Wisdom of Micro-Crowds in Evaluating Solutions to Esoteric Engineering Problems

Nurcan Gecer Ulu, Michael Messersmith, Kosa Goucher-Lambert, Jonathan Cagan, Levent Burak Kara
Department of Mechanical Engineering
Carnegie Mellon University
Pittsburgh, PA 15213
Email: lkara@cmu.edu

The wisdom of crowd effect has been studied in political and economic forecasting where the crowds can be remarkably accurate in estimating true answers. In this study, we investigate the wisdom of crowd in esoteric engineering problems. We used various statistical techniques ranging from averaging to more sophisticated Bayesian networks for aggregating the crowd answers. Our results suggest that the wisdom of crowd effect is valid for these esoteric engineering problems in practitioner crowds, where people have exposure or familiarity to the domain of the esoteric problem, but are not necessarily domain experts. On the other hand, we did not obtain any accurate estimation from diverse groups of ordinary people gathered through Amazon Mechanical Turk. Since practitioners are less prevalent than ordinary crowds, we investigate wisdom of micro-crowds. For micro-crowds of practitioners, consisting of 4-15 practitioners, we were able to produce crowd estimates that were more accurate than the individual estimates of the majority of the people in the micro-crowd. We also observed that the wisdom of crowd effect is maintained for less intuitive problems in addition to problems that require more intuitive evaluations. This indicates that challenging engineering problems that lack effective computational solutions can vastly benefit from the wisdom of crowd effect in practitioner groups combining different perspectives of individuals

1 Introduction

Wisdom of crowds is the idea that the collective estimate of a group can be more accurate than estimates of individuals, even those who are experts in the domain of the problem. Demonstration of wisdom of crowds dates back to early 20th century [1, 2]. Recently, this phenomenon has been shown to be useful in many fields such as popular culture, social studies, behavioral economics, psychology and politics [3]. This paper investigates how the wisdom of crowd effect performs in micro-crowds containing a small number of individuals. For brevity, we refer to the wisdom of micro-crowds as WoMC.

Crowdsourcing is a process that can engage the wisdom of crowds. By definition, crowdsourcing is the assignment of a task to a large population of workers [4]. With the rise of crowdsourcing platforms such as Amazon Mechanical Turk (AMT), harnessing human resources have provided ideal solutions to high-speed and low cost data generation for many problems, especially machine learning. Traditional crowdsourcing focuses on assigning tasks that are human easy and computer hard. One popular example is computer vision applications where crowds may be given a set of images and asked to identify certain objects [5]. Such a question is extremely intuitive for all people, and so the correct answer can be inferred from a crowd consensus. However, complications arise when the question presented requires expertise in specific domain [6]. In the context of this paper, we define such questions as esoteric problems.

This paper focuses on the esoteric domain of engineering problems. While engineering tasks are greatly facilitated by computers, they are limited by the computational tools available. Technology is improving everyday and we are faced with new challenges where no established solutions exist. The emergence of crowdsourcing offers a potential new method to reduce the cost of engineering design, development and evaluation through the application of the wisdom of crowds. In addition, the wisdom of crowds may provide a useful new tool for approaching these challenging new problems that lack ground truth answers.

One recent study asked the AMT population to evaluate the structural strength of various bracket designs [7]. The paper concluded that crowdsourcing evaluation fails for engineering tasks, in part due to the limited population of experts in the sampled crowd. This suggests that it may be impractical to rely on a sufficient number of experts being present in the sampled population in order to derive an accurate consensus.

On the other hand, collaboration of experts have been studied in [8–12]. Lorenz et al. [13] shows how social influence during the interactions of a group can weaken the wisdom of crowd effect in general knowledge questions. Moreover, Hong et al. [14] theoretically and computationally demonstrate specialists must become similar and can be

outperformed by diverse groups. These studies motivate and leave open an important question: can we use crowdsourcing in esoteric domain of engineering design evaluation to achieve wisdom of crowds? In addition to non-experts and experts, which have been the primary focus of other research on crowdsourcing in engineering, we introduce a third category of individuals known as practitioners. We define a practitioner as having exposure or familiarity to the domain of the esoteric problem in question.

In this work, we investigate the WoMC concept for esoteric engineering problems. To gather empirical data, we have conducted multiple crowdsourcing experiments with different difficulty and intuitiveness levels on crowds with ranging expertise. To analyze this empirical data using a computational framework, we developed a Bayesian model and benchmarked it against other statistical aggregation methods such as arithmetic mean. We further discuss the WoMC concept through investigating the effect of crowd size as well as its extension to conceptual design problems. With an understanding of these relationships, we aim to provide an explanation of how wisdom of crowd can be achieved in engineering tasks.

This paper demonstrates that the wisdom of crowd can be effective in practitioner groups, with the aggregated crowd result outperforming a majority of the individual practitioners sampled. For micro-crowds of practitioners, consisting of 4-15 practitioners, we were able to produce crowd estimates that were more accurate than the individual estimates of the majority of the people in the micro-crowd. These promising results suggest that the WoMC may provide a powerful tool for answering difficult problems in which computational methods have not been established and ground truth answers do not exist yet. In addition, these results support the establishment of online communities for practitioners, which could facilitate future applications of the wisdom of crowds in esoteric domains such as engineering design.

2 Background

Our work builds on crowdsourcing with a focus on esoteric engineering design evaluation in practitioner micro-crowds. Below, we cover crowdsourcing platforms, some recent developments in crowdsourcing, consensus through collaboration and use of crowdsourcing in engineering.

A critical component in crowdsourcing is access to crowds. Since we investigate the effect of expertise level in a crowd, choice of crowdsourcing platform is a key to reach the desired workers. The most prominent crowdsourcing platform is Amazon Mechanical Turk (AMT) [15], with more recent emergences of CrowdFlower (CF) [16] and Prolific Academic (ProA) [17]. While these platforms are designed for diverse crowds of general public, each platform attracts different workers [18]. A study on these found CF and ProA to have more diverse populations than AMT and that ProA developed as a platform for conducting research [19].

AMT allows the surveys to be targeted to selected demographics. Yet, esoteric engineering problems require a more specific group that we define as practitioners and there are no platforms to reach such a crowd. Our practitioner crowds are gathered through the university community. We believe our work demonstrates the need for new platforms that enable the access to practitioner crowds for esoteric problems.

With the rise of crowdsourcing platforms, now there is an easy access to a large population of potential workers that can participate in design evaluations quickly [20]. These developments lead to many research in how to create successful crowdsourcing studies [21]. Some resulting techniques include use of explicitly verifiable questions to identify malicious users and encourage honest responses as well as task fingerprinting that logs completion time, mouse movements, key presses, and scroll movements, which can be used as metrics to vet suspect responses [22]. Here, we focus on use of crowdsourcing in esoteric domains rather than general knowledge or preference type studies.

Wikipedia has been intensely studied as a prominent success story of the application of wisdom of crowds. It has been shown that collaboration between thousands of users have developed an encyclopedia of comparable accuracy to encyclopedias written by a collection of experts [23]. Special topics in Wikipedia can be considered as esoteric domains. However, due to the direct collaboration between editors on Wikipedia, the fundamental assumption of independent evaluations for wisdom of crowds is violated. Further research found that the core of Wikipedia's articles were developed by a small elite, and that more diverse populations began contributing later [24].

While use of collaboration is a common practice in engineering [10, 11], we are interested in WoMC that can be generated through independent diverse crowds. Surowiecki [3] states that one requirement for a good crowd judgement is that peoples decisions are independent of one another, which is further validated by the results from [13] which demonstrate that the same crowd of individuals produce a more accurate crowd consensus when there is no collaboration during the process.

Crowdsourcing has also been used in engineering design for incorporating customer preferences. Crowdsourcing is a powerful tool for collecting customer input on their preferences for a product, particularly in regard to balancing style with brand recognition [25]. Analyzing customer preferences for different fundamental geometric designs of products can aid engineers in evoking the desired response in customers [26]. In contrast to such research into crowdsourcing subjective engineering problems, this paper focuses on esoteric engineering problems with objective attributes.

Use of crowdsourcing in design evaluation have been studied in [7]. This paper calls for more research into the subject motivates our work to investigate wisdom of crowd in design evaluations to find when and how it works. Identifying experts in a crowd through demographics and testing mechanical reasoning with check questions has been studied in [27]. While these works focus on solving engineering problems with ordinary people, we investigate a new crowd concept, practitioner micro-crowds. We demonstrate collective estimates of practitioner crowds can be accurate although the practitioners may have significant estimation er-

rors individually.

Another use of crowdsourcing in engineering is to collect design proposals. One of the notable examples of this was GE Aviations use of a GrabCAD design challenge, in which they tasked users with designing a titanium 3D printed bracket [28]. Open call challenges such as these can attract too many entries than can be easily managed by experts. In a way, our work is complimentary to these challenges as it paves the way to use crowdsourcing for evaluation of the crowdsourced design proposals.

Another open question within the engineering design research community relates to the consistent evaluation of conceptual designs. In contrast to esoteric engineering problems, conceptual design problems have no *true* solution. Researchers often utilize conceptual design studies to explore characteristics of the design process, such as the impact of analogical stimuli on solution characteristics [29–31]. Typically, design output from such studies is evaluated qualitatively; trained experts rate defined metrics, such as the novelty or quality, across a wide design space [32]. Unsurprisingly, the process of both training and rating design solutions can be incredibly time consuming and costly. This is particularly true for cognitive studies requiring hundreds of design concepts to be evaluated at a given time. Another challenge with the current approach to evaluating conceptual design solutions is that when multiple experts are used, they do not always agree upon the particular merits of a given design concept. This can lead to low inter-rater reliability metrics, and require researchers to retrain experts prior to having them re-evaluate designs. With this in mind, a combined human-computational framework that removes the necessity of training experts could greatly improve and expedite the conceptual design evaluation process. In this work, we also explore WoMC for evaluation of conceptual designs.

3 Experimental Design

In order to investigate the wisdom of crowd in esoteric engineering applications, it is necessary to understand its relationship between crowds and problem types. Here, we explain the characteristics of crowds and the problems used in this study.

3.1 Crowd Types

Two important key factors in the wisdom of crowds are diversity of opinion and independence. Therefore, a crowd should include people with a variety of opinions rather than a group of elites or experts that may create bubbles and conform to each other's opinions [3]. To support independence, we collected survey results through a web-based survey tool providing anonymity and avoiding communication between participants.

In this work, we consider two types of crowds as AMT workers and practitioners. AMT crowds represent the opinion of ordinary people which may include anyone on any skill level. On the other hand, practitioner group represents the people who have knowledge in the discipline however

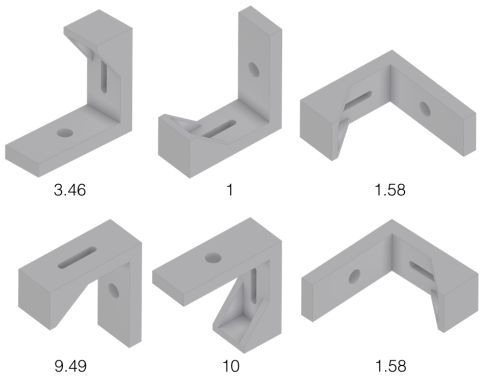

Fig. 1. 3D printing-1: support material question.

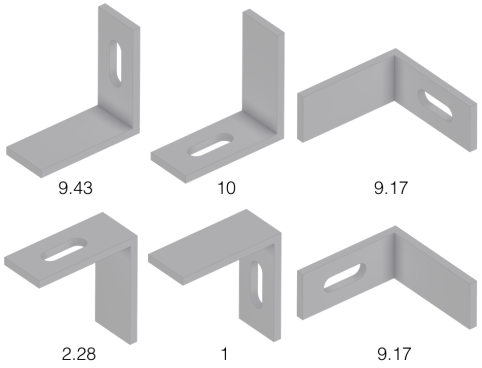

Fig. 2. 3D printing-2: surface finish question.

is not necessarily a domain expert for a given task. For example, practitioner can represents a person that has an understanding of mechanical engineering discipline in general but does not specialize in the field of given task such as heat transfer, structural mechanics, manufacturing etc.

For the practitioner group study, a total of 15 participants were recruited from a sample of more than 300 graduate students at Mechanical Engineering Department of Carnegie Mellon University. Note that these students do not represent experts and estimation error in their answers vary significantly (Fig. 8). For AMT surveys, we gathered groups of 100 people through Amazon Mechanical Turk who recieved monetary compensation. In order to represent general public, we did not set any specific demographic groups. Only for the AMT survey of the structural mechanics question, we used the data provided by Burnap et al. [7].

3.2 Design of Surveys

In this study, we investigated the wisdom of crowd with four different surveys that range in the difficulty and intuitiveness. While all surveys require familiarity with engineering problem at hand, 3D printing questions (Fig. 1, 2, 3) focus on more intuitive concepts such as area/volume evaluations. On the other hand, the structural mechanics problem (Fig. 4) is less intuitive since estimating strength of an arbitrary topology provides a more challenging task even for experts.

Although engineering problems are often *computer*

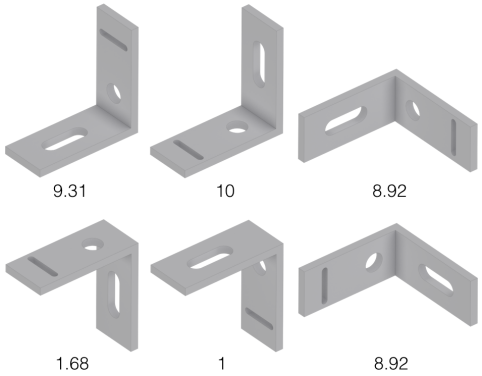

Fig. 3. 3D printing-3: surface finish question.

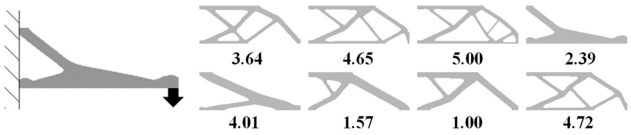

Fig. 4. The structural mechanics problem [7].

easy, human hard, they are solved using expert intuition when no computational tools are available. To assess whether such situations could benefit from wisdom of crowd, we conducted a series of surveys for problems we already know how to solve so that we can evaluate the crowd answers. Structural mechanics survey (Fig. 4) provides a very good example of such problems since these optimal mechanical design problems have been solved by expert intuition until the introduction of topology optimization techniques in 1990s [33]. Therefore, we now have the tools to evaluate the aggregated crowd answer and how well it performs. As a result, our study on this survey demonstrates why computer easy world of engineering can benefit from the wisdom of crowds.

We developed survey questions such that participants can produce a reasonable estimation even though they may not have the knowledge of exact answers. We also used a rating based approach on a predefined scale to eliminate problems in choosing the right order of magnitude. Each survey consist of multiple questions to facilitate expertise inference later in crowd aggregation stage. In all of the surveys, participants are presented with the question statement and all alternative shapes to be analyzed altogether. Then, they are asked to give their ratings for each individual question consecutively. Next paragraphs explain the details of each survey.

Figure 1 presents 3D printing-1 survey where participants are asked to rate the amount of support material required to print the same object in different orientations with an FDM type printer. For each of the given orientations, participants are required to evaluate the amount of support material needed on a scale from 1 to 10, 1 being very little and 10 being a lot of support material. We compute the required support material as the volume that is created by the projection of overhang areas to the base with zero overhang angle. Then, the scores are scaled linearly between 1 and 10.

3D printing-2 survey is about evaluating the surface finish quality of the same object in particular orientations (Fig 2). The participants are asked to rate the quality of the printed object considering the amount of surfaces in contact with support material for each given orientation. Surface quality rating is between 1 and 10, 1 being very poor finish and 10 being very good finish. To find the true surface finish ratings, we compute the overhangs areas with zero overhang angle and scale the areas inversely between 1 and 10. 3D printing-3 survey (Fig. 3) asks the same question on the same objects with more features that increase the difficulty of evaluation.

Figure 4 shows the structural mechanics problem that is presented in [7]. In this survey, participants are presented with eight different bracket designs intended to support a downward force at the end of the bracket. Then, they are asked to rate the strength of each bracket on a scale from 1 to 5. On this scale, 1 corresponds to a very weak design whereas 5 is very strong. One particular reason why we use the structural mechanics survey is that estimating strength of arbitrary topologies is less intuitive to humans compared to volume/area evaluations. While humans are exposed to volume/area computations in daily life, rating strength of an arbitrary design requires a specific experience [34].

4 Crowd estimate aggregation

Choice of aggregation measure effects the collective estimate of the group thereby understanding of wisdom of crowd. In this section, we explain the statistical methods we used to aggregate the crowd answer.

In a crowd of n participants with a set of estimates $y_1, y_1, ..., y_n$, we first compute the crowd estimate using arithmetic mean as $y^{agg} = \frac{1}{n}\sum_{j=1}^{n} y_i$. Some studies discuss that the median or geometric mean can result in more accurate estimates to demonstrate the wisdom of crowd [1, 13]. Geometric mean of an estimate set can be calculated as $\exp(\frac{1}{n}\sum_{j=1}^{n}\ln(y_i))$. Another approach is to use majority voting which is selecting the score that has the most repetitions in the data. Our approach is suitable for this type of aggregation since we have discrete set of ratings for our evaluation questions. The majority vote is found as the mode of the estimate set.

Bayesian networks have been implemented in a variety of crowdsourcing applications in order to mitigate the noise from biased responses. These studies model the sources of bias using models that consider problem difficulty and competence of participants [5–7, 35–38]. Similar to these approaches, we develop a Bayesian model as presented in Fig. 5. Note that the Bayesian networks approach does not require prior knowledge of true answers, participant expertise or problem difficulty. The only observed variable is the participant answer for each question.

Participants with high expertise provide accurate answers with very small errors while non-experts can give answers with large errors. There is another level on the skill spectrum that corresponds to adversarial participants who in-

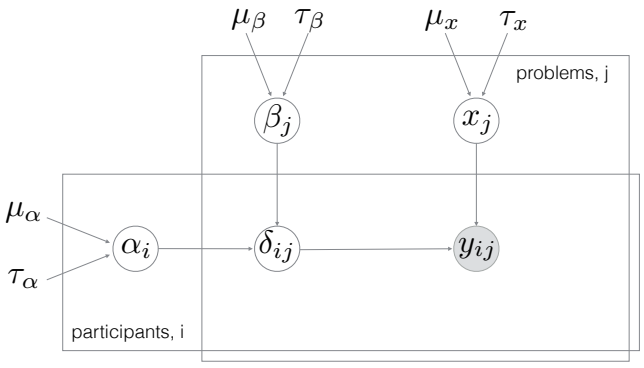

Fig. 5. The Bayesian network model.

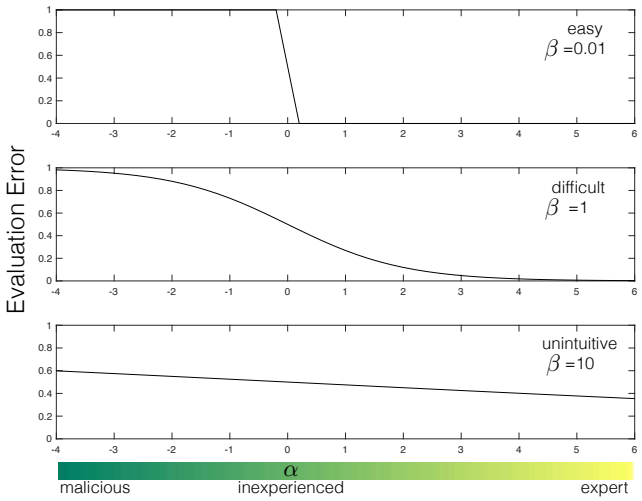

Fig. 6. The error variance with participant expertise and problem difficulty.

tentionally give the wrong answers. Since the answers are maliciously wrong, the amount of error is even more than that of a non-expert that randomly guesses the answers. Figure 6 explains the effect of question difficulty with the varying participant level. For a very easy question, even unskilled participants can give answers with small error and anyone adversarial can make the most damage. As the questions get difficult, expertise affect the accuracy of answers more. On the other hand, unintuitive questions are answered through random guesses by participant at all skill levels resulting in similar error values for all. We model this rating process as follows:

$$\delta_{ij} = \frac{exp(-\alpha_i/\beta_j)}{1+exp(-\alpha_i/\beta_j)} \tag{1}$$

where the participant expertise is modeled by the parameter $\alpha_i \in (-inf, +\inf)$ and the problem difficulty is $\beta_j \in (0, +\inf)$. The resulting evaluation error becomes $\delta_{ij} \in [0,1]$. We model the evaluation process as a random variable with a truncated Gaussian distribution around the true score ($\mu = x_j$) with a variance as evaluation error, δ_{ij}. To bring everything into the same scale, evaluations, y_{ij}, are scaled to $[0,1]$ from the original survey scale. The true scores are also represented as $x_j \in [0,1]$.

Considering all of our assumptions, we obtain the graphical model as shown in 5. Assuming a Bayesian treatment with priors on all parameters, , the joint probability distribution can be written as

$$\begin{aligned} p(\mathbf{y},\mathbf{x},\delta,\alpha,\beta) = &\prod_i p(\alpha_i)\prod_j p(\beta_j)p(x_j) \\ &\prod_{ij} p(y_{ij}|\delta_{ij},x_j)p(\delta_{ij}|\alpha_i,\beta_j) \end{aligned} \tag{2}$$

Note that we exclude hyper-parameters for brevity. In our implementation, we use Gaussian priors for α with mean, $\mu_\alpha = 1$, and precision, $\tau_{alpha} = 1$. Since the value of β needs to be positive, we impose truncated Gaussian prior with mean, $\mu_\beta = 1$, and precision, $\tau_{beta} = 1$, with a lower bound as $+\varepsilon$. For the true scores, x_j, we use a truncated Gaussian with bounds $[0,1]$, mean $\mu_x = 0.5$ and precision $\tau_x = 0.1$.

We use Markov chain Monte Carlo (MCMC) simulations to infer the results. In the MCMC simulations, we utilize a Metropolis step method. Empirically, we observe that using thinning interval of 3 and burn-in length of 10^5 works well with $5x10^5$ iterations.

5 Results

To demonstrate the WoMC in esoteric engineeing problems, we have conducted four experiments with two crowds having different skill levels. Here, we analyze the results of surveys and investigate when and how we observe WoMC.

Summary of surveys outcomes. The results of surveys with different crowds and aggregation methods are summarized in Table 1. We also provide the RMS error values for each question in the surveys (Fig. 7). We scale all scores between 0 and 1 for direct comparison across surveys. While for a single question, the collective error can be defined as the difference between the true answer and the aggregated answer, $(y^t - y^{agg})$, we compute the root mean square (RMS) error for a survey containing m questions, $\sqrt{\frac{1}{m}\sum_{j=1}^{m}(y_j^t - y_j^{agg})^2}$ since it gives a performance measure in the same scale. Note that answers of participants are gathered as discrete ratings rather than continuous variables. While arithmetic mean, geometric mean and Bayesian networks aggregate a continuous value for the discrete inputs, median and majority voting results are discrete. For this reason, we used the true continuous answers for the methods with continuous aggregates and we rounded the scores to compare with discrete aggregates.

Crowd expertise and aggregation methods. We did not observe any accurate estimations in AMT groups. The resulting RMS error values are very high in the context of the data being scaled from 0 to 1. On the other hand, re-

Table 1. The wisdom of crowd effect exist in engineering problems with expert groups and Bayesian model gives the best estimate in most cases. While AMT groups result in high errors that suggest poor accuracy, no statistical aggregation method consistently performs better.

	RMS error in crowd estimation				
Question	Arithmetic mean	Geometric mean	Median	Majority voting	Bayesian model
3D printing-1, practitioner	0.111	0.091	0.136	0.079	0.055
3D printing-1, AMT	0.403	0.378	0.430	0.336	0.363
3D printing-2, practitioner	0.202	0.236	0.197	0.163	0.113
3D printing-2, AMT	0.438	0.462	0.473	0.540	0.600
3D printing-3, practitioner	0.196	0.198	0.136	0.111	0.116
3D printing-3, AMT	0.402	0.431	0.363	0.453	0.561
Structural Mech., practitioner	0.197	0.217	0.198	0.342	0.173
Structural Mech., AMT	0.339	0.352	0.385	0.395	0.392

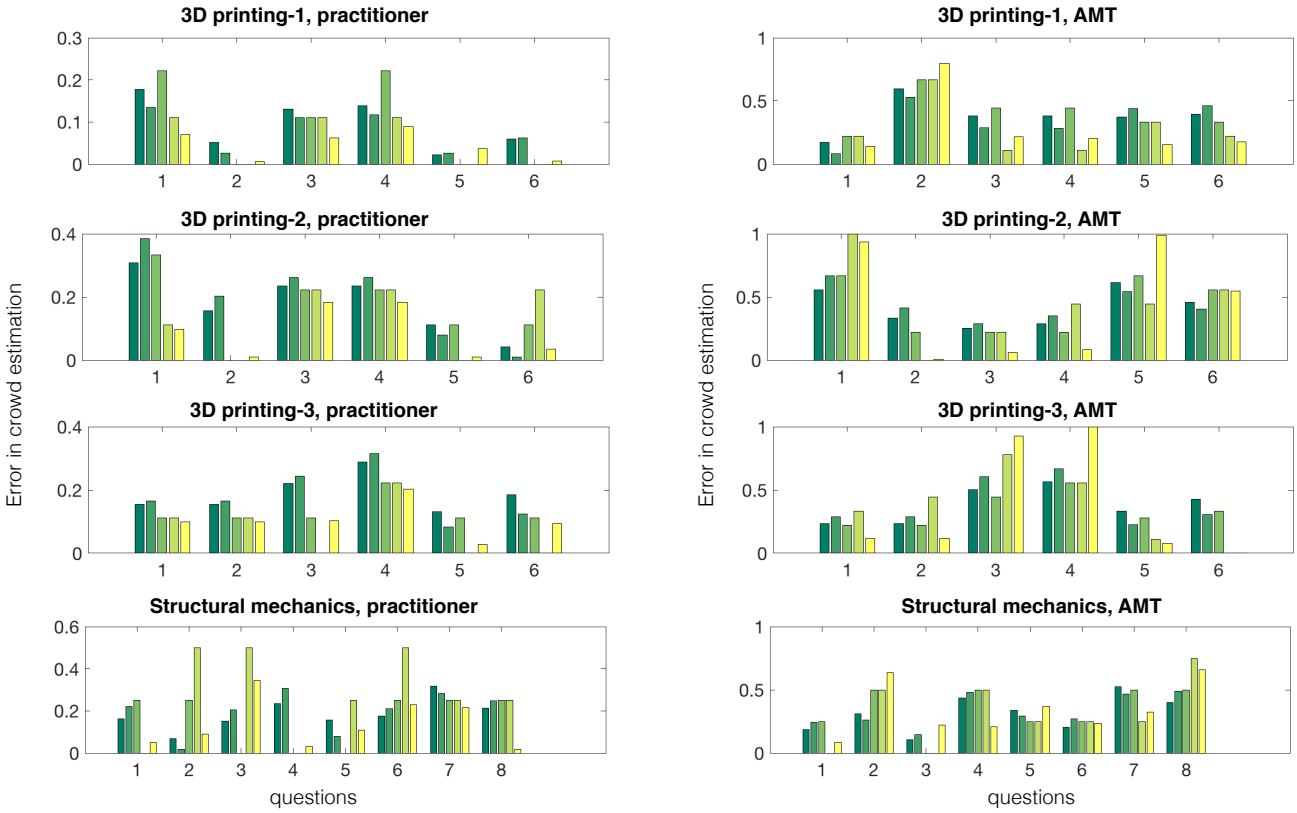

Fig. 7. Error of crowd estimation for each question in the four survey groups. Each bar group represents the error of the crowd estimation aggregated through arithmetic mean, geometric mean, median, majority voting and Bayesian model, respectively.

sults of practitioner crowd studies suggest that crowdsourcing can be useful for esoteric problems. When the aggregation methods are compared, our results demonstrate that Bayesian networks can provide consistently good estimations in practitioner groups. However, Bayesian networks are outperformed by other methods in all of our AMT studies. This matches previous findings on how crowdsourcing with AMT populations fails for engineering design evaluations [7].

While arithmetic mean can give better estimation results in AMT crowds, it results in almost double the RMS of the best Bayesian network results in practitioner groups. This can be explained with the data distributions. While estimates of AMT crowds are closer to normal distribution, distribution of practitioner crowd estimates are skewed. Due to this property, median or majority voting can outperform arith-

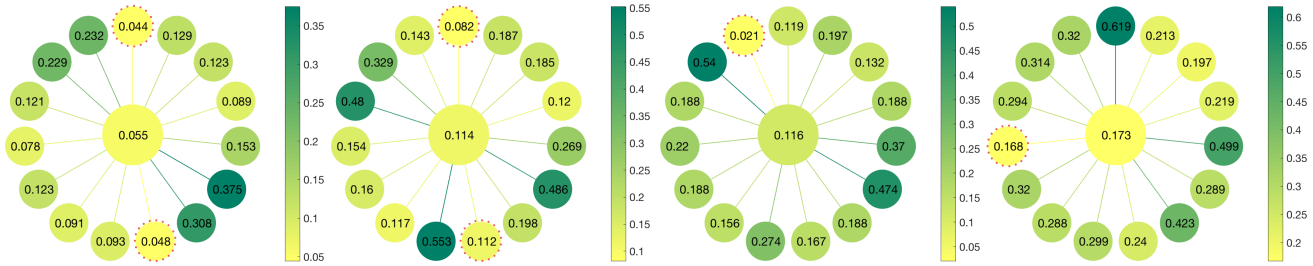

Fig. 8. Estimation error significantly varies in the practitioner group. Collective estimate of the practitioner crowd is more accurate than vast majority of individual practitioners. Collective error of the crowd and errors of individual practitioners in the crowd are given in the center node and surrounding nodes,respectively. The color of the circles represents the error and better performing individuals are marked with a dashed circle. The results for 3d printing-1,2,3 and structural mechanics questions are given from left to right.

Table 2. Percentile rank of crowd estimation in individual estimations for the practitioner crowd.

	Percentile rank of crowd estimation	
Question	Continuous	Discrete
3D printing-1	87%	100%
3D printing-2	87%	93%
3D printing-3	93%	93%
Structural Mech.	93%	100%

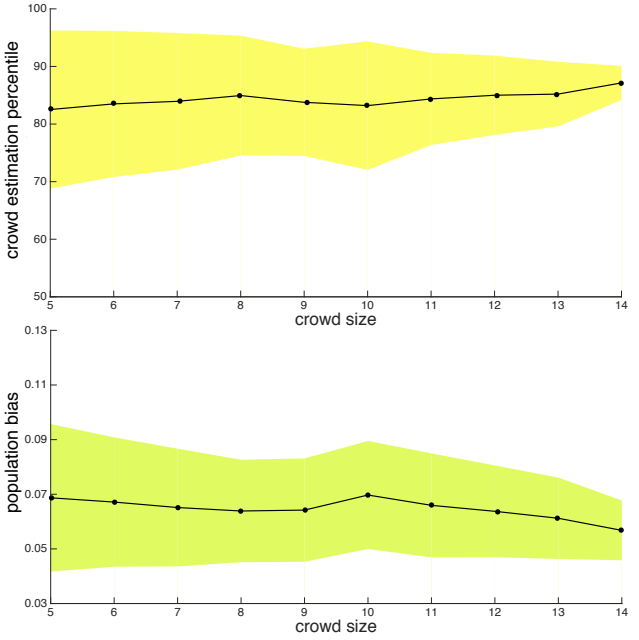

Fig. 9. Effect of crowd size on the success of crowd answer success as percentile and population bias. We observe a slightly increasing trend in the percentiles whereas a significant decrease in the standard deviation (yellow shaded) as the crowd size increases. This suggests that higher percentile ranks can be achieved with higher probability in larger crowds.

metic averaging in practitioner groups. However we also observed majority voting to result in very poor performance in the structural mechanics survey with practitioner group. In any of our studies, we have not observed the best performance with geometric mean. This can be explained with the fact that we ask our questions as ratings in specified ranges whereas geometric mean is useful when data/responses vary in order of magnitude [13].

WoMC and individuals. To analyze the wisdom of crowds effect, we compare the performance of the aggregated crowd estimation with answers of individuals in the crowd (Fig. 8). Since we only observe reasonable accuracy in practitioner groups, we only present this data for these crowds. We used the collective answers aggregated with Bayesian networks as it consistently performed well in practitioner group studies.

Figure 8 shows that the collective estimation of the crowd is more accurate than most of the individuals. Note that the practitioner group is composed of individuals with different skill levels and estimation errors significantly vary in the group. This confirms that Bayesian networks can give an accurate measure of the wisdom of crowds for the problems of our study with esoteric nature. This can be explained through the participant expertise and problem difficulty based inference that considers all answers of an individuals rather than a single answer. Our results suggest that the Bayesian networks approach does not undermine the wisdom of crowd effect by picking out only elite group of experts in the group but allow the diverse perspectives to be incorporated. This can be explained by the fact that expertise concept here is not asserted but rather inferred as a latent variable in the Markov Chain Monte Carlo simulations.

Statistically, we quantify the wisdom of crowds using percentile rank that represents the part of people that are outperformed by the collective answer. High percentile values in Table 2 suggest that the wisdom of crowd is achieved. We compute the percentile rank of crowd using two error metrics as continuous and discrete. In the continuous version, we compute the distance between true answers and participant ratings. Since the participants only evaluate the design options using integer ratings, we use round the true answers to the nearest integer in the discrete measure while computing the individual estimate errors. Although it might be assumed that the discrete measure would result in less error in

individual estimates, we results in suggest that the continuous metric actually produce smaller error values. This can be explained with the fact that discrete measures can also increase the error due to round off. It is important to note the distinction between percentile rank and accuracy of the collective estimate. The percentile rank denotes the relative performance of the collective estimate compared to the individual estimates while the accuracy refers to the size of the RMS error between the estimate and ground truth value.

Effect of crowd size. Platforms such as AMT enable the collection of answers from large groups of diverse people. However, one can most likely reach only a limited number of people when practitioners are required for an esoteric problem. For this purpose, we investigate the wisdom of crowd effect with different group sizes (Fig. 9) and observed that WoMC can still be observed in smaller groups.

We analyzed crowd size with the 3D printing-1 survey and computed crowd estimation using Bayesian networks since we observe consistent good performance (Table1). Initially, our practitioner studies are conducted with 15 participants. To simulated micro-crowds with less participants, we generate combinations of 5 to 14 individuals of the 15 participant set. Since there can be many combinations for some crowd sizes, we limit the number of combinations to 500 for each group size by random selection. The results suggest that the wisdom of crowd effect can still be observed for smaller group sizes. We also observe that the probability of obtaining crowd estimations with higher success (percentile) increase with larger crowds. The estimation error of the micro-crowds and its standard deviation also decrease as the number of people increase. We also investigate the effect of population bias as the crowd size change. Population bias is defined as the error of aggragated guess accross the crowd [39]. Figure 9 demonstrates that both mean and standard deviation of population bias has a decreasing trend with the increasing crowd size.

6 Discussions

6.1 Insights for Design

From the analysis we conducted, we identified some key insights on how WoMC can be achieved in esoteric engineering problems. Here, we highlight these key points.

Problem difficulty and intuitiveness. We designed our surveys in ranging difficulty and intuitiveness. Specifically, all 3D printing questions are based on volume and area estimations whereas structural mechanics question is not as intuitive and acquainted. While we observe more accurate estimates in 3D printing questions, we do not observe a significant difference in the wisdom of crowds (*i.e.*, number of individuals outperformed by collective estimation). Moreover, we also do not observe a significant difference between the results of 3D printing-2 and 3D printing-3 surveys while we ask the same question on similar objects with increased number of features. This suggests that the wisdom of crowds work for more difficult or unintuitive questions as much as it works for easier and intuitive ones. Perhaps the different perspectives of individuals in the practitioner groups help to mitigate the impact of intuitiveness.

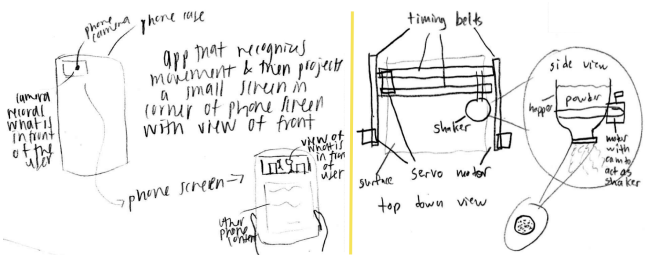

Fig. 10. Example conceptual designs.

Level of expertise. Populations of ordinary people, collected through platforms such as AMT, perform poorly on esoteric engineering problems. However, our results indicate that the wisdom of crowds can be engaged in practitioner crowds of the domain of the esoteric problem. This conclusion demonstrates that experts may not necessarily be required to solve complex engineering problems. Groups of practitioners who are still gaining experience in the domain may prove to be a valuable asset. Furthermore, practitioner crowds may be more accessible than experts.

Aggregation methods. In the context of practitioner populations, we determined that the most effective aggregation method was the Bayesian network. This technique is most valid given a minimum level of expertise of the group, since the Bayesian network was outperformed by the arithmetic mean for the AMT populations. For practitioner groups, the exposure to the domain of the esoteric problem allows the Bayesian network to mitigate the mistakes made by individual practitioners who have comparatively less experience than the other practitioners.

Size of crowd. The standard deviation of the percentile rank of the collective estimate from the crowd decreases as the crowd size increases, indicating that larger practitioner crowds will produce more accurate estimates. However, with as few as 5-14 practitioners, we produced collective estimates in the 90 percentile. This demonstrates that WoMC can still be effective in crowdsourcing applications, which is particularly important since practitioner crowds are less prevalent than diverse crowds.

6.2 Conceptual Design Evaluations

As an extension of the methods presented in this paper, the feasibility of using a Bayesian network model within the context of conceptual designs was explored. To accomplish this, a practitioner evaluation study was run in which each individual practitioner evaluated a pre-existing set of conceptual design solutions that had also previously been evaluated by two trained experts. Fifteen practitioners were recruited from Carnegie Mellon University, each specializing in Mechanical Engineering (Design focus), or Product Development. Participants were allowed a maximum of 120 minutes to complete the ratings, and were monetarily compensated for their time.

Each practitioner evaluated 114 conceptual designs, corresponding to one of four design problems. These prob-

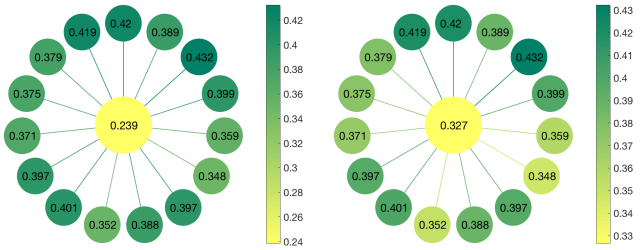

Fig. 11. For conceptual design survey, we observe significant estimation errors for each individual practitioner. Individual estimation errors of practitioners are given at the surrounding nodes and the collective estimation error is the center node. Left: arithmetic mean, Right: Bayesian model.

Table 3. RMS error in crowd estimation for the conceptual design evaluations.

Aggregation method	RMS error
Arithmetic mean	0.2388
Geometric mean	0.6028
Median	0.3256
Majority voting	0.3652
Bayesian model	0.3268

lems are as follows: a device that disperses a light coating of a powdered substance over a surface [40], a way to minimize accidents from people walking and texting on a cell phone [41], a device to immobilize a human joint [42] and a device to remove the shell from a peanut in areas with no electricity [43]. This set of conceptual design solutions was taken from a solution set collected for prior work by Goucher-Lambert and Cagan [44]. In that study, inter-rater reliability was assessed using the 114 solution concepts included here. Each design was evaluated across four metrics: usefulness, feasibility, novelty, and quality. Practitioners were provided with one-sentence criteria for each metric (including scoring), and did not see any example solutions prior to rating designs. Example concepts for two of the problems are shown in Figure 10. The goal here is to determine the accuracy of the Bayesian network model for class of problems with extremely low structural and functional similarity.

Table 3 summarizes the collective estimation errors aggregated with different methods. Here, we observe that Bayesian model does not perform well and is outperformed by arithmetic mean. Looking into individual estimation errors gives an insight into why Bayesian model is not performing well for these conceptual designs that lack the structural and functional similarity. Figure 11 demonstrates that every individual in the practitioner group makes significant estimation error. Even though the estimation aggregated through Bayesian model is better than all individuals, it is still very high due to large estimation errors of each practitioner. In contrast to our previous esoteric engineering problems, conceptual design problems have no *true* solution. We believe open ended nature of conceptual design problems creates a challenge for consistent evaluation and requires further exploration.

7 Conclusions and Future Work

As engineering and technology expand, there are new problems that emerge which may not have an analytic or computational solution. Such problems are difficult for any one person to solve, even among experts of the domain. Therefore, it is necessary to aggregate the perspectives of multiple people to obtain a holistic understanding of the task. This paper demonstrates that crowdsourcing with practitioner crowds is an applicable tool for collecting these different perspectives due to WoMC.

When the problem is properly framed, such that the responses to the questions can be easily quantified and thus rated objectively, WoMC can allow the collective estimate of the crowd to be more accurate than the majority of individuals in the crowd. This demonstrates a potential for future expansion of crowdsourcing in engineering, by integrating both design generation/innovation with design evaluation. Crowds may be utilized to develop large quantities of designs to solve an esoteric engineering task, and then the wisdom of crowds may be used to evaluate the proposed designs.

Acknowledgements

We would like to thank authors of [7] for making their data publicly available.

References

[1] Galton, F., 1907. "The ballot-box". *Nature,* **75**(1952), p. 509.
[2] Hooker, R. H., 1907. "Mean or median". *Nature,* **75**, pp. 487–488.
[3] Surowiecki, J., 2005. *The wisdom of crowds*. Anchor.
[4] Howe, J., 2006. "The rise of crowdsourcing". *Wired magazine,* **14**(6), pp. 1–4.
[5] Wah, C., 2006. "Crowdsourcing and its applications in computer vision". *University of California, San Diego*.
[6] Wauthier, F. L., and Jordan, M. I., 2011. "Bayesian bias mitigation for crowdsourcing". In Advances in neural information processing systems, pp. 1800–1808.
[7] Burnap, A., Ren, Y., Gerth, R., Papazoglou, G., Gonzalez, R., and Papalambros, P. Y., 2015. "When crowdsourcing fails: A study of expertise on crowdsourced design evaluation". *Journal of Mechanical Design,* **137**(3), p. 031101.
[8] Yang, M. C., 2010. "Consensus and single leader decision-making in teams using structured design methods". *Design Studies,* **31**(4), pp. 345–362.
[9] Gurnani, A., and Lewis, K., 2008. "Collaborative, decentralized engineering design at the edge of ra-

tionality". *Journal of Mechanical Design,* **130**(12), p. 121101.

[10] Summers, J. D., and Shah, J. J., 2010. "Mechanical engineering design complexity metrics: size, coupling, and solvability". *Journal of Mechanical Design,* **132**(2), p. 021004.

[11] Takai, S., 2010. "A game-theoretic model of collaboration in engineering design". *Journal of Mechanical Design,* **132**(5), p. 051005.

[12] Cabrerizo, F. J., Ureña, R., Pedrycz, W., and Herrera-Viedma, E., 2014. "Building consensus in group decision making with an allocation of information granularity". *Fuzzy Sets and Systems,* **255**, pp. 115–127.

[13] Lorenz, J., Rauhut, H., Schweitzer, F., and Helbing, D., 2011. "How social influence can undermine the wisdom of crowd effect". *Proceedings of the National Academy of Sciences,* **108**(22), pp. 9020–9025.

[14] Hong, L., and Page, S. E., 2004. "Groups of diverse problem solvers can outperform groups of high-ability problem solvers". *Proceedings of the National Academy of Sciences of the United States of America,* **101**(46), pp. 16385–16389.

[15] Amazon mechanical turk. `https://www.mturk.com`. Accessed: 2017-01-30.

[16] CrowdFlower. `https://www.crowdflower.com`. Accessed: 2017-01-30.

[17] Prolific academic. `https://www.prolific.ac/`. Accessed: 2017-01-30.

[18] Berinsky, A. J., Huber, G. A., and Lenz, G. S., 2012. "Evaluating online labor markets for experimental research: Amazon. com's mechanical turk". *Political Analysis,* **20**(3), pp. 351–368.

[19] Peer, E., Brandimarte, L., Samat, S., and Acquisti, A., 2017. "Beyond the turk: Alternative platforms for crowdsourcing behavioral research". *Journal of Experimental Social Psychology,* **70**, pp. 153–163.

[20] Brabham, D. C., 2008. "Crowdsourcing as a model for problem solving: An introduction and cases". *Convergence,* **14**(1), pp. 75–90.

[21] Kittur, A., Chi, E. H., and Suh, B., 2008. "Crowdsourcing user studies with mechanical turk". In Proceedings of the SIGCHI conference on human factors in computing systems, ACM, pp. 453–456.

[22] Rzeszotarski, J. M., and Kittur, A., 2011. "Instrumenting the crowd: using implicit behavioral measures to predict task performance". In Proceedings of the 24th annual ACM symposium on User interface software and technology, ACM, pp. 13–22.

[23] Kittur, A., and Kraut, R. E., 2008. "Harnessing the wisdom of crowds in wikipedia: quality through coordination". In Proceedings of the 2008 ACM conference on Computer supported cooperative work, ACM, pp. 37–46.

[24] Kittur, A., Chi, E., Pendleton, B. A., Suh, B., and Mytkowicz, T., 2007. "Power of the few vs. wisdom of the crowd: Wikipedia and the rise of the bourgeoisie". *World wide web,* **1**(2), p. 19.

[25] Burnap, A., Hartley, J., Pan, Y., Gonzalez, R., and Papalambros, P. Y., 2015. "Balancing design freedom and brand recognition in the evolution of automotive brand styling". In ASME 2015 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference, American Society of Mechanical Engineers, pp. V007T06A047–V007T06A047.

[26] Orbay, G., Fu, L., and Kara, L. B., 2015. "Deciphering the influence of product shape on consumer judgments through geometric abstraction". *Journal of Mechanical Design,* **137**(8), p. 081103.

[27] Burnap, A., Gerth, R., Gonzalez, R., and Papalambros, P. Y., 2017. "Identifying experts in the crowd for evaluation of engineering designs". *Journal of Engineering Design,* **28**(5), pp. 317–337.

[28] Morgan, H., Levatti, H., Sienz, J., Gil, A., and Bould, D., 2014. "Ge jet engine bracket challenge: A case study in sustainable design". *Sustainable Design and Manufacturing 2014 Part 1*, p. 95.

[29] Fu, K., Chan, J., Cagan, J., Kotovsky, K., Schunn, C., and Wood, K., 2013. "The meaning of ?near? and ?far?: the impact of structuring design databases and the effect of distance of analogy on design output". *Journal of Mechanical Design,* **135**(2), p. 021007.

[30] Murphy, J., Fu, K., Otto, K., Yang, M., Jensen, D., and Wood, K., 2014. "Function based design-by-analogy: a functional vector approach to analogical search". *Journal of Mechanical Design,* **136**(10), p. 101102.

[31] Moreno, D. P., Hernandez, A. A., Yang, M. C., Otto, K. N., Hölttä-Otto, K., Linsey, J. S., Wood, K. L., and Linden, A., 2014. "Fundamental studies in design-by-analogy: A focus on domain-knowledge experts and applications to transactional design problems". *Design Studies,* **35**(3), pp. 232–272.

[32] Shah, J. J., Kulkarni, S. V., and Vargas-Hernandez, N., 2000. "Evaluation of idea generation methods for conceptual design: effectiveness metrics and design of experiments". *Journal of mechanical design,* **122**(4), pp. 377–384.

[33] Bendsøe, M. P., 1989. "Optimal shape design as a material distribution problem". *Structural optimization,* **1**(4), pp. 193–202.

[34] Nobel-Jørgensen, M., Malmgren-Hansen, D., Bærentzen, J. A., Sigmund, O., and Aage, N., 2016. "Improving topology optimization intuition through games". *Structural and Multidisciplinary Optimization,* **54**(4), pp. 775–781.

[35] Whitehill, J., Wu, T.-f., Bergsma, J., Movellan, J. R., and Ruvolo, P. L., 2009. "Whose vote should count more: Optimal integration of labels from labelers of unknown expertise". In Advances in neural information processing systems, pp. 2035–2043.

[36] Bachrach, Y., Graepel, T., Minka, T., and Guiver, J., 2012. "How to grade a test without knowing the answers—a bayesian graphical model for adaptive crowdsourcing and aptitude testing". *arXiv preprint arXiv:1206.6386*.

[37] Welinder, P., Branson, S., Belongie, S. J., and Perona,

P., 2010. “The multidimensional wisdom of crowds.”. In NIPS, Vol. 23, pp. 2424–2432.

[38] Lakshminarayanan, B., and Teh, Y. W., 2013. “Inferring ground truth from multi-annotator ordinal data: a probabilistic approach”. *arXiv preprint arXiv:1305.0015*.

[39] Vul, E., and Pashler, H., 2008. “Measuring the crowd within probabilistic representations within individuals”. *Psychological Science,* **19**(7), pp. 645–647.

[40] Linsey, J. s., Wood, K. l., and Markman, A. b., 2008. “Modality and representation in analogy”. *Artif. Intell. Eng. Des. Anal. Manuf.,* **22**(2), Jan., pp. 85–100.

[41] Miller, S. R., Bailey, B. P., and Kirlik, A., 2014. “Exploring the utility of bayesian truth serum for assessing design knowledge”. *Hum.-Comput. Interact.,* **29**(5-6), Aug., pp. 487–515.

[42] Wilson, J. O., Rosen, D., Nelson, B. A., and Yen, J., 2010. “The effects of biological examples in idea generation”. *Design Studies,* **31**(2), pp. 169 – 186.

[43] Viswanathan, V. K., and Linsey, J. S., 2013. “Design fixation and its mitigation: a study on the role of expertise”. *Journal of Mechanical Design,* **135**(5), p. 051008.

[44] Goucher-Lambert, K., and Cagan, J., 2017. “Using crowdsourcing to provide analogies for designer ideation in a cognitive study”. In DS 87-8 Proceedings of the 21st International Conference on Engineering Design (ICED 17) Vol 8: Human Behaviour in Design, Vancouver, Canada.