HybridEval

A Human-AI Collaborative Approach for Evaluating Design Ideas at Scale

Mesbah, Sepideh; Arous, Ines; Yang, Jie; Bozzon, Alessandro

DOI
10.1145/3543507.3583496

Publication date
2023

Document Version
Final published version

Published in
ACM Web Conference 2023 - Proceedings of the World Wide Web Conference, WWW 2023

HybridEval: A Human-AI Collaborative Approach for Evaluating Design Ideas at Scale

Sepideh Mesbah*
sepideh.mesbah@booking.com
Booking
Amsterdam, Netherlands

Ines Arous*
ines@exascale.info
University of Fribourg
Fribourg, Switzerland

Jie Yang+, Alessandro Bozzon
{j.yang-3,a.bozzon}@tudelft.nl
Delft University of Technology
Delft, Netherlands

ABSTRACT

Evaluating design ideas is necessary to predict their success and assess their impact early on in the process. Existing methods rely either on metrics computed by systems that are effective but subject to errors and bias, or experts' ratings, which are accurate but expensive and long to collect. Crowdsourcing offers a compelling way to evaluate a large number of design ideas in a short amount of time while being cost-effective. Workers' evaluation is, however, less reliable and might substantially differ from experts' evaluation.

In this work, we investigate workers' rating behavior and compare it with experts. First, we instrument a crowdsourcing study where we asked workers to evaluate design ideas from three innovation challenges. We show that workers share similar insights with experts but tend to rate more generously and weigh certain criteria more importantly. Next, we develop a hybrid human-AI approach that combines a machine learning model with crowdsourcing to evaluate ideas. Our approach models workers' reliability and bias while leveraging ideas' textual content to train a machine learning model. It is able to incorporate experts' ratings whenever available, to supervise the model training and infer worker performance. Results show that our framework outperforms baseline methods and requires significantly less training data from experts, thus providing a viable solution for evaluating ideas at scale.

CCS CONCEPTS

• **Human-centered computing** → **Human computer interaction (HCI)**; • **Computing methodologies** → **Machine learning**.

KEYWORDS

Idea evaluation, crowdsourcing, human-AI collaboration, scalability

ACM Reference Format:
Sepideh Mesbah*, Ines Arous*, and Jie Yang+, Alessandro Bozzon. 2023. HybridEval: A Human-AI Collaborative Approach for Evaluating Design Ideas at Scale. In *Proceedings of the ACM Web Conference 2023 (WWW '23), April 30–May 04, 2023, Austin, TX, USA.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3543507.3583496

* Both authors contributed equally to this research..
+ Corresponding author.

WWW '23, April 30–May 04, 2023, Austin, TX, USA

ACM ISBN 978-1-4503-9416-1/23/04.
https://doi.org/10.1145/3543507.3583496

1 INTRODUCTION

The evaluation of design ideas is a critical process impacting the development of any new business. It allows designers to move from the ideation phase, where they create new concepts, to the execution phase in case of a positive evaluation or alter the idea in case of a negative one. Since the evaluation process has a crucial impact on developing a new idea into a business, many efforts have been dedicated to defining relevant criteria for evaluation (e.g., desirability, feasibility, and viability) [12, 13, 22].

To obtain evaluations over these criteria, designers often seek judgment from domain experts. While their evaluation is accurate, involving experts in filtering ideas is expensive and time-consuming [29]. Moreover, experts with the requisite knowledge are scarce because specializing in a particular innovation subfield takes a substantial amount of time. In real-world scenarios where companies had to select design ideas, relying on in-house experts has proven to be prohibitively slow and lead to bottlnecks in the production and decision making process [29]. For instance, Google, who promised to reward the best five ideas for using technology to improve the world [15], had to recruit 3000 employees to evaluate the collected ideas, a process that put them over nine months behind schedule.

Crowdsourcing offers a compelling way for design idea evaluation. In addition to being fast and cost-effective [39], crowd workers possess a large diversity of knowledge in design ideation. Considering the benefits of using the "wisdom" of a crowd [55], researchers started investigating how to leverage crowdsourcing for evaluating design ideas [14, 30, 36]. The most common approach consists of using a voting system, where workers go through the candidate ideas and upvote their preferred ones. Once all votes are collected, the selected ideas are those with the highest number of votes. This consensus-based approach suffer from two main limitations: 1) the crowd is a heterogenous group of experts and nonexperts evaluators, hence considering all votes equally reduces the accuracy of the results [14]; and 2) this voting mechanism renders the rich-get-richer problem where workers tend to vote for only the few ideas that have already received good ratings [2].

A further complication of crowdsourced ideas evaluation is the multitude of evaluation criteria. In the current methods, participants vote for ideas with a unique value indicating their preference. While this voting mechanism is easy to use by participants, it does not consider the multi-criteria evaluation developed by the design-thinking discipline (i.e., desirability, feasibility and viability), although participants are more likely to correctly estimate certain criteria more than others. As a matter of fact, some studies have shown that workers tend to underestimate the cost needed to implement an idea, i.e. its feasibility [16, 27, 44]. For instance, the Fiat Mio car was

a fully crowdsourced vehicle design concept, where participants suggested ideas and voted for their peers. Although the collected ideas were innovative, many of them had to be filtered or altered by experts due to their lack of feasibility [16]. Similarly, Dell found through their crowdsourced ideation initiative "IdeaStorm", that participants tend to underestimate the costs needed to implement an idea [27]. Therefore, systems using a unique value for voting ideas are inaccurate in simulating experts' ratings.

In our work, we set out to measure worker's ability to provide a multi-criteria evaluation of design ideas extracted from three open innovation challenges in the OpenIDEO website[1]. We conduct a rigorous study to compare experts and workers evaluation using the methodologies developed by the design thinking community [12, 13]. These methodologies allow for a general overview of evaluation ability characteristics, but also a more detailed understanding through the individual criteria assessment. We collect ratings and free-text justifications for each criteria from both workers and experts for each idea in the selected challenges. By juxtaposing numerical ratings with their justifications, we aim to gather a better understanding of workers' strengths and weaknesses on ideas' evaluation.

Our findings indicate that workers share similar viewpoints with experts in assessing aspects related to the desirability of an idea while having divergent views on the resources needed to implement it. We also find that workers generally tend to rate higher than experts and are more enthusiastic to original ideas even with unclear sources of funding. To better understand the overlap between workers and experts assessment, we conduct a series of thematic analysis of experts and workers justifications. We find that workers and experts tend to highlight similar concerns in an idea but assess them differently. For instance, when an idea has been implemented in other countries and is proposed to be adapted to a local market, experts focus on the lack of innovation while workers applaud its feasibility, and often merit the desirability for communities.

We leverage these findings to develop a human-AI collaborative approach to evaluate ideas by combining machine learning algorithms with feedback from crowd workers and experts (when available). We present `HybridEval`, a Bayesian framework that models the innovation value of an idea based on multi-criteria. To evaluate design ideas, `HybridEval` jointly learns a feature-based idea rating model and workers' modeling, with supervision from a small number of expert labels. Given our findings of workers' tendency of rating higher than experts, we model their performance by jointly measuring their reliability, i.e., the trustworthiness of their answers, and their bias, i.e., their tendency to underestimate or overestimate the rating for an idea. The machine learning model parameters and workers' performance are updated iteratively, allowing their learning processes to benefit from each other until an agreement on the rating of the ideas is reached. Eventually, for each idea an innovation value is estimated by `HybridEval`, which would allow designers evaluate ideas quickly and in large quantities.

To evaluate our method, we investigate: 1) the quality of crowd ratings; 2) the quality of ratings estimated by our proposed human-AI approach; and 3) finally, the impact of experts' ratings on the performance of our technique. Our results show that despite the disagreement between workers and experts on certain aspects in idea evaluation, using the `HybridEval` technique, we can approximate experts' ratings. In addition, by exploiting workers' ratings while modeling their reliability and bias, our approach is able to match the performance of experts with one-third of the experts' ratings. In summary, we make the following key contributions:

- We conduct a rigorous study to compare workers' and experts' evaluation of design ideas based on multiple criteria;
- We introduce a Bayesian framework that integrates a machine learning model with ratings from both experts and workers to collaboratively assess design ideas;
- We evaluate our framework on real-world datasets collected from three open innovation challenges and show that it can reach expert performance with significantly fewer experts' ratings, thus providing a scalable solution to ideas evaluation at scale.

2 RELATED WORK

In this section, we discuss the state of the art in idea evaluation with a focus on existing methods that leverage crowdsourcing in their modeling. Then, we review existing human-AI collaborative approaches, which are methodologically related to our work.

2.1 Idea Evaluation

Ideation is the creation of original ideas through a cycle of exploring multiple solutions (divergent thinking) and selecting the best one (convergent thinking). [24, 31]. Once an idea is clearly defined, designers need to assess its potential before moving forward and prototyping them. A common approach is to seek an expert evaluation, which is reliable but costly and time-consuming.

To circumvent the need for experts' evaluation, several methods have been developed for ideas' evaluation. These methods fall into two broad categories: model-based and rating-based evaluation. Model-based evaluation methods consist of metrics that mathematically describe the innovation value of a set of designs. For instance, Shah et al. [50] define a hierarchical model to decompose the concepts proposed in an idea into a tree, measure their similarity with others, and compute the number and variety of all collected ideas. These metrics cover different evaluation aspects but do not automatically allow the comparison of many ideas; therefore, Nelson et al. [42] introduce a scaling factor to existing metrics such that they diminish the value of a set of designs if they have low novelty. Another improvement consists of considering the uniformness degree of the ideas' distribution in measuring the variety of ideas [54]. An alternative approach consists of finding a trade-off between the diversity and quality of a design idea [1], where the authors propose a volume-based coverage method, namely the Determinantal Point Processes, with a multi-objective function to maximize the score attributed to high-quality and diverse design ideas. More recently, a hierarchical topic modeling approach [5] was used to build a failure taxonomy and predict the failure and its cause of design ideas.

Rating-based evaluation leverages ratings from a crowd of evaluators to select the most innovative ideas. A standard method consists of ranking ideas by the number of collected votes, which renders two main problems: 1) crowd workers have different levels of expertise; hence, their evaluations are not necessarily reliable; and 2) crowd workers tend to upvote the ideas that already have received

[1] https://www.openideo.com/

positive ratings which limits the number of ideas receiving votes. To address those limitations, Ahmed and Fuge [2] combine ratings, the idea features, and its uniqueness within a machine learning classifier to predict winning ideas in open innovation challenges. Another approach consists of using pairwise comparison to assess the similarity between ideas [3] or to infer the collective preference [8, 36]. A different methodology named "Bag of Lemons" aims to filter low-quality ideas [30, 39], where participants are given a fixed amount of "lemons," and they are asked to distribute them to the ideas they feel are the least likely to be selected as winners by an expert committee. Our approach is different in that it considers the reliability and bias of workers in evaluating ideas. Furthermore, we judge all submitted ideas based on a set of design-thinking criteria (desirability, feasibility, and viability), while [2] relies on votes from online platforms, which may result in biased outcomes as workers tend to only vote for ideas that have already received high ratings.

A separate line of research in human computation and crowdsourcing has investigated the task design and the results of soliciting participants in idea evaluation. For instance, one study [18] compared crowd-rated with expert-rated contests and the criteria for winning both. Other studies entangle the ideation process with its evaluation [17, 40, 45], where crowds create ideas and are instantly shown new phrases to simulate their creativity. Other collaborative ideation tools were designed to cooperate with users on drawing [37, 43], music creation [38] and brainstorming [4].

Compared with these methods, our study establishes the differences between workers and experts in providing a multi-criteria evaluation of design ideas. Our findings indicate workers' tendency to rate higher than experts. We integrate these findings in the design of a human-AI framework where we model workers' performance and bias. To the best of our knowledge, we are the first to compare the multi-criteria evaluation between workers and experts and inject the extracted results in the design of a human-AI framework.

2.2 Human-AI Collaborative Approaches

Human-AI systems aim to benefit from the complementary strengths of humans and machines intelligence by having humans and AI collaborate closely with each other, in order to solve tasks that are complex for AI models or humans alone [6, 28]. Traditionally, human computation has been used *before* training an AI model for data annotation [41, 52, 57] or feature selection [20, 47, 58]. Alternatively, it has been used *after* model training for model selection [46, 49], evaluation [19, 32] and debugging [9, 33, 51]. The aforementioned methods tend to consider the human computation and artificial intelligence as disentangled processes. Recent efforts have sought for a deep integration between human computation and training AI models [6, 7, 11, 25, 56]. These methods propose to enhance the collaboration between humans and AI by iteratively learning human characteristics (e.g., reliability and bias) and model parameters in a mutually boosting manner until the desired result is achieved. Our work can be seen as a development of this recent line of work in the context of multi-criteria ideas evaluation. We use workers' evaluation only, to train a machine learning model for each criteria, which in turns uses a small amount of experts labels to rectify workers modeling. These two processes alternate until an agreement on the multi-criteria evaluation is reached.

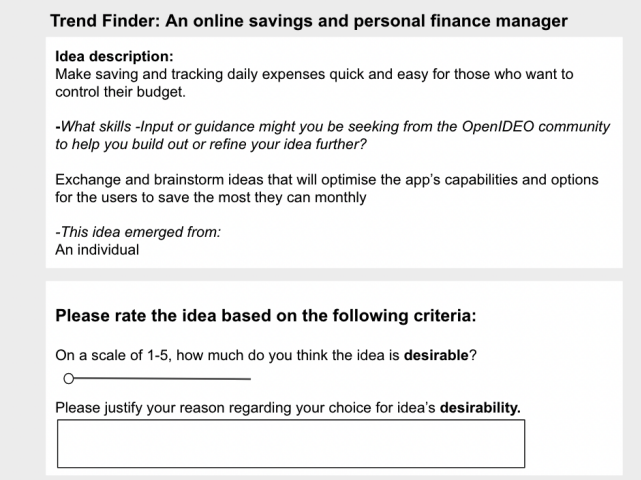


Figure 1: Interface of the idea rating task.

3 METHODOLOGY

In this section, we first describe our crowdsourcing task for ideas rating and then introduce HybridEval to learn the rating of each criteria from workers' ratings and a small set of experts' ratings.

3.1 Crowdsourcing Study

We first present our design for the idea rating task, which is used to collect data to evaluate our proposed framework. In the following, we describe a set of criteria to assess the innovation value of an idea and then describe the crowdsourcing task. We present an analysis of the effectiveness of rating ideas based on multi-criteria from both crowd workers and experts in the next section.

3.1.1 Innovation Criteria for Rating Ideas. In our work, we rely on a design thinking methodology that originated from IDEO [23] to test ideas and determine their innovation value. It consists in evaluating the co-existence of three main criteria, which are desirability, feasibility and viability [12, 13, 22]. We also conducted a series of interviews with two experts in the field of design engineering, who confirmed the validity of these criteria in evaluating the innovation value of design ideas and their success potential.

- **Desirability.** A test for desirability measures the usefulness of an idea and whether it addresses an urgent need for end-users.
- **Feasibility.** A test for feasibility assess the resources needed to implement an idea. These resources mainly include the monetary funding and the technical knowledge required to implement the idea.
- **Viability.** A test for viability measures whether an idea is sustainable on the long-term.

We additionally add another criteria which is "Overall Feeling", to get the overall impression of workers and experts of an idea. All the criteria including the overall feeling were rated using a 5-point Likert scale.

3.1.2 Dataset. We collected our dataset from the OpenIDEO website [35], which provides access to ideas submitted in challenges to

Dataset	No. Ideas	Challenge Title
Dementia	120	How might we better support family caregivers as they care for a loved one with dementia?
Financial	75	How might we use the power of communities to financially empower those who need it most?
Energy	86	How might communities lead the rapid transition to renewable energy?

Table 1: Challenges selected from OpenIDEO with their corresponding number of ideas.

solve social issues. We collected 281 ideas from three randomly selected challenges summarized in Table 1. Each idea in the challenge consists of a title and a long text describing the idea.

3.1.3 Task Design. We published a task on Mturk, where workers rate ideas. Note that our task was approved by the ethical committee of our institute and that we include an informed consent at the beginning of the task with an explanation on how to rate the ideas based using the innovation criteria. Then, we ask participants to provide ratings for each design idea on the four innovation criteria. We additionally ask workers to justify their rating for each criteria [41] to ensure they spend enough time on the task. The interface of our task is depicted in Figure 1. For the crowdsourcing task, we recruited 49 crowdworkers from Amazon MTurk workers, with Masters Qualification, who have demonstrated excellent performance across different tasks, with a HIT approval rate above 70%. Each idea is rated by two or three workers and each task took 8 minutes to complete on average and those who finished the task were given a 1 (USD) reward, as per the US federal government's minimum wage mandate. We also collect expert labels that allow us to bootstrap and evaluate our model. We advertised the project and interviewed candidates with respect to expertise and dedication. To that end, we hired two designers from the industrial design faculty (same experts we interviewed) and asked each of them to rate 281 ideas on a 5-point Likert scale on the same criteria.

3.2 `HybridEval` for Idea Rating

We now introduce our `HybridEval` approach. We first formally define our problem and then describe our overall framework.

3.2.1 Problem Formulation. Throughout this paper, we use boldface lowercase letters to denote vectors and boldface uppercase letters to denote matrices. For an arbitrary matrix $\mathbf{M}$, we use $\mathbf{M}_{i,j}$ to denote the entry at the i-th row and j-th column. We use capital letters (e.g., $\mathcal{P}$) in calligraphic math font to denote sets and $|\mathcal{P}|$ to denote the cardinality of a set $\mathcal{P}$.

We denote $\mathcal{I}$ a set of ideas, C the set of criteria for idea evaluation, and $\mathcal{J}$ a set of workers rating the ideas. Each idea $i \in \mathcal{I}$ is represented with a feature vector $\mathbf{x}_i$ and $\mathcal{I}_E$ is a subset of ideas rated by experts. We use $\mathbf{A}^c$ to denote the worker-idea matrix where each element $\mathbf{A}^c_{i,j}$ is a rating between 1 and 5 given by a worker $j \in \mathcal{J}$ to an idea i for a certain criterion $c \in C$.

Problem Definition. Given $\mathcal{I}$ (the set of ideas), $\mathbf{x}_i$ (the feature vector of $i \in \mathcal{I}$), $\mathcal{J}$ (the set of workers rating the ideas), $\mathcal{A}$ (the worker-idea rating matrices for all criteria), and $\mathcal{I}_E$ (the subset of ideas rated by experts), we aim at inferring the rating of all criteria in C for each idea in $\mathcal{I} \setminus \mathcal{I}_E$ using $\mathbf{x}_i$ and $\mathbf{A}^c$.

3.2.2 The `HybridEval` Framework. We model the score of idea i for a criterion z_i with a Gaussian distribution, the worker's reliability r_j with a Gamma distribution and their bias b_j with a Gaussian distribution, as given next:

$$z_i \sim \mathcal{N}(\mu_i, \sigma_i),\ r_j \sim \Gamma(A, B),\ b_j \sim \mathcal{N}(m, \frac{1}{\alpha}). \tag{1}$$

The Gamma distribution allows us to quantify our confidence in estimating the worker's reliability. Note that for ease of reading, we do not distinguish the variables for different criteria – the same framework applies to each criterion, with a distinct set of variable settings (i.e., parameters of the framework).

The parameter μ_i of the idea's score z_i is predicted from the idea features $\mathbf{x}_i$ through an arbitrary machine learning model:

$$\mu_i = \text{softmax}(f^{\mathcal{W}}(\mathbf{x}_i)), \tag{2}$$

3.2.3 Variational Inference for `HybridEval`. Learning the parameters of `HybridEval` resorts to maximizing the following likelihood function:

$$p(\mathbf{A}) = \int p(\mathbf{A}, \mathbf{z}, \mathbf{r}, \mathbf{b}|\mathbf{X}; \mathcal{W})\, d\mathbf{z}, \mathbf{r}, \mathbf{b}, \tag{3}$$

where $\mathbf{z}$ is the latent score for one of the criteria used to evaluate ideas, and $\mathbf{r}$ and $\mathbf{b}$ are the latent reliability scores and biases for all workers. $\mathbf{X}$ represents the feature matrix of all ideas and $\mathcal{W}$ is the set of machine learning parameters.

Since Eq. (3) contains more than one latent variable, it is computationally infeasible to optimize [53]. Therefore, we consider the log of the likelihood function, which could be optimized using the variational expectation-maximization algorithm [53] in two steps: 1) the E-step, where we approximate $p(\mathbf{z}, \mathbf{r}, \mathbf{b}|\mathbf{A}, \mathbf{X}; \mathcal{W})$ with a variational distribution $q(\mathbf{z}, \mathbf{r}, \mathbf{b})$; and 2) the M-step, where we learn the parameters $\mathcal{W}$ given the newly inferred latent variables. In the following, we describe both steps.

E-step. In the E-step, we iterate between updating the criterion score of an idea z_i, the worker's reliability and bias r_j and b_j. To update the criterion score $q(z_i)$, we use:

$$q(z_i) \sim \mathcal{N}(\frac{W}{V}, \frac{1}{V}), \text{where} \begin{cases} W = \sum_j \frac{A_j}{B_j}(\mathbf{A}_{i,j} - m_j) + \frac{\mu_i}{\sigma_i^2}, \\ V = (\sum_j \frac{A_j}{B_j} + \frac{1}{\sigma_i^2}). \end{cases} \tag{4}$$

To update the worker's reliability and bias, we use respectively Eq. (5) and Eq. (6).

$$q(r_j) \sim Gamma(X, Y), \text{where} \\ \begin{cases} X = A_j + \frac{|\mathcal{I}_j|}{2}, \\ Y = B_j + \frac{1}{2}(\frac{|\mathcal{I}_j|}{\alpha_j} + \sum_i [\mathbf{A}^2_{i,j} + \sigma_i^2 + 2\mu_i(m_j - \mathbf{A}_{i,j}) - 2\mathbf{A}_{i,j} m_j]). \end{cases} \tag{5}$$

$$q(b_j) \sim \mathcal{N}(\frac{L}{K}, \frac{1}{K}), \text{where} \begin{cases} K = \frac{A_j|\mathcal{I}_j|}{B_j} + \alpha_j, \\ L = \alpha_j m_j + \frac{A_j}{B_j} \sum_i (\mathbf{A}_{i,j} - \mu_i). \end{cases} \quad (6)$$

M-step. Given the criterion score of an idea, the worker's reliability and bias inferred in the E-step, the M-step learns the parameters $\mathcal{W}$ of the machine learning model. This step requires to optimize the inverse of the cross-entropy between $q(z_i)$ and $p(z_i|\mathbf{x}_i;\mathcal{W})$, which is widely used as the loss function for many classifiers and can be optimized using standard model training methods (e.g., back-propagation for neural networks). For ideas labeled by experts ($i \in \mathcal{I}_E$), the labels can be used to fix $q(z_i)$, thereby being incorporated into the learning process.

4 COMPARING WORKERS AND EXPERTS

We conduct a series of quantitative and thematic analysis of experts and workers ratings. For our thematic analysis, we follow the process indicated by Braun and Clark [10]. We start by defining general categories of codes based on the reasons mentioned for ratings. We then cluster categories based on common themes with a focus on our goal to identify main sources of agreement and disagreement between the different groups (experts and workers). Finally, we analyse the main themes and report our results.

4.1 Inter-Experts Agreement

Overall, we observe that experts have a higher agreement on assigning low ratings than high ratings. As an example, Figure 2 shows the inter-rater agreement between the two experts with a root mean squared error (RMSE) on the Energy dataset. Low values of RMSE indicate high agreement between experts and vice versa.

4.1.1 When do experts agree? The agreement between experts on low ratings is mainly due to vague ideas or when their funding source is undetermined. For instance, we find both experts agree when an idea is shortly described. One expert commented: "This idea needs more meat to it to be more of an idea. now it just feels like a proposal." The other expert had a similar comment, where he said: "The idea is not really well explained or elaborate for me to judge on desirability." We also observe that experts agree on assigning a low rate for feasibility when an idea requires the involvement of many parties. Take the example of an idea that requires multi-stakeholders to get involved, both experts assigned 2 for feasibility and one of them commented that "it sounds like a difficult task to bring all these stakeholders together."

Experts rarely assign high ratings, with less than 15% of ideas receiving high ratings from both experts. High ratings are typically given when a prototype has been developed and tested or when the idea addresses a real problem with a clear solution. The experts agreed on high ratings when the idea is a "fully working concept" or the funding schema is well defined and when crowdfunding has been done.

4.1.2 When do experts disagree? Some of the proposed ideas in the energy challenge present novel solutions but are not sustainable over time or require different parties' involvement. These ideas are the ones where experts tend to disagree most on their rating: While one expert highlights its impact and novelty, the other expert questions its ability to sustain itself in the long term. Another source of disagreement between experts is when an idea has already been implemented: While one expert applauds its feasibility saying "It already exist so it is feasible.", the other expert criticizes its lack of novelty saying "it is not clear to me what the added value is for all stakeholders."

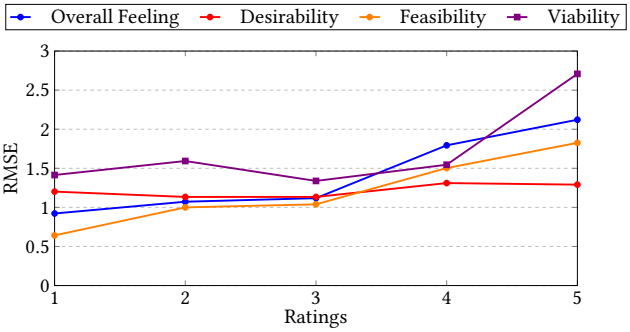


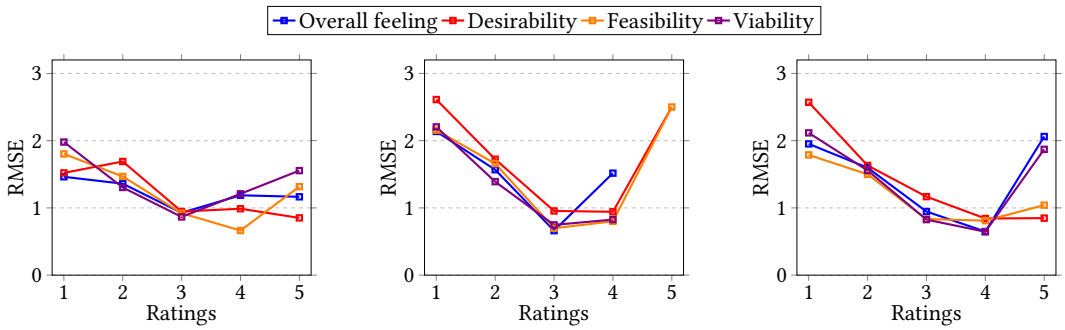
Figure 2: Inter-rater agreement between experts.

4.2 Inter-Workers Agreement

Overall, we observe that workers tend to give high ratings more frequently than low ratings specially when rating the desirability of an idea. We find that ratings above three are at least two times the ratings under three for all three criteria. We show the distribution of ratings collected from experts and crowd workers in Figure 5 in the appendix.

4.2.1 When do workers agree? By conducting our thematic analysis, we extract three main reasons for which workers tend to provide high ratings: 1) when an idea benefits a large community, for instance a large rural area or women in developing countries; 2) if the idea helps raising awareness. For instance an idea about teaching children financial knowledge, workers found that it can "help children learn to make better financial decisions as adults" and "make good financial choices from a young age"; 3) ideas that require low resources to be implemented. Take the example of an idea of solar-powered classrooms for students, both workers assessed that "all the tech is there" to implement the idea. We also extract reasons for which workers agree on assigning low ratings. These reasons are mainly due to lack of clarity, complexity of the idea or if there are safety and sustainability concerns.

4.2.2 When do workers disagree? We analyse the justifications workers provide when they disagree on ideas' ratings. We find the main sources of disagreement are the feasibility of an idea and its potential attractiveness. Take the example of an idea about a platform to teach financial knowledge in 30 seconds or less: while a worker thinks that "people would be willing to help in the development", another worker disagreed on one's ability to "learn everything they need in that short of time". Another example is about adding the living wage of parties involved in making a product in its package. One worker found the idea viable, another worker judged it provides "unnecessary information".



(a) Dementia (b) Financial (c) Energy

Figure 3: Disagreement in the ratings between expert and crowd in Dementia (a), Financial (b) and Energy (c) datasets. Note several curves in the Financial challenge do not have RMSE when ratings=5 because no experts rated 5.

4.3 Experts-Workers Agreement

4.3.1 Overall Comparison. Overall, we observe workers tend to assign the highest score more often than the expert. On the one hand, workers and experts have the same tendency to give ratings of two and four. On the other hand, the expert assigns ratings one and three more frequently compared to workers. These observations are consistent across the different criteria and datasets. They showcase the importance of modeling workers' bias and tendency to assign specific ratings more than others. We also measure the disagreement between the workers' average ratings and experts' ratings in RMSE as depicted in Figure 3. For all datasets, we observe high disagreement between experts and workers in assigning extreme ratings (one and five) and lower disagreement for ratings between two and four.

4.3.2 When do workers agree with experts? Our analysis of the reasons supporting experts' and workers' rating show that workers and experts tend to agree on assigning high scores to ideas with a clear funding schema and when they tackle real problems. They also tend to agree on giving low scores for three prominent cases: 1) when an idea is unclear about its funding source and its action plan; 2) when an idea has no added value and requires a lot of investment from people and experts and 3) when an idea is hard to implement. For instance, workers and experts give a low score to ideas that have "too narrow of a focus", "not clear" or "quite complex".

4.3.3 When do workers disagree with experts? We identified four primary sources of disagreement: 1) experts discard ideas that have been already implemented for their lack of novelty, while workers tend to find them valuable and relevant for many people. For instance, for an idea about adding some charging stations for electronic vehicles, the expert assigned a low score because it "already exists in Europe", while workers found it a "convenient solution"; 2) experts discard ideas for their lack of clarity while workers focus on their attractiveness and ease of implementation. For one idea about reducing a community's carbon footprint, experts questioned how the idea can be implemented and advertised while workers found it to be "a good plan to roll back emissions." 3) experts discard ideas out of topic, while workers would still look into them. Take the example of one idea presented in the energy challenge to help rural women and youth alleviate poverty by supporting them in developing their own business, experts assigned a low score because 'there is nothing here about renewable energy.', while workers assigned a high score saying "Gender equality throughout the world is a bold task" 4) experts encourage solutions that tackle a real problem while workers might rate them low if they require a lot of funding. An example of such case is an idea for a renewable electricity generation for cold storage facilities for fish, the expert found the idea feasible and will greatly improve the quantity of sellable fish, while workers found the project "expensive to implement".

5 HYBRIDEVAL EXPERIMENTS AND RESULTS

This section presents the empirical validation of `HybridEval` [2]. We first evaluate the performance of `HybridEval` against several baseline methods (Section 5.2), and then investigate the effect of experts' labels on the performance of the framework (Section 5.3).

5.1 Experimental Setup

5.1.1 Dataset & Features. We use the same data as the crowdsourcing study including ideas from three challenges, namely Financial, Dementia and Energy (statistics shown in Table 1). We feed our machine learning model with the ideas' description. To represent them, we test pre-trained GoogleNews word2vec word-embedding, contextual embedding from BERT and TF-IDF features. We empirically find TF-IDF to work best on our dataset. This can be due to the long sentences used in the ideas description. In some preliminary experiments, we used the length of the ideas as an input feature of the machine learning model. However, it did not impact the overall performance. Thus, we removed it from the input features.

5.1.2 Models & Comparison Methods. For the machine learning part of our framework although our main focus was not on the performance of different ML models, we tested a set of machine learning models (linear, tree-based, and neural networks) and observed that the AdaboostRegressor performed the best (due to the effectiveness of boosting in combining multiple weak learners). We set n-estimators = 200 and learning-rate=1 after a randomized search for hyperparameters, to train a model for each of the criteria we are interested in (i.e., described in Section 3.1.1). We brought the ratings of both the expert and crowd into the range of [0, 1] using the function t(x)=(x-1)/4.

[2]The source code and dataset is available at: https://github.com/mesbahs/HybridEval

We compare our approach with the following methods: 1) Crowd-Rating, where we compute the average of worker's rating for each criterion; 2) ML-Expert, where we use experts' ratings to train a machine learning model; 3) ML-Crowd, where we use the average of workers' ratings in addition to the experts ratings to train a machine learning model.

5.1.3 Evaluation Protocol. For the given datasets (Financial, Dementia and Energy), we split them into subsets of the training (i.e., 20%, 40%, 60% and 80%) and test data (i.e., 80%, 60%, 40%, 20%) to simulate the effect of limited expert rating availability. The subsets are randomly selected, and experiments are repeated 10 times for each size setting. We evaluate the performance of our approach by comparing it to experts' rating using Root Mean Square Error (RMSE). Low RMSE indicate high performance.

5.2 HybridEval Performance

Figures 4(a-c) show the performance of the compared methods on the three datasets. We observe that among the comparison methods, Crowd-Rating outperforms other techniques when rating the *feasibility* criterion across all three datasets and has lower performance than other methods for other criteria. This performance difference can be explained by crowd's ability to assess idea's ease of implementation which reflects their ability to accurately rate idea's feasibility. We observe ML-Expert outperforms ML-Crowd and Crowd-Rating in the *desirability*, *viability* and *overall feeling* criteria across all datasets, which indicates the effectiveness of expert labels for training the machine learning model, as well as the consistency of expert labels in the training and test set.

Most importantly, HybridEval outperforms all baseline methods on the criteria of *desirability*, *viability* and *overall feeling*, and achieves comparable results with Crowd-Rating on *feasibility*. Comparing the performance of HybridEval with those of ML-Expert and ML-Crowd, we find that while mixing expert labels with crowd labels by averaging crowd ratings does not result in a high performance model (i.e., ML-Crowd) and can even degrade the performance compared with using expert labels only (i.e., ML-Expert), carefully mixing experts' and workers' labels can lead to a model that performs even better than using experts' labels only. The result indicates the effectiveness of HybridEval in combining workers' and experts' labels for model training, which can be attributed to HybridEval's ability in inferring workers' performance characteristics (reliability and bias). The results on ML-Crowd verifies the effectiveness of the *e* step in our approach. In HybridEval, the worker reliability and bias are updated incrementally at each iteration. By doing so, we keep ratings from the workers that are more reliable and close to the true ratings of the experts and reduce the error caused by unreliable workers.

5.3 Impact of Supervision Degree

As shown in Figures 4(a-c), HybridEval consistently outperforms ML-Expert using 40% of expert labels. Moreover, we observe that HybridEval trained with only 20% of expert labels obtains comparable results to ML-Expert trained on 60% of the expert labels on the Energy dataset across different criteria. This shows that by leveraging the human-AI approach we can reduce the amount of expert label needed for rating the ideas. In Figure 4a-4c, when increasing the number of training samples, the performance decreases. This is due to the small size of the dataset used for the evaluation in the Financial and Energy datasets. For instance when we use 80% of the data for training purpose only a small subset of the data (i.e., 20%) will remain for testing which is only 15 samples for the Financial dataset.

6 DISCUSSION AND CONCLUSION

In this paper, we developed a human-AI system for evaluating ideas by combining machine learning algorithms and ratings from crowd workers, with the goal of reducing the cost of multi-criteria idea assessment. In this section, we discuss the ratings collected from the crowd and the results produced by our human-AI approach, along with the design implications, limitations, and future work opportunities.

6.1 Crowd Idea Ratings

We collected workers' ratings for ideas on three IDEO challenges. We made several observations. First, we observed that workers are able to rate the *desirability* of an idea, which can be explained by their ability to assess the novelty of an idea and its attractiveness. Workers' ability to spot innovative ideas has been observed in previous studies [2] where ideas selected by experts in OpenIdeo challenges received more comments than other ideas even before the evaluation announcement. Second, we observed a significant disagreement between workers and experts in giving low ratings, which can be interpreted by workers' tendency to be more generous in rating than experts and show the importance of modeling worker's bias. Finally, we analyzed the rationales behind the crowd and expert ratings and found that the primary source of disagreement between experts and crowd workers resides in weighing the importance of specific criteria. For instance, experts discard ideas with unclear funding schema while workers might applaud their desirability. In these cases, workers provide a different perspective from experts, which can be valuable in rating ideas. These analyses are aligned with the research in [8] that takes into account different perspectives in idea prioritization. Workers are capable of assessing *feasibility* when the necessary resources are clear and accessible, but tend to underestimate the cost and accessibility of investors for appealing ideas. This highlights the importance of using a limited number of expert labels to improve the performance of the human-AI framework in modeling worker's performance.

6.2 HybridEval for Idea Evaluation

We explored the behavior of our approach on three different datasets with varying supervision degrees. Our framework achieves the best trade-off compared with baseline methods in estimating the rating for different criteria across the three datasets. When appropriately modeled, workers' ratings provide an important contribution to the model training. We also observed that HybridEval has comparable performance with only 20% of training data to a machine learning model trained with 60% of the training data across different datasets and criteria. Our experimental findings suggest that our human-AI approach can reduce the need for expert labels for rating ideas. Moreover, we find that our modeling strategy for workers' performance allows to adjust their rating with experts' rating and identifies ways to learn effectively from workers' ratings.

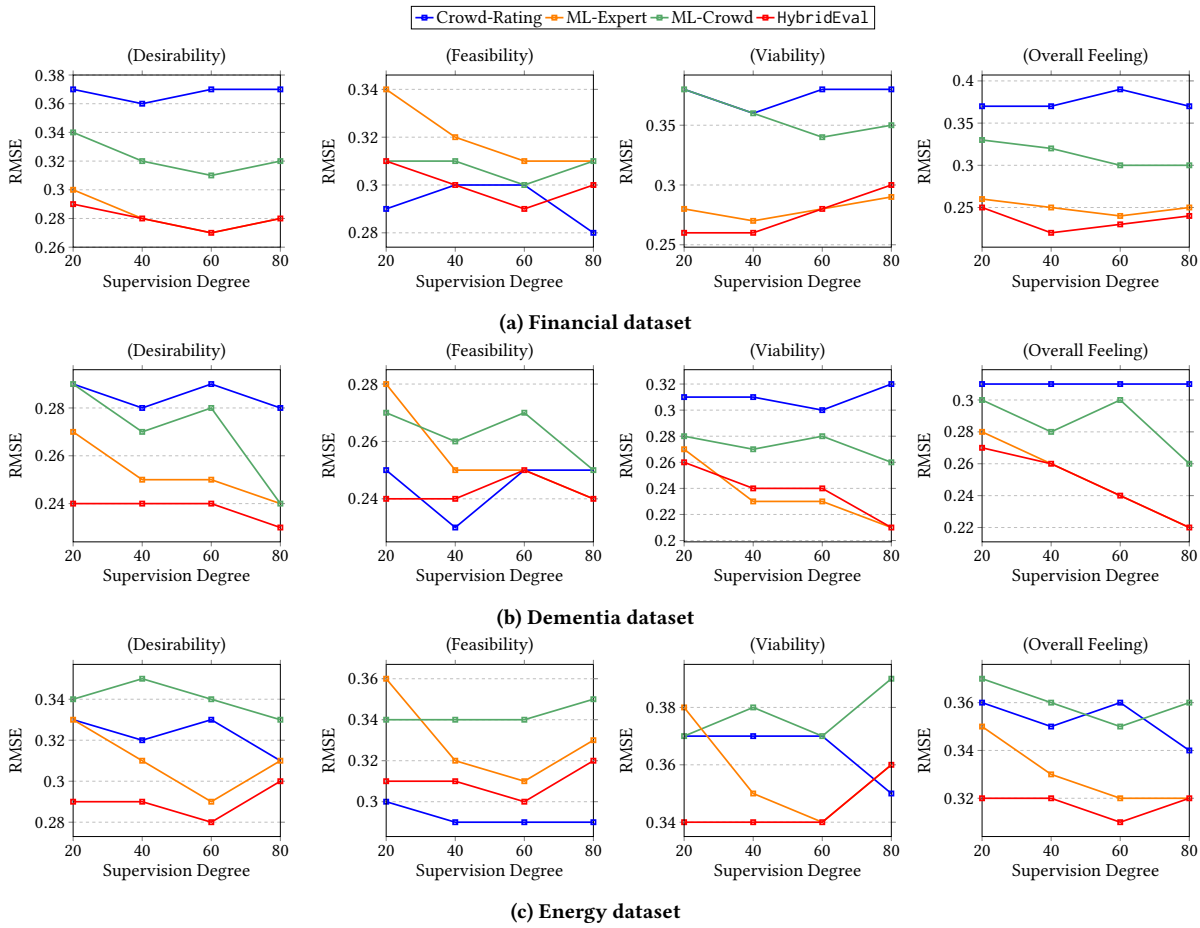


Figure 4: Comparison between the performance of our approach `HybridEval` and the baseline methods `Crowd-Rating`, `ML-Expert` and `ML-Crowd`, across all the criteria and the three datasets, measured by RMSE. See Tables 2, 3 and 4 in appendix for the exact number.

6.3 Implications for Design and Beyond

Our approach can be used in various applications where evaluation for open-ended proposals or answers is needed. In the design context, these applications include not only the evaluation of design ideas, but also the elicitation of design insights – an essential step to support the ideation process – from large crowds [26, 48]. Our approach allows to address a major bottleneck of scalability in existing design practice, where design insights are usually obtained in small-scale studies (e.g., interviews). Our approach also applies to evaluating answers from crowdsourcing in general, not only as a means to support design through e.g., crowd ideation or insights elicitation, but also beyond, e.g., content creation like translation or writing product reviews. For example, by incorporating our approach into online platforms for e-commerce, we can assess effectively textual data quickly and in large-scale. Specifically, it can also be used for evaluating online reviews on a set of criteria (e.g., helpfulness, sentiment) to show the most relevant ones to customers [21]. Our approach can also be applied in education and scientific contexts, e.g., to evaluate assignments from a large group of students on MOOC platforms, or to evaluate the conformity of scholarly reviews from a large pool of reviewers in attractive scientific conferences [7, 34].

6.4 Limitations and Future Work

We evaluated our framework on three challenges of the OpenIDEO. Further experiments are needed on larger datasets and from different websites to obtain a comprehensive understanding of our approach's full capabilities and limitations. For future work, we plan to investigate ways to decompose the crowdsourcing task to obtain more annotations on the ideas. We also plan to use natural language processing techniques to summarize the ideas for workers such that they accomplish the crowdsourcing task in a shorter amount of time.

ACKNOWLEDGMENTS

This research has been supported by the TU Delft Design@Scale AI Lab. The research has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 874724. Alessandro Bozzon acknowledges the Dutch Research Council (NWO) support under Grant No. 314–99-300.

REFERENCES

[1] Faez Ahmed and Mark Fuge. 2018. Ranking ideas for diversity and quality. *Journal of Mechanical Design* 140, 1 (2018), 011101.

[2] Faez Ahmed and Mark D. Fuge. 2017. Capturing Winning Ideas in Online Design Communities. In *CSCW*. ACM, 1675–1687.

[3] Faez Ahmed, Sharath Kumar Ramachandran, Mark Fuge, Samuel Hunter, and Scarlett Miller. 2019. Interpreting idea maps: Pairwise comparisons reveal what makes ideas novel. *Journal of Mechanical Design* 141, 2 (2019), 021102.

[4] Salvatore Andolina, Khalil Klouche, Diogo Cabral, Tuukka Ruotsalo, and Giulio Jacucci. 2015. InspirationWall: Supporting Idea Generation Through Automatic Information Exploration. In *Creativity & Cognition*. ACM, 103–106.

[5] Sequoia R Andrade and Hannah S Walsh. 2022. Discovering a Failure Taxonomy for Early Design of Complex Engineered Systems Using Natural Language Processing. *Journal of Computing and Information Science in Engineering* 23, 3 (2022), 031001.

[6] Ines Arous, Jie Yang, Mourad Khayati, and Philippe Cudré-Mauroux. 2020. Opencrowd: A human-ai collaborative approach for finding social influencers via open-ended answers aggregation. In *Proceedings of The Web Conference 2020*. 1851–1862.

[7] Ines Arous, Jie Yang, Mourad Khayati, and Philippe Cudré-Mauroux. 2021. Peer Grading the Peer Reviews: A Dual-Role Approach for Lightening the Scholarly Paper Review Process. In *WWW*. ACM / IW3C2, 1916–1927.

[8] Yukino Baba, Jiyi Li, and Hisashi Kashima. 2020. CrowDEA: Multi-View Idea Prioritization with Crowds. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 8. 23–32.

[9] Agathe Balayn, Panagiotis Soilis, Christoph Lofi, Jie Yang, and Alessandro Bozzon. 2021. What do you mean? Interpreting image classification with crowdsourced concept extraction and analysis. In *Proceedings of the Web Conference 2021*. 1937–1948.

[10] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative research in psychology* 3, 2 (2006), 77–101.

[11] Alexander Braylan and Matthew Lease. 2020. Modeling and aggregation of complex annotations via annotation distances. In *Proceedings of The Web Conference 2020*. 1807–1818.

[12] Tim Brown. 2008. Design Thinking. *harvard business review* (2008), 1.

[13] Tim Brown. 2009. *Change by design: How design thinking creates new alternatives for business and society*. Collins Business.

[14] Alex Burnap, Yi Ren, Richard Gerth, Giannis Papazoglou, Richard Gonzalez, and Panos Y Papalambros. 2015. When crowdsourcing fails: A study of expertise on crowdsourced design evaluation. *Journal of Mechanical Design* 137, 3 (2015), 031101.

[15] EV Buskirk. 2010. Google struggles to give away $10 million. *Wired Magazine* (2010).

[16] Flaviano Celaschi, Manuela Celi, and Laura Mata García. 2011. The extended value of design: an advanced design perspective. *Design Management Journal* 6, 1 (2011), 6–15.

[17] Kimmy Wa Chan, Stella Yiyan Li, and John Jianjun Zhu. 2015. Fostering customer ideation in crowdsourcing community: The role of peer-to-peer and peer-to-firm interactions. *Journal of Interactive Marketing* 31 (2015), 42–62.

[18] Liang Chen and De Liu. 2012. Comparing strategies for winning expert-rated and crowd-rated crowdsourcing contests: First findings. In *18th Americas Conference on Information Systems 2012, AMCIS 2012*. 97–107.

[19] Yuyan Chen, Yanghua Xiao, and Bang Liu. 2022. Grow-and-Clip: Informative-yet-Concise Evidence Distillation for Answer Explanation. *arXiv preprint arXiv:2201.05088* (2022).

[20] Alvaro HC Correia and Freddy Lecue. 2019. Human-in-the-loop feature selection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 2438–2445.

[21] Gerardo Ocampo Diaz and Vincent Ng. 2018. Modeling and prediction of online product review helpfulness: a survey. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 698–708.

[22] Arkin Efeoglu, Charles Møller, Michel Sérié, and Harry Boer. 2013. Design thinking: characteristics and promises. In *Proceedings of 14th International CINet Conference on Business Development and Co-creation*. 241–256.

[23] Joe Gerber. 2000. How to Prototype a New Business. https://www.ideou.com/blogs/inspiration/how-to-prototype-a-new-business. Accessed: 2022-05-31.

[24] Milene Gonçalves, Carlos Cardoso, and Petra Badke-Schaub. 2016. Inspiration choices that matter: the selection of external stimuli during ideation. *Design Science* 2 (2016).

[25] Mitchell L. Gordon, Michelle S. Lam, Joon Sung Park, Kayur Patel, Jeff Hancock, Tatsunori Hashimoto, and Michael S. Bernstein. 2022. Jury Learning: Integrating Dissenting Voices into Machine Learning Models. In *CHI Conference on Human Factors in Computing Systems*. ACM. https://doi.org/10.1145/3491102.3502004

[26] Kosa Goucher-Lambert and Jonathan Cagan. 2019. Crowdsourcing inspiration: Using crowd generated inspirational stimuli to support designer ideation. *Design Studies* 61 (2019), 1–29.

[27] Yan Huang, Param Vir Singh, and Kannan Srinivasan. 2014. Crowdsourcing new product ideas under consumer learning. *Management science* 60, 9 (2014), 2138–2159.

[28] Jialun Aaron Jiang, Kandrea Wade, Casey Fiesler, and Jed R Brubaker. 2021. Supporting serendipity: Opportunities and challenges for Human-AI Collaboration in qualitative analysis. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–23.

[29] Mark Klein and Ana Cristina Garcia. 2014. The bag of stars: High-speed idea filtering for open innovation. *Available at SSRN 2387180* (2014).

[30] Mark Klein and Ana Cristina Bicharra Garcia. 2015. High-speed idea filtering with the bag of lemons. *Decision Support Systems* 78 (2015), 39–50.

[31] Janin Koch, Nicolas Taffin, Michel Beaudouin-Lafon, Markku Laine, Andrés Lucero, and Wendy E Mackay. 2020. ImageSense: an intelligent collaborative ideation tool to support diverse human-computer partnerships. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW1 (2020), 1–27.

[32] Rafal Kocielnik, Saleema Amershi, and Paul N Bennett. 2019. Will you accept an imperfect ai? exploring designs for adjusting end-user expectations of ai systems. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–14.

[33] Todd Kulesza, Margaret Burnett, Weng-Keen Wong, and Simone Stumpf. 2015. Principles of explanatory debugging to personalize interactive machine learning. In *Proceedings of the 20th international conference on intelligent user interfaces*. 126–137.

[34] Igor Labutov and Christoph Studer. 2017. JAG: a crowdsourcing framework for joint assessment and peer grading. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 31.

[35] Karim R Lakhani, Anne-Laure Fayard, Natalia Levina, and Stephanie Healy Pokrywa. 2012. OpenIDEO. *Harvard Business School Technology & Operations Mgt. Unit Case* 612-066 (2012).

[36] Jiyi Li. 2022. Context-based Collective Preference Aggregation for Prioritizing Crowd Opinions in Social Decision-making. In *Proceedings of the ACM Web Conference 2022*. 2657–2667.

[37] Yuyu Lin, Jiahao Guo, Yang Chen, Cheng Yao, and Fangtian Ying. 2020. It Is Your Turn: Collaborative Ideation With a Co-Creative Robot through Sketch. In *CHI*. ACM, 1–14.

[38] Ryan Louie, Andy Coenen, Cheng Zhi Huang, Michael Terry, and Carrie J. Cai. 2020. Novice-AI Music Co-Creation via AI-Steering Tools for Deep Generative Models. In *CHI*. ACM, 1–13.

[39] Ioanna Lykourentzou, Faez Ahmed, Costas Papastathis, Irwyn Sadien, and Konstantinos Papangelis. 2018. When crowds give you lemons: Filtering innovative ideas using a diverse-bag-of-lemons strategy. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 1–23.

[40] Maximilian Mackeprang, Abderrahmane Khiat, and Claudia Müller-Birn. 2018. Concept validation during collaborative ideation and its effect on ideation outcome. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–6.

[41] Tyler McDonnell, Matthew Lease, Mucahid Kutlu, and Tamer Elsayed. 2016. Why is that relevant? Collecting annotator rationales for relevance judgments. In *HCOMP*. AAAI Press, USA, 139–148.

[42] Brent A Nelson, Jamal O Wilson, David Rosen, and Jeannette Yen. 2009. Refined metrics for measuring ideation effectiveness. *Design Studies* 30, 6 (2009), 737–743.

[43] Changhoon Oh, Jungwoo Song, Jinhan Choi, Seonghyeon Kim, Sungwoo Lee, and Bongwon Suh. 2018. I Lead, You Help but Only with Enough Details: Understanding User Experience of Co-Creation with Artificial Intelligence. In *CHI*. ACM, 649.

[44] Marion K Poetz and Martin Schreier. 2012. The value of crowdsourcing: can users really compete with professionals in generating new product ideas? *Journal of product innovation management* 29, 2 (2012), 245–256.

[45] Samuel Rhys Cox, Yunlong Wang, Ashraf Abdul, Christian Von Der Weth, and Brian Y. Lim. 2021. Directed diversity: Leveraging language embedding distances for collective creativity in crowd ideation. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–35.

[46] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. Why should I trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*. ACM, San Francisco, CA, USA, 1135–1144.

[47] Md Abdus Salam, Mary E Koone, Saravanan Thirumuruganathan, Gautam Das, and Senjuti Basu Roy. 2019. A human-in-the-loop attribute design framework for classification. In *The World Wide Web Conference*. 1612–1622.

[48] Brita Schemmann, Andrea M Herrmann, Maryse MH Chappin, and Gaston J Heimeriks. 2016. Crowdsourcing ideas: Involving ordinary users in the ideation phase of new product development. *Research Policy* 45, 6 (2016), 1145–1154.

[49] Bruno Schneider, Dominik Jäckle, Florian Stoffel, Alexandra Diehl, Johannes Fuchs, and Daniel Keim. 2018. Integrating data and model space in ensemble learning by visual analytics. *IEEE Transactions on Big Data* 7, 3 (2018), 483–496.

[50] Jami J Shah, Steve M Smith, and Noe Vargas-Hernandez. 2003. Metrics for measuring ideation effectiveness. *Design studies* 24, 2 (2003), 111–134.

[51] Shahin Sharifi Noorian, Sihang Qiu, Ujwal Gadiraju, Jie Yang, and Alessandro Bozzon. 2022. What Should You Know? A Human-In-the-Loop Approach to Unknown Unknowns Characterization in Image Recognition. In *Proceedings of the ACM Web Conference 2022*. 882–892.

[52] Rion Snow, Brendan O'connor, Dan Jurafsky, and Andrew Y Ng. 2008. Cheap and fast–but is it good? evaluating non-expert annotations for natural language tasks. In *Proceedings of the 2008 conference on empirical methods in natural language processing*. 254–263.

[53] Dimitris G Tzikas, Aristidis C Likas, and Nikolaos P Galatsanos. 2008. The variational approximation for Bayesian inference. *IEEE Signal Processing Magazine* 25, 6 (2008), 131–146.

[54] Paul-Armand Verhaegen, Dennis Vandevenne, Jef Peeters, and Joost R Duflou. 2013. Refinements to the variety metric for idea evaluation. *Design Studies* 34, 2 (2013), 243–263.

[55] Peter Welinder, Steve Branson, Pietro Perona, and Serge Belongie. 2010. The multidimensional wisdom of crowds. *Advances in neural information processing systems* 23 (2010).

[56] Jie Yang, Thomas Drake, Andreas Damianou, and Yoelle Maarek. 2018. Leveraging crowdsourcing data for deep active learning an application: Learning intents in alexa. In *Proceedings of the 2018 World Wide Web Conference*. 23–32.

[57] Omar Zaidan, Jason Eisner, and Christine Piatko. 2007. Using "annotator rationales" to improve machine learning for text categorization. In *Human language technologies 2007: The conference of the North American chapter of the association for computational linguistics; proceedings of the main conference*. The Association for Computational Linguistics, 260–267.

[58] James Zou, Kamalika Chaudhuri, and Adam Kalai. 2015. Crowdsourcing feature discovery via adaptively chosen comparisons. In *Third AAAI Conference on Human Computation and Crowdsourcing*.

A APPENDIX

In this section, we present the results of some additional experiments. First, we show the distribution of ratings collected from experts and crowd workers in Figure 5 and the performance comparison of idea evaluation techniques in Tables 2, 3 and 4.

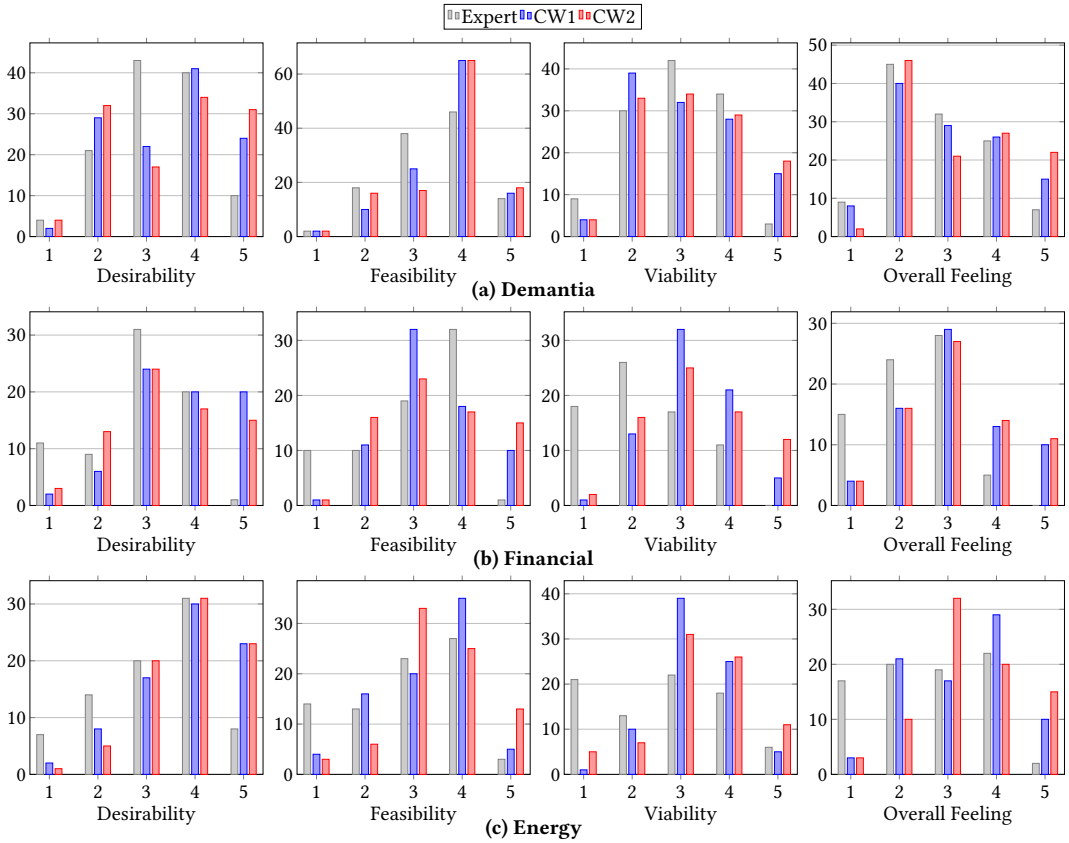


Figure 5: The distribution of ratings collected from experts and crowd in three challenges: Dementia (a), Financial (b) and Energy (c). Legend: CW1 - the first crowd worker; CW2 - the second crowd worker.

	Desirability				Feasibility				Viability				Overall Feeling			
	20%	40%	60%	80%	20%	40%	60%	80%	20%	40%	60%	80%	20%	40%	60%	80%
Crowd-Rating	.37	.36	.37	.37	.29	.30	.30	.28	.38	.36	.38	.38	.37	.37	.39	.37
ML-Expert	.30	.28	.27	.28	.34	.32	.31	.31	.28	.27	.28	.29	.26	.25	.24	.25
ML-Crowd	.34	.32	.31	.32	.31	.31	.30	.31	.38	.36	.34	.35	.33	.32	.30	.30
HybridEval	.29	.28	.27	.28	.31	.30	.29	.30	.26	.26	.28	.30	.25	.22	.23	.24

Table 2: Financial dataset: Comparison between our approach (`HybridEval`) and `Crowd-Rating`, `ML-Expert` and `ML-Crowd` measured by RMSE.

	Desirability				Feasibility				Viability				Overall Feeling			
	20%	40%	60%	80%	20%	40%	60%	80%	20%	40%	60%	80%	20%	40%	60%	80%
Crowd-Rating	.29	.28	.29	.28	.25	.23	.25	.25	.31	.31	.30	.32	.31	.31	31	.31
ML-Expert	.27	.25	.25	.24	.28	.25	.25	.24	.27	.23	.23	.21	.28	.26	.24	.22
ML-Crowd	.29	.27	.28	.24	.27	.26	.27	.25	.28	.27	.28	.26	.30	.28	.30	.26
HybridEval	.24	.24	.24	.23	.24	.24	.25	.24	.26	.24	.24	.21	.27	.26	.24	.22

Table 3: Dementia dataset: Comparison between our approach (`HybridEval`) and `Crowd-Rating`, `ML-Expert` and `ML-Crowd` measured by RMSE.

	Desirability				Feasibility				Viability				Overall Feeling			
	20%	**40%**	**60%**	**80%**	**20%**	**40%**	**60%**	**80%**	**20%**	**40%**	**60%**	**80%**	**20%**	**40%**	**60%**	**80%**
Crowd-Rating	.33	.32	.33	.31	.30	.29	.29	.29	.37	.37	.37	.35	.36	.35	36	.34
ML-Expert	.33	.31	.29	.31	.36	.32	.31	.33	.38	.35	.34	.36	.35	.33	.32	.32
ML-Crowd	.34	.35	.34	.33	.34	.34	.34	.35	.37	.38	.37	.39	.37	.36	.35	.36
HybridEval	.29	.29	.28	.30	.31	.31	.30	.32	.34	.34	.34	.36	.32	.32	.31	.32

Table 4: Energy dataset: Comparison between our approach (`HybridEval`) and `Crowd-Rating`, `ML-Expert` and `ML-Crowd`, measured by RMSE.