Proceedings of the ASME 2013 International Design Engineering Technical Conferences
IDETC 2013
August 4-7, 2013, Portland, Oregon, USA

DETC2013- 13020

A SIMULATION BASED ESTIMATION OF CROWD ABILITY AND ITS INFLUENCE ON CROWDSOURCED EVALUATION OF DESIGN CONCEPTS

Alex Burnap*
Yi Ren
Panos Y. Papalambros
Optimal Design Laboratory
University of Michigan
Ann Arbor, MI

Richard Gonzalez
Department of Psychology
Department of Statistics
University of Michigan
Ann Arbor, MI

Richard Gerth
National Automotive Center
TARDEC-NAC
Warren, MI

ABSTRACT

Crowdsourced evaluation is a promising method for evaluating attributes of design concepts that require human input. One factor in obtaining good evaluations is the ratio of high-ability to low-ability participants within the crowd. In this paper we introduce a Bayesian network model capable of finding participants with high design evaluation ability, so that their evaluations may be weighted more than those of the rest of the crowd. The Bayesian network model also estimates a score of how well each design concept performs with respect to a design attribute without knowledge of the true scores. Monte Carlo simulation studies tested the quality of the estimations on a variety of crowds consisting of participants with different evaluation ability. Results suggest that the Bayesian network model estimates design attribute performance scores much closer to their true values than simply weighting the evaluations from all participants in the crowd equally. This finding holds true even when the group of high ability participants is a small percentage of the entire crowd. **Keywords:** Crowdsourcing, Design Concept Evaluation, Machine Learning.

1 Introduction

Suppose we wish to evaluate a set of military vehicle design concepts with respect to a set of mission performance attributes. For many attributes, detailed engineering simulations are used to obtain accurate evaluations, such as finite-element analysis to evaluate blast resistance or human mobility modeling to evaluate ergonomics. However, for some attributes, physics-based simulation is difficult and evaluation requires human input.

To obtain human evaluations on these perceptual design attributes [10], one may ask a number of specialists to evaluate the vehicle concepts. The ability to make an evaluation is likely scattered over the “collective intelligence" of a large number of people with diverse backgrounds [5] and viewpoints.

Crowdsourced evaluation, or the delegation of an evaluation task to a large and unknown group of people, is a promising approach to obtain evaluations on perceptual attributes. This approach draws on lessons from online communities, like Wikipedia, which have shown that accuracy and comprehensiveness is possible in large crowdsourced settings.

An important lesson from such community efforts is the need to implement a consistent method of filtering "signal" from "noise;" namely, obtaining valid contributions from those that are not. This general “signal to noise problem" manifests itself in any large crowd with heterogenous abilities, and in a crowdsourced evaluation it is desirable to identify high ability participants from the rest of the crowd. Consequently, identifying evaluations from high-ability participants will make the crowdsourcing process more effective and efficient.

In this paper, we explore the identification of high-ability participants through simulation of the crowdsourced evaluation process. The goal is to test the identification process prior to conducting experiments and collecting data with an actual human crowd. Clearly, simulation results depend on the modeling

*Address all correspondence to this author. Email: aburnap@umich.edu

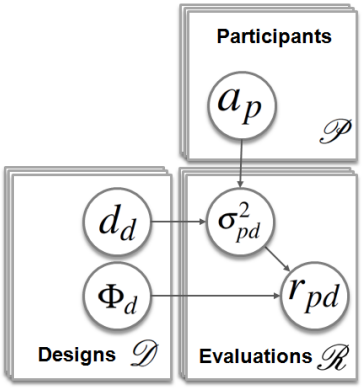

FIGURE 1. Graphical representation of the Bayesian network model. This model describe a crowd of participants making evaluations r_{pd} that have error from the true score Φ_d. Each participant has an evaluation ability a_p and each design has an evaluation difficulty d_d.

assumptions made, which may not hold true in reality. As with modeling of physical phenomena, we need to identify clearly such assumptions and plan to study what happens if these assumptions do not hold at a later time.

In the proposed simulation we use a Bayesian network model to find the subgroup of high-ability participants in a large crowd made up of a heterogeneous mixture of participant abilities. In addition, the proposed Bayesian network is used to estimate the performance scores of the designs on a perceptual attribute. The model does not possess knowledge of the true performance score of the design, which in reality is unknown. Instead, it builds off the assumption that participants with high evaluation ability tend to rate designs consistently closer to the design's true score whereas low ability participants rate designs with more random variation.

In the Bayesian network model, we make the important assumption that the evaluation of a design given by a participant is a random variable centered at the unknown true score of that design. This random variable then follows a distribution parameterized by the participants ability (i.e., knowledge or experience in the specific requirements and attributes to be evaluated) and the difficulty of evaluating the 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). A graphical representation of the Bayesian network showing these relationships is shown in Figure 1.

The performance of the proposed model is compared to a "majority vote" of design evaluations, where the evaluations of all participants are weighted equally. Specifically, we run Monte Carlo simulations to investigate how crowds consisting of different mixtures of high-ability and low-ability participants affect the crowd's overall evaluation error. The results suggest that the Bayesian network produces estimated design scores that are much closer to their true design scores than majority voting, because it can identify the high-ability participants when they exist, even in small numbers, and give their design evaluations greater weight. When experts are missing, the two methods have the same performance.

The remainder of this paper is organized as follows. Section 2 reviews related work from statistics and psychology. Section 3 describes the Bayesian network model. Section 4 details the statistical inference scheme of the Bayesian network. Section 5 presents the simulation studies and discusses results. We conclude in Section 6 with limitations of this work and opportunities for future development.

2 Related Work

We extend an earlier model of crowdsourced design evaluation by incorporating participant ability and design difficulty in a probabilistic framework [11, 3]. Relevant 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 [13]. We build on this work by additionally modeling the interaction effect between participant ability and design difficulty, similar to recent work dealing with standardized testing on multiple choice tests [1].

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 inspired by recent work in hierarchical Bayesian test theory [7].

3 Bayesian Network Model for Evaluation Ability

Let the crowdsourced evaluation contain D designs and P participants. We denote the true (normalized) performance score of design d as $\Phi_d \in [0,1]$, and the evaluation 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 this paper: (1) We assume that participants evaluate designs without systematic biases, i.e., given infinite chances of evaluating one specific design, the average score by a participant will meet the true score of that design, regardless of their ability [6]; note that this assumption also implies that no participants purposely give bad evaluations; (2) we assume that evaluations 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; neither will it be affected by the evaluation given by a different user; (3) we assume that the evaluation ability of participants is constant during the entire evaluation process; (4) we assume that all participants are fully incentivized

and do not exhibit fatigue; and (5) 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 Equation 1 and shown in Figure 3b.

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

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

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

We want the average evaluation error for a given participant on a given design to be a function of the participant's evaluation ability a_p and the design's evaluation difficulty d_d. In addition, this function should be sigmoidal to capture the notion that there exists a threshold of necessary background knowledge to make an accurate evaluation. Figure 2a illustrates this function, and from these physical justifications, we set the first requirement on the participant's evaluation error random variable using the expectation operator $\mathbb{E}$ in Equation (3).

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

The unknown parameters 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 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.

The next requirement we set is that the variance $\mathbb{V}$ of the participant evaluation error is 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.

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

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

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

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

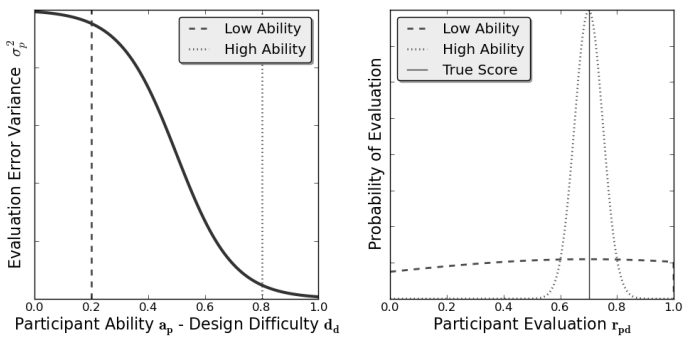

FIGURE 2. (a) Plot showing the relationship between the evaluation error variance σ_{pd}^2, the participant evaluation ability a_p, and the design evaluation difficulty d_d. (b) Low evaluation ability relative to the design evaluation difficulty results in an almost uniform distribution of observed evaluation score, while high ability results in participants making evaluations closer to the true score.

The hierarchical random variables of the participant 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 hyperparameters μ_a, σ_a^2, μ_d, σ_d^2 set globally for all participants and designs as shown in Equations (7) and (8).

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

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

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

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

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

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

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

Case	Type of Crowd	Varied Parameter	Figure	Number of Crowd Simualtions
I	Homogeneous Crowd	Average Crowd Evaluation Ability	4	75
II	Mixed Crowd with Low Ability	Variance of Crowd Evaluation Ability	5	478
III	Mixed Crowd with High Ability	Variance of Crowd Evaluation Ability	5	146
IV	Crowd with Low and High Ability	Mixture Coefficients of Delta Functions	6	250

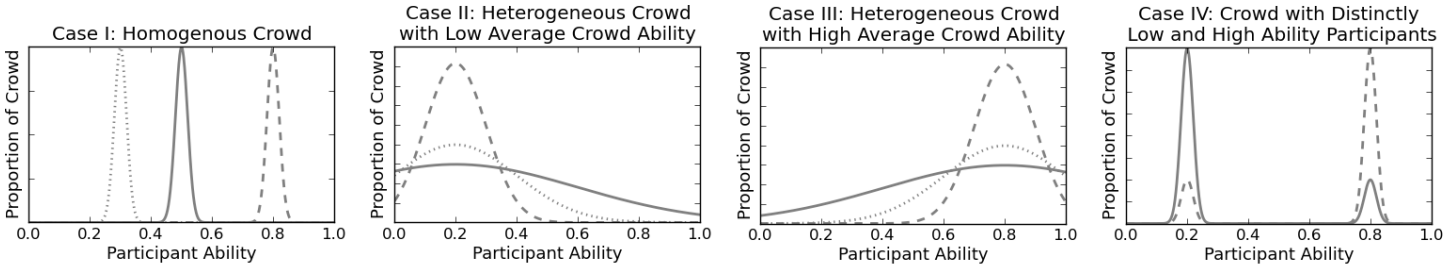

FIGURE 3. Crowd ability distributions for the four cases tested. Each plot shows possible randomly drawn crowds with the varied parameter detailed in the table.

4 Parameter Estimation of the Bayesian Network

To summarize, the proposed Bayesian network model is built upon the following unknown parameters: Participants' abilities $\{a_p\}$, designs' difficulties $\{d_d\}$, true performance scores of designs $\{\Phi_d\}$, and hyper-parameters - α, β, μ_a, σ_a^2, μ_d, σ_d^2. This section elaborates on how these parameters can be estimated using the observed evaluations of the participants $\mathbf{R} = \{r_{pd}\}$.

Two techniques are used in sequence for estimation. First, Maximum A Posteriori estimation is performed using a derivative-free minimization method, Powell's conjugate direction algorithm [9], to get a point estimate (initial guess) of the parameter values that maximize Equation (11). This point estimate is then used to initiate an adaptive Metropolis-Hastings Markov Chain Monte Carlo algorithm [4, 2, 8] that allows us to determine the estimates of each unknown parameter.

5 Simulation Studies for Score Estimation

Recall that the Bayesian network model detailed in Section 3 is used to find the group of crowd participants with high evaluation ability on a perceptual design attribute, so that their design evaluations may be weighted higher for a more accurate and efficient overall crowdsourced evaluation.

We now study how well the Bayesian network model estimates scores by contrasting it with majority voting, i.e. simply averaging the observed scores from the crowd by giving every participant equal weight. These studies are done with a variety of crowd compositions, namely different mixtures of high-ability and low-ability participants. The quality metric for these tests is defined to be the mean-squared error between the crowd's estimated design attribute performance score and the true design attribute performance scores over all designs, as shown in Equation (12).

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

Figure 3 summarizes the four cases studied, with each case capturing a different mixture of evaluation abilities within the crowd. Case I studies how homogeneous crowds, those composed of participants with similar abilities, perform on the evaluation task. The study varies the average ability of the homogeneous crowd to see its effect. Case II studies how a heterogeneous crowd with low average ability performs on the evaluation task, while varying the relative heterogeneity within the crowd. Case III is similar to Case II, except that the crowd has high average ability. Case IV is an extreme version of Case II/III, where the crowd is distinctly composed of only very high or very low ability participants, with the varied parameter being the relative proportion between the two.

Each case has a crowd of 60 participants with individual abilities independently sampled from a probability density representing the distribution of abilities within the crowd as shown in Figure 3. We simulate 8 different designs with evaluation difficulties d_d fixed at 0.5, and true design attribute performance scores Φ_d drawn independently from a uniform density over the interval $[0,1]$. We simulate the evaluation task for each participant as giving a score on how well she/he believes a given design will perform on the perceptual attribute. Each participant completes 60 evaluation tasks, each time scoring a randomly selected

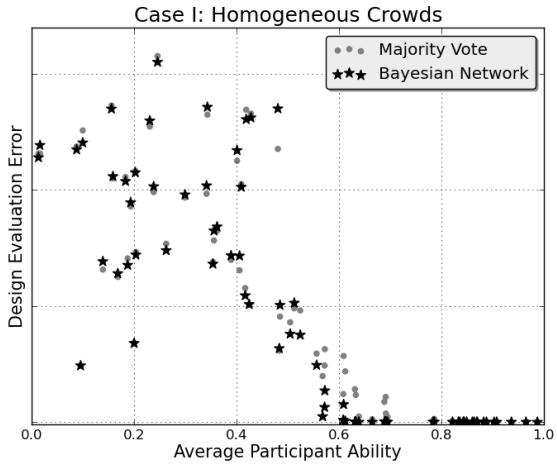

FIGURE 4. Case I: Design evaluation error from the majority vote and the Bayesian network methods as a function of average participant ability for homogeneous crowds.

design from the set of 8 designs. For numerical simplicity, the participant gives a design evaluation using 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 at 0.1.

After evaluation tasks are concluded and the evaluation matrix $\mathbf{R}$ populated, the crowd's overall estimated design attribute performance scores are calculated using both the Bayesian network and the majority vote scheme. The mean-squared errors between estimated and true scores are then obtained to capture the quality of the two schemes.

For the four cases, we compare the mean-squared errors from the Bayesian model and majority vote, varying the test parameters of interest. Each test parameter controls the distribution of the participant ability within the crowd as described in Table 1. In the first case, overall crowd evaluation errors with respect to the average participant ability are plotted in Figure 4. Here all data points were generated using the same narrow crowd evaluation ability variance $\sigma_a = 0.1$, allowing us to compare relatively homogeneous crowds with varying abilities. We note that the Bayesian model approaches low design evaluation error earlier than majority vote. Further, if the average participant evaluation ability is relatively high, both majority vote and the graphical model perform equally well with small design evaluation error. When the average ability is relatively low, neither majority vote nor the Bayesian model 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 weighting of participants in the overall crowd evaluation. Similarly, a group of participants where "everyone has the ability" to evaluate a set of designs should perform just as well with majority vote as with other weightings. The latter case has been shown in current crowdsourcing microtasks market where tasks commonly include annotation of images or taking surveys [12].

The next two cases investigate a crowd with varying ability variance σ_a and constant mean ability ($\mu_a = \text{constant}$). We set high mean ability in Case II ($\mu_a = 0.8$) and low in Case III ($\mu_a = 0.2$). Figure 5 for Case II shows the mean-squared error trend

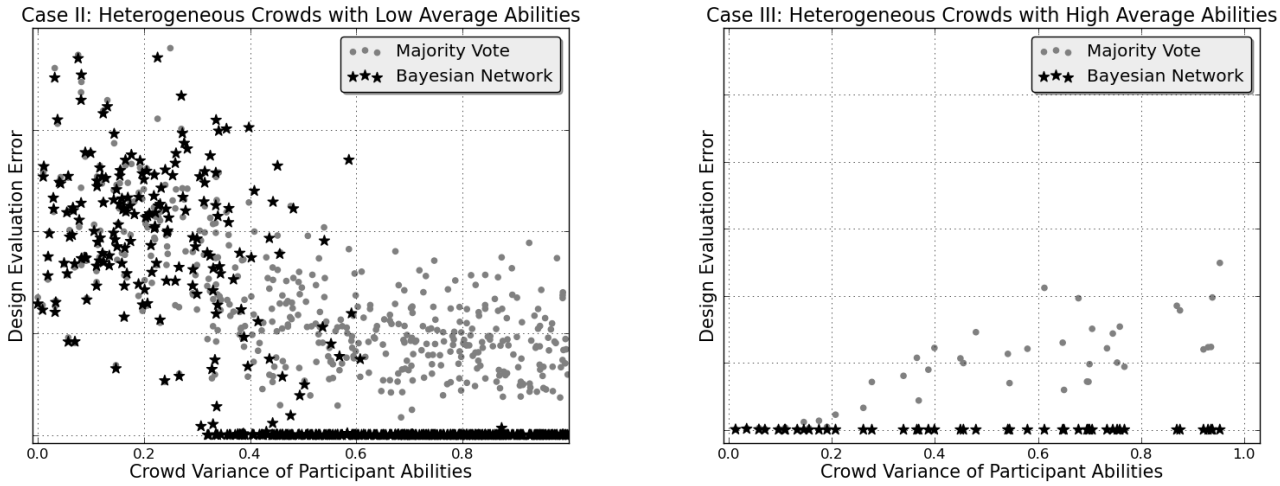

FIGURE 5. Case II/III: Design evaluation error over a set of designs for a mixed crowd with a unimodal evaluation ability distribution as a function of varying ability variance. Case II plot is for low average evaluation ability (left) while Case III is for high average evaluation ability (right).

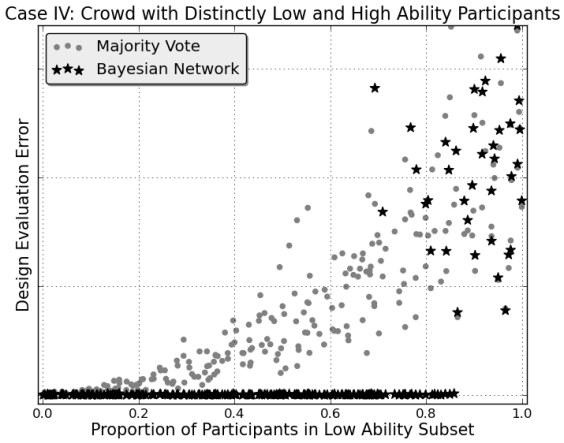

FIGURE 6. Case IV: Design evaluation error for a crowds made up of two sub-populations, one at high evaluation ability and the other at low evaluation ability.

with increasing ability variance in the crowd when the crowd on average does not have the ability to make good evaluations. Interestingly, the proposed Bayesian method performs much better than the majority vote scheme under high variance. This is because the Bayesian model identifies the small group of experts from the less competent crowd and weighs their responses heavier than the rest. This functionality of the Bayesian model deteriorates with decreasing variance in ability as the chance of having experts in the crowd is lowered.

Figure 5 shows Case III with high average evaluation ability. At low variance we recreate a situation from Figure 4 where "everyone has the ability." As the variance of crowd evaluation ability is increased, we add relatively more low evaluation ability participants into the crowd. Majority vote gets progressively worse, as expected, but the Bayesian model remains robust and consistently outputs low error evaluations over a set of designs.

These experiments suggest a transition point where the Bayesian model goes from performing similarly to majority vote to dramatically better. This transition point is a function of the ratio of high ability participants to low ability participants. As soon as the Bayesian model is able to correctly "find" a critical number of participants with high evaluation ability, it can surmise the correct design performance scores, and vice-versa. When the ratio of high evaluation ability participants to low evaluation participants is too low, the Bayesian model incorrectly determines that some low evaluation ability participants have high evaluation ability leading to incorrect criteria scores, and vice-versa.

We test for this transition point in the last case by modeling the crowd ability density as a bimodal mixture model of two delta functions, at low and high evaluation ability, respectively. As shown in Figure 6, when the evaluation ability is consistently high in the crowd, we will have low evaluation error. Adding low evaluation participants, the majority vote design evaluation error increases. The transition we are looking for is evident in this plot around 0.6, or when over 40% of the crowd has high evaluation ability. At this point, the Bayesian model is able to correctly identify the sub-population of high evaluation ability participants, leading to very low design evaluation error. One must note that this transition point exists for the crowd parameters in this simulation and may not be a general finding.

6 Conclusion

Crowdsourcing is a promising method of increasing the quality of evaluations on design attributes that require human input due to a wide diversity of needed abilities for true evaluation. A Monte Carlo simulation was developed to understand how crowds with different abilities may affect the quality of a crowdsourced design evaluation. For the simulation setup tested, we found that if no participants in the crowd had the requisite evaluation ability, a crowdsourced evaluation would not work well regardless of the weighting scheme used to create the crowd's estimated performance score of design alternatives. Conversely, in the case where every participant in the crowd had the requisite ability, a simple majority voting scheme works very well.

When the crowd is relatively homogeneous, majority vote is a good method of combining design evaluations from participants within a crowd. When the crowd is not homogeneous, but rather a mixture of high ability and low ability participants, having a method that weighs design evaluations from participants according to their ability can significantly reduce design evaluation error. In particular, the Bayesian network model performs well when experts exist, even if they make up only a small percentage of the overall crowd.

The modeling assumption that participants in the crowd evaluate without systematic biases is significant. Relaxing this assumption would allow group level clustering, which could help identify latent groups of participants with various backgrounds and experiences.

Future directions for this work include increasing the fidelity of the simulation by incorporating an evaluation bias vector with variable intervals, expanding to multidimensional participant evaluation abilities and design evaluation difficulties, and explicitly modeling psychological phenomena contributing to evaluation task biases. In addition to these simulation improvements, obtaining design evaluations from human participants is an obvious next step to enable model verification and to develop potential heuristics for an effective crowdsourced design evaluation process.

Acknowledgement

This research was partially supported by the Automotive Research Center, a U.S. Army Center of Excellence in Modeling of Simulation of Ground Vehicle Systems headquartered at the University of Michigan. This support is gratefully acknowledged.

References

[1] Y. Bachrach, T. Graepel, T. Minka, and J. Guiver. How to grade a test without knowing the answers—a bayesian graphical model for adaptive crowdsourcing and aptitude testing. In *Proceedings of the 29th International Conference on Machine Learning*, 2012.

[2] A. E. Gelfand and A. F. Smith. Sampling-based approaches to calculating marginal densities. *Journal of the American statistical association*, 85(410):398–409, 1990.

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

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

[5] L. Hong and S. E. Page. 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):16385–16389, 2004.

[6] J. C. Nunnally. *Psychometric Theory 3E*. Tata McGraw-Hill Education, 2010.

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

[8] A. Patil, D. Huard, and C. J. Fonnesbeck. Pymc: Bayesian stochastic modelling in python. *Journal of statistical software*, 35(4):1, 2010.

[9] M. J. Powell. An efficient method for finding the minimum of a function of several variables without calculating derivatives. *The computer journal*, 7(2):155–162, 1964.

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

[11] Y. Ren and P. Y. Papalambros. On design preference elicitation with crowd implicit feedback. In *Proceedings of the ASME International Design Engineering Technical Conferences*. ASME, 2012.

[12] V. S. Sheng, F. Provost, and P. G. Ipeirotis. Get another label? improving data quality and data mining using multiple, noisy labelers. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 614–622. ACM, 2008.

[13] P. Welinder, S. Branson, S. Belongie, and P. Perona. The multidimensional wisdom of crowds. *Advances in Neural Information Processing Systems*, 23:2424–2432, 2010.