Alex Burnap[1]
Design Science,
University of Michigan,
Ann Arbor, MI 48109
e-mail: aburnap@umich.edu

Yi Ren
Research Fellow
Department of Mechanical Engineering,
University of Michigan,
Ann Arbor, MI 48109
e-mail: yiren@umich.edu

Richard Gerth
Research Scientist
National Automotive Center,
TARDEC-NAC,
Warren, MI 48397
e-mail: richard.j.gerth.civ@mail.mil

Giannis Papazoglou
Department of Mechanical Engineering,
Cyprus University of Technology,
Limassol 3036, Cyprus
e-mail: papazogl@umich.edu

Richard Gonzalez
Professor
Department of Psychology,
University of Michigan,
Ann Arbor, MI 48109
e-mail: gonzo@umich.edu

Panos Y. Papalambros
Professor
Fellow ASME
Department of Mechanical Engineering,
University of Michigan,
Ann Arbor, MI 48109
e-mail: pyp@umich.edu

When Crowdsourcing Fails: A Study of Expertise on Crowdsourced Design Evaluation

Crowdsourced evaluation is a promising method of evaluating engineering design attributes that require human input. The challenge is to correctly estimate scores using a massive and diverse crowd, particularly when only a small subset of evaluators has the expertise to give correct evaluations. Since averaging evaluations across all evaluators will result in an inaccurate crowd evaluation, this paper benchmarks a crowd consensus model that aims to identify experts such that their evaluations may be given more weight. Simulation results indicate this crowd consensus model outperforms averaging when it correctly identifies experts in the crowd, under the assumption that only experts have consistent evaluations. However, empirical results from a real human crowd indicate this assumption may not hold even on a simple engineering design evaluation task, as clusters of consistently wrong evaluators are shown to exist along with the cluster of experts. This suggests that both averaging evaluations and a crowd consensus model that relies only on evaluations may not be adequate for engineering design tasks, accordingly calling for further research into methods of finding experts within the crowd. [DOI: 10.1115/1.4029065]

Keywords: crowdsourcing, design evaluation, crowd consensus, evaluator expertise

1 Introduction

Suppose we wish to evaluate a set of vehicle design concepts with respect to attributes that have objective answers. For many of these objective attributes, the "true score" may be determined using detailed physics-based simulations, such as finite-element analysis to evaluate crashworthiness or human mobility modeling to evaluate ergonomics; however, for some objective attributes such as maintainability, physics-based simulation is difficult or not possible at all. Instead, these objective attributes require human input for accurate evaluation.

To obtain evaluations over these objective attributes, one may ask a number of specialists to evaluate the set of vehicle design concepts. This assumes that the requisite expertise is imbued within this group of specialists. Oftentimes though, the expertise to make a comprehensive evaluation is instead scattered over the "collective intelligence" of a much larger crowd of people with diverse backgrounds [1].

Crowdsourced evaluation, or the delegation of an evaluation task to a large and possibly unknown group of people through an open call [2,3], is a promising approach to obtain such design evaluations. Crowdsourced evaluation draws from the pioneering works of online communities, such as Wikipedia, which have shown that accuracy and comprehensiveness are possible in a large crowdsourced setting requiring expertise. Although crowdsourcing has seen recent success in both academic studies [4] and industry applications [5,6], there are limited reference materials on the use of crowdsourced evaluation for engineering design.

In this study, we explore how the expertise of evaluators in the crowd affects crowdsourced evaluation for engineering design, where expertise is defined as the probability that an evaluator gives an evaluation close to the design's true score. The choice of exploring expertise comes from an important lesson in managing successful online community efforts, namely, the need to implement a systematic method of filtering "signal" from "noise" [7]. In a crowdsourced evaluation process, this manifests itself as a need of screening good evaluations from bad evaluations, in particular when we are given a heterogeneous crowd made up of a mixture of expert and nonexpert evaluators. In this case, averaging evaluations from all participants with equal weight will reduce the

[1] Corresponding author.

Contributed by the Design Theory and Methodology Committee of ASME for publication in the Journal of Mechanical Design. Manuscript received April 29, 2014; final manuscript received November 6, 2014; published online January 9, 2015. Assoc. Editor: Jonathan Cagan.

accuracy of the crowd's combined evaluation, also called the *crowd consensus* [8], due to incorrect design evaluations from low-expertise evaluators. Accordingly, a desirable goal is to identify the experts from the rest of the crowd, thus allowing a more accurate crowd consensus by giving their evaluations more weight.

With this goal in mind, we developed and benchmarked a crowd consensus model of the crowdsourced evaluation process using a Bayesian network that does not require prior knowledge of the true scores of the designs or the expertise of each evaluator in the crowd, yet still aims to estimate accurate design scores by identifying the experts within the crowd and overweighing their evaluations. This statistical model links the expertise of evaluators in the crowd (i.e., knowledge or experience for the design being evaluated), the evaluation difficulty of each design (e.g., a detailed 3D model provides more information than a 2D sketch and may therefore be easier for an expert to evaluate accurately), and the true score of each of the designs. It must be noted that this model relies *only* on evaluations from the crowd; i.e., we do not explicitly measure expertise or difficulty; these variables are latent and only implicitly inferred.

This crowd consensus model rests on the key assumption that low-expertise evaluators are more likely to "guess," and are thus more likely to give random evaluations to designs. This assumption is modeled by defining an evaluation as a random variable centered at the true score of the design being evaluated [9]. A graphical representation of the Bayesian network showing these relationships is given in Fig. 1.

The performance of the Bayesian network crowd consensus model versus the baseline method of averaging evaluations is explored through two studies on the same "simple" engineering design evaluation task of rating the strength of a load-bearing bracket [10]. First, we created simulated crowds to generate evaluations for a set of designs. These crowds had a homogeneous or heterogeneous expertise distribution, representing two cases that may be found in a human crowd. Second, we used a human crowd recruited from the crowdsourcing platform Amazon Mechanical Turk [11] and performed a crowdsourced evaluation with the same crowd and task properties as in the simulation.

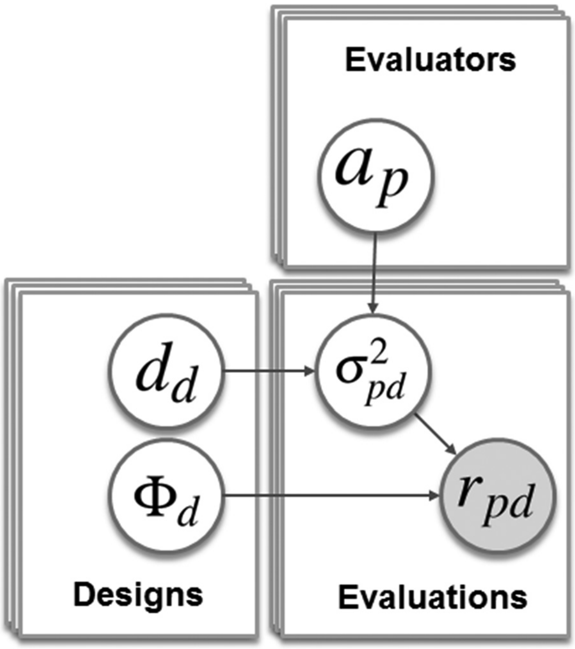

Fig. 1 Graphical representation of the Bayesian network crowd consensus model. This model describes a crowd of evaluators making evaluations r_{pd} that have error from the true score Φ_d. Each evaluator has an expertise a_p and each design has an difficulty d_d. The gray shading on the evaluation r_{pd} denotes that it is the only observed data for this model.

Our results show that we are *not* able to achieve a more accurate design evaluation using the crowd consensus model than just averaging all evaluations. Even for the simple engineering design evaluation task in this study, the modeling assumption that low-expertise evaluators guess more randomly was found not to hold. Upon further investigation, it was found that there exist numerous clusters of "consistently wrong" evaluators that wash out the evaluations from the cluster of experts.

The main contribution of this paper is this finding; namely, that crowdsourced evaluation can fail for even a simple engineering design evaluation task due to the expertise distribution of the crowd. Averaging already gives a low-accuracy estimate of design scores due to the large number of low-expertise evaluators, and a crowd consensus model relying *only* on information about evaluations may not be able to find the experts in the crowd. This study thus serves as justification for further research into methods of finding experts within crowds, particularly when they are shrouded by numerous clusters of consistently wrong nonexperts.

The remainder of this paper is organized as follows. Section 2 reviews relevant research within the engineering design, psychometrics, and crowdsourcing literature, as well as research motivations from industry. Section 3 presents the Bayesian network crowd consensus model and modeling assumptions. Section 4 details the statistical inference scheme of the Bayesian network. Section 5 describes the simulated crowd study and results. Section 6 describes the human crowd study and discusses its results. We conclude in Sec. 7 with implications of this work and opportunities for future research.

2 Related Work

Within the engineering design community, attention is being drawn to the use of crowdsourcing for informing design decisions [12]. Design preferences have been captured using crowdsourced data on social media sites [13,14], as well as through more directed crowdsourced elicitation using online surveys for preference learning [15,16]. Our work differs from these works in that we focus on design evaluation with an objective answer, thus necessitating the estimation of evaluator expertise. Within design evaluation for objective attributes, recent research has used crowdsourcing for idea evaluation [17,18] and creativity evaluation [19]. There also exists much research studying the effect of a single decision maker versus crowd consensus decisions [20,21]. Our work is relevant to these research efforts in that we extend previous findings of the potential limitations of using the entire crowd for design evaluation.

Modeling the crowdsourced evaluation process exists in the literature extending at least back to Condorcet [22], with foundational contributions from the psychometrics community under item response theory [23] and Rasch models [24]. These models have been applied to standardized tests, with several extensions to include hierarchical structure [25] similar to the crowd consensus model in this work. Additional foundational literature from econometrics includes "mechanism design" such as prediction markets and peer prediction [26,27]. For simplicity, we do not consider important findings and approaches from this econometrics literature, instead assuming all evaluators give truthful evaluations and are similarly incentivized by a fixed-sum payment.

More recently, the crowdsourcing community has developed numerous crowd consensus models capturing the expertise of evaluators in a crowdsourced evaluation process [8]. Many of these models are qualitatively similar, with differences in the treatment of evaluator bias [28–30], form of the likelihood function (e.g., ordinal, ranking, binary) [31], extent to which the true score is known [32], and methods of scaling up to larger data sets [30,33]. These models are most often applied to tasks that are "human easy, computer hard," such as image annotation [30,34], planning and scheduling [35], and natural language processing

[36,37]. Our study is also qualitatively similar to this literature, but with a key difference on the application to an engineering design task and the subsequent distribution of expertise in the crowd.

Specifically, many of these recent crowdsourced evaluation efforts are applied to tasks in which a majority of evaluators within the crowd have the expertise to give an accurate evaluation (e.g., does this image contain a "duck"?) [8]. As a result, either averaging or taking a majority vote of the crowd's evaluators is often already quite accurate [38]. For these cases, expertise may often represent the notion of task consistency and attentiveness, with low-expertise evaluators being more spammy or malicious [30].

In contrast, many engineering design tasks may require expertise that only exists in a sparse minority of the crowd. This notion is supported by prior industrial applications of crowdsourced evaluation for engineering design. The Fiat Mio was a fully crowdsourced vehicle design concept, yet the large number of low-expertise submissions resulted in Fiat using its design and engineering teams as a filter without the use of algorithmic assistance [39]. Local Motors Incorporated developed the Rally Fighter using a crowdsourced evaluation system similar to this study, but heavily weighted evaluations of the internal design team [40]. For these engineering design tasks, the notion of expertise may instead represent specialized knowledge and heuristics necessary to give an accurate evaluation.

3 A Bayesian Network Model for Crowd Consensus

We introduce a crowd consensus model that statistically aggregates the evaluations from the set of evaluators using a Bayesian network to estimate the true design scores. More formally, let the crowdsourced evaluation contain D designs and P evaluators. We denote the true score of design d as $\Phi_d \in [0,1]$, and the evaluation from evaluator p for design d as $\mathbf{R} = \{r_{\text{pd}}\}$, where $r_{\text{pd}} \in [0,1]$. Each design d has an evaluation difficulty d_d, and each evaluator p has an evaluation expertise a_p.

Some significant assumptions we make are highlighted here: (1) Evaluators evaluate designs without systematic biases, i.e., given infinite chances of evaluating one specific design, the average score of the evaluators will converge to the true score of that design regardless of their expertise [9,41]; note that this assumption also implies that no evaluators purposely give bad evaluations; (2) Evaluations are independent, i.e., the evaluation on one design from one evaluator will not be affected by the evaluation made by that evaluator for any other design nor will be affected by the evaluation given by a different evaluator. (3) The expertise of evaluators is constant during the entire evaluation process. (4) All evaluators are fully incentivized and do not exhibit fatigue. Without loss of generality, we consider human evaluations real-valued in the range of zero to one.

The evaluation r_{pd} is modeled as a random variable following a truncated Gaussian distribution around the true performance score Φ_d as detailed by Eq. (1) and shown in Fig. 2(*a*)

$$r_{\text{pd}} \sim \text{Truncated} - \text{Gaussian}\left(\Phi_d, \sigma_{\text{pd}}^2\right), \quad r_{\text{pd}} \in [0,1] \tag{1}$$

The variance of density σ_{pd}^2 is interpreted as the error an evaluator makes when using his or her cognitive processes while evaluating the design and is described by a random variable taking an Inverse-Gamma distribution

$$\sigma_{\text{pd}}^2 \sim \text{Inverse} - \text{Gamma}\left(\alpha_{\text{pd}}, \beta_{\text{pd}}\right) \tag{2}$$

The average evaluation error for a given evaluator on a given design is a function of the evaluator's expertise a_p and the design's difficulty d_d. In addition, this function is sigmoidal to capture the notion that there exists a threshold of necessary background knowledge to make an accurate evaluation. Figure 2(*b*) illustrates this function. We set the first requirement on the evaluator's error random variable using the expectation operator $\mathbb{E}$ as shown below

$$\mathbb{E}\left[\sigma_{\text{pd}}^2\right] = \frac{1}{1 + e^{\theta(d_d - a_p) - \gamma}} \tag{3}$$

The random variables θ and γ are introduced as model parameters to allow more flexibility in modeling evaluation tasks and are assumed to be the same for all evaluators and designs: A high value of the scale parameter θ will sharply bisect the crowd into good evaluators with negligible errors and bad evaluators that evaluate almost randomly; the location parameter γ captures evaluation losses intrinsic to the system, such as those stemming from the human–computer interaction.

Next, the variance $\mathbb{V}$ of the evaluator error is considered constant, capturing the notion that, while we hope the major variability in the evaluation error to be captured by Eq. (3), other reasons exist to spread this error, represented by constant C as shown below

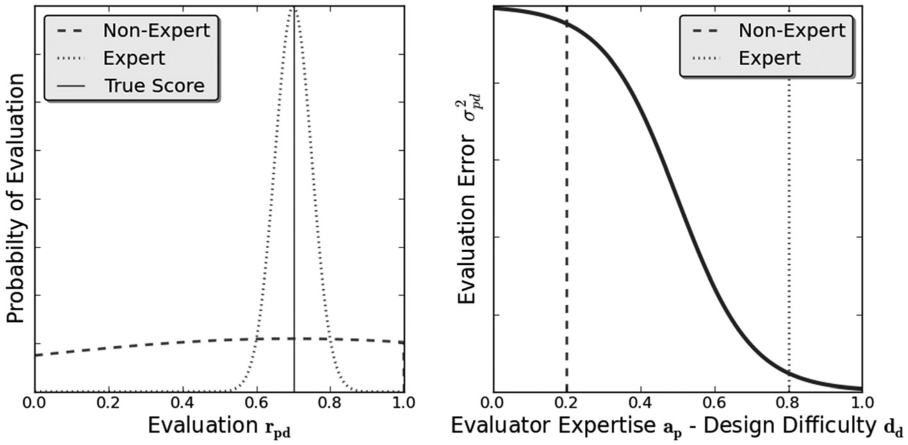

Fig. 2 (*a*) Low evaluation expertise (dashed) relative to the design evaluation difficulty results in an almost uniform distribution of an evaluator's evaluation response, while high evaluation expertise (dotted) results in evaluators making evaluations closer to the true score. (*b*) An evaluator's evaluation error variance σ_{pd}^2 as a function of that evaluator's expertise a_p given some fixed design difficulty d_d and crowd-level parameters θ and γ.

$$\mathbb{V}\left[\sigma_{\mathrm{pd}}^2\right] = \mathrm{C} \tag{4}$$

Following the requirements given by Eqs. (3) and (4), we reparameterize the Inverse-Gamma of Eq. (2) to obtain in the following equations:

$$\alpha_{\mathrm{pd}} = \frac{1}{C\left(1+e^{\theta(d_d-a_p)-\gamma}\right)^2} + 2 \tag{5}$$

$$\beta_{\mathrm{pd}} = \left(\frac{1}{e^{\theta(d_d-a_p)-\gamma}}\right)\left(\frac{1}{Ce^{2\theta(d_d-a_p)-2\gamma}} + 1\right) \tag{6}$$

The hierarchical random variables of the evaluator's evaluation expertise a_p and the design's evaluation difficulty d_d are both restricted to the range [0,1]. We let their distributions be truncated Gaussians with parameters μ_a, σ_a^2, μ_d, σ_d^2 set globally for all evaluators and designs as shown in the following equations:

$$a_p \sim \mathrm{Truncated} - \mathrm{Gaussian}\left(\mu_a, \sigma_a^2\right), \quad a_p \in [0,1] \tag{7}$$

$$d_d \sim \mathrm{Truncated} - \mathrm{Gaussian}\left(\mu_d, \sigma_d^2\right), \quad d_d \in [0,1] \tag{8}$$

The probability densities over θ and γ are assumed as Gaussian with parameters $\mu_\theta, \sigma_\theta^2, \mu_\gamma, \sigma_\gamma^2$ as shown in the following equations:

$$\theta \sim \mathrm{Gaussian}\left(\mu_\theta, \sigma_\theta^2\right) \tag{9}$$

$$\gamma \sim \mathrm{Gaussian}\left(\mu_\gamma, \sigma_\gamma^2\right) \tag{10}$$

Finally, by combining all random variables described in this section, we obtain the joint probability density function as shown below

$$p(\mathbf{a}, \mathbf{d}, \Phi, \mathbf{R}, \theta, \gamma) = p(\theta)p(\gamma)\prod_{p=1}^{P} p(a_p) \times \prod_{d=1}^{D} p(r_{\mathrm{pd}}|a_p, d_d, \theta, \gamma, \Phi_d)p(d_d)p(\Phi_d) \tag{11}$$

Note that all hyperparameters are implicitly included.

4 Estimation and Inference of the Bayesian Network

The Bayesian network crowd consensus model is built upon the following random variables: Evaluators' expertises $\{a_p\}_{p=1}^P$, designs' difficulties $\{d_d\}_{d=1}^D$, true scores of designs $\{\Phi_d\}_{d=1}^D$, and parameters θ, γ, μ_a, σ_a^2, μ_d, σ_d^2. This section explains the settings for inferring the random variables and estimating the parameters using the observed evaluations of the evaluators $\mathbf{R} = \{r_{\mathrm{pd}}\}_{p=1,\ldots,P;d=1,\ldots,D}$.

Two techniques are used in sequence. Maximum a posteriori estimation is performed using Powell's conjugate direction algorithm [42], a derivative-free optimization method, to get initial estimates of the parameters that maximize Eq. (11). These point estimates are then used to initiate an adaptive Metropolis–Hastings Markov Chain Monte Carlo (MCMC) algorithm [43–45] that determines the estimates of all unknown parameters and infers posterior distributions of the random variables. The posterior sample size of the single-chained MCMC simulation is set to 2×10^5, thinned by a factor of 2, with the first half discarded as burn-in.

5 Simulated Crowd Study

We now study how the expertise distribution of the crowd affects the crowdsourced evaluation process using Monte Carlo simulations. There are two main goals of this study. First, we want to understand how crowds made up of different mixtures of high and low-expertise evaluators affect the crowd's combined scores of designs and the subsequent evaluation error from the true scores of the designs. Second, we want to understand the performance differences between the Bayesian network and by averaging. Specifically of interest are the conditions under which the Bayesian network is able to find the subset of high-expertise evaluators within the crowd so that it can give greater weight to their responses.

There are two crowd expertise distribution cases we test, as shown in Fig. 3. Case I is that of a homogeneous crowd, where all evaluators making up the crowd have similar expertise. The varied parameter in the homogenous case is the average expertise of the crowd, thus testing the question: How well can a crowd perform if no individual evaluator can evaluate correctly or, alternatively, if every evaluator can evaluate correctly? Case II is that of a heterogeneous crowd, where the crowd is made up of a mixture of high and low-expertise evaluators. In this case, we fix the average expertise of the crowd to be low, so that most evaluators cannot evaluate designs correctly. Instead, the varied parameter in the heterogeneous case is the variance of the crowd's expertise distribution. This tests the question: How well can a crowd perform as a function of its proportion of high-expertise to low-expertise evaluators?

The procedure for these studies is to use the Monte Carlo simulation environment to: (1) generate a crowd made up of evaluators with expertise drawn from the tested expertise distribution (Case I or II), and a set of designs with true scores unknown to the crowd; (2) simulate the evaluation process by generating a rating between

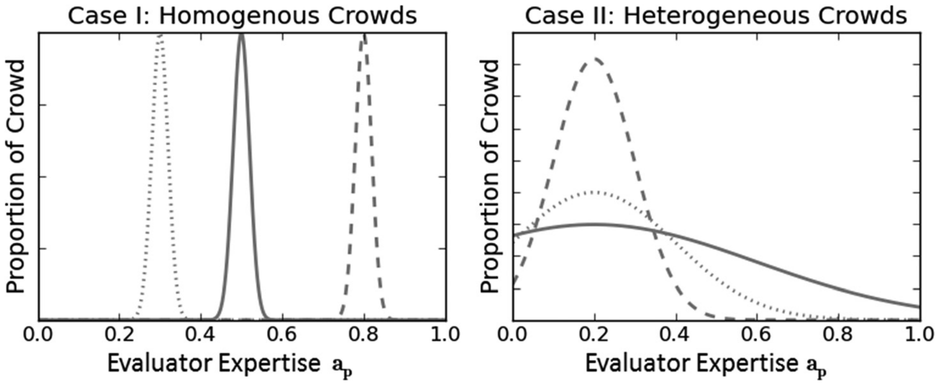

Fig. 3 Crowd expertise distributions for Cases I and II that test how the expertise of evaluators within the crowd affect evaluation error for homogeneous and heterogeneous crowds, respectively. Three possible sample crowds are shown for both cases.

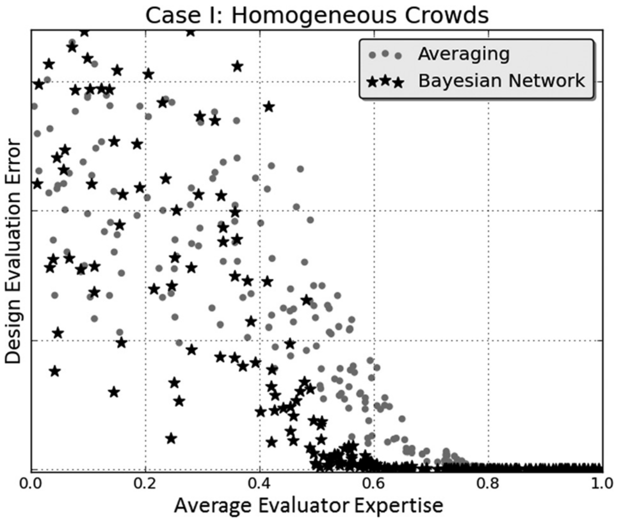

Fig. 4 Case I: Design evaluation error from the averaging and Bayesian network methods as a function of average evaluator expertise for homogeneous crowds. This plot shows that, when dealing with homogeneous crowds, aggregating the set of evaluations into the crowd's consensus score only sees marginal benefits from using the Bayesian network around 0.4–0.7 range of evaluator expertise.

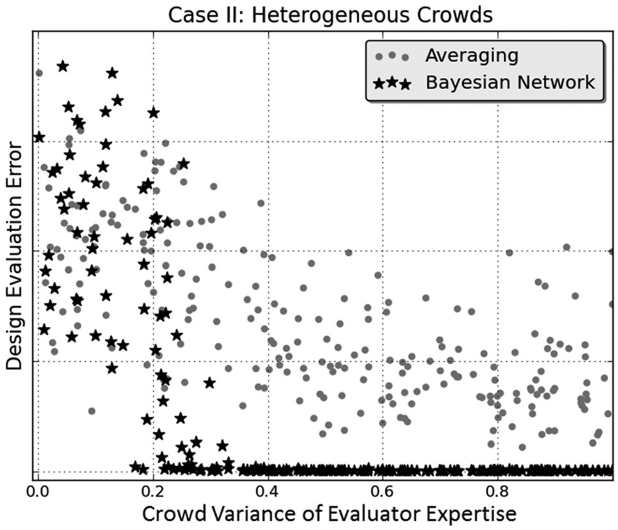

Fig. 5 Case II: Design evaluation error over a set of designs for a mixed crowd with low average evaluation expertise. With increasing crowd variance of expertise there is an increasingly higher proportion of high-expertise evaluators present within the crowd. This leads to a point where the Bayesian network is able to identify the cluster of high-expertise evaluators, upon which evaluation error drops to zero.

1 and 5 that each evaluator within the crowd gives to each design; (3) combine the evaluator-level ratings into the crowd's combined score for each design using either the Bayesian network or by averaging; and (4) calculate the evaluation error between the true scores of the designs and the combined scores from either the Bayesian network or by averaging.

The simulation setup for these studies consisted of 60 evaluators per crowd, as well as eight designs with scores drawn uniformly from the range [0,1] and evaluation difficulties $\{d_d\}$ set at 0.5 for all designs. The evaluation process for each evaluator is to rate all eight designs in the continuous interval [1,5] according to a deterministic equation given by the right hand side of Eq. (3), with the location parameter γ set at 0 and the scale parameter θ set at 0.1. After the crowd's combined scores are obtained, either by the Bayesian network or by averaging, the evaluation error between the combined scores $\hat{\Phi}_d$ and the true scores is calculated using the mean-squared error (MSE) metric as shown below

$$\mathrm{MSE} = \frac{1}{D}\sum_{d=1}^{D}\left(\hat{\Phi}_d - \Phi_d\right)^2 \tag{12}$$

The results of Case I are shown in Fig. 4. Each data point represents a distinct simulated crowd with average expertise given on the x-axis, and associated design evaluation error between the overall estimated score and the true scores on the y-axis. All crowds in Case I were generated using the same narrow crowd expertise variance $\sigma_a = 0.1$ to create homogeneous crowds. The results show that if the average evaluator expertise is relatively high, both averaging and the Bayesian network perform similarly with small design evaluation error. In contrast, when the average expertise is relatively low, neither averaging nor the Bayesian network can estimate the true scores very well. Note that around an average evaluator expertise of 0.4–0.7, the Bayesian network performs marginally better than averaging.

This observation agrees with intuition. A group of evaluators where "no one has the expertise" to evaluate a set of designs should not collectively have the expertise to evaluate a set of designs just by changing the relative weightings of evaluators and their individual evaluation responses upon combination when determining the crowd's combined score. Similarly, a group of evaluators where "everyone has the expertise" to evaluate a set of designs should perform well regardless of the relative weighting between evaluators. The key result for Case I is this: When the crowd has a homogeneous distribution of evaluator expertise, it does not significantly matter which weighting scheme one assigns between various evaluators and their evaluations; the Bayesian network and averaging will perform similarly to each other.

The results of Case II are shown in Fig. 5. Contrary to Case I, distinct crowds represented by each data point have on average the same expertise $\mu_a = 0.2$. Moving right along the x-axis designates crowds with increasingly higher proportions of high-expertise evaluators within the crowd. We observe that the Bayesian network performs much better than averaging after a certain point on the x-axis; the point where a sufficient number of high-expertise evaluators is contained within the crowd. Under these conditions, the Bayesian network identifies the small group of experts from the less competent crowd and weighs their evaluation more than the rest, thus leading to combined scores much closer to the true scores of the designs. This observation is not present when the crowd does not have the sufficient number of high-expertise evaluators within the crowd. When this occurs, as is shown on the left side of the x-axis, the situation of no one has the expertise is recreated from Case I.

In summary, we created simulated crowds to test the influence of crowd expertise on the crowdsourced evaluation process. Two cases were tested, representing homogeneous and heterogeneous expertise distributions. Under the modeling assumptions described in Sec. 3, we find that: (1) when the crowd is homogeneous, it does not matter what weighting scheme is used, as both averaging and the Bayesian network give similar results; (2) when the crowd is heterogeneous, the Bayesian network is able to output the crowd's combined score much closer to the true scores under the condition that a sufficient number of "expert" evaluators exist within the crowd.

6 Human Crowd Study

In this section, we test the performance of the Bayesian network crowd consensus model as compared with averaging using an engineering design evaluation task and a real human crowd. The evaluation task was chosen to be a simple classic structural design problem for a load-bearing bracket [10], in which evaluators are

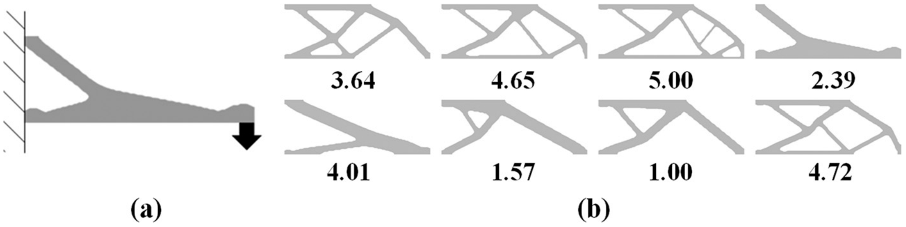

Fig. 6 (*a*) Boundary conditions for bracket strength evaluation and (*b*) the set of all eight bracket designs

asked to rate the capabilities of bracket designs to carry a vertical load as shown in Fig. 6.

Participants: The human crowd consisted of 181 evaluators recruited using the crowdsourcing platform Amazon Mechanical Turk [11]. For the bracket designs, eight bracket topologies were generated using the same amount of raw material. The deformation induced by tensile stress upon vertical loading of each bracket was calculated in OptiStruct [46]. The strength of a bracket was defined as the amount of deformation under a common load and was subsequently scaled linearly between 1 and 5 as labeled in Fig. 6. The scaled strength values were considered as the true scores, which were later used to calculate evaluation errors from the estimations from either the Bayesian network or averaging methods.

Procedure: The evaluation process for each evaluator was as follows: The eight bracket designs were first presented all together to the evaluator, who was then asked to review these designs to get an overall idea of their strengths. After at least 20 s, the evaluator was allowed to continue to the next stage where the designs were presented sequentially and in random order. For each design, the evaluator was asked to evaluate its strength using a rating between 1 and 5, with 1 being "Very Weak" and 5 being "Very Strong." To gather these data, a website with a database backend was set up that recorded when an evaluator gave an evaluation to a particular bracket design [47].

Data analysis: A preprocessing step was carried out before the data were fed into either the Bayesian network or averaging crowd consensus methods. Specifically, since some evaluators would give ratings all above three while some others tended to give ratings all around three, all evaluations were linearly rescaled to a range of 1–5. It should be noted that while this mapping ensures that everyone gives "1s and 5s," it does not help to remove nonlinear biases in between an evaluator's most extreme evaluations. To calculate design evaluation error, the same MSE metric was used as in the simulated crowd study and as given in Eq. (12).

6.1 Human Crowd Study Results.

The Bayesian network crowd consensus model did *worse* than averaging when estimating the true scores of the bracket designs as shown in Table 1.

According to the simulation results, the Bayesian network can only do worse than averaging if it is not able to find the experts in the crowd. This could happen under either of the following two situations: (1) the modeling assumption made in Sec. 3 holds, namely, that low-expertise evaluators are less consistent (more random) in their evaluations, but there are just no high-expertise evaluators; (2) the modeling assumption is violated, in that there exist low-expertise evaluators consistently wrong in their evaluations. In this situation, the Bayesian network crowd consensus model would mistakenly identify evaluators as having high expertise due to their consistency and overweigh their incorrect evaluations.

Table 1 Mean-squared evaluation error and standard deviation from entire human crowd using averaging and Bayesian network estimation

	Design evaluation error (std.)
Averaging	1.001 (N/A)
Bayesian network	1.728 (0.006)

Visualizing the crowd's expertise distribution: We now show that situation (2) above has occurred; namely, there are indeed consistently wrong evaluators that exist in the human crowd. To show this, we cluster the eight-dimensional human evaluation data to find clusters of similar evaluators, and then flatten these clustered data to two dimensions for visualization. This clustering finds groups of evaluators who give consistent evaluation, regardless of whether such evaluations are correct or incorrect. In other words, members of a cluster were consistent in their evaluations not necessarily to the right or wrong answer, but consistent to others in the cluster.

The clustering algorithm we used is density-based and uses the Euclidean distance metric to identify clusters of evaluators who gave similar evaluations [48]. This clustering method was chosen as it can account for varying clustering sizes, as well as not necessitating that every evaluator belongs to a cluster. The flattening from eight dimensions to two dimensions was done using metric multidimensional scaling.

We see in Fig. 7 that five clusters of similar evaluators were found, while Table 2 gives the evaluation error of each cluster. We find that the cyan cluster is made up of high-expertise expert evaluators, as evidenced by their evaluation error. In contrast, the other four clusters were consistently wrong in their evaluations.

This analysis suggests that finding expert evaluators through an open call is possible even for a task like structural design, in which expertise is sparsely distributed through the crowd. However, while the Bayesian network crowd consensus is a theoretical

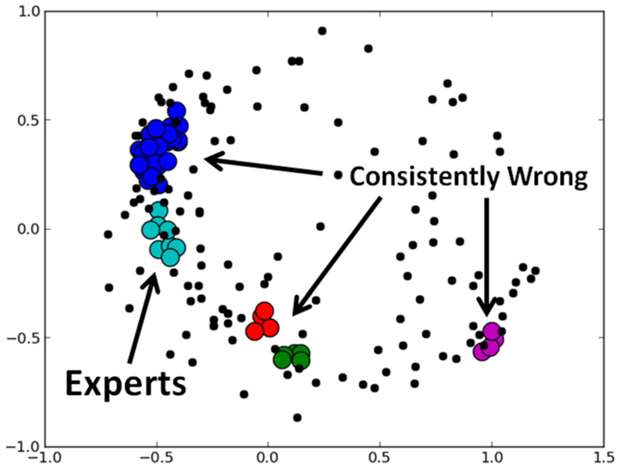

Fig. 7 Clustering of evaluators based on how similar their evaluations are across all eight designs. Each black or colored point represents an individual evaluator, where colored points represent evaluators who were similar to at least 3 other evaluators, and black points represent evaluators who tended to evaluate more uniquely

Table 2 Mean-squared evaluation errors from the five clusters of similarly evaluators

Cluster color	Design evaluation error
Blue	1.415
Cyan experts	0.544
Red	1.652
Green	2.203
Magenta	6.031

way to identify these evaluators, its application in reality is limited by the fact that there exist other (more numerous) clusters of evaluators who are just as consistent yet wrong in their evaluations.

6.2 Followup to Human Crowd Study. For completeness of the human crowd study, we conducted three followup experiments to capture the differences between the simulated crowd assumptions and results and the human crowd results. The first followup experiment augments the human crowd data with simulated experts, in order to offset the consistently wrong evaluators with a larger cluster of experts. The second followup experiment tests the effect of removing the consistently wrong evaluators from the human crowd study. The third followup experiment remains entirely in simulation and shows that the existence of enough consistently wrong evaluators will also cause the Bayesian network crowd consensus to fail to find experts in simulation as well, thus mimicking the results of the human study.

6.2.1 Human Crowd Augmented With Simulated Experts. We show in Fig. 8 how the design evaluation error would be reduced if extra expert evaluators, i.e., evaluators with evaluations exactly the same as true scores, were collected in addition to the original 181 evaluators from the human study. Notice that the error should be reduced monotonically as the number of experts increases. However, the stochastic nature of the estimation process of a Bayesian network could cause suboptimal estimations. Similar to the simulations in Fig. 5, one can observe the phase-changing phenomenon in the change of the design evaluation error. This phase change represents when the Bayesian network is indeed able to find the experts in the crowd. Notice that although adding ten additional experts does not make a majority of the crowd as expert, it is sufficient for the Bayesian network crowd consensus model to locate the experts and subsequently overweigh their evaluations.

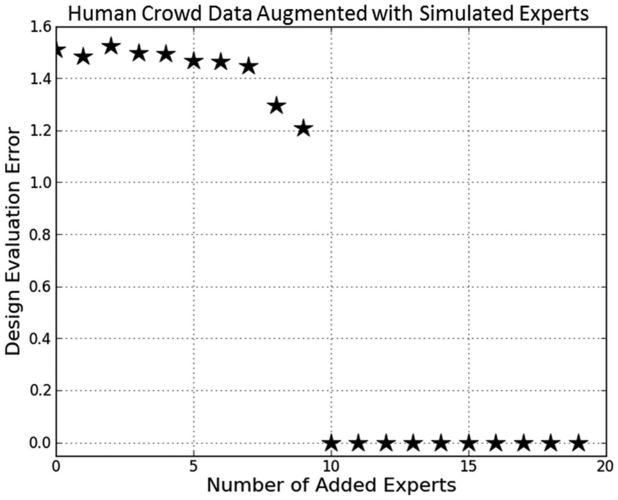

Fig. 8 Design evaluation error with respect to additional experts

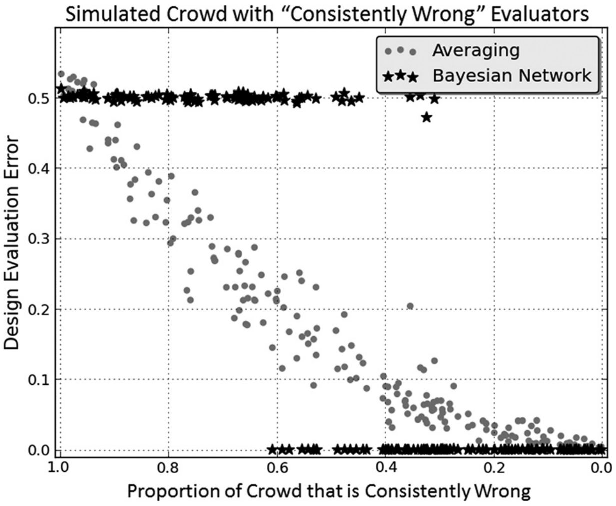

Fig. 9 Design evaluation error with respect to the proportion of the expert group

6.2.2 Human Crowd With Consistently Wrong Evaluators Removed. We address how removing the consistently wrong evaluators affects the crowd's evaluation error, in which the consistently wrong evaluators are those found by clustering as shown in Fig. 7. As reference, averaging the evaluations of the entire crowd results in a MSE of 1.001 as given earlier in Table 1.

Removing the consistently wrong evaluators resulted in a *worse* evaluation error at 1.228 than averaging the entire crowd. This finding suggests that either the consistently wrong evaluators are not as wrong as the nonconsistent nonexperts (i.e., the humans that were not clustered as represented black dots in Fig. 7) or that nonexpert evaluation errors at the design level tend to cancel each other out.

It is found that indeed evaluation errors are being canceled at the design level. This is suggested by finding that the evaluation error of only the nonconsistently wrong (black dots) is 1.339, while the evaluation error of both the consistent and nonconsistently wrong (i.e., all but the experts) is 1.060. Note that the nonconsistently wrong evaluators have an average evaluation error lower than that of any of the consistently wrong evaluators.

This analysis suggests that it is not sufficient, at least as far as this sample goes, to use the Bayesian network crowd consensus model to identify consistently wrong evaluators and simply omit them from the evaluation task. While their evaluations may obscure identification of the experts, they may be useful as they may be also canceling out errors from other evaluators.

6.2.3 Simulated Crowd With Consistently Wrong Evaluators. In this scenario, we tested a set of simulations in which the crowd contained two clusters of evaluators. One cluster, the experts, can always evaluate correctly; the other cluster is almost the same, except that evaluators in this cluster always rate one design consistently wrong by 0.5. We vary the crowd proportion of consistently wrong evaluators from 100% to 0% and calculate the corresponding evaluation errors as shown in Fig. 9. While the error from averaging changes linearly with respect to the proportion, that from the Bayesian network takes only two phases. The result mimics what we saw with the human study; the Bayesian network simply considers one of the clusters as the experts based on the cluster size and spread, regardless of whether the cluster is consistently correct or consistently wrong.

7 Conclusion

Crowdsourcing is a promising method to evaluate engineering design concepts that require human input, due to the possibility of leveraging evaluation expertise distributed over a large number of people. For engineering design tasks, a common characteristic of typical crowdsourced design evaluation processes is that the crowd is composed of a heterogeneous mixture of high and low-expertise evaluators. Simply averaging all evaluations from the crowd results in inaccurate crowd consensus scores for the set of designs, due to the large number of low-expertise evaluators. Consequently, a key challenge in such crowdsourced evaluation processes is to find the subset of expert evaluators in the crowd so that their evaluations may be given more weight.

In this paper, we developed and benchmarked a crowd consensus model in the form of a Bayesian network that aims to find the expert evaluators and subsequently give their evaluations more weight. The key modeling assumption for this crowd consensus model is that low-expertise evaluators tend to guess, resulting in more random evaluations than expert evaluators.

We tested, using both simulated crowds and a human crowd, how the Bayesian network crowd consensus model performs compared to averaging all evaluations for a simple engineering design evaluation task. We showed in simulation that when assumptions hold, the Bayesian network is able to find the experts in the crowd and outperform averaging. However, the results of the human crowd study show that we were *not* able to achieve a more accurate design evaluation using the Bayesian network crowd consensus model than just averaging all evaluations. It was found that there were numerous clusters of consistently wrong evaluators in the crowd, causing the Bayesian network to believe they were the experts, and consequently overweighing their (wrong) evaluations. These results suggest that crowd consensus models that *only* observe evaluations may not be suitable for crowdsourced evaluation tasks for engineering design, contrasting with many of the recent successes from the crowdsourcing literature.

Crowdsourced evaluation can fail for even a simple engineering design evaluation task due to the expertise distribution of the crowd; averaging already gives a low-accuracy estimate of design scores due to the large number of low-expertise evaluators, and crowd consensus models relying only on evaluations may not be able to find the experts in the crowd. Consequently, further research is needed into practical methods to find experts when they are only a small subset of the crowd as well as shrouded by numerous clusters of consistent yet incorrect evaluators.

Promising avenues in this direction may be in extending crowd consensus model to include relevant information to the engineering design evaluation task as has been done with item features [49], evaluator confidence [50], evaluator behavioral measures [51], and expertise assessed over longitudinal tasks [52]. Another useful direction may be in analytic conditions for when experts in the crowd may be found [53–56], possibly in the form of practical questions or tests to run before setting up an entire crowdsourced evaluation process. While this initial step displays potential challenges for crowdsourced evaluation for even simple engineering design tasks, such extended crowd consensus models are likely to benefit a multitude of research communities.

Acknowledgment

This research was partially supported by the National Science Foundation Grant No. CMMI 1266184 and by the Automotive Research Center, a U.S. Army Center of Excellence in Modeling and Simulation of Ground Vehicle Systems headquartered at the University of Michigan. This support was gratefully acknowledged. We additionally thank the four anonymous reviewers for comments and suggestions regarding the contribution of this work.

References

[1] 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.

[2] Estellés-Arolas, E., and González-Ladrón-de Guevara, F., 2012, "Towards an Integrated Crowdsourcing Definition," J. Inf. Sci., **38**(2), pp. 189–200.

[3] Gerth, R. J., Burnap, A., and Papalambros, P., 2012, "Crowdsourcing: A Primer and its Implications for Systems Engineering," 2012 NDIA Ground Vehicle Systems Engineering and Technology Symposium, Troy, MI, Aug. 14–16.

[4] 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, Florence, Italy, Apr. 5–10, pp. 453–456.

[5] Von Ahn, L., Maurer, B., McMillen, C., Abraham, D., and Blum, M., 2008, "Recaptcha: Human-Based Character Recognition via Web Security Measures," Science, **321**(5895), pp. 1465–1468.

[6] Warnaar, D. B., Merkle, E. C., Steyvers, M., Wallsten, T. S., Stone, E. R., Budescu, D. V., Yates, J. F., Sieck, W. R., Arkes, H. R., Argenta, C. F., Shin, Y., and Carter, J. N., 2012, "The Aggregative Contingent Estimation System: Selecting, Rewarding, and Training Experts in a Wisdom of Crowds Approach to Forecasting," Proceedings of the 2012 AAAI Spring Symposium: Wisdom of the Crowd, Palo Alto, CA, Mar. 26–28.

[7] Ipeirotis, P. G., and Paritosh, P. K., 2011, "Managing Crowdsourced Human Computation: A Tutorial," Proceedings of the 20th International World Wide Web Conference Companion, Hyderabad, India, Mar. 28–Apr. 1, pp. 287–288.

[8] Sheshadri, A., and Lease, M., 2013, "Square: A Benchmark for Research on Computing Crowd Consensus," Proceedings of the 1st AAAI Conference on Human Computation and Crowdsourcing, Palm Springs, CA, Nov. 7–9.

[9] Nunnally, J., and Bernstein, I., 2010, *Psychometric Theory 3E*, McGraw-Hill Series in Psychology, McGraw-Hill Education, New York.

[10] Papalambros, P. Y., and Shea, K., 2005, "Creating Structural Configurations," *Formal Engineering Design Synthesis*, E. K. Antonsson, and J. Cagan, eds., Cambridge University, Cambridge, UK, pp. 93–125.

[11] Amazon, 2005, "Amazon Mechanical Turk," http://www.mturk.com

[12] Van Horn, D., Olewnik, A., and Lewis, K., 2012, "Design Analytics: Capturing, Understanding, and Meeting Customer Needs Using Big Data," ASME Paper No. DETC2012-71038.

[13] Tuarob, S., and Tucker, C. S., 2013, "Fad or Here to Stay: Predicting Product Market Adoption and Longevity Using Large Scale, Social Media Data," ASME Paper No. DETC2013-12661.

[14] Stone, T., and Choi, S.-K., 2013, "Extracting Consumer Preference From User-Generated Content Sources Using Classification," ASME Paper No. DETC2013-13228.

[15] Ren, Y., and Papalambros, P. Y., 2012, "On Design Preference Elicitation With Crowd Implicit Feedback," ASME Paper No. DETC2012-70605.

[16] Ren, Y., Burnap, A., and Papalambros, P., 2013, "Quantification of Perceptual Design Attributes Using a Crowd," Proceedings of the 19th International Conference on Engineering Design (ICED13), Design for Harmonies, Vol. 6, Design Information and Knowledge, Seoul, Korea, Aug. 19–22.

[17] Kudrowitz, B. M., and Wallace, D., 2013, "Assessing the Quality of Ideas From Prolific, Early-Stage Product Ideation," J. Eng. Des., **24**(2), pp. 120–139.

[18] Grace, K., Maher, M. L., Fisher, D., and Brady, K., 2014, "Data-Intensive Evaluation of Design Creativity Using Novelty, Value, and Surprise," Int. J. Des. Creativity and Innovation, pp. 1–23.

[19] Fuge, M., Stroud, J., and Agogino, A., 2013, "Automatically Inferring Metrics for Design Creativity," ASME Paper No. DETC2013-12620.

[20] Yang, M. C., 2010, "Consensus and Single Leader Decision-Making in Teams Using Structured Design Methods," Des. Stud., **31**(4), pp. 345–362.

[21] Gurnani, A., and Lewis, K., 2008, "Collaborative, Decentralized Engineering Design at the Edge of Rationality," ASME J. Mech. Des., **130**(12), p. 121101.

[22] de Caritat, M. J. A. N., 1785, *Essai sur l'application de l'analyse à la probabilité des décisions rendues à la pluralité des voix*, L'imprimerie royale, Paris, France.

[23] Lord, F. M., 1980, *Applications of Item Response Theory to Practical Testing Problems*, Erlbaum, Mahwah, NJ.

[24] Rasch, G., 1960/1980, "Probabilistic Models for Some Intelligence and Achievement Tests, Expanded Edition (1980) With Foreword and Afterword by B. D. Wright," *Copenhagen*, Danish Institute for Educational Research, Denmark.

[25] Oravecz, Z., Anders, R., and Batchelder, W. H., 2013, "Hierarchical Bayesian Modeling for Test Theory Without an Answer Key," Psychometrika (published online), pp. 1–24.

[26] Miller, N., Resnick, P., and Zeckhauser, R., 2005, "Eliciting Informative Feedback: The Peer-Prediction Method," Manage. Sci., **51**(9), pp. 1359–1373.

[27] Prelec, D., 2004, "A Bayesian Truth Serum for Subjective Data," Science, **306**(5695), pp. 462–466.

[28] Wauthier, F. L., and Jordan, M. I., 2011, "Bayesian Bias Mitigation for Crowdsourcing," Adv. Neural Inf. Process. Syst., pp. 1800–1808.

[29] 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," Proceedings of the 29th International Conference on Machine Learning, Edinburgh, Scotland, UK, June 26–July 1.

[30] Welinder, P., Branson, S., Belongie, S., and Perona, P., 2010, "The Multidimensional Wisdom of Crowds," Adv. Neural Inf. Process. Syst., **10**, pp. 2424–2432.

[31] Lakshminarayanan, B., and Teh, Y. W., 2013, "Inferring Ground Truth From Multi-Annotator Ordinal Data: A Probabilistic Approach," preprint arXiv 1305.0015.

[32] Tang, W., and Lease, M., 2011, "Semi-Supervised Consensus Labeling for Crowdsourcing," Special Interest Group on Information Retrieval 2011 Workshop on Crowdsourcing for Information Retrieval, Beijing, China, July 28, pp. 1–6.

[33] Liu, Q., Peng, J., and Ihler, A. T., 2012, "Variational Inference for Crowdsourcing," Adv. Neural Inf. Process. Syst., pp. 701–709.

[34] Whitehill, J., Ruvolo, P., Wu, T., Bergsma, J., and Movellan, J. R., 2009, "Whose Vote Should Count More: Optimal Integration of Labels From Labelers of Unknown Expertise," Adv. Neural Inf. Process. Syst., **22**, pp. 2035–2043.

[35] Kim, J., Zhang, H., André, P., Chilton, L. B., Mackay, W., Beaudouin-Lafon, M., Miller, R. C., and Dow, S. P., 2013, "Cobi: A Community-Informed Conference Scheduling Tool," Proceedings of the 26th Annual ACM symposium on User Interface Software and Technology, St Andrews, UK, Oct. 8–11, pp. 173–182.

[36] Snow, R., O'Connor, B., Jurafsky, D., and Ng, A. Y., 2008, "Cheap and Fast—but Is It Good?: Evaluating Non-Expert Annotations for Natural Language Tasks," Proceedings of the Conference on Empirical Methods in Natural Language Processing, Honolulu, HI, pp. 254–263.

[37] Zaidan, O. F., and Callison-Burch, C., 2011, "Crowdsourcing Translation: Professional Quality From Non-Professionals," Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Portland, OR, pp. 1220–1229.

[38] Sheng, V. S., Provost, F., and Ipeirotis, P. G., 2008, "Get Another Label? Improving Data Quality and Data Mining Using Multiple, Noisy Labelers," Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Association for Computing Machinery, Las Vegas, NV, Aug. 24–27, pp. 614–622.

[39] Celaschi, F., Celi, M., and García, L. M., 2011, "The Extended Value of Design: An Advanced Design Perspective," Des. Manage. J., **6**(1), pp. 6–15.

[40] Bommarito, M. F. R., Gong, A., and Page, S., 2011, "Crowdsourcing Design and Evaluation Analysis of DARPA's XC2V Challenge," University of Michigan Technical Report.

[41] Caragiannis, I., Procaccia, A. D., and Shah, N., 2013, "When Do Noisy Votes Reveal the Truth?," Proceedings of the Fourteenth ACM Conference on Electronic Commerce, Philadelphia, PA, June 16–20, pp. 143–160.

[42] Powell, M. J., 1964, "An Efficient Method for Finding the Minimum of a Function of Several Variables Without Calculating Derivatives," Comput. J., **7**(2), pp. 155–162.

[43] Haario, H., Saksman, E., and Tamminen, J., 2001, "An Adaptive Metropolis Algorithm," Bernoulli, **7**(2), pp. 223–242.

[44] Gelfand, A. E., and Smith, A. F., 1990, "Sampling-Based Approaches to Calculating Marginal Densities," J. Am. Stat. Assoc., **85**(410), pp. 398–409.

[45] Patil, A., Huard, D., and Fonnesbeck, C. J., 2010, "PyMC: Bayesian Stochastic Modelling in Python," J. Stat. Software, **35**(4), pp. 1–81.

[46] Schramm, U., Thomas, H., Zhou, M., and Voth, B., 1999, "Topology Optimization With Altair Optistruct," Proceedings of the Optimization in Industry II Conference, Banff, Canada.

[47] University of Michigan—Optimal Design Laboratory, 2013, "Turker Design—Crowdsourced Design Evaluation," http://www.turkerdesign.com.

[48] Ester, M., Kriegel, H.-P., Sander, J., and Xu, X., 1996, "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases With Noise," Knowl. Discovery Data Min., **96**, pp. 226–231.

[49] Raykar, V. C., Yu, S., Zhao, L. H., Valadez, G. H., Florin, C., Bogoni, L., and Moy, L., 2010, "Learning From Crowds," J. Mach. Learn. Res., **11**, pp. 1297–1322.

[50] Prelec, D., Seung, H. S., and McCoy, J., 2013, "Finding Truth Even If the Crowd Is Wrong, Technical Report, Working Paper," MIT.

[51] 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, Santa Barbara, CA, pp. 13–22.

[52] Budescu, D. V., and Chen, E., 2014, "Identifying Expertise to Extract the Wisdom of the Crowds," Management Science (published online) pp. 1–34.

[53] Della Penna, N., and Reid, M. D., 2012, "Crowd & Prejudice: An Impossibility Theorem for Crowd Labelling Without a Gold Standard," Proceedings of 2012 Collective Intelligence Conference, Cambridge, MA, Apr. 18–20.

[54] Waggoner, B., and Chen, Y., 2013, "Information Elicitation Sans Verification," Proceedings of the 3rd Workshop on Social Computing and User Generated Content, Philadelphia, PA, June 16.

[55] Davis-Stober, C. P., Budescu, D. V., Dana, J., and Broomell, S. B., 2014, "When Is a Crowd Wise?," Decision, **1**(2), pp. 79–101.

[56] Kruger, J., Endriss, U., Fernández, R., and Qing, C., 2014, "Axiomatic Analysis of Aggregation Methods for Collective Annotation," Proceedings of the 2014 International Conference on Autonomous Agents and Multi-Agent Systems, Paris, France, May 5–9, pp. 1185–1192.