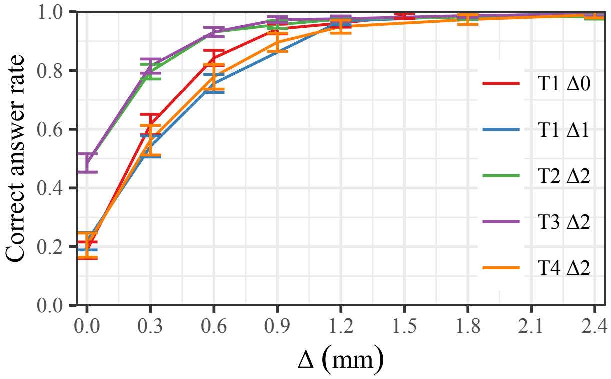


Figure 7: Correct answer rates for the different values of the augmented size of the tallest segment.

Indeed, participants selected more segments when the task was difficult than when it was easy. We expected participants to select all 5 segments in the $\delta = 0$ case since it was impossible for them to distinguish them. However, we observe that they selected only 2.5 segments for that value of δ in both Tasks 2 and 3 (Figure 6). It explains why the error rate is 50%, greater than the chance level (20%), but much smaller than 100% (Figure 7). Therefore we can validate H3: allowing participants to be imprecise increases their certainty. In addition, Mann–Whitney U tests did not reveal significant differences (p>0.1) between the tasks 2 and 3 for each δ value (Figure 6), suggesting that asking for a confidence self-assessment does not influence the number of answers participants select.

The correct answer rates as discussed above (Figure 7) are the average correct response rates for each contributor. In crowdsourcing campaigns, the employer is interested in the results obtained by aggregating the answers. In the next section, we explain how to aggregate the answers, by taking into account imprecision and certitude.

3.2.3 *Answer aggregation leveraging imprecision and certainty.*

The underlying idea behind answer aggregation is to leverage the quantity of answers in order to minimize the odds of collecting incorrect information.

We compare the majority voting traditionally used for the answer aggregation in crowdsourcing platforms to the Expectation-Maximisation algorithm and an approach using the theory of belief function. The theory of belief function allows modelling the imprecision and uncertainty of the answers, so by using this theory for the answer aggregation we should obtain better results than the majority voting that does not consider the certainty of the contributor in their answer.

Majority Voting (MV). MV is an aggregation method commonly used in crowdsourcing platforms because it is simple to implement. This method is generally used with precise data only, that is with questions where contributors had to provide a precise answer. We adapt the majority vote here so that it can also be used for imprecise answers and to compare the results of the three tasks. To do this, the segments are modeled by indicator functions r_{cq}, with c a contributor and q a question. Let Ω be the set of possible answers to the question q and $X \subset \Omega$ be the contributor's answer c to the question, then:

$$\begin{cases} r_{cq}(X) = 1, X \subset \Omega \text{ if the contributor chooses the answer } X \\ r_{cq}(Y) = 0, Y \in \Omega \backslash X, \text{ otherwise} \end{cases} \tag{4}$$

The answers are aggregated by summing the indicators, and the answer that was given by the largest number of contributors is selected.

Expectation-Maximisation algorithm (EM). The algorithm proposed by Dempster *et al.* [9] allows the estimation of missing data. It is iterative and composed of two phases: "Expectation" which estimates the unknown data thanks to the current parameters, and "Maximization" which calculates the new parameters according to the current data. Both are repeated until the algorithm converges. David and Skene [7] apply EM in a framework similar to that of crowdsourcing, we draw inspiration from them for our experiments.

Belief Function Theory. The theory of belief functions [8] can be used to model the uncertainty and imprecision of imperfect sources of information. Applied to crowdsourcing, contributors are sources of information and their answers, in the context of our experiment, may be imprecise and/or uncertain. The frame of discernment Ω is here the set of proposed answers to the contributors.

For this study we use simple mass functions m_{cq}^{Ω} (equation (2)) to model a contributor's answer to a question with α the numerical value associated with the certainty that the contributor c has provided for their answer to question q. All these numerical values are given in Table 1. The function m_{cq} characterizes the fact that the contributor partially believes in his answer X, which may be imprecise, with a mass α but no more. The mass functions are then aggregated per question by an averaging operator for all the contributors. In order to make a decision on the answer, the aggregated mass function m_q^{Ω} is transformed into a pignistic probability according to the equation:

$$betP(X) = \sum_{Y \in 2^{\Omega}, Y \neq \emptyset} \frac{|X \cap Y|}{|Y|} \frac{m_q^{\Omega}(Y)}{1 - m_q^{\Omega}(\emptyset)}, X \in \Omega \tag{5}$$

The segment with the highest probability is considered the largest segment.

Table 1: Numerical values α associated to the certainty scale.

Totaly uncertain	0.2
Uncertain	0.3
Slightly uncertain	0.4
Neither certain nor uncertain	0.5
Slightly certain	0.6
Certain	0.7
Totaly certain	0.8

The combination operators were applied to the three tasks with all the crowd. For tasks 2 and 4, since the certainty of the contributor is not known, we therefore arbitrarily chose certainty: "Neither certain, nor uncertain" for the belief function method. Answers obtained after aggregation are compared to expected correct answers in order to calculate correct answer rates for each data aggregation method. These correct answer rates are listed for each task in Tables 2, 3, 4, 5, and 6.

For all the tasks it is only when the 5 segments are similar in size that the correct response rate is not equal to 1, except for the Δ_1 of the Task 1 for which a segment size increased by $\delta = 0.3$ mm get a correct answer rate lower than 1. The possible imprecision of the contributors is not penalizing for the quality of the aggregated data since the rate of good response remains maximum. The lowest correct answer rate for $\delta = 0$ is due to the fact that the largest segment is indistinguishable. Moreover, this correct answer rate is equal to the chance rate of 0.2 for tasks 2 and 3.

In our final analysis of the results illustrated in Figure 8, we gradually increase the number of contributors selected for the aggregation of responses, starting with 2 contributors to the whole crowd. For this aggregation, we are only interested in questions for which $\delta = 0.3$. We have chosen this difficulty value because it is common to all tasks and is the highest difficulty after $\delta = 0$. We have chosen not to use the questions for which $\delta = 0$, because for this difficulty value the 5 segments being of identical size.

Whereas for $\delta = 0.3$ it is possible to obtain a correct answer rate of 1 for the majority of the tasks as shown in Tables 2, 3, 4, and 5. The correct answer rates presented in Figure 8 are calculated as follows. A group of $n \in \{2, 4, 5, 8, 10, 15, 20, 25, 30, 35, 40, 45, 50, 55, 60, 70, 80, 90, 100\}$ contributors is formed, the contributors being randomly selected. For this

Table 2: Aggregation xp1 (Δ_0)

Δ_0 (mm)	0	0.3	0.6	0.9	1.2	1.5
MV	**0.13**	1.00	1.00	1.00	1.00	1.00
EM	**0.13**	1.00	1.00	1.00	1.00	1.00
Belief	**0.13**	1.00	1.00	1.00	1.00	1.00

Table 3: Aggregation xp1 (Δ_1)

Δ_1 (mm)	0	0.3	0.6	1.2	1.8	2.4
MV	**0.27**	**0.93**	1.00	1.00	1.00	1.00
EM	**0.33**	**0.93**	1.00	1.00	1.00	1.00
Belief	**0.13**	**0.93**	1.00	1.00	1.00	1.00

Table 4: Aggregation xp2

Δ_2 (mm)	0	0.3	0.6	0.9	1.2	1.8	2.4
MV	**0.20**	1.00	1.00	1.00	1.00	1.00	1.00
EM	**0.10**	1.00	1.00	1.00	1.00	1.00	1.00
Belief	**0.20**	1.00	1.00	1.00	1.00	1.00	1.00

Table 5: Aggregation xp3

Δ_2 (mm)	0	0.3	0.6	0.9	1.2	1.8	2.4
MV	**0.20**	1.00	1.00	1.00	1.00	1.00	1.00
EM	**0.10**	1.00	1.00	1.00	1.00	1.00	1.00
Belief	**0.20**	1.00	1.00	1.00	1.00	1.00	1.00

Table 6: Aggregation xp4

Δ_2 (mm)	0	0.3	0.6	0.9	1.2	1.8	2.4
MV	**0.33**	**0.97**	**0.97**	1.00	1.00	1.00	1.00
EM	**0.27**	1.00	1.00	1.00	1.00	1.00	1.00
Belief	**0.4**	**0.93**	**0.93**	1.00	1.00	1.00	1.00

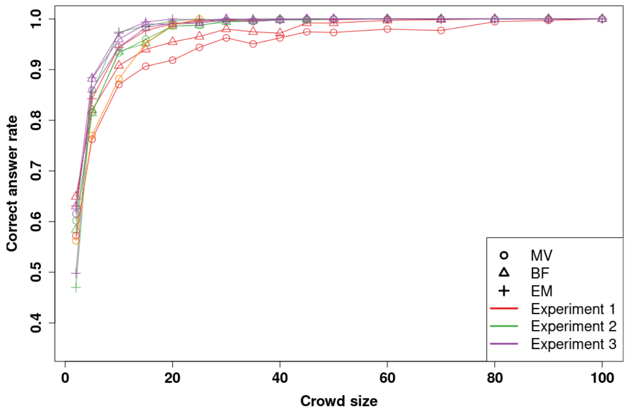


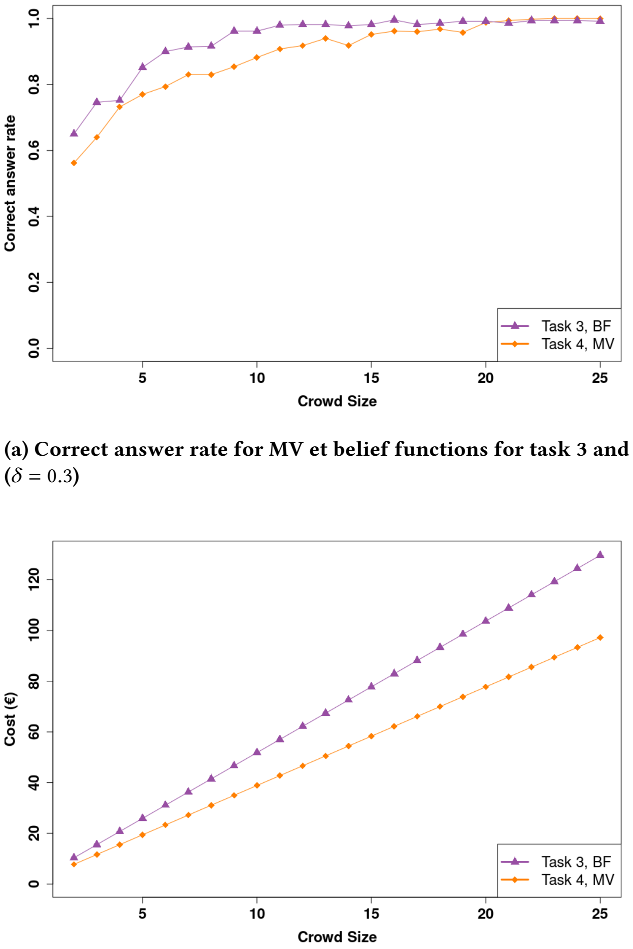
Figure 8: Correct answer rate for $\delta = 0.3$ questions increasing the number of contributors selected to constitute the crowd.

group of n contributors the responses for which $\delta = 0.3$ are aggregated, by MV and by belief function, then the correct answer rate on the aggregated data is calculated for both aggregation methods. This process of random selection of contributors and calculation of correct answer rates is repeated 50 times. Figure 8 shows, for a group of n contributors, the average of the 50 correct answer rates obtained for questions where $\delta = 0.3$. We thus wish to see if it was really necessary to constitute a crowd of 100 contributors in order to obtain 100% correct answers, or if we could have reduced the size of the crowd. Another interest of this analysis is to know if the proposed interface has an impact on the number of contributors needed in the crowd.

According to Figure 8 the rate of good response increases with the size of the crowd until the maximum value of 1 is reached. On this graph, the correct response rate is always higher than the chance rate of 0.2 even when the crowd is composed of only two contributors, so there is no randomness here. This graph also shows higher correct answer rates for task 3 than for task 1. The interface we propose allowing the contributor to be imprecise reaches a high correct answer rate with fewer contributors than a more traditional interface requiring the contributor to be precise in their answer. We can thus see by graphical reading that 30 contributors are sufficient to have a good response rate of 1 with the interface of task 3 and an aggregation by belief function. To obtain the same rate with the interface of task 1, a minimum of 55 contributors is required with an aggregation using belief functions and 90 for the MV. This gain in the number of contributors solicited is interesting for the employer because the number of contributors and the type of task directly impact the cost of a crowdsourcing campaign. It should be noted that the difference in correct answer rate between the aggregation by belief function and the aggregation by MV that we have just observed for task 1 is also present for the other tasks. Thus, for the three tasks, the aggregation of results using belief functions obtains a better correct answer rate than the MV.

The interfaces allowing the contributor to be imprecise offer better results compared to those constraining the contributor to give a precise answer, which tend to validate hypothesis H4. Moreover, by asking the contributor certainty in the interface of the task 3 in addition to the imprecision (as describe section 3.1) the results also improve. Considering the imprecision and the uncertainty of answers allow obtaining better results with fewer contributors. Moreover, aggregation using belief functions is therefore more relevant for the interface using uncertainty and imprecision propose.

The interface in task 3 allows the contributor to be imprecise while giving certainty about their answer is attractive to the employer because the crowd size required to complete the campaign is smaller than in traditional interfaces that do not consider imprecision. Nevertheless, the time required for the contributor to be imprecise and give his certainty (Task 3) is longer than for precise answers without certainty requested (Task 4). Yet time is another significant variable in the cost of the campaign to the employer, so



(a) Correct answer rate for MV et belief functions for task 3 and 4 ($\delta = 0.3$)

(b) Crowdsourcing campaign cost for different crowd sizes

Figure 9: Comparison of correct answer rates ($\delta = 0.3$) and campaign costs according to crowd size

we will now compare the campaign costs for Tasks 3 and 4. Figure 9a compares the average correct response rates for $\delta = 0.3$ for the contributions of task 3 modeled by the theory of belief functions and those of task 4 for which the MV is used. The method followed to obtain this figure is the same as for figure 8. A set of contributors is chosen 50 times and the average of the correct answer rates is performed. For task 4, the crowd is smaller and composed of 25 contributors. The size of the crowd therefore varies between 2 and 25 contributors for the graphs in Figure 9. Figure 9b shows the costs of the crowdsourcing campaigns associated with tasks 3 and 4 as a function of the number of contributors that make up the crowd. Task 3 lasts 16 minutes and costs the employer €5.18 per contributor compared to €3.88 for 12 minutes for task 4. Task 4 is much faster for the contributor and less costly for the employer for the same number of contributors compared to task 3. However, the aggregation of the data from task 3 makes it possible to obtain a better rate of correct answers more quickly than task 4 and this with fewer co-contributors. The contributor therefore takes more time to answer the questions with the possibility of being imprecise while giving his certainty, contrary to the traditional interfaces without imprecision and certainty. On the other hand, thanks to the modeling of the answers by the theory of belief functions, the employer can call upon a smaller crowd for an identical or even better correct answer rate, which reduces the cost of the campaign while improving the user experience.

4 DISCUSSION AND CONCLUSION

In this paper we investigate how contributors may leverage a more expressive user interface for crowdsourcing. More precisely, we conducted a study relying on four tasks: one where contributors were asked to report a self-esteemed degree of certainty in their answer, one where they could provide more or less precision by selected several of the suggested options when in doubt, one where they both had to report their certainty and can provide answers of various precision, the last one where they have to provide a precise answer and cannot report their certainty.

The tests carried out in this paper were conducted on the Crowdpanel platform and leveraged a task based on visual perception in order to know the ground truth and control the level of difficulty of the task, independently of participants’ prior knowledge. The analysis of the results obtained provides several valuable insights.

First, it confirms that contributors degree of certainty in their answer increases when the task gets easier.

Second, we observe that the contributors make good use of the possibility to be imprecise, and that imprecision increases with the difficulty of the task as contributors do not hesitate to select several options when the task is of high difficult. They, however, rarely select all possible answers even when the task is impossible to answer. It is also worth noting that this use imprecision allows contributors to be more certain about their answers. Indeed, a more traditional interface in crowdsourcing platforms requires the contributors to be precise in their answer, forcing them sometimes to make a risky choice if they have a doubt between several answers, or even to choose randomly if they are completely unaware of the answer. These answers given under constrain diminish the contributors’ certainty in their work without being able to point this out to the employer. Thanks to the interface that we propose when it is imprecise, the contributors are thus more certain of their contribution than if they had to make a choice.

Then, we used the theory of belief functions to model the uncertainty and imprecision using our dataset. This allowed us to consider imprecision and certainty in provided answers, which allowed us to obtain a better good answer rate with the contributions during aggregation than the MV traditionally used in crowdsourcing platforms. This confirms the benefits of giving the opportunity to contributors to self-assess their certainty and provide imprecise answers, which could be a real advantage for employers because, thanks to it and the use of the theory of belief functions for the aggregation of the collected data, the crowd required to perform the task is smaller than for a traditional interface. In this way, it is possible to obtain the information faster and from fewer contributors compared to with more traditional interfaces that do not provide these possibilities.

Compared to [10, 12, 15] we do not introduce imprecision and uncertainty in an educational setting but in closed questionnaires used in the field of surveys and more precisely on crowdsourcing platforms. Considering the results of our studies, we recommend giving to the contributor the possibility to be imprecise and uncertain in survey questionnaires.

REFERENCES

[1] Lina Abassi and Imen Boukhris. 2019. A Worker Clustering-Based Approach of Label Aggregation under the Belief Function Theory. *Applied Intelligence* 49, 1 (Jan. 2019), 53–62. https://doi.org/10.1007/s10489-018-1209-z

[2] Yael Amsterdamer, Yael Grossman, Tova Milo, and Pierre Senellart. 2013. Crowd Mining. In *Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data* (New York, New York, USA) *(SIGMOD '13)*. Association for Computing Machinery, New York, NY, USA, 241–252. https://doi.org/10.1145/2463676.2465318

[3] AMT [n.d.]. Amazon Mechanical Turk. https://docs.aws.amazon.com/AWSMechTurk/latest/RequesterUI/CreatingYourBatchofHITs.html Accessed: 17/12/2020.

[4] Yoram Bachrach, Thore Graepel, Gjergji Kasneci, Michal Kosinski, and Jurgen Van Gael. 2012. Crowd IQ: Aggregating Opinions to Boost Performance. In *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems - Volume 1* (Valencia, Spain) *(AAMAS '12)*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 535–542.

[5] Rubi Boim, Ohad Greenshpan, Tova Milo, Slava Novgorodov, Neoklis Polyzotis, and Wang-Chiew Tan. 2012. Asking the right questions in crowd data sourcing. In *2012 IEEE 28th International Conference on Data Engineering*. IEEE, 1261–1264.

[6] Géry Casiez and Nicolas Roussel. 2011. No More Bricolage!: Methods and Tools to Characterize, Replicate and Compare Pointing Transfer Functions. In *Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology* (Santa Barbara, California, USA) *(UIST '11)*. ACM, New York, NY, USA, 603–614. https://doi.org/10.1145/2047196.2047276

[7] Alexander Philip Dawid and Allan M Skene. 1979. Maximum likelihood estimation of observer error-rates using the EM algorithm. *Applied statistics* (1979), 20–28.

[8] Arthur P. Dempster. 2008. *Upper and Lower Probabilities Induced by a Multivalued Mapping*. Springer Berlin Heidelberg, Berlin, Heidelberg, 57–72. https://doi.org/10.1007/978-3-540-44792-4_3

[9] Arthur P Dempster, Nan M Laird, and Donald B Rubin. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the royal statistical society. Series B (methodological)* (1977), 1–38.

[10] Javier Diaz, Maria Rifqi, Bernadette Bouchon-Meunier, Sandra Jhean-Larose, and Guy Denhiére. 2008. Imperfect Answers in Multiple Choice Questionnaires. In *Times of Convergence. Technologies Across Learning Contexts*, Pierre Dillenbourg and Marcus Specht (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 144–154.

[11] Steven Dow, Anand Kulkarni, Scott Klemmer, and Björn Hartmann. 2012. Shepherding the Crowd Yields Better Work. In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work* (Seattle, Washington, USA) *(CSCW '12)*. Association for Computing Machinery, New York, NY, USA, 1013–1022. https://doi.org/10.1145/2145204.2145355

[12] Graham Farrell. 2006. A comparison of an innovative web-based assessment tool utilizing confidence measurement to the traditional multiple choice, short answer and problem solving questions. https://hdl.handle.net/2134/4587

[13] Olusegun Folorunso and Olusegun Afeez Mustapha. 2015. A fuzzy expert system to Trust-Based Access Control in crowdsourcing environments. *Applied Computing and Informatics* 11, 2 (2015), 116 – 129. https://doi.org/10.1016/j.aci.2014.07.001

[14] G. Kazai, J. Kamps, and N. Milic-Frayling. 2013. An analysis of human factors and label accuracy in crowdsourcing relevance judgments. *Inf Retrieval* 16 (2013), 138–178. https://doi.org/10.1007/s10791-012-9205-0

[15] K.S. Khan, D.A. Davies, and J.K. Gupta. 2001. Formative self-assessment using multiple true-false questions on the Internet: feedback according to confidence about correct knowledge. *Med Teach.* 23, 2 (2001), 158–163. https://doi.org/10.1080/01421590031075

[16] Aniket Kittur, Jeffrey V. Nickerson, Michael Bernstein, Elizabeth Gerber, Aaron Shaw, John Zimmerman, Matt Lease, and John Horton. 2013. The Future of Crowd Work. In *Proceedings of the 2013 Conference on Computer Supported Cooperative Work*. Association for Computing Machinery, New York, NY, USA, 1301–1318. https://doi.org/10.1145/2441776.2441923

[17] Dalila Koulougli, Allel HadjAli, and Idir Rassoul. 2016. Handling query answering in crowdsourcing systems: A belief function-based approach. In *2016 Annual Conference of the North American Fuzzy Information Processing Society, NAFIPS 2016*. IEEE, El Paso, TX, USA, 1–6. https://doi.org/10.1109/NAFIPS.2016.7851590

[18] Adam Marcus and Aditya Parameswaran. 2015. Crowdsourced data management: Industry and academic perspectives. *Foundations and Trends in Databases* 6, 1-2 (2015), 1–161.

[19] Arnaud Martin. 2019. Conflict Management in Information Fusion with Belief Functions. In *Information quality in information fusion and decision making*. Springer, 79–97.

[20] Amal Ben Rjab; Mouloud Kharoune; Zoltan Miklos; Arnaud Martin. 2016. Characterization of experts in crowdsourcing platforms. *Belief Functions: Theory and Applications.* 9861 (2016).

[21] A. T. Nguyen, Byron C. Wallace, and Matthew Lease. 2015. Combining Crowd and Expert Labels Using Decision Theoretic Active Learning. In *Proceedings of the Third AAAI Conference on Human Computation and Crowdsourcing, HCOMP 2015*. https://www.aaai.org/ocs/index.php/HCOMP/HCOMP15/paper/view/11567

[22] Bahareh Rahmanian and Joseph G. Davis. 2014. User Interface Design for Crowdsourcing Systems. In *Proceedings of the 2014 International Working Conference on Advanced Visual Interfaces* (Como, Italy) *(AVI '14)*. Association for Computing Machinery, New York, NY, USA, 405–408. https://doi.org/10.1145/2598153.2602248

[23] Vikas C Raykar and Shipeng Yu. 2012. Annotation models for crowdsourced ordinal data. *Journal of Machine Learning Research* 13 (2012).

[24] Glenn Shafer. 1976. *A mathematical theory of evidence*. Vol. 42. Princeton university press.

[25] Aaron D. Shaw, John J. Horton, and Daniel L. Chen. 2011. Designing Incentives for Inexpert Human Raters. In *Proceedings of the ACM 2011 Conference on Computer Supported Cooperative Work* (Hangzhou, China) *(CSCW '11)*. Association for Computing Machinery, New York, NY, USA, 275–284. https://doi.org/10.1145/1958824.1958865

[26] C. Thierry, J.-C. Dubois, Y. Le Gall, and A. Martin. 2019. Modeling Uncertainty and Inaccuracy on Data from Crowdsourcing Platforms: MONITOR. In *2019 IEEE 31st International Conference on Tools with Artificial Intelligence (ICTAI)*. 776–783. https://doi.org/10.1109/ICTAI.2019.00112

[27] Peter Welinder, Steve Branson, Serge Belongie, and Pietro Perona. 2010. The Multidimensional Wisdom of Crowds. In *Proceedings of the 23rd International Conference on Neural Information Processing Systems - Volume 2* (Vancouver, British Columbia, Canada) *(NIPS'10)*. Curran Associates Inc., Red Hook, NY, USA, 2424–2432. https://doi.org/10.5555/2997046.2997166

[28] Jacob Whitehill, Ting-fan Wu, Jacob Bergsma, Javier Movellan, and Paul Ruvolo. 2009. Whose Vote Should Count More: Optimal Integration of Labels from Labelers of Unknown Expertise. In *Advances in Neural Information Processing Systems*, Y. Bengio, D. Schuurmans, J. Lafferty, C. Williams, and A. Culotta (Eds.), Vol. 22. Curran Associates, Inc., 2035–2043. https://proceedings.neurips.cc/paper/2009/file/f899139df5e1059396431415e770c6dd-Paper.pdf

5 ACKNOWLEDGMENTS

This work is supported by the Conseil Départemental des Côtes-d'Armor and PEPS funding from the CNRS.