Nurcan Gecer Ulu[1]
Department of Mechanical Engineering,
Carnegie Mellon University,
Pittsburgh, PA 15213
e-mail: ngu@cmu.edu

Michael Messersmith
Department of Mechanical Engineering,
Carnegie Mellon University,
Pittsburgh, PA 15213
e-mail: mmessers@andrew.cmu.edu

Kosa Goucher-Lambert
Department of Mechanical Engineering,
Carnegie Mellon University,
Pittsburgh, PA 15213
e-mail: kgoucher@andrew.cmu.edu

Jonathan Cagan
Department of Mechanical Engineering,
Carnegie Mellon University,
Pittsburgh, PA 15213
e-mail: cagan@cmu.edu

Levent Burak Kara
Department of Mechanical Engineering,
Carnegie Mellon University,
Pittsburgh, PA 15213
e-mail: lkara@cmu.edu

Wisdom of Microcrowds in Evaluating Solutions to Esoteric Engineering Problems

A multitude of studies in economics, psychology, political and social sciences have demonstrated the wisdom of crowds (WoC) phenomenon, where the collective estimate of a group can be more accurate than estimates of individuals. While WoC is observable in such domains where the participating individuals have experience or familiarity with the question at hand, it remains unclear how effective WoC is for domains that traditionally require deep expertise or sophisticated computational models to estimate objective answers. This work explores how effective WoC is for engineering design problems that are esoteric in nature, that is, problems (1) whose solutions traditionally require expertise and specialized knowledge, (2) where access to experts can be costly or infeasible, and (3) in which previous WoC studies with the general population have been shown to be highly ineffective. The main hypothesis in this work is that in the absence of experts, WoC can be observed in groups that consist of practitioners who are defined to have a base familiarity with the problems in question but not necessarily domain experts. As a way to emulate commonly encountered engineering problem-solving scenarios, this work studies WoC with practitioners that form microcrowds consisting of 5–15 individuals, thereby giving rise to the term the wisdom of microcrowds (WoMC). Our studies on design evaluations show that WoMC produces results whose mean is in the 80th percentile or better across varying crowd sizes, even for problems that are highly nonintuitive in nature.
[DOI: 10.1115/1.4042615]

1 Introduction

Crowdsourcing is emerging as a cost-effective, rapid approach to problem solving in a variety of disciplines where the collective estimate of a group can outperform the individuals, even in the presence of domain experts. This phenomenon is known as the wisdom of crowds (WoC) and has been demonstrated across a range of problem domains [1–3]. Traditional crowdsourcing naturally focuses on tasks that are human easy and computer hard, such as vision problems where crowds are asked to identify and label objects in large sets of images [4]. In such problems, the task is typically very intuitive for humans, and thus, the correct answer can be inferred from a crowd consensus. In engineering problems requiring domain expertise, however, crowdsourcing has proven to be significantly less effective, in part due to the limited number of experts in the sampled crowd [5]. This suggests that extending traditional crowdsourcing to tasks requiring expertise is nontrivial, especially if experts are scarce. As an alternative to crowdsourcing, expert collaboration has been extensively studied [6–10]. However, interactions among group members have been shown to lead to similarity of experts [11], which may result in experts being outperformed by diverse groups [12]. As such, it remains unclear how conventional crowdsourcing can be made truly effective for engineering design problems, especially for tasks that require expertise.

As one step toward addressing this gap, this work explores the effectiveness of WoC for engineering design problems that are esoteric in nature. Esoteric problems are defined as those (1) that traditionally require expertise and specialized knowledge, (2) where access to experts can be costly or infeasible, and (3) in which previous WoC studies with the general population have been shown to be highly ineffective [5]. The main hypothesis in this work is that in the absence of experts, WoC can be observed in groups that consist of *practitioners* who are defined to have a base familiarity with the domain and the problems in question, even though no single individual may have the expertise to correctly solve the problem. With this definition, experts are a subset of practitioners. However, in this work, in contrast to purely expert crowds, practitioner crowds are characterized by individual responses that exhibit both significant accuracy (deviation from the ground truth) and precision errors (variation among the responses). This new definition and focus on practitioners stand in contrast to previous studies that explore WoC in design that rely either on the general population crowds where experts are extremely scarce and unknown [5] or on teams of experts [7] as the basis of crowds. Additionally, as a way to emulate commonly encountered engineering problem-solving scenarios, this work studies WoC with practitioners that form microcrowds consisting of 5–15 individuals (rather than tens or hundreds of individuals), thereby giving rise to the term the *wisdom of microcrowds* (WoMC) which is central to the presented work.

As part of this study, four design assessment questions with varying levels of difficulty and intuitiveness were deployed where the participants were asked to assess the quality of the candidate design solutions. Several data aggregation methods were developed and tested on the acquired data. The results suggest that WoMC with practitioners can indeed be observed, where the crowd estimate outperforms the individuals in the vast majority of instances. To facilitate benchmarking, these results have been obtained for problems in which there already exists an objectively true solution (i.e., benchmark results obtained through optimization). As such, it could be argued that crowdsourcing is remarkably unnecessary for such problems where solution methods already exist. However, the most significant conclusion of the presented work is that for current or future engineering design problems where algorithmic solutions may currently not exist, small groups of practitioners may in fact provide very effective solutions. Note that, in this

[1]Corresponding author.

Contributed by the Design Theory and Methodology Committee of ASME for publication in the Journal of Mechanical Design. Manuscript received April 2, 2018; final manuscript received January 9, 2019; published online April 16, 2019. Assoc. Editor: Katja Holtta-Otto.

context, the current lack of solution methods implies a lack of experts, which reinforces the importance of practitioners.

An interesting limitation of the presented work, however, is that when applied to open-ended, conceptual design problems where no objectively true solution exists, the performance of WoMC declines significantly. The results indicate that in such cases, the individuals in the crowd tend to make significant estimation errors when benchmarked against expert ratings. Nevertheless, it remains unclear whether these estimation errors are due to the practitioners' inability to accurately assess candidate solutions or whether there exist issues even with expert ratings of such open-ended problems.

1.1 Crowdsourcing Scenarios in Esoteric Domains.

In engineering, crowdsourcing is often used in the form of grand challenges to gather candidate solutions where the crowdsourced solutions are assessed by experts juries. However, the use of experts to assess these solutions may not be ideal since these grand challenges are usually created for very complex problems where experts cannot solve optimally. The idea of wisdom of microcrowds with practitioners that is presented here can be an alternative to experts juries in such problems. This way, both the generation and assessment of the candidate solutions can be crowdsourced through practitioner crowds that are exposed to the esoteric domain. Note that microcrowds are composed of anonymous and noninteracting people with individually unknown expertise levels unlike traditional design teams.

Another esoteric crowdsourcing scenario includes online communities in advancing fields. Thingiverse, an online community for 3D printing designs, or GrabCAD, an online community for sharing computer aided design designs, can be examples of communities in advancing esoteric fields. These communities already include large groups of people that are familiar with their respective domains, practitioners. Such communities can benefit from the utility of our work to asses candidate designs that may in fact produce successful outcomes, which would be critically important in cases where no appropriate computational evaluation techniques exist. One such problem might be planning for hybrid manufacturing [13]. While hybrid manufacturing pushes the boundaries of production processes, its use is limited to manually created suboptimal plans since there are no established computational solutions to handle such complex planning. Our studies in this paper could help selecting the best possible plan among these manually created candidate solutions.

2 Background

Amazon Mechanical Turk (AMT)[2], CrowdFlower[3], and Prolific Academic[4] are among the most prominent crowdsourcing platforms. These platforms have a wide reach and are designed to be representative of the general public consisting of diverse crowds [14]. While AMT allows the surveys to be targeted toward specific demographics, it is difficult to identify crowds that share a prescribed technical background. By contrast, our work focuses on solving esoteric problems via microcrowds that consist of practitioners.

Previous studies have developed task design and response quality detection methods as a way to maximize the useful information content in crowdsourcing [15,16]. Example methods include the use of explicitly verifiable questions to identify malicious users and to encourage honest responses, and task fingerprinting to monitor completion time, mouse movements, key presses, and scroll movements, which can all be used as indicator attributes for detecting suspect responses [17]. The effect of incentives and competition in crowdsourcing data quality has been investigated in Ref. [18]. The presented work uses the response speed as one such indicator to vet data quality and monetary compensation as incentive.

Consensus through collaboration is a widely used approach in engineering [6,9]. However, driven by the previous observations that there is a danger of expert collaboration to result in a singular thought pattern that could be outperformed by diverse groups [11], this work explores WoMC with individuals who remain independent and form crowds that are more diverse than collaborating experts. Surowiecki [3] argues that one requirement for a good crowd judgment is that people's decisions remain independent of one another. This was further validated by Lorenz et al. [12] where individuals were observed to produce collectively more accurate crowd estimations over cases where the same individuals were informed by others' estimates. Independence of opinion (no contact between the individuals) constitutes one of the major differences between the traditional collaborative design teams and microcrowds. Team network structure in mass collaboration design projects and the effect of individual's characteristics have been studied in Ref. [19]. In contrast to our work, aforementioned study assumes that each individual in the group has known levels of expertise and ability. Yet, in the context of our work, a practitioner's expertise level is unknown albeit the group is assumed to have an exposure to the problem domain. Thus, unlike traditional design teams, microcrowds are composed of anonymous and noninteracting people with individually unknown expertise levels.

Burnap et al. [5,20] explored the use of crowdsourcing in engineering design assessment as well as techniques for identifying the experts in a crowd. These studies do not assume an a priori knowledge of the individuals' background and are thus greatly suited for studies involving large crowds. Our work builds on and complements these studies by focusing on a small group of practitioners, none of whom may be an expert but whose technical familiarity with the problem domain is significantly higher and more homogeneous compared to crowds extracted from the general population.

Crowdsourcing has also been used in design for identifying customer preferences to balance the style with brand recognition [21] or to study the relationship between the product geometry and consumer judgment of style [22]. Ghosh et al. [23] modeled user preferences by considering perceptions estimated by user–product interaction data. While these works primarily focus on eliciting subjective judgments of preference and perception, the main focus of the presented work is to crowdsource solutions to engineering problems where an objectively true solution must exist (albeit unknown).

Another popular use of crowdsourcing involves the discovery of diverse solutions to complex technical problems involving very high-dimensional design spaces, such as the GE bracket design challenge [24]. While the generation of solutions is typically the core challenge (hence crowdsourced), candidate solutions can be rather easily assessed using computational analysis tools. However, the main hypothesis and the utility of our work are that further crowdsourcing to assess candidate designs may in fact produce successful outcomes, which would be critically important in cases where no appropriate computational evaluation technologies exist.

Another open problem within the engineering design research community where crowdsourcing could provide value relates to the consistent evaluation of conceptual designs. In contrast to engineering problems with known solutions (i.e., structural mechanics), conceptual design problems have no *true* solution. When studying the conceptual design process, researchers often utilize cognitive studies to explore specific process characteristics, such as the impact of analogical stimuli on solution output [25–27]. 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 [28]. 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

[2]https://www.mturk.com
[3]https://www.crowdflower.com
[4]https://www.prolific.ac/

hundreds of design concepts to be evaluated at a given time [29]. 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. Recently, evaluation of creativity in conceptual designs by crowdsourcing and impact of expertise on creative concept selection have been studied [30,31]. Toh et al. [32] developed a machine learning approach for computing design creativity of large sets of design ideas. In this work, we also explore WoMC for the evaluation of conceptual designs.

3 Experimental Design

In order to study the wisdom of crowd in esoteric engineering applications, it is necessary to understand the relationship between crowds and problem types. This section explains the characteristics of the crowd participants and the design problems used in this work.

3.1 Crowd Population. Two key factors in the WoC 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 providing anonymity and independence across participants. To support diversity of opinion, we collected crowds through AMT or students specializing various topics in mechanical engineering.

This work considers two types of crowds: AMT workers and practitioners. AMT crowds consist of individuals from the population at large, with no explicit control over an individual's level of expertise. On the other hand, the practitioner group represents individuals who have familiarity and knowledge within the target domain, however are not necessarily domain *experts* for the given task. For example, a practitioner would be an individual who has studied or currently practices mechanical engineering but does not necessarily specialize in the field of a given task such as heat transfer, structural mechanics, or manufacturing. For a practitioner group, performance of individuals may have significant variation yet the base domain knowledge pushes the estimation method to accurate levels. Note that with this definition, experts are a subset of practitioners.

For the practitioner group, 15 mechanical engineering graduate students at Carnegie Mellon University were recruited to participate. Each participant was compensated monetarily for their time. The 15 practitioners were recruited from an available pool of over 300 graduate students. It is important to note that these students have different skill levels. As will be shown later, this can be observed by large individual estimation errors and significant performance variation among the group members. Students in our study are graduate students who already have engineering degrees as well as engineering experience through internships and possible full-time jobs. In that sense, they also represent engineers not just students. For the AMT surveys, groups of 100 people were gathered through Amazon Mechanical Turk, receiving monetary compensation. In order to remain true to the notion of general public as closely as possible, no specific demographic groups were targeted. For the structural mechanics questions (discussed in detail below), the study used the data provided by Burnap et al. [5].

3.2 Survey Design and Questions. This study investigates the WoC with four different surveys that range in the challenge they present to a human. All surveys require the respondents to be knowledgeable about the terminology used in the questions. 3D printing questions (Figs. 1, 2, and 3) aim to probe broadly intuitive

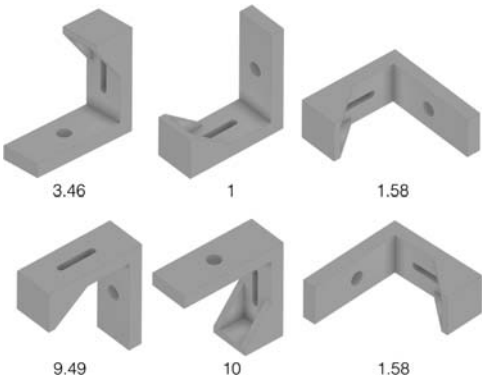

Fig. 1 3D printing-1: support material question. Numbers indicate the amount of support material required to print the object at the given orientation on a scale from 1 (very little) to 10 (a lot).

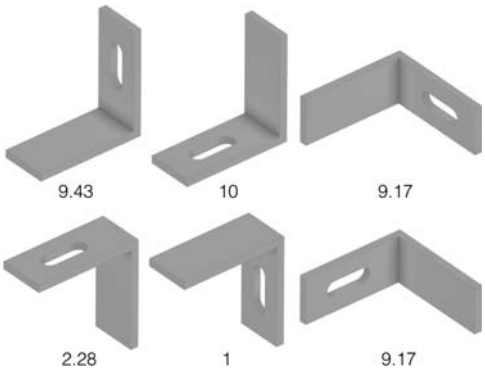

Fig. 2 3D printing-2: surface finish question. Numbers indicate the surface quality rating between 1 (poor) and 10 (excellent).

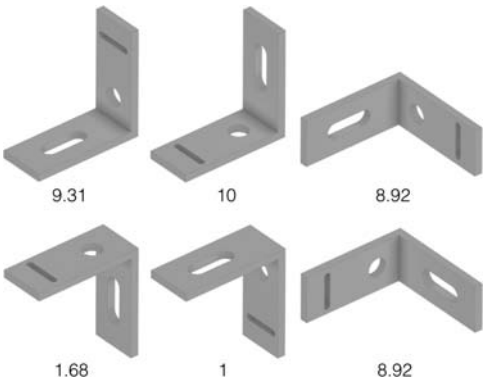

Fig. 3 3D printing-3: surface finish question. Numbers indicate the surface quality rating between 1 (poor) and 10 (excellent).

perception skills involving visual estimations of areas and volumes. However, they are designed to be increasingly more challenging. Conversely, the structural mechanics problem that involves estimating shape deformations (Fig. 4) presents a much greater challenge to humans, even for experts.

Although engineering problems are often computer easy and human hard, they are solved using expert intuition when no computational tools are available. A series of surveys for problems with known solutions such that the crowd evaluation accuracy could be determined, assessed whether such situations could benefit from WoC. The structural mechanics problem (Fig. 4) provides a good example, as such structural design problems had been solved primarily by experts' knowledge and intuition until the introduction of topology optimization techniques in the 1990s [33].

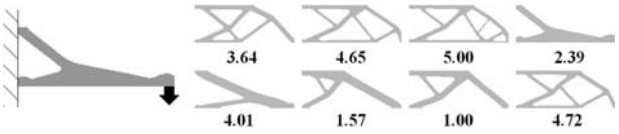

Fig. 4 The structural mechanics problem [5]. Numbers indicate the strength of each bracket between 1 (weak) and 5 (strong). (Permission to reprint from ASME copyright 2015.)

Therefore, there now exist tools to computationally evaluate the aggregated crowd evaluations and benchmark the performance against true values. As such, practitioners' performance on such problems (which can now be objectively assessed) may provide insights into whether crowd evaluations of design proposals may yield successful outcomes especially for engineering challenges for which computational modeling and analysis tools may not yet exist.

A rating-assignment approach within a predefined scale is utilized. Each survey consists of multiple questions (e.g., rating the amount of support material for six different orientations) to facilitate expertise inference later in the crowd aggregation stage. In all surveys, participants are presented with the problem statement and the candidate solutions to be rated.

Figure 1 shows the 3D printing-1 survey where participants are asked to rate the amount of support material required to print an object at various orientations using a fused-deposition printer. For each of the given orientations, participants are required to evaluate the amount of support material needed on a scale from 1 (very little) to 10 (a lot). The benchmark analysis computes the required support material as the volume that is created by the projection of overhangs to the base with a zero overhang angle [34]. Then, the scores are scaled linearly between 1 and 10 to create the benchmark values.

The 3D printing-2 survey is about evaluating the surface finish quality of an object in various orientations (Fig. 2). The participants are asked to rate the quality of the printed object considering the amount of surfaces in contact with the support material for each presented orientation. Surface quality rating is between 1 (poor) and 10 (excellent). To compute the true surface finish, the overhang areas are computed with the zero overhang angle. Then, the overhang areas are scaled inversely between 1 and 10 such that 1 represents large support material contact with poor finish and 10 is very good finish with the least amount of support material contact. The 3D printing-3 survey (Fig. 3) asks the same question on an object with more features that increase the difficulty of evaluation.

In the structural design survey, participants are presented with eight different bracket designs intended to support a downward force at the end of the bracket (Fig. 4). Then, they are asked to rate the strength of each bracket on a scale from 1 (weak) to 5 (strong), where strength is defined to be the amount of deformation under the given load [5]. The main reason we use this problem is that estimating the strength of arbitrary shapes is significantly more demanding compared to volume/area evaluations. While humans are exposed to volume/area computations in daily life, rating the strength of an arbitrary design requires a specific experience [35], which is highly unlikely to be prevalent in the general population.

4 Crowd Estimate Aggregation Techniques

The choice of the aggregation method affects the collective estimate of the group. For instance, previous studies show that the median or geometric mean can result in estimates that are more accurate over the arithmetic mean [1,12]. This section explains the different aggregation methods used in this work.

The following metrics are used: arithmetic mean, geometric mean, median, majority voting, and Bayesian networks. In a crowd of n participants with a set of estimates $Y: y_1, \ldots, y_n$ where $y_i \in \mathbb{Z}: 1 \le y_i \le 10$ for all i, the arithmetic mean is $y^{agg} = (1/n)\sum_{j=1}^{n} y_i$. The geometric mean is $\exp((1/n)\sum_{j=1}^{n} \ln(y_i))$. The median is the median value in Y. The majority vote is the mode of Y.

Bayesian networks have been widely used in crowdsourcing to mitigate the noise from biased responses. Relevant studies model the sources of bias using models that consider problem difficulty and the competence of participants [4,5,36–40]. Similar to these approaches, this work adopts a Bayesian model as shown in Fig. 5. The evaluation process is modeled such that for participant i working on problem j, participant expertise, α_i, and problem difficulty, β_j, result in variance, δ_{ij}. Thus, the evaluation of participant i on problem j, y_{ij}, is obtained when the true score of the problem, x_j, is combined with the variance, δ_{ij}. Note that the Bayesian model does not require prior knowledge of the true answers, participant expertise, or problem difficulty. The only observed variable is the participant answer for each question.

The variance is obtained using participant expertise and problem difficulty. This work assumes that a participant may be malicious, inexperienced, or experienced. Also, a problem can be easy, difficult, or unintuitive. Defining both parameters on a continuous range, the variance is modeled 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 $\alpha_i \in (-\text{inf}, +\text{inf})$ and the problem difficulty is $\beta_j \in (0, +\text{inf})$. The resulting variance becomes $\delta_{ij} \in [0, 1]$. The evaluation process is modeled as a random variable with a truncated Gaussian distribution around the true score ($\mu = x_j$) with a variance 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 $x_j \in [0, 1]$.

The relationship between the evaluation variance with participant expertise and problem difficulty is further explained in Fig. 6. The variance indicates how far the evaluations may be spread apart from the true score. Therefore, a high variance implies the probability of sampling far away from the true score, resulting in a high evaluation error. From the perspective of precision, i.e., reciprocal of variance, small variance means high precision, meaning higher chance of getting the correct evaluation. As anticipated, smaller variance is observed as expertise increases as shown in Fig. 6 for three problem difficulty levels that correspond to easy, difficult, and unintuitive. On the other hand, nonexperts can give answers with large variance. Yet, there is a potential for malicious participants who intentionally give the wrong answers. Since the answers are maliciously wrong, the amount of variance (thus the evaluation error) is even more than that of a nonexpert that randomly guesses the answers. On the other hand, for a very easy question, even unskilled participants can give answers with a small variance and anyone malicious can make the most damage (Fig. 6, top). As the questions get more difficult, expertise affects the variance of answers more (Fig. 6, mid). Yet, an unintuitive question cannot be evaluated with small variance (high precision) by participants at any skill level and evaluated with similar variance since all participants evaluate the problem with random guesses (Fig. 6, bottom).

The structure explained above leads to the graphical model shown in Fig. 5. In the model, participant expertise, α_i, problem

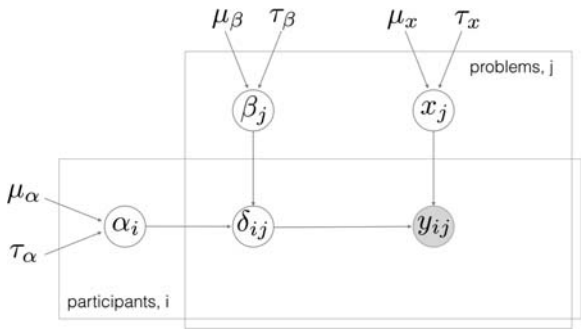

Fig. 5 The Bayesian network model

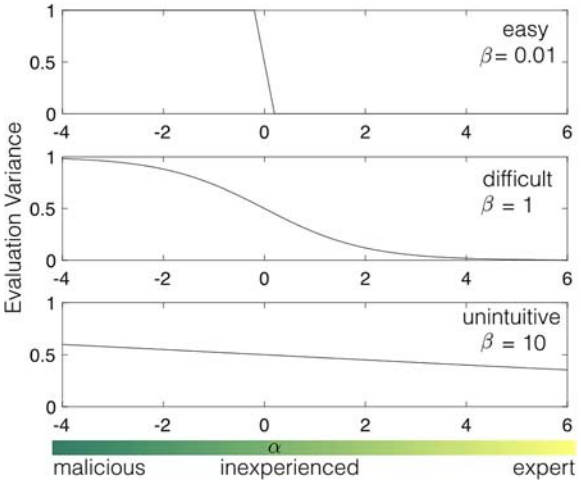

Fig. 6 The evaluation variance with participant expertise shown at three different problem difficulty levels as easy, difficult, and unintuitive

difficulty, β_j, and true scores, x_j, are sampled from a known prior distribution and these determine the observed evaluations, y_{ij}. Given a set of observed evaluations, the task is to infer the most likely values of true scores, x_j, together with the participant expertise, α_i, and problem difficulty, β_j, parameters. Assuming a Bayesian treatment with priors on all parameters, the joint probability distribution can be written as

$$p(\mathbf{y}, \mathbf{x}, \delta, \alpha, \beta) = \prod_i p(\alpha_i) \prod_j p(\beta_j) p(x_j) \times \prod_{ij} p(y_{ij}|\delta_{ij}, x_j) p(\delta_{ij}|\alpha_i, \beta_j) \quad (2)$$

Equation (2) excludes hyperparameters for brevity. In our implementation, we use Gaussian priors for α with mean, $\mu_\alpha = 1$, and precision, $\tau_{\text{alpha}} = 1$. Since the value of β needs to be positive, the implementation imposes a truncated Gaussian prior with mean, $\mu_\beta = 1$, and precision, $\tau_{\text{beta}} = 1$, 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 $\tau_x = 0.1$.

Markov chain Monte Carlo simulations are employed to infer the results utilizing the Metropolis-Hastings method. Empirically, we observe that using thinning interval of 3 and burn-in length of 10^5 works well with 5×10^5 iterations.

5 Results

To demonstrate the WoMC in esoteric engineering problems, we conducted four surveys on two sets of crowds (practitioners and AMT workers) having different skill levels as explained in Secs. 1–4. This section presents the results of the surveys and compares the performance of the aggregation methods.

5.1 Survey Results. The results of the surveys with different crowds and aggregation methods are summarized in Table 1. All scores are scaled between 0 and 1 for direct comparison across surveys. In addition to the overall survey results, Fig. 7 includes estimation errors for each question in the surveys. While the collective error can be defined as the difference between the true answer and the aggregated answer ($y^t - y^{\text{agg}}$) for a single question, this work uses the root mean square (RMS) error for multiquestion surveys since it provides a performance measure in the same scale as the individual questions. For a survey containing m questions, the collective error can be computed as $\sqrt{(1/m)\sum_{j=1}^{m}(y_j^t - y_j^{\text{agg}})^2}$. Note that the participant responses are discrete scores rather than continuous variables. While arithmetic mean, geometric mean, and Bayesian networks produce a real number from discrete inputs, median and majority voting remain discrete values. For consistency, we compare continuous and discrete aggregates with true continuous answers and their rounded values, respectively.

5.2 Crowd Expertise and Aggregation Methods. As shown in Table 1, with the AMT groups, there is no accurate estimation with any of the aggregation methods, with RMS errors around 40% and as high as 60%. Moreover, the Bayesian network method is outperformed by the other methods in all of the AMT studies. This outcome is consistent with previous findings that argue crowdsourcing AMT populations for engineering design evaluations may produce unreliable results [5]. On the other hand, the results of the practitioner studies suggest that crowdsourcing can indeed be useful for the same kinds of problems, where consistently more accurate estimations are obtained relative to the AMT groups.

When the aggregation methods are compared, no single method appears to be best in the AMT studies. On the other hand, for the practitioner groups, the results indicate that the Bayesian network consistently produces accurate crowd estimations. Of note, for both the practitioner and the AMT groups, the geometric mean method never emerges as the best approach. This can be explained by the fact that the responses are constrained within particular upper and lower bounds (1–10 for the 3D printing and 1–5 for the structural design problems) where the range spans only one order of magnitude, whereas the geometric mean is most useful when input data vary in orders of magnitude [12].

Table 1 In practitioner groups, the WoC effect is observable as evidenced by the low RMS errors (over the scale 0–1). The Bayesian model gives the best estimate in most cases for practitioners. For the AMT groups, however, the high RMS errors suggest poor estimation accuracy hence much weaker WoC. Note that for the AMT groups, no single 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 mechanics, practitioner	0.197	0.217	0.198	0.342	0.173
Structural mechanics, AMT	0.339	0.352	0.385	0.395	0.392

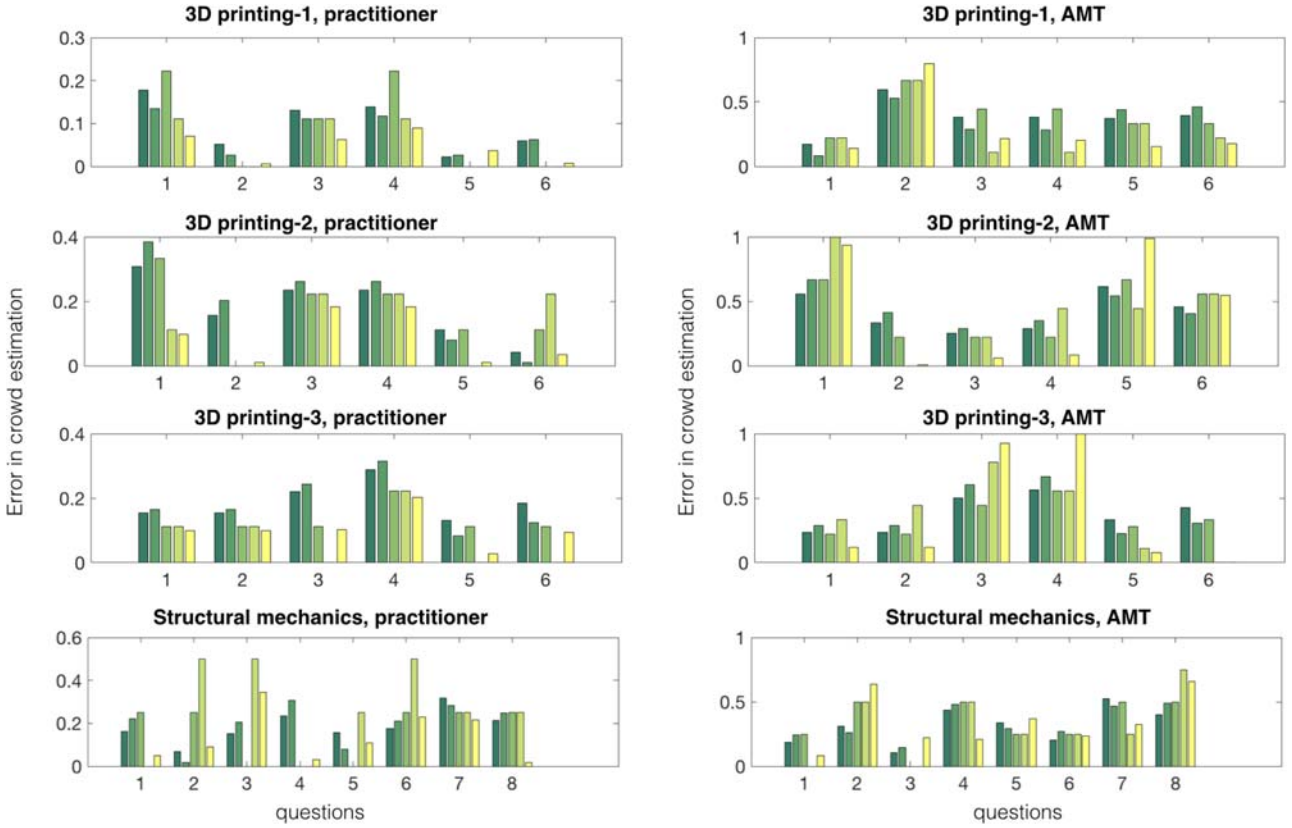

Fig. 7 Error of crowd estimation for each question in the four survey sets. Each bar group represents the RMS of the crowd aggregated through arithmetic mean, geometric mean, median, majority voting, and Bayesian model. Note that for each of the 3D printing surveys, there are six questions. For the structural design survey, there are eight questions.

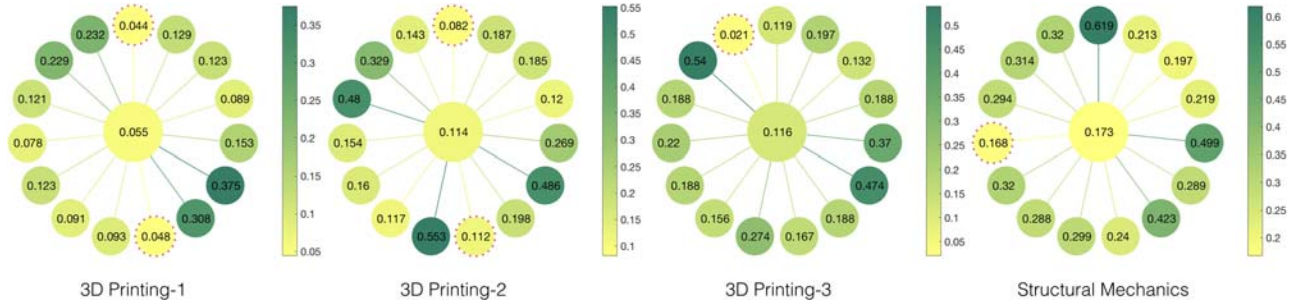

Fig. 8 Estimation error significantly varies in the practitioner group. Collective estimate of the practitioner crowd is more accurate than the vast majority of individual practitioners. Collective error of the crowd and errors of individual practitioners in the crowd are given in the center node and the surrounding nodes, respectively. The color of the circles represents the error with dark green representing a high error and light yellow a low error. Individuals who perform better than the collective answer are marked with a dashed circle.

5.3 WoMC and Individuals. To analyze the WoC effect, the performance of the aggregated crowd estimation is compared against the individuals (Fig. 8). Only practitioner crowds are included in this analysis as we do not observe a reasonable accuracy in AMT surveys. The collective answers aggregated with Bayesian networks are employed as they consistently perform well in practitioner group studies.

Figure 8 shows that the collective estimation of the crowd is more accurate than most of the individuals.[5] 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 produce an accurate measure of the WoC for the problems that are of esoteric nature. This can be explained by the participant expertise and problem difficulty-based inference that considers all answers of an individual to multiple questions collectively rather than a single one. Moreover, these results suggest that the Bayesian networks' approach does not undermine the WoC effect by erroneously honing in on only an elite group of experts in the group and instead allows diverse perspectives to be incorporated. This can be explained by the fact that the level of expertise is not prescribed but rather inferred as a latent variable in the Markov chain Monte Carlo simulations.

Table 2 further quantifies the WoC effect by revealing the fraction of people that are outperformed by the collective answer. A higher percentile suggests that a higher fraction of individuals are

[5]WoC is not expected to outperform all individuals. Rather, its effectiveness is proportional to the fraction of individuals it is able to outperform. In actual use, which individuals have the best answer is unknown.

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 mechanics	93	100

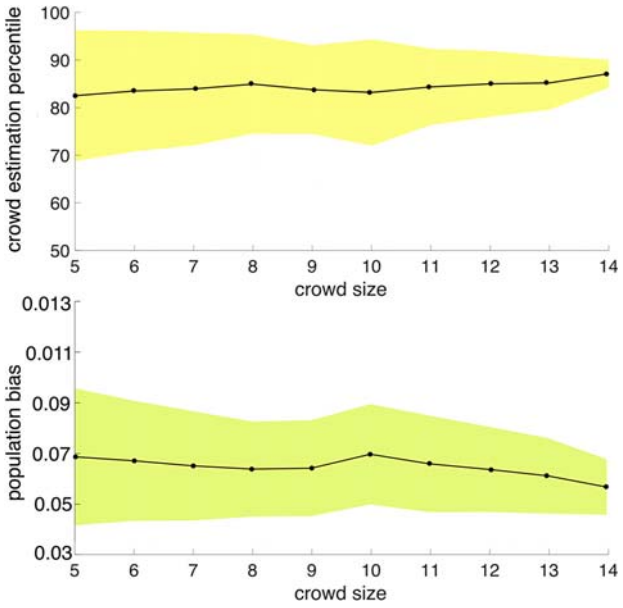

Fig. 9 The effect of crowd size on the performance of the crowd estimate represented as the percentile and population bias. A slightly upward trend in the percentiles and a significant decrease in the standard deviation (yellow shaded) as the crowd size increases suggest that higher percentile ranks can be achieved with stronger certainty in larger crowds. For population bias, both mean and standard deviation slightly decrease as crowd size increases.

outperformed, hence a stronger WoC effect is achieved. The percentile rank of the crowd is computed using two error metrics as continuous and discrete: the continuous percentile rank computed as the distance between the true answers and participant ratings; the discrete measure rounding the true answers to the nearest integer while computing the individual estimate errors. Note that the discrete measure can be significantly affected by these round off errors. The difference between continuous and discrete percentile ranks can be explained by this fact. Of note is the distinction between the percentile rank and the accuracy of the collective estimate. The percentile rank reveals the relative performance of the collective estimate compared to the individual estimates, while the accuracy refers to the RMS error between the estimate and ground truth benchmark.

5.4 Effect of Crowd Size. Platforms such as AMT enable access to large and diverse groups. However, in most practical problem-solving settings, only a limited number of practitioners are likely to be accessible for the solution of the engineering challenge. To gain insight into the impact of small-sized practitioner groups, we analyze the WoC effect across even smaller group sizes, leading to the term microcrowds (WoMC).

Figure 9 shows that WoMC can still be observed in smaller groups. The crowd size is analyzed with the 3D printing-1 survey and crowd estimation is computed using Bayesian networks (Table 1). Initially, practitioner studies are conducted with 15 participants. To simulate microcrowds with smaller number of participants, a subset of 500 randomly generated combinations of 5–14 individuals were generated from the original 15 participant set. The results suggest that the WoC effect can still be observed in diminishing group sizes. The probability of obtaining crowd estimations with higher success (percentile) increases with larger crowds. An approximately 6% increase in percentile rank with 10% decrease in standard deviation is observed as the crowd size is increased from 5 to 14. Figure 9 also shows the effect of crowd size on population bias, defined as the error of aggregated estimate across the crowd [41]. Both the mean and standard deviation slightly decrease with the increasing crowd size.

5.5 Evaluation Metric. One might argue that internal scales may play an important role in people's ratings. In other words, different individuals may use different internal scales and their definition of *very weak, very strong*, or *very little* may differ. For example, a strict grader may score the surface quality between 0 and 5 (instead of 0–10) while another rater gives scores between 5 and 10. For this reason, we compare our RMS error metric with Kendall's tau coefficient [42] which is a correlation metric that measures the similarity of the orderings.

Figure 10 shows the correlation of the collective estimate and each individual participant's answers to ground truth similar to Fig. 8. Note that the RMS error and Kendall's tau correlation have an inverse relationship (i.e., better performance is indicated by a smaller error or a larger correlation value). When two metrics are compared, in general, we observe similar wisdom of crowds in terms of the number of people that the aggregated result has outperformed. 3D printing-1 and structural mechanics surveys result in the

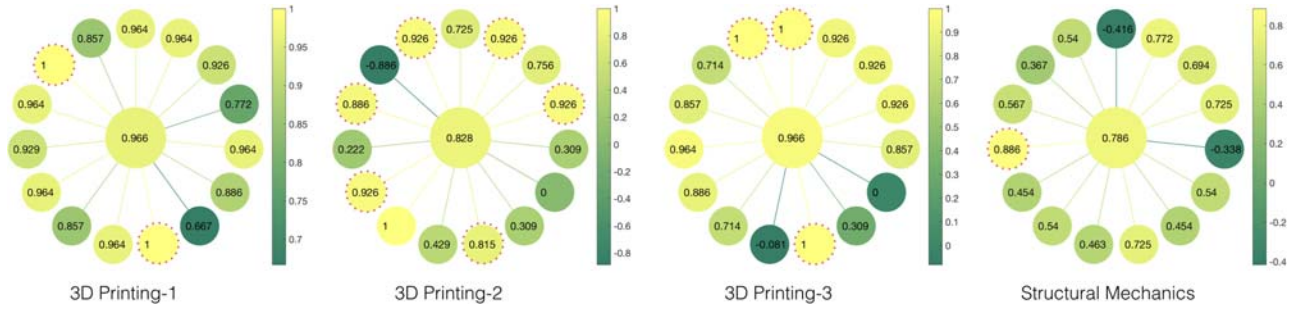

Fig. 10 Kendall's tau coefficient as evaluation metric: correlation of collective estimation and individual practitioner ratings to ground truth is given in the center node and the surrounding nodes, respectively. The color of the circles represents the correlation with dark green representing low correlation and light yellow representing high correlation. Individuals who perform better than the collective answer are marked with a dashed circle. Note that higher values indicate better correlation meaning better performance.

same performance with both evaluation metrics. When the Kendall's tau coefficient is used as the evaluation metric, we still observe that the collective estimate of the practitioner crowd is more accurate than the vast majority of individual practitioners which is a key observation in this study. We believe that the reason for this similar behavior may be the preconditioning in our questions where we ask users to scale their rating between predefined boundaries as well as having access to all candidate questions before rating each question. This way, the participants are preconditioned to use the given scale rather than their internal scales.

5.6 Conceptual Design Evaluations. As an extension of the methods presented in this paper, the feasibility of using a practitioner-sourced 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 min to complete the ratings and were monetarily compensated for their time.

Each practitioner evaluated 114 conceptual designs, corresponding to one of the four design problems. These problems are as follows: a device that disperses a light coating of a powdered substance over a surface [43], a way to minimize accidents from people walking and texting on a cell phone [44], a device to immobilize a human joint [45], and a device to remove the shell from a peanut in areas with no electricity [46]. This set of conceptual design solutions was taken from a solution set collected for prior work by Goucher-Lambert and Cagan [29]. Each design was evaluated across four metrics: usefulness, feasibility, novelty, and quality. In the previous study, consistency of the two trained experts was assessed using the intraclass correlation coefficient (ICC). ICC correlations have been reported as ICC > 0.65, ICC > 0.77, ICC > 0.71, and ICC > 0.50 for usefulness, feasibility, novelty, and quality, respectively. While three of the four metrics demonstrate strong correlation and the other metric (quality) was fair, all inter-reliability levels are within the range of values typically found in behavioral studies with human raters [47]. During our experiments, 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 Fig. 11. The goal here is to determine the accuracy of the Bayesian

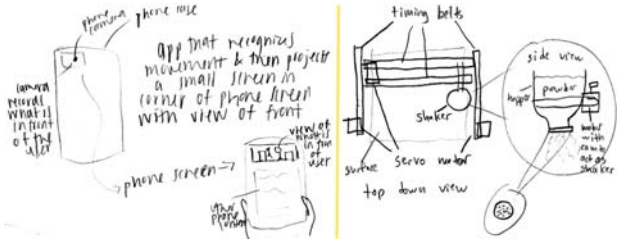

Fig. 11 Example conceptual designs

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

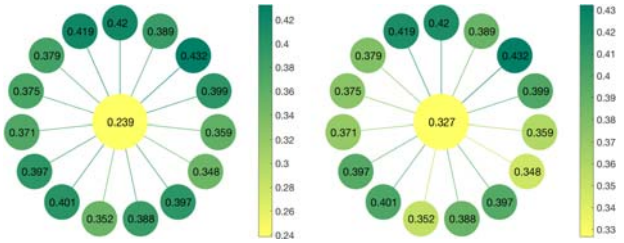

Fig. 12 The conceptual design survey illustrates significant estimation errors for each individual practitioner. Individual estimation errors of practitioners are given at the surrounding nodes, and the collective estimation error is the center node. Left: arithmetic mean, right: Bayesian model.

network model for a class of problems with extremely low structural and functional similarity.

Table 3 summarizes the collective estimation errors aggregated with different methods. Here, the Bayesian model does not perform well and is outperformed by arithmetic mean. In addition to the large collective estimation errors, Fig. 12 illustrates that individual estimation errors of practitioners are also significantly large.

6 Discussions

The analyses conducted identified some key insights on how WoMC can be achieved in esoteric engineering problems, highlighted as follows.

6.1 Problem Intuitiveness and Difficulty. All of the surveys require specific knowledge about the engineering problem at hand but they range in intuitiveness and difficulty levels. 3D printing questions are based on qualitative area/volume estimations in 3D scenes, in which humans are expected to be relatively comfortable with. On the other hand, the structural design problem is significantly more demanding since estimating the strength of complex geometries requires a deeper familiarity and experience within the domain [35]. While individuals are able to make more accurate estimations in the 3D printing questions than they can in the structural design question, an interesting observation is that no significant difference in the wisdom of crowds (i.e., percentile rank) is observed implying that crowdsourcing works equally effective in both cases. Moreover, no significant difference between the results of 3D printing-2 and 3D printing-3 surveys occur, even though the latter is more demanding with a larger number of geometrical features. These results suggest that even for problems that are demanding, the WoC is attainable at levels comparable to those attained in less demanding problems.

6.2 Level of Expertise. Populations of ordinary people (e.g., AMT crowds) perform poorly on esoteric engineering problems. Results indicate that the wisdom of crowds can be achieved in practitioner microcrowds of the domain of such problems. This suggests that people who are still gaining experience in the domain may prove to be a valuable asset as problem solvers. This is especially important as practitioner crowds may be more accessible than experts.

6.3 Aggregation Methods. In the context of practitioner populations, the most effective aggregation method found in this work is the Bayesian network. For practitioner groups, the exposure to the domain of the esoteric problem builds *true* consistency in the data and allows the Bayesian network to mitigate the mistakes made by individual practitioners. In the AMT groups, however, we observe consistently wrong answers due to lack of expertise. For that reason, the Bayesian network method performs worse than

arithmetic mean here for AMT populations as also discussed in Ref. [5]. This work indicates that the Bayesian network method is more effective given a minimum level of expertise in the group.

6.4 Crowd Size. As shown in Fig. 9, as the crowd size increases, the mean percentile performance increases (albeit modestly) while the standard deviation of the percentile rank of the group estimates decreases over sets of different microcrowds. This indicates that larger practitioner crowds will likely lead to better and more consistent outcomes. On the performance of WoMC on an absolute scale, our results indicate that group estimates in the 90th percentile can be achieved with as few as 5–14 practitioners. This suggests that in cases where computational tools are not readily available, high quality assessments on engineering problems can be gleaned from small groups of practitioners.

6.5 Conceptual Design Evaluations. When assessing solutions to a set of open-ended, conceptual problems, practitioner crowds struggle to give answers at a level that experts do. For these problems, the estimation error in crowd estimation aggregated with the Bayesian model is significant and it is outperformed by arithmetic mean. Looking into individual estimation errors gives an insight into why the Bayesian model is not performing well for these conceptual designs that lack the structural and functional similarity. Figure 12 demonstrates that every individual in the practitioner group makes a significant estimation error. Even though the estimation aggregated through the Bayesian model is better than all individuals, it is still very high due to large estimation errors of each practitioner. In contrast to the previous esoteric engineering problems, conceptual design problems have no *true* solution. We believe that the open-ended nature of conceptual design problems creates a challenge for consistent evaluation in crowd sourced environments and requires further exploration.

7 Conclusion

This work explored the ability of crowdsourced populations to estimate accurate values for a variety of esoteric problems within the domain of engineering design. Results demonstrate that the wisdom of crowd is most effective in practitioner groups, or groups of individuals who possess some level of domain knowledge, but are not necessarily experts. Aggregated crowd results of practitioners achieve high accuracy across a range of problems. By simulating small groupings of 5–15 practitioners, called microcrowds, it is found that crowd estimates perform more accurately than individual estimates across the majority of the studies. These results suggest that the WoMC can provide a powerful tool for answering difficult problems in which computational methods have not been established. In addition, these results argue for the establishment of online communities of practitioners, which could facilitate the solution of future engineering challenges. However, the results also suggest that the practitioner crowds struggle to evaluate open-ended conceptual design problems at a level that experts do. An open research question is thus the utility of crowdsourcing for problems involving open-ended synthesis.

Acknowledgment

We would like to thank authors of [5] for making their data publicly available.

References

[1] Galton, F., 1907, "The Ballot-Box," Nature, **75**(1952), pp. 509–510.
[2] Hooker, R. H., 1907, "Mean or Median," Nature, **75**, pp. 487–488.
[3] Surowiecki, J., 2005, *The Wisdom of Crowds*, Anchor, New York.
[4] Wah, C., 2006, *Crowdsourcing and Its Applications in Computer Vision*, University of California, San Diego, CA.
[5] 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," J. Mech. Des., **137**(3), p. 031101.
[6] Summers, J. D., and Shah, J. J., 2010, "Mechanical Engineering Design Complexity Metrics: Size, Coupling, and Solvability," J. Mech. Des., **132**(2), p. 021004.
[7] Yang, M. C., 2010, "Consensus and Single Leader Decision-Making in Teams Using Structured Design Methods," Des. Stud., **31**(4), pp. 345–362.
[8] Gurnani, A., and Lewis, K., 2008, "Collaborative, Decentralized Engineering Design at the Edge of Rationality," J. Mech. Des., **130**(12), 121101.
[9] Takai, S., 2010, "A Game-Theoretic Model of Collaboration in Engineering Design," J. Mech. Des., **132**(5), p. 051005.
[10] 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 Syst., **255**, pp. 115–127.
[11] Hong, L., and Page, S. E., 2004, "Groups of Diverse Problem Solvers Can Outperform Groups of High-Ability Problem Solvers," Proc. Natl. Acad. Sci. U.S.A., **101**(46), pp. 16385–16389.
[12] Lorenz, J., Rauhut, H., Schweitzer, F., and Helbing, D., 2011, "How Social Influence Can Undermine the Wisdom of Crowd Effect," Proc. Natl. Acad. Sci. U.S.A., **108**(22), pp. 9020–9025.
[13] Lorenz, K., Jones, J., Wimpenny, D., and Jackson, M., 2015, "A Review of Hybrid Manufacturing," Solid Freeform Fabrication Symposium (SFF), Austin, TX, Aug., pp. 10–12.
[14] Berinsky, A. J., Huber, G. A., and Lenz, G. S., 2012, "Evaluating Online Labor Markets for Experimental Research: Amazon.com's Mechanical Turk," Polit. Anal., **20**(3), pp. 351–368.
[15] Kittur, A., Chi, E. H., and Suh, B., 2008, "Crowdsourcing User Studies With Mechanical Turk," Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, ACM, pp. 453–456.
[16] Kittur, A., Nickerson, J. V., Bernstein, M., Gerber, E., Shaw, A., Zimmerman, J., Lease, M., and Horton, J., 2013, "The Future of Crowd Work," Proceedings of the 2013 Conference on Computer Supported Cooperative Work, ACM, pp. 1301–1318.
[17] Rzeszotarski, J. M., and Kittur, A., 2011, "Instrumenting the Crowd: Using Implicit Behavioral Measures to Predict Task Performance," Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology, ACM, pp. 13–22.
[18] Panchal, J. H., Sha, Z., and Kannan, K. N., 2017, "Understanding Design Decisions Under Competition Using Games With Information Acquisition and a Behavioral Experiment," J. Mech. Des., **139**(9), 091402.
[19] Ball, Z., and Lewis, K., 2018, "Observing Network Characteristics in Mass Collaboration Design Projects," Des. Sci., **4**, p. e4.
[20] Burnap, A., Gerth, R., Gonzalez, R., and Papalambros, P. Y., 2017, "Identifying Experts in the Crowd for Evaluation of Engineering Designs," J. Eng. Des., **28**(5), pp. 317–337.
[21] 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," ASME 2015 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference, American Society of Mechanical Engineers.
[22] Orbay, G., Fu, L., and Kara, L. B., 2015, "Deciphering the Influence of Product Shape on Consumer Judgments Through Geometric Abstraction," J. Mech. Des., **137**(8), 081103.
[23] Ghosh, D. D., Olewnik, A., and Lewis, K. E., 2017, "An Integrated Framework for Predicting Consumer Choice Through Modeling of Preference and Product Use Data," ASME 2017 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference, American Society of Mechanical Engineers.
[24] Morgan, H., Levatti, H., Sienz, J., Gil, A., and Bould, D., 2014, "Ge Jet Engine Bracket Challenge: A Case Study in Sustainable Design," Sustain. Des. Manuf. Part 1), pp. 95–107.
[25] 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," J. Mech. Des., **135**(2), 021007.
[26] 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," J. Mech. Des., **136**(10), 101102.
[27] 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," Des. Stud., **35**(3), pp. 232–272.
[28] 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," J. Mech. Des., **122**(4), pp. 377–384.
[29] Goucher-Lambert, K., and Cagan, J., 2019, "Crowdsourcing Inspiration: Using Crowd Generated Inspirational Stimuli to Support Designer Ideation," Des. Stud., **61**, pp. 1–29.
[30] Green, M., Seepersad, C. C., and Hölttä-Otto, K., 2014, "Crowd-Sourcing the Evaluation of Creativity in Conceptual Design: A Pilot Study," ASME 2014 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference, American Society of Mechanical Engineers, pp. V007T07A016–V007T07A016.
[31] Gosnell, C. A., and Miller, S. R., 2016, "But Is It Creative? Delineating the Impact of Expertise and Concept Ratings on Creative Concept Selection," J. Mech. Des., **138**(2), 021101.

[32] Toh, C. A., Starkey, E. M., Tucker, C. S., and Miller, S. R., 2017, "Mining for Creativity: Determining the Creativity of Ideas Through Data Mining Techniques," ASME 2017 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference, American Society of Mechanical Engineers, pp. V007T06A010–V007T06A010.
[33] Bendsøe, M. P., 1989, "Optimal Shape Design as a Material Distribution Problem," Struct. Optim., **1**(4), pp. 193–202.
[34] Alexander, P., Allen, S., and Dutta, D., 1998, "Part Orientation and Build Cost Determination in Layered Manufacturing," Comput. Aided Des., **30**(5), pp. 343–356.
[35] Nobel-Jørgensen, M., Malmgren-Hansen, D., Bærentzen, J. A., Sigmund, O., and Aage, N., 2016, "Improving Topology Optimization Intuition Through Games," Struct. Multidiscipl. Optim., **54**(4), pp. 775–781.
[36] 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," Advances in Neural Information Processing Systems, pp. 2035–2043.
[37] 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," preprint arXiv:1206.6386.
[38] Welinder, P., Branson, S., Belongie, S. J., and Perona, P., 2010, "The Multidimensional Wisdom of Crowds," NIPS, Vol. **23**, pp. 2424–2432.
[39] Lakshminarayanan, B., and Teh, Y. W., 2013, "Inferring Ground Truth From Multi-Annotator Ordinal Data: A Probabilistic Approach," preprint arXiv:1305.0015.
[40] Wauthier, F. L., and Jordan, M. I., 2011, "Bayesian Bias Mitigation for Crowdsourcing," Advances in Neural Information Processing Systems, pp. 1800–1808.
[41] Vul, E., and Pashler, H., 2008, "Measuring the Crowd Within Probabilistic Representations Within Individuals," Psychol. Sci., **19**(7), pp. 645–647.
[42] Kendall, M. G., 1955, *Rank Correlation Methods*, Charles Griffin, London.
[43] Linsey, J. S., Wood, K. L., and Markman, A. B., 2008, "Modality and Representation in Analogy," Artif. Intell. Eng. Des. Anal. Manuf., **22**, pp. 85–100.
[44] 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**, pp. 487–515.
[45] Wilson, J. O., Rosen, D., Nelson, B. A., and Yen, J., 2010, "The Effects of Biological Examples in Idea Generation," Des. Stud., **31**(2), pp. 169–186.
[46] Viswanathan, V. K., and Linsey, J. S., 2013, "Design Fixation and Its Mitigation: A Study on the Role of Expertise," J. Mech. Des., **135**(5), 051008.
[47] Cicchetti, D. V., 1994, "Guidelines, Criteria, and Rules of Thumb for Evaluating Normed and Standardized Assessment Instruments in Psychology," Psychol. Assess. **6**(4), pp. 284–290.