A Study of Crowd Ability and its Influence on Crowdsourced Evaluation of Design Concepts

Alex Burnap[*]
Ph.D. Candidate
Design Science
University of Michigan
Ann Arbor, Michigan
Email: aburnap@umich.edu

Yi Ren
Research Fellow
Department of Mechanical Engineering
University of Michigan
Ann Arbor, Michigan
Email: yiren@umich.edu

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

Giannis Papazoglou
Visiting Scholar
Department of Mechanical Engineering
Cyprus University of Technology
Limassol, Cyprus
Email: papazoglou@umich.edu

Rich Gonzalez
Professor
Department of Psychology
University of Michigan
Ann Arbor, Michigan
Email: gonzo@umich.edu

Panos Y. Papalambros
Professor, Fellow of ASME
Department of Mechanical Engineering
University of Michigan
Ann Arbor, Michigan
Email: pyp@umich.edu

Crowdsourced evaluation is a promising method of evaluating attributes of a design that require human input, such as maintainability of a vehicle. The challenge is to correctly estimate the design scores using a massive and diverse crowd, particularly when only a minority of evaluators give correct evaluations. As an alternative to simple averaging, this paper introduces a Bayesian network approach that models the human evaluation process and estimates design scores, taking human abilities in evaluating the design into account. Simulation results indicate that the proposed method is preferred to averaging since it identifies the experts from the crowd, under the assumptions that (1) experts do exist and (2) only experts have consistent evaluations. These assumptions, however, do not always hold as indicated by the results of a human study. Clusters of consistent yet incorrect human evaluators are shown to exist along with the cluster of experts. This suggests that additional data such as evaluators' background are needed to isolate the correct clusters of experts for design evaluation tasks .

Keywords: *crowdsourcing, design evaluation, sparse evaluation ability, machine learning*

1 Introduction

Suppose we wish to evaluate a set of military vehicle design concepts with respect to objective mission performance attributes. For many objective attributes, the "true score" may be determined using detailed physics-based simulations, such as finite-element analysis to evaluate blast resistance or human mobility modeling to evaluate ergonomics; however, for some objective attributes, such as situational awareness, 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 the requisite ability is imbued within this group of specialists. Oftentimes though, the ability 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 unknown group of people, is a promising approach to obtain such design evaluations. Crowdsourced evaluation draws from the pioneering works of online communities, like Wikipedia, which have shown that accuracy and comprehensiveness are possible in a large

[*] Corresponding author.

Report Documentation Page	*Form Approved* *OMB No. 0704-0188*

1. REPORT DATE **01 MAY 2014**	2. REPORT TYPE **Journal Article**	3. DATES COVERED **09-02-2014 to 16-04-2014**

4. TITLE AND SUBTITLE **A Study of Crowd Ability and its Influence on Crowdsourced Evaluation of Design Concepts**	5a. CONTRACT NUMBER **W56HZV-04-2-0001**
	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S) **Alex Burnap; Yi Ren; Richard Gerth; Giannis Papazoglou; Rich Gonzalex**	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Department of Mechanical Engineering,University of Michigan,2300 Hayward St,Ann Arbor,Mi,48109**	8. PERFORMING ORGANIZATION REPORT NUMBER **; #24704**
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES) **U.S. Army TARDEC, 6501 East Eleven Mile Rd, Warren, Mi, 48397-5000**	10. SPONSOR/MONITOR'S ACRONYM(S) **TARDEC**
	11. SPONSOR/MONITOR'S REPORT NUMBER(S) **#24704**

12. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

Submitted to Journal of Mechanical Design 2014

14. ABSTRACT

Crowdsourced evaluation is a promising method of evaluating attributes of a design that require human input, such as maintainability of a vehicle. The challenge is to correctly estimate the design scores using a massive and diverse crowd, particularly when only a minority of evaluators give correct evaluations. As an alternative to simple averaging, this paper introduces a Bayesian network approach that models the human evaluation process and estimates design scores, taking human abilities in evaluating the design into account. Simulation results indicate that the proposed method is preferred to averaging since it identifies the experts from the crowd, under the assumptions that (1) experts do exist and (2) only experts have consistent evaluations. These assumptions, however, do not always hold as indicated by the results of a human study. Clusters of consistent yet incorrect human evaluators are shown to exist along with the cluster of experts. This suggests that additional data such as evaluators- background are needed to isolate the correct clusters of experts for design evaluation tasks.

15. SUBJECT TERMS

crowdsourcing, design evaluation, sparse evaluation ability, machine learning

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**	**Public Release**	**10**	

crowdsourced setting. Although many successful online communities exist, there are limited reference materials on the use of crowdsourced evaluation for engineering design.

In this study, we explore how the ability of *evaluators* in the crowd affects the crowdsourced evaluation process, where *ability* is defined as the probability that a participant gives an *evaluation* close to the design's *true score*. The choice of exploring ability comes from an important lesson from successful online community efforts, namely, the need to implement a systematic method of filtering "signal" from "noise" [2]. 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 high-ability and low-ability participants. In this case, averaging evaluations from all participants with equal weight will reduce the accuracy of the crowd's combined evaluation due to bad evaluations from low-ability participants. Accordingly, a desirable goal is to identify the high-ability participants from the rest of the crowd, as their "signal" will be closer to the true scores of the designs, and their evaluations may be subsequently given more weight.

To achieve this goal, we statistically model the crowdsourced evaluation process with a Bayesian network that does not require prior knowledge of the true scores of the designs or of the ability of each evaluator in the crowd, yet still aims to identify the high-ability participants within the crowd. This model links the ability 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. The model rests on the key assumption that low-ability evaluators are more likely to "guess," and while guessing, to evaluate designs more randomly. This assumption is modeled by defining an evaluation be a random variable centered at the true score of the design being evaluated [3]. A graphical representation of the Bayesian network showing these relationships is given in Figure 1.

The performance of the Bayesian network versus the baseline method of Averaging were explored through two studies. First, we created simulated crowds to generate evaluations for a set of designs. These crowds had a homogeneous or heterogeneous ability distribution, representing two cases that may be found in a human crowd. Second, we used a human crowd recruited from the crowdsourcing platform Amazon's Mechanical Turk [4], and performed a crowdsourced evaluation with the same crowd and task properties as in the simulation.

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

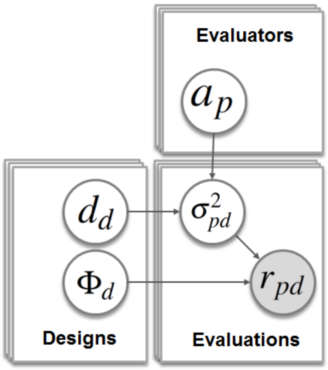


Fig. 1. Graphical representation of the Bayesian network model. This model describes a crowd of evaluators making evaluations r_{pd} that have error from the true score Φ_d. Each evaluator has an ability a_p and each design has an difficulty d_d. The gray shading on the evaluation r_{pd} denotes that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

2 Related Work

Within the engineering design community, attention is being drawn to the use of crowdsourcing for better informing subjective design decisions [5]. Methods using publically accessible crowdsourced data from social media sites have been used for preference learning [6, 7]. More directed crowdsourced evaluation with online surveys have been also used for idea evaluation [8], creativity evaluation [9], and aesthetic preference learning [10]. Our work differs from these works in that we focus on an objective task, thus necessitating the estimation of evaluator ability.

Much literature modeling the ability of evaluators in a crowd exists from the psychometrics community under Item Response Theory [11] and Rasch Models [12]. These models have been applied to standardized tests, with several extensions to include hierarchical structure [13] similar to this study's model. More recently, the human-computer interaction, machine learning, and crowdsourcing communities have modeled the ability of evaluators in a crowdsourced evaluation process for various tasks. These tasks are typically "human easy, computer hard," such as image annotation [14, 15], planning and scheduling [16], and natural language processing [17, 18].

Many of these models are qualitatively similar, with differences in the treatment of evaluator bias [15, 19, 20], form of the likelihood function (e.g., ordinal, ranking, binary) [21], extent to which the true score is known [22], and methods of scaling up to larger datasets [15, 23]. Our study is also qualitatively similar to this literature, but with a key difference on the application to an engineering design task and its subsequent distribution of ability in the crowd.

Specifically, many of these recent crowdsourced evaluation tasks have a majority of evaluators with the ability to give an accurate evaluation (e.g., how many animals are in this image?) [24]. As a result, either averaging or taking a majority vote of the crowd's evaluators is often already quite accurate [25]. For these cases, ability often represents the notion of task consistency and attentiveness, with low-ability evaluators being more spammy or malicious [15].

In contrast, engineering design tasks may require ability that is more sparsely scattered amongst the crowd. This 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-ability submissions resulted in Fiat using its design and engineering teams as a filter without the use of algorithmic assistance [26]. Local Motors Incorporated developed the Rally Fighter using a crowdsourced evaluation system similar to this study, but strongly weighted evaluations of the internal design team [27]. For these engineering design tasks, the notion of ability instead may represent specialized knowledge and heuristics necessary to give an accurate evaluation.

3 A Bayesian Network Model for Human Evaluations

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 $\mathbf{R} = \{r_{pd}\}$ where $r_{pd} \in [0,1]$. Each design d has an evaluation difficulty d_d, and each evaluator p has an evaluation ability a_p. Some significant assumptions we made shall be highlighted here: (1) We assume that 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 ability [3]; note that this assumption also implies that no evaluators purposely give bad evaluations; (2) we assume that evaluation responses are independent, i.e., the evaluation on one design from one user will not be affected by the evaluation made by that user for any other design, nor will it be affected by the evaluation given by a different user; (3) we assume that the ability of evaluators is constant during the entire evaluation process; (4) we assume that 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 evaluator 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 Figure 2a.

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

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:

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

The average evaluation error for a given evaluator on a given design is a function of the evaluator's ability 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 2b illustrates this function. We set the first requirement on the evaluator's error random variable using the expectation operator $\mathbb{E}$ in Eq. (3).

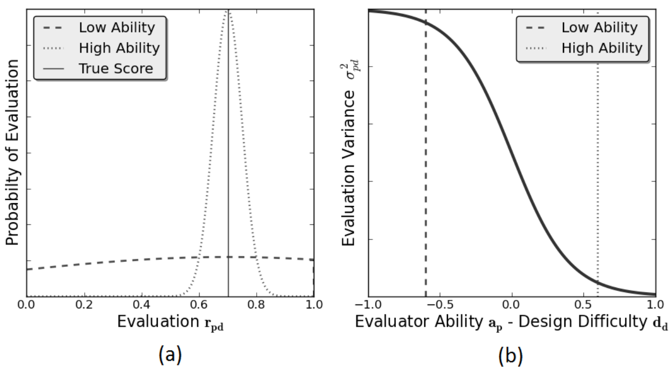


Fig. 2. (a) Low evaluation ability (dashed) relative to the design evaluation difficulty results in an almost uniform distribution of an evaluator's evaluation response, while high evaluation ability (dotted) results in evaluators making evaluations closer to the true score. (b) An evaluator's evaluation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

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

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.

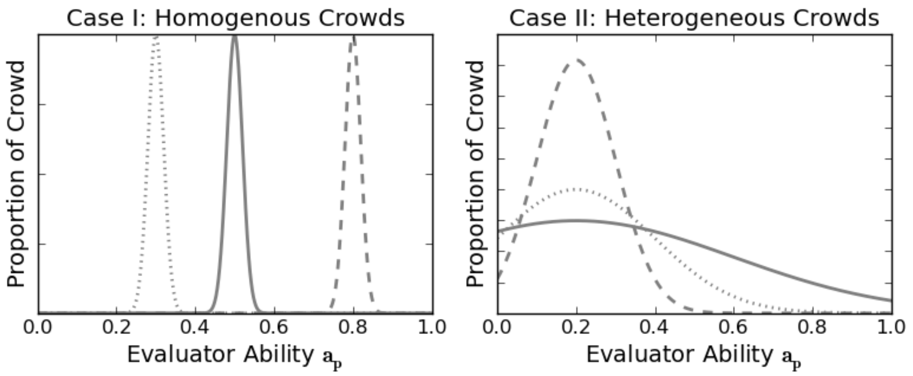
Next, the variance $\mathbb{V}$ of the evaluator error is considered constant, capturing the notion that, while we hope the major variability in the evaluation error to be captured by Equation (3), other reasons exist to spread this error, represented by constant C in Eq. (4).

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

Following the requirements given by Eq. (3) and (4), we reparameterize the Inverse-Gamma of Eq. (2) to obtain Eq. (5) and (6).

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

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



Case	Type of Crowd	Varied Parameter	Figure	Number of Crowd Simualtions
I	Homogeneous Crowd	Average Crowd Ability	4	250
II	Heterogeneous Crowd	Variance of Crowd Ability	5	250

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

The hierarchical random variables of the evaluator's evaluation ability 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 Eq. (7) and (8).

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

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

The probability densities over θ and γ are assumed as Gaussian with parameters μ_θ, σ_θ^2, μ_γ, σ_γ^2 as shown in Eq. (9) and (10).

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

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

Finally, by combining all random variables described in this section, we obtain the joint probability density function shown in Eq. (11). Note that all hyperparameters are implicitly included.

$$p\left(\mathbf{a}, \mathbf{d}, \Phi, \mathbf{R}, \theta, \gamma\right) = \qquad (11)$$

$$p(\theta)p(\gamma)\prod_{p=1}^{P} p(a_p) \prod_{d=1}^{D} p(r_{pd}|a_p, d_d, \theta, \gamma, \Phi_d)p(d_d)p(\Phi_d)$$

4 Estimation and Inference of the Bayesian Network

The proposed Bayesian network model is built upon the following random variables: evaluators' abilities $\{a_p\}_{p=1}^{P}$, 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 infering the random variables and estimating the parameters using the observed evaluations of the evaluators $\mathbf{R} = \{r_{pd}\}_{p=1,\ldots,P;d=1,\ldots,D}$.

Two techniques are used in sequence. Maximum a posteriori estimation is performed using Powell's conjugate direction algorithm [28], a derivative-free optimization method, to get initial estimates of the parameters that maximize Equation (11). These point estimates are then used to initiate an adaptive Metropolis-Hastings Markov Chain Monte Carlo (MCMC) algorithm [29–31] 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 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 ability 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-ability 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 Averaging. Specifically of interest are the conditions under which the Bayesian network is able to find the subset of high-ability evaluators within the crowd so that it can give greater weight to their responses.

There are two crowd ability distribution cases we test, as shown in Figure 3. Case I is that of a homogeneous crowd, where all evaluators making up the crowd have similar abilities. The varied parameter in the homogenous case is the average ability 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-ability evaluators. In this case we fix the average ability 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 ability distribu-

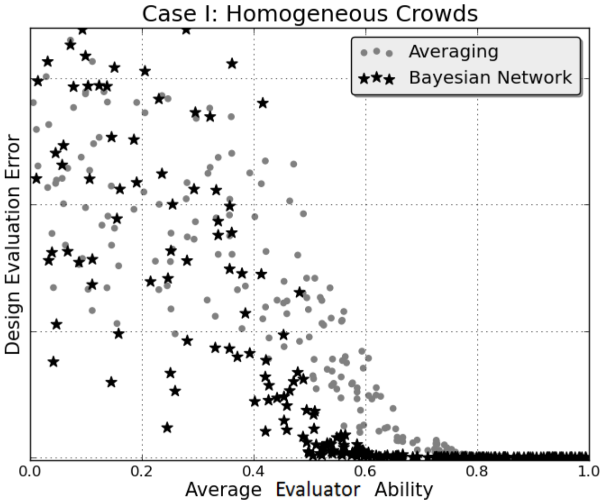


Fig. 4. Case I: Design evaluation error from the Averaging and the Bayesian network methods as a function of average evaluator ability for homogeneous crowds. This plot shows that when dealing with homogeneous crowds, combining the set of evaluator responses into the crowd's combined score is invariant to the combination method used.

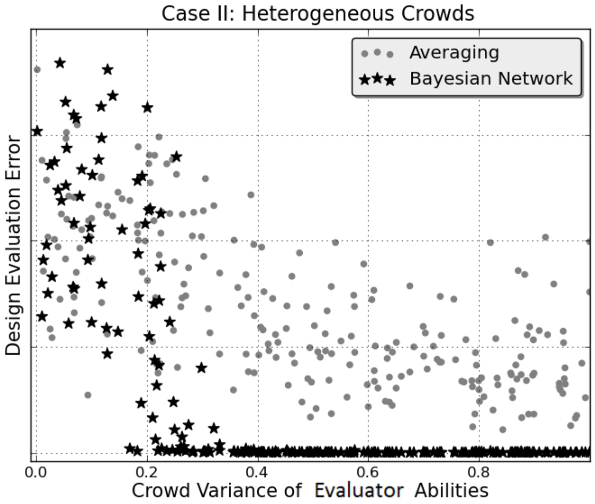


Fig. 5. Case II: Design evaluation error over a set of designs for a mixed crowd with low average evaluation ability. With increasing crowd variance of ability there is an increasingly higher proportion of high-ability evaluators present within the crowd. This leads to a point where the Bayesian network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

tion. This tests the question: How well can a crowd perform as a function of its proportion of high-ability to low-ability evaluators?

The procedure for these studies is to use the Monte Carlo simulation environment to: (1) Generate a crowd made up of evaluators with abilities drawn from the tested ability 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 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 Averaging; and (4) calculate the evaluation error between the true scores of the designs and the combined scores from either the Bayesian network or 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 Equation (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 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 in Equation (12).

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

The results of Case I are shown in Figure 4. Each data point represents a distinct simulated crowd with average ability 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 ability variance $\sigma_a = 0.1$ to create homogeneous crowds. The results show that if the average evaluator evaluation ability is relatively high, both Averaging and the Bayesian network perform equally well with small design evaluation error. In contrast, when the average ability is relatively low, neither Averaging nor the Bayesian network can estimate the true scores very well.

This observation agrees with intuition. A group of evaluators where "no one has the ability" to evaluate a set of designs should not collectively have the ability 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 ability" 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 abilities, it does not matter what 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 Figure 5. Contrary to Case I, distinct crowds represented by each data point have on average the same ability $\mu_a = 0.2$. Instead, moving right along the x-axis designates crowds with increasingly higher proportions of high-ability evaluators within the crowd. In this case, 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-ability evaluators is contained within the crowd. Under these conditions,

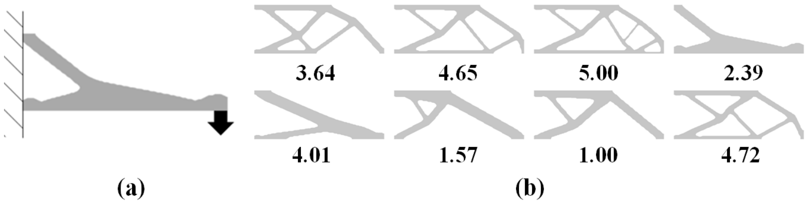


Fig. 6. (a) Boundary conditions for bracket strength evaluation, (b) the set of all eight bracket designs

the Bayesian network identifies the small group of experts from the less competent crowd and weighs their evaluation more so 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-ability 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 ability" is recreated from Case I.

In summary, we have created simulated crowds to test the influence of crowd ability on the crowdsourced evaluation process. Two cases were tested, representing homogeneous and heterogeneous ability distributions. Under the modeling assumptions described in Section 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 high-ability evaluators exist within the crowd.

6 Human Crowd Study

In this section we set up a design evaluation task for a real human crowd to test our modeling assumptions. The evaluation task was chosen to be a classic structural design problem for a load-bearing bracket [32], in which evaluators are asked to rate the capabilities of bracket designs to carry a vertical load as shown in Figure 6.

Participants

The human crowd consisted of 181 evaluators recruited using the crowdsourcing platform Amazon Mechanical Turk. 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 [33]. 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 Figure 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 user, who was then asked to review these designs to get an overall idea of their strengths. After at least 20 seconds, the user 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 "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 [34].

Data analysis

A preprocessing step was carried out before the data were fed into either the Bayesian network or Averaging techniques. Specifically, since some evaluators would give ratings all above 3 while some others tended to give ratings all around 3, all evaluations were linearly rescaled to a range of 1-5. It should be noted that while this mapping ensures that everyone gives '1's and '5's, it does not help to remove nonlinear biases in between an evaluator's most extreme evaluations. To calculate design evaluation error, the same mean-squared error metric was used as in the simulated crowd study and as given in Equation (12).

6.1 Results

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

	Design Evaluation Error (std.)
Averaging	1.001 (N/A)
Bayesian Network	1.728 (0.006)

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

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

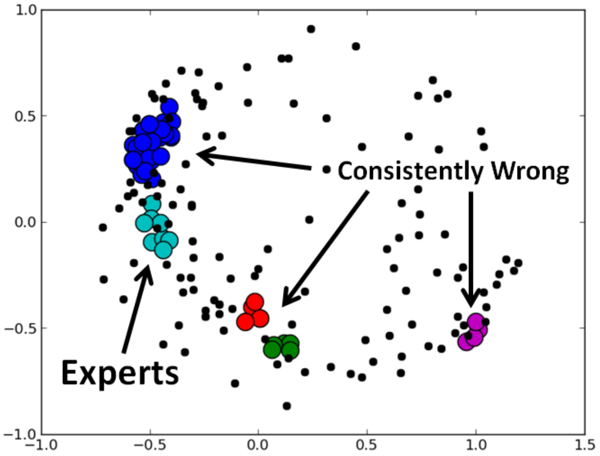


Fig. 7. Clustering of evaluators based on how similar their evaluations are across all eight designs. Each black or colored point represents an individual evaluator, where colored points represent evaluators who were similar to at least 3 other evaluators, and black points represent evaluators who tended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

abilities due to their consistency and overweigh their incorrect evaluations.

Visualizing the crowd's ability 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 have used is density-based and uses the Euclidean distance metric to identify clusters of evaluators who gave similar evaluations [35]. This clustering method was chosen as it can account for varying clustering sizes, as well as not necessitating that every evaluator belong to a cluster. The flattening from eight dimensions to two dimensions was done using multidimensional scaling.

We see in Figure 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 ability "expert" evaluators, as evidenced by their evaluation error. In contrast, the other four clusters were consistent but wrong in their evaluations.

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

Cluster Color	Design Evaluation Error
Blue	1.826
Cyan "Experts"	0.796
Red	1.805
Green	2.394
Magenta	6.275

Table 2. Mean-squared evaluation errors from the 5 clusters of similarly evaluators.

6.2 Follow-up to human crowd study

For completeness of the human study, we conducted two follow-up experiments to capture the differences between the simulated crowd assumptions and results, and the human crowd results. The first follow-up 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 follow-up experiment remains entirely in simulation, and shows that the existence of enough "consistently wrong" evaluators will also cause the Bayesian network to fail in simulation as well, thus mimicking the results of the human study.

6.2.1 Human crowd augmented with simulated experts

We show in Figure 8 how the design evaluation error would be reduced if extra expert evaluations, i.e., evaluations exactly the same as true scores, were collected in addition to the original 181 responses 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 sub-optimal estimations. Similar to the simulations in Figure 5, one can observe the phase-changing phenomenon in the change of the design evaluation error.

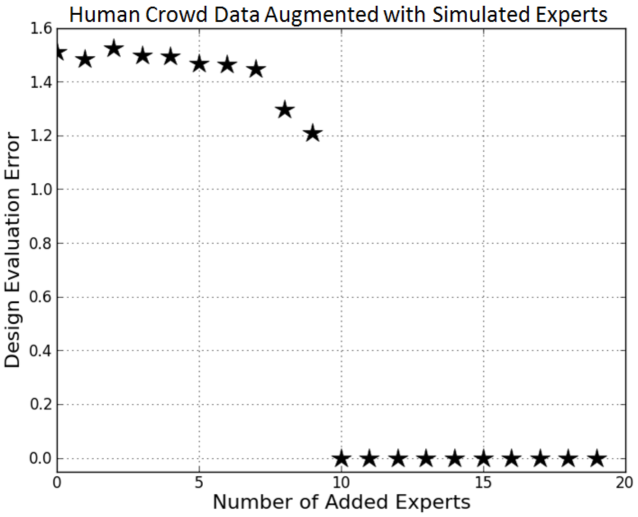


Fig. 8. Design evaluation error with respect to additional experts.

6.2.2 Simulation of “consistently wrong” evaluators

In this scenario, we tested a set of simulations in which the crowd contained two clusters of evaluations. 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 off by 0.5. We vary the crowd proportion of “experts” from 0 to 1 and calculate the corresponding evaluation errors, as shown in Figure 9. While the error from Averaging changes linearly with respect to the proportion, that from the Bayesian network takes only two phases. The result mimicks what we saw with the human study; the Bayesian network simply considers one of the two groups as the experts and trusts its evaluations, and that decision is made based on the group sizes.

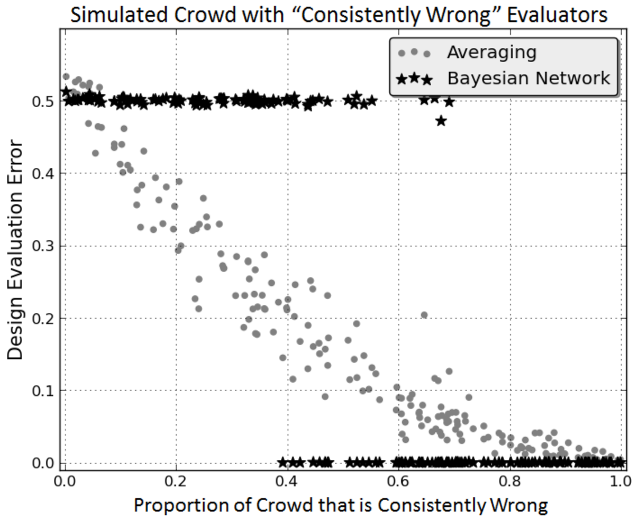


Fig. 9. Design evaluation error with respect to the proportion of the expert group.

7 Conclusion

Crowdsourcing is a promising method to evaluate engineering design concepts that require human input, due to the possibility of leveraging evaluation ability that is distributed over a large number of people. A common characterisitic of crowdsourced design evaluation processes is that the crowd is composed of a heterogeneous mixture of high and low-ability evaluators. A key challenge in such crowdsourced evaluation processes is to find the subset of high ability, or expert, evaluators in the crowd such that their evaluations may be given more weight.

In this paper we proposed a Bayesian network to model human evaluations. The key modeling assumption is that low-ability evaluators tend to give less consistent (more random) evaluations than expert evaluators. We tested using simulated crowds how both the Averaging and the Bayesian network can be affected by the distribution of evaluator abilities and showed that, when assumptions hold, the Bayesian network approach is preferable to simple Averaging and requires fewer experts to achieve a good estimation of the true design scores across all simulation settings.

A human crowd study on bracket strength evaluation was then conducted. Evaluators recruited through Amazon Mechanical Turk gave evaluations on eight bracket designs according to how strong the brackets were under load. The result of this study was that the Bayesian network model did worse at estimating the true strengths of the bracket designs. Upon further investigation, it was found that there were numerous clusters of “consistently wrong” evaluators in the crowd. These clusters caused the Bayesian network to believe they were the experts, and consequently overweigh their (wrong) evaluations.

While the human study did not showcase the superiority of Bayesian network over Averaging, it does reveal the challenges of performing such crowdsourced evaluations when dealing with even a simple engineering design task. The distribution of evaluation ability in this study sharply contrasts many of the recent successes within the human-computer interaction, computer vision, and crowdsourcing communities; namely, we show that only a minority of the crowd are experts and that there exist numerous clusters of consistent yet incorrect evaluators.

Further study into methods to find experts in settings in which they are the minority is justified. These methods may generalize our definition of evaluator ability by incorporating relevant information about the evaluation process, as well as setup analytic conditions under which it is impossible to find experts. This study is thus a first step in showing that extra information in the form of evaluator variables, design variables, and task variables may be needed to find expert evaluators for even simple engineering design tasks, as such experts are otherwise overshadowed by the crowd.

Acknowledgement

This research was partially supported by the National Science Foundation Grant 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 is gratefully acknowledged.

References

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

[2] Ipeirotis, P. G., and Paritosh, P. K., 2011. “Managing crowdsourced human computation: a tutorial”. *Proceedings of the 20th international conference companion on World wide web*, pp. 287–288.

[3] Nunnally, J., and Bernstein, I., 2010. *Psychometric Theory 3E*. McGraw-Hill series in psychology. McGraw-Hill Education.

[4] Amazon, 2005. Amazon mechanical turk. http://www.mturk.com.

[5] Van Horn, D., Olewnik, A., and Lewis, K., 2012. “Design analytics: Capturing, understanding, and meeting customer needs using big data”. *ASME 2012 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*, pp. 863–875.

[6] 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 2013 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*, p. V02BT02A012.

[7] Stone, T., and Choi, S.-K., 2013. "Extracting consumer preference from user-generated content sources using classification". *ASME 2013 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*, p. V03AT03A031.

[8] Kudrowitz, B. M., and Wallace, D., 2013. "Assessing the quality of ideas from prolific, early-stage product ideation". *Journal of Engineering Design,* **24**(2), pp. 120–139.

[9] Fuge, M., Stroud, J., and Agogino, A., 2013. "Automatically inferring metrics for design creativity". *ASME 2012 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*, p. V005T06A010.

[10] Ren, Y., and Papalambros, P. Y., 2012. "On design preference elicitation with crowd implicit feedback". *ASME 2012 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*, pp. 541–551.

[11] Lord, F. M., 1980. *Applications of item response theory to practical testing problems*. Mahwah, NJ: Erlbaum.

[12] Rasch, G., 1960/1980. "Probabilistic models for some intelligence and achievement tests, expanded edition (1980) with foreword and afterword by b.d. wright". *Copenhagen, Denmark: Danish Institute for Educational Research*.

[13] Oravecz, Z., Anders, R., and Batchelder, W. H., 2013. "Hierarchical bayesian modeling for test theory without an answer key". *Psychometrika*, pp. 1–24.

[14] 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.". *Advances in Neural Information Processing Systems,* **22**, pp. 2035–2043.

[15] Welinder, P., Branson, S., Belongie, S., and Perona, P., 2010. "The multidimensional wisdom of crowds.". *Advances in Neural Information Processing Systems,* **10**, pp. 2424–2432.

[16] 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*, pp. 173–182.

[17] 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*, pp. 254–263.

[18] 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-Volume 1*, pp. 1220–1229.

[19] Wauthier, F. L., and Jordan, M. I., 2011. "Bayesian bias mitigation for crowdsourcing". *Advances in Neural Information Processing Systems*, pp. 1800–1808.

[20] 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*.

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

[22] 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*.

[23] Liu, Q., Peng, J., and Ihler, A. T., 2012. "Variational inference for crowdsourcing.". *Advances in Neural Information Processing Systems*, pp. 701–709.

[24] Sheshadri, A., and Lease, M., 2013. "Square: A benchmark for research on computing crowd consensus". *First AAAI Conference on Human Computation and Crowdsourcing*.

[25] 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*, pp. 614–622.

[26] Celaschi, F., Celi, M., and García, L. M., 2011. "The extended value of design: An advanced design perspective". *Design Management Journal,* **6**(1), pp. 6–15.

[27] 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*.

[28] Powell, M. J., 1964. "An efficient method for finding the minimum of a function of several variables without calculating derivatives". *The Computer Journal,* **7**(2), pp. 155–162.

[29] Haario, H., Saksman, E., and Tamminen, J., 2001. "An adaptive metropolis algorithm". *Bernoulli,* **7**(2), pp. 223–242.

[30] Gelfand, A. E., and Smith, A. F., 1990. "Sampling-based approaches to calculating marginal densities". *Journal of the American Statistical Association,* **85**(410), pp. 398–409.

[31] Patil, A., Huard, D., and Fonnesbeck, C. J., 2010. "Pymc: Bayesian stochastic modelling in python". *Journal of Statistical Software,* **35**(4), p. 1.

[32] Papalambros, P. Y., and Shea, K., 2005. "Creating structural configurations". In *Formal engineering design synthesis*, E. K. Antonsson and J. Cagan, eds. Cambridge University Press, pp. 93–125.

[33] Schramm, U., Thomas, H., Zhou, M., and Voth, B.,

1999. “Topology optimization with altair optistruct”. *ASME Proceedings of the Optimization in Industry II Conference*.

[34] University of Michigan - Optimal Design Laboratory, 2013. Turker design - crowdsourced design evaluation. http://www.turkerdesign.com.

[35] Ester, M., Kriegel, H.-P., Sander, J., and Xu, X., 1996. “A density-based algorithm for discovering clusters in large spatial databases with noise.”. *Knowledge Discovery and Data Mining,* **96**, pp. 226–231.