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

Alex Burnap*
Ph.D. Student
Design Science Progam
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
Cyprus City, 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 maintenability of a vehicle. The challenge is to correctly estimate the design scores based on massive and diverse crowdsourced evaluations. As an alternative to simple averaging, this paper introduces a Bayesian network approach that models the human evaluation process and estimates 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 and incorrect human evaluations exist along with the clusters of experts. This suggests that data on participants' background and a more sophisticated statistical model are needed to isolate the correct clusters of experts.

1 Introduction

Suppose we wish to evaluate a set of military vehicle design concepts with respect to mission performance attributes. For many 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 attributes, such as situational awareness, physics-based simulation is difficult or not possible at all. Instead, these "perceptual attributes" require human input for accurate evaluation [1].

To obtain evaluations over these perceptual attributes, one may ask a number of specialists to evaluate the 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 [2].

Crowdsourced evaluation, or the delegation of an evaluation task to a large and unknown group of people, is a promising approach to obtain design evaluations over perceptual attributes. Crowdsourced evaluation draws from the pioneering works of online communities, like Wikipedia, which have shown that accuracy and comprehensiveness are possible in a large crowdsourced setting. Although many successful online communities exist, there are limited reference materials on how to optimally setup a crowdsourced

*Corresponding author.

evaluation process. For example, which crowd structure parameters, like crowd hierarchy and crowd collaboration, are most important to manage when setting up a crowdsourced evaluation process? Similarly, at the level of the *participant* within the crowd, which task structure parameters, like rating vs. ranking or time between successive evaluations, influence the crowdsourced evaluation process?

To help answer these questions, an agent-based Monte Carlo simulation of the crowdsourced evaluation process has been developed. Imbuing this simulation with various crowd or task structure properties allows one the opportunity to learn optimal setup and management heuristics, as well as fundamental limitations of the crowdsourced evaluation process. Naturally, the modeling assumptions going into the simulation environment must be empirically tested for validity. The general workflow is first to use the simulation environment to test the influence of parameters and modeling assumptions on the crowdsourced evaluation process, and then to validate the simulation results with data from human studies.

The first parameter we explored using the simulation environment is the "ability" of participants in the crowd, where ability is defined as the probability that a participant gives an evaluation "response" close to the design's "true score". The choice of exploring participant ability comes from an important lesson from successful online community efforts, namely, the need to implement a systematic method of filtering "signal" from "noise" [3]. In a crowdsourced evaluation process, this manifests itself as a need of screening good responses from bad responses, 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 responses from all participants with equal weight will reduce the accuracy of the crowd's "combined score" due to bad responses 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 responses may be subsequently given more weight.

To achieve this goal, we developed a Bayesian network model that does not require prior knowledge of the true scores of the designs or of the ability of each participant in the crowd, yet still aims to identify the high-ability participants within the crowd. This model links the ability of participants in the crowd (i.e., knowledge or experience over the perceptual attribute 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 participants are more likely to "guess," and while guessing, to evaluate designs more randomly. This assumption is modeled by making the participant's response be a random variable centered at the true score of the design being evaluated [4]. A graphical representation of the Bayesian network showing these relationships is given in Figure 1.

The simulation model and its modeling assumptions, as well as the performance of the Bayesian network versus averaging were explored through two studies. First, the simulation environment was used to generate participant evaluation responses for a set of designs by simulated crowds of participants. These crowds had a heterogeneous or homogeneous 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 [5], and performed a crowdsourced evaluation with the same crowd and task properties as in the simulation.

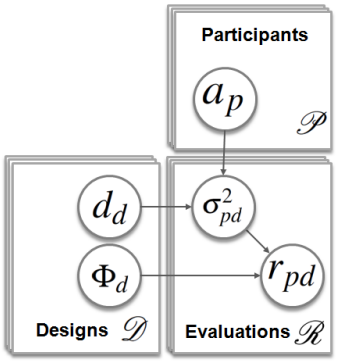


Fig. 1. Graphical representation of the Bayesian network model. This model describes a crowd of participants making evaluations r_{pd} that have error from the true score Φ_d. Each participant has an evaluation ability a_p and each design has an evaluation difficulty d_d.

The remainder of this paper is organized as follows. Section 2 reviews related work from statistics, psychology, and crowdsourcing literature, as well as research motivations from practice. 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.

2 Related Work

Many of the modeling decisions and objectives of this work were informed by prior crowdsourcing applications in design practice. 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 [6]. Local Motors Incorporated developed the Rally Fighter using a rating system similar to our work, but with non-adaptive weightings over evaluations [7].

This work extends an earlier model of the crowdsourced design evaluation process by incorporating participant ability and design difficulty in a probabilistic framework [8, 9]. Research from statistical machine learning applied to modeling crowdsourcing environments has studied techniques to learn participant ability and design problem difficulty for binary tasks such as image annotation [10]. We build on this work by modeling the interaction effect between participant ability and design difficulty, similar to recent work dealing with standardized testing on multiple choice tests [11]. Instead

of binary or multiple choice tasks, we structure our model to work with data on a bounded and continuous range. The parameterization for this derivation was built using methods from hierarchical Bayesian test theory [12].

3 A Bayesian Network Model for Human Evaluations

Let the crowdsourced evaluation contain D designs and P participants. We denote the true score of design d as $\Phi_d \in [0,1]$, and the evaluation response from participant 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 participant p has an evaluation ability a_p. Some significant assumptions we made shall be highlighted here and potential assumption relaxations will be discussed at the end of the paper: (1) We assume that participants evaluate designs without systematic biases, i.e., given infinite chances of evaluating one specific design, the average score of the participants will converge to the true score of that design regardless of their ability [4]; note that this assumption also implies that no participants 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 participants is constant during the entire evaluation process; (4) we assume that all participants 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 participant 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.

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

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

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

The average evaluation error for a given participant on a given design is a function of the participant's evaluation ability a_p and the design's evaluation 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 participant's evaluation error random variable using the expectation operator $\mathbb{E}$ in Eq. (3).

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

The random variables θ and γ are introduced to allow more flexibility in modeling evaluation tasks and are assumed to be the same for all participants and designs: A high

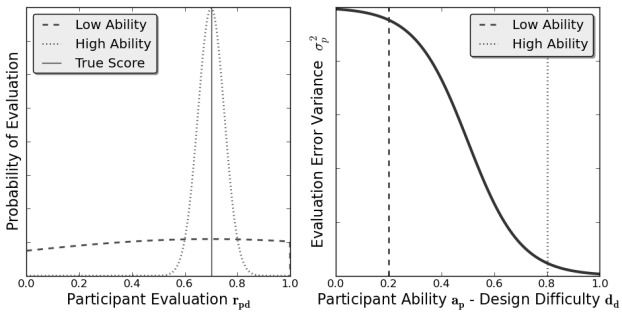


Fig. 2. (a) Low evaluation ability (dashed) relative to the design evaluation difficulty results in an almost uniform distribution of a participant's evaluation response, while high evaluation ability (dotted) results in participants making evaluations closer to the true score. (b) Participant's evaluation response variance σ^2_{pd} as a function of the participant's evaluation ability a_p, given some fixed design difficulty d_d and hyperparameters θ and γ.

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 participant evaluation 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).

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

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

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

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

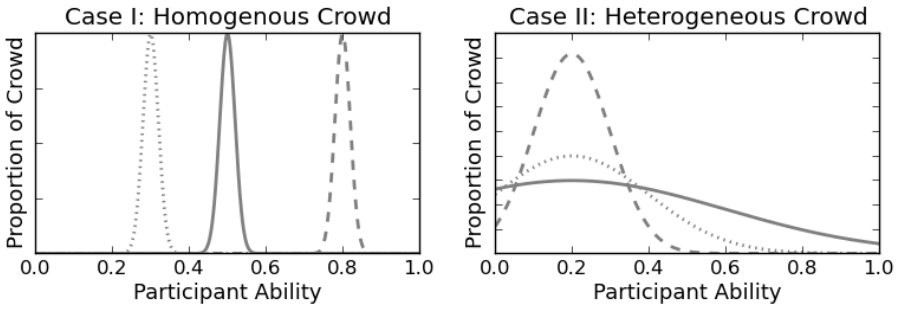
The hierarchical random variables of the participant's evaluation ability a_p and the design's evaluation difficulty d_d are both restricted to the range [0,1]. We let their distributions be truncated Gaussians with parameters μ_a, σ^2_a, μ_d, σ^2_d set globally for all participants and designs as shown in Eq. (7) and (8).

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

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

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

$$\theta \sim \text{Gaussian}\left(\mu_\theta, \sigma^2_\theta\right) \tag{9}$$



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

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

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

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

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

4 Estimation and Inference of the Bayesian Network

The proposed Bayesian network model is built upon the following random variables: Participants' 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 participants $\mathbf{R} = \{r_{pd}\}_{p=1,\ldots,P;d=1,\ldots,D}$.

Two techniques are used in sequence. Maximum A Posteriori estimation is performed using Powell's conjugate direction algorithm [13], 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 [14–16] 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 participants 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 participants 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 participants 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 participant can evaluate correctly or, alternatively, if every participant 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 participants. In this case we fix the average ability of the crowd to be low, so that most participants cannot evaluate designs correctly. Instead, the varied parameter in the heterogeneous case is the variance of the crowd's ability distribution. This tests the question: How well can a crowd perform as a function of its proportion of high-ability to low-ability participants?

The procedure for these studies is to use the Monte Carlo simulation environment to: (1) Generate a crowd made up of participants 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 participant within the crowd gives to each design; (3) combine the participant-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 participants per crowd, as well as eight designs with scores drawn uniformly from the range [0,1] and evaluation diffi-

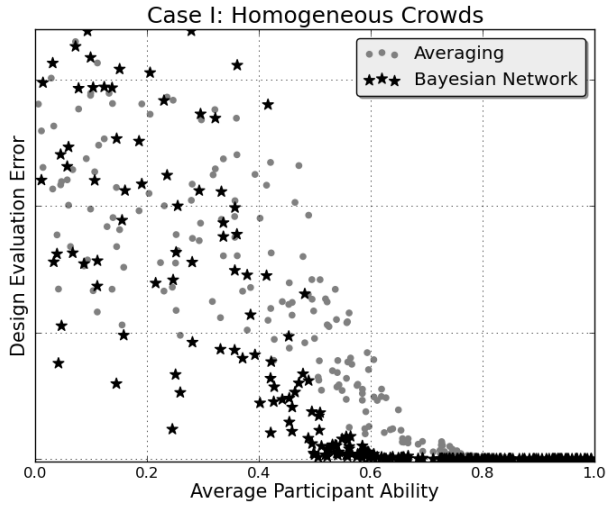


Fig. 4. Case I: Design evaluation error from the Averaging and the Bayesian network methods as a function of average participant ability for homogeneous crowds. This plot shows that when dealing with homogeneous crowds, combining the set of participant responses into the crowd's combined score is invariant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

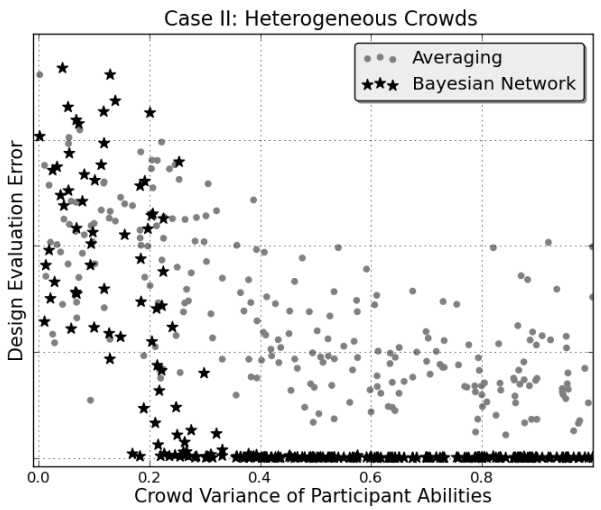


Fig. 5. Case II: Design evaluation error over a set of designs for a mixed crowd with low average evaluation ability. With increasing crowd variance of ability there is an increasingly higher proportion of high-ability participants present within the crowd. This leads to a point where the Bayesian network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

culties $\{d_d\}$ set at 0.5 for all designs. The evaluation process for each participant 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 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 metric as shown in Equation (12).

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

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 participant 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 participants 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 participants and their individual evaluation responses upon combination when determining the crowd's combined score. Similarly, a group of participants where "everyone has the ability" to evaluate a set of designs should perform well regardless of the relative weighting between participants. The key result for Case I is this: When the crowd has a homogeneous distribution of participant abilities, it does not matter what weighting scheme one assigns between various participants 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 participants 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 participants 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 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 participants 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 used the Monte Carlo simulation environment 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 participants 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 [17]. We will show the difficulties in applying the Bayesian network to a design evaluation practice and discuss the value of using this method compared to simple averaging.

6.1 Study setup

Eight bracket topologies were generated using the same amount of raw material, and participants in the human crowd were asked to rate the capabilities of these designs to carry a vertical load as shown in Figure 6. The strength of each bracket was calculated in OptiStruct [18]. All bracket strength were then scaled to between 1 and 5 and labeled in the figure. These scaled strength values were considered as the true scores, which were later used to calculate evaluation errors, and were not used in either the Bayesian network or Averaging methods.

The evaluation process for participants was designed as follows: The eight designs were first presented all together to the user, who was asked to review these designs and get an overall idea of their strengths. The user was then allowed to continue after at least 20 seconds to the next stage where the designs were presented sequentially and in random order. For each design, the participant was asked to evaluate its strength using a rating between 1 and 5, with 1 being "Very Weak" and 5 "Very Strong."

6.2 Results and analysis

A website was set up to gather responses [19] and a total number of 181 participants were recruited using Amazon Mechanical Turk.

One difficulty in analyzing these data is the existence of evaluation bias, which is not modeled in the Bayesian network. For example, some people would give ratings all above 3 while some others tend to give ratings all around 3. A simple treatment of such biased data is to rescale ratings of each participant to the 1-5 scale. It should be noted that while this mapping ensures that everyone gives '1's and '5's, it does not help to remove biases for people who only give extreme ratings.

The Bayesian network and Averaging methods performed using this preprocessed data showed that both methods can correctly identified the best three brackect designs. However, the Bayesian network is slightly worse than Averaging under the MSE criterion of Equation (12). This could happen under either of the following two situations: (1) There could be insufficient number of high-ability participants. According to the simulation results, this is the case where the Bayesian network is not able to correctly identify the group of high ability people and therefore will not perform well. (2) The critical assumption we made states that

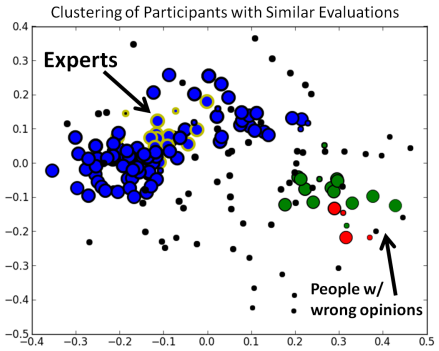


Fig. 7. Clustering using DBSCAN [20] on the collected data. Circle thickness represents the similarity between the true ratings and that of each participant. The individuals with highest similarities to the true ratings are highlighted in yellow circles.

any rating given by a participant is a random variable centered at the true score, while the ratings given by a low ability person will have more variance. The Bayesian network model will not work well if there exists a group of low ability participants who have consistent but wrong evaluations on one design, e.g., everyone evaluates a design as 1 while the true score is 5. In fact, if the low ability people are the majority and all consistent, the Bayesian network would mistakenly take these people as having high abilities and derive estimates of scores worse than simple Averaging.

We now show that both of these two situations exist in the recruited crowd. For better exposition, we use Figure 7 to visualize the data, which consists of eight ratings for each of the 181 participants. Multi-dimensional scaling is applied to map these data down to a two-dimensional space for visualization, where we used the Euclidean distance as the similarity measure. A clustering is then performed using the same similarity measure so that we can highlight the few major clusters of participants. In addition, we also denote by the circle thickness the similarity between the true ratings and that of each participant. The few individuals with highest similarities to the true ratings are highlighted in yellow circles.

A few things can be observed from this figure: Overall, the group in blue can be seen as the "high ability" group while others as "low ability" ones since most blue points have ratings relatively similar to the truth. In fact, a close look at the data shows that the blue crowd on average can identify the top three designs and the others rarely do so. Considering that the blue crowd is the majority, this explains why both methods can identify the top three designs successfully. Nonetheless, while the majority of the blue crowd can evaluate some designs fairly well, they fail on the others. Meanwhile, we notice that the groups of low-ability participants are closely clustered, indicating high similarity within the groups. This result also violates our modeling assumption. These observations together explain why the Bayesian network fails to perform better than Averaging.

The data analysis suggests that finding high-ability participants through an open call is possible, even for a task that

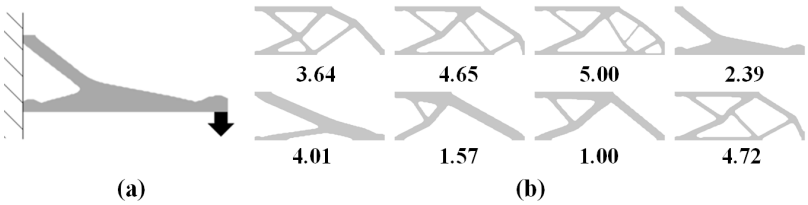


Fig. 6. (a) Boundary conditions for bracket strength evaluation, (b) The set of all eight bracket designs

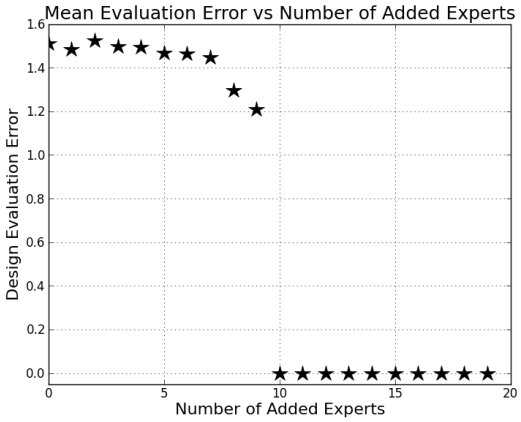


Fig. 8. Design evaluation error with respect to additional experts.

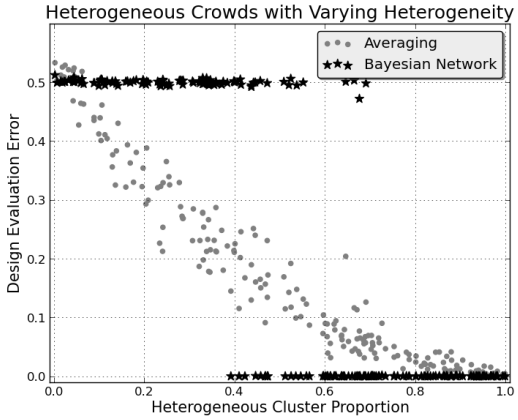


Fig. 9. Design evaluation error with respect to the proportion of the expert group.

requires expertise such as structural design. However, while the Bayesian network is a theoretical way to identify these participants, its application in reality is limited by factors such as (1) the existence of experts and (2) the clustering phenomenon we observed. For completeness of the study, we investigated the following "What-if" scenarios to understand what the result would be (1) if we were able to get some more experts in the human study and (2) if all evaluations were accurate except for one group being consistently wrong.

6.2.1 The effect of 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 esitmation error.

6.2.2 The effect of clustered evaluations

In this scenario, we test a set of simulations where the crowd contains two clusters of evaluations. One cluster, "the experts", can always evaluate correctly; and the other cluster is almost the same, except that people there always rate one design off by 0.5. We vary the proportion of "experts" from 0 to 1 and calculate the corresponding evaluation errors, as shown in Figure 9. The result is as expected but revealing. While the error from Averaging changes linearly with respect to the proportion, that from the Bayesian network takes only two phases. The 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.

6.3 Discussion

While the human study did not showcase the superiority of Bayesian network over Averaging, it does reveal the challenges of performing such crowdsourced evaluations and the directions for future research. Good evaluation estimations could be derived based not only on the estimation method we pick, but also on the proportions of people with various abilities within the crowd. While it is true that neither Averaging or Bayesian network would provide good estimations when participants are observed to be clustered, such an observation, however, suggest that more participants should be recruited or a more sophisticated statistical model should be used to account for the clustering phenomenon. For example, one could associate the clustering with demographic backgrounds of the participants and try to understand what common features among people make them think in similar ways and how these features would affect their abilities in this evaluation task. The insights we gain could help us to target better a specific crowd in future studies.

This study is a first step in understanding how to setup and manage a crowdsourced evaluation, as it highlighted

some of the key modeling challenges when dealing with a crowd of participants with diverse abilities.

7 Conclusion

Crowdsourcing is a promising method to evaluate designs over perceptual attributes requiring human input, yet further research must be done before a crowdsouced evaluation process can be setup and managed properly.

In this paper we proposed a Bayesian network to model human evaluations. The key modeling assumption is that any human evaluation on any design is a random variable centered at the true score of the design, regardless of the evaluation ability of the participant. We tested using simulations how the resulting estimations of the Bayesian network can be affected by the distribution of participant 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.

The human crowd study on bracket strength evaluation was then conducted to verify the simulation results. Evaluations on eight bracket designs were recorded using Amazon MTurk. The intriguing results from analyzing the data are that (1) while crowdsourcing allowed us to find particular experts without setting any criterion on people, the size of the expert group might not be sufficient for the Bayesian network to identify the experts; and (2) clustered consistent but incorrect evaluations will violate the modeling assumptions and weaken the strength of the Bayesian network. The findings from this study suggest that a mechanism to explain the clusering of human evaluations is essential to a successful crowdsourced evaluation task.

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] Reid, T. N., Gonzalez, R. D., and Papalambros, P. Y., 2010. "Quantification of perceived environmental friendliness for vehicle silhouette design". *ASME Journal of Mechanical Design,* **132**, p. 101010.

[2] 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.

[3] Grossman, W., 1997. *Net wars*. New York University Press, New York.

[4] Nunnally, J., and Bernstein, I., 2010. *Psychometric Theory 3E*. McGraw-Hill series in psychology. McGraw-Hill Education (India) Pvt Limited, Delhi.

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

[6] 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.

[7] Bommarito, M., F. R. G. A., and Page, S., 2011. Crowdsourcing design and evaluation analysis of darpa's xc2v challenge.

[8] Ren, Y., and Papalambros, P., 2012. "On design preference elicitation with crowd implicit feedback". In Proceedings of the ASME International Design Engineering Technical Conferences & Computers and Information in Engineering Conference.

[9] Gerth, R. J., Burnap, A., and Papalambros, P., 2012. "Crowdsourcing: A primer and its implications for systems engineering". In 2012 NDIA Ground Vehicle Systems Engineering and Technology Symposium.

[10] Welinder, P., Branson, S., Belongie, S., and Perona, P., 2010. "The multidimensional wisdom of crowds". In *Advances in Neural Information Processing Systems 23*, J. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. Zemel, and A. Culotta, eds. pp. 2424–2432.

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

[12] Oravecz, Z., Anders, R., and Batchelder, W. H., 2012. "Hierarchical bayesian modeling for test theory without an answer key".

[13] 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.

[14] Haario, H., Saksman, E., and Tamminen, J., 2001. "An adaptive metropolis algorithm". *Bernoulli*, pp. 223–242.

[15] 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.

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

[17] 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, Cambridge.

[18] Schramm, U., Thomas, H., Zhou, M., and Voth, B., 1999. "Topology optimization with altair optistruct". In ASME Proceedings of the Optimization in Industry II Conference.

[19] Turker design - crowdsourced design evaluation. `http://www.turkerdesign.com`.

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