Comparing Pineapples with Lilikois: An Experimental Analysis of the Effects of Idea Similarity on Evaluation Performance in Innovation Contests

Victoria Banken
University of Innsbruck
victoria.banken@uibk.ac.at

Isabella Seeber
University of Innsbruck
isabella.seeber@uibk.ac.at

Ronald Maier
University of Innsbruck
ronald.maier@uibk.ac.at

Abstract

Identifying promising ideas from large innovation contests is challenging. Evaluators do not perform well when selecting the best ideas from large idea pools as their information processing capabilities are limited. Therefore, it seems reasonable to let crowds evaluate subsets of ideas to distribute efforts among the many. One meaningful approach to subset creation is to draw ideas into subsets according to their similarity. Whether evaluation based on subsets of similar ideas is better than compared to subsets of random ideas is unclear. We employ experimental methods with 66 crowd workers to explore the effects of idea similarity on evaluation performance and cognitive demand. Our study contributes to the understanding of idea selection by providing empirical evidence that crowd workers presented with subsets of similar ideas experience lower cognitive effort and achieve higher elimination accuracy than crowd workers presented with subsets of random ideas. Implications for research and practice are discussed.

1. Introduction

The more ideas are generated in open spaces, such as in open innovation contests, the more likely it is that truly good ideas are contributed [7]. Yet, even if the contest phase has resulted in a huge number of ideas, the success of an innovation contest is dependent on the idea selection phase and whether the best opportunity can actually be identified [12, 25]. Idea selection is cognitively demanding [3], time-consuming [2], and individuals often fail to identify the best ideas [12]. One way to ease the cognitive demand is to prompt idea evaluators towards excluding ideas instead of including ideas into a consideration set [17]. In addition, to curb selection duration, more and more organizations outsource idea evaluation from a small team to a large crowd [4, 5].

There exists first evidence that prompting crowd evaluators towards excluding bad ideas results in higher evaluation accuracy [23]. Yet large innovation contests often produce many duplicate or similar ideas [11] that do not enrich the solution space, but consume additional time and resources during idea screening [18]. Hence, several studies emphasize the importance of organizing the large pool of ideas and categorizing them for idea evaluation [37, 42]. This becomes particularly relevant for crowd-based idea evaluation in large innovation contests, in which it could be unfeasible to let each crowd worker assess all ideas. Given that crowd tasks are rather brief [35] and individuals' information processing capabilities are limited [46], effective selection techniques need to be found that allow to meaningfully distribute a subset of ideas to crowd workers for idea evaluation.

The creation of subsets according to idea similarity could be such a crowdsourcing technique for more effective crowd-based idea evaluation. There exists empirical evidence, that idea similarity is indicative of idea quality, yet with heterogeneous findings [18, 24, 47, 49]. Idea similarity can be established by organizing ideas into the same category [24]. Following this, eliminating ideas from within the same category should make the evaluation cognitively easier [1]. It remains unclear if the theorized effects of idea similarity on cognitive demand can also improve the performance of crowd-based idea evaluation. Consequently, more empirical evidence needs to be provided to understand the role of idea similarity when crowds are tasked to evaluate submissions from crowd initiatives like innovation contests [49]. We see a research gap in our understanding of how the provision of crowd workers with subsets created according to idea similarity affects evaluation performance and cognitive demand. Hence, our research question is: *How do evaluation performance and perceived cognitive demand differ between crowd workers that eliminate low quality ideas from subsets of similar ideas and crowd workers that eliminate low quality ideas from subsets of random ideas?*

We aim to study whether the provision of subsets of similar ideas will lead to higher evaluation performance in terms of accuracy, false positive and false negative rates and lower cognitive demand in terms of perceived cognitive effort and information overload.

URI: https://hdl.handle.net/10125/59483
ISBN: 978-0-9981331-2-6

2. Theoretical Background

2.1. Elimination in Idea Selection

Idea selection starts after the conclusion of an idea contest and ends with winner determination. At IBM's Innovation Jam, for example, 10 out of 45,000 ideas were funded to create new businesses [1]. In Cisco's I-Prize competition one winner was chosen out of 1,200 distinct ideas [20]. In these exemplary cases, contest organizers first assembled a shortlist of high quality ideas in a screening phase from which a jury of experts determined the winner in a final phase. For idea screening, they usually turn to evaluators that are not the ideators themselves to reduce a potential bias due to the endowment effects [33], the likelihood to select an idea in which you already invested a lot, that is one's own idea [22]. This different set of evaluators could be an internal or external small team, but also another crowd.

But people are often not able to discern the best ideas [12]. Their ability to make high-quality decision is limited due to high information load [9, 44] and high cognitive effort [48]. Cognitive load represents the load that is imposed by the task and its representation on the human cognitive system when performing this task [46]. Cognitive effort is the amount of resources that humans need to allocate to the task to meet the tasks' information processing demands [34]. Studies on idea selection found that prompting evaluators towards exclusion (eliminate the bad ideas) requires less effort than prompting evaluators towards inclusion (select the best idea). [26]. An exclusion strategy describes eliminating the less likely alternatives from an initial set [17, 39]. Eliminating or excluding ideas is believed to ease the cognitive effort of decision-makers as they tend to engage in attribute-based processing instead of alternative-based processing [26]. Under attribute-based processing, evaluators would consider a single attribute of an idea for comparison with other ideas before other attributes are considered. Under alternative-based processing, evaluators would consider an idea (alternative) with all its attributes before moving to the next alternative [38]. Yet, it is unclear how the performance or quality of idea selection can be evaluated.

2.2. Assessing the Quality of Idea Selection

Many idea selection procedures rely on a binary assessment of idea quality (good vs. bad ideas) as it is less time-consuming and cognitively demanding than applying e.g., multiple quality criteria to be assessed on rating scales. The binary nature allows to measure overall accuracy of an evaluators' idea assessment by using common metrics from the field of Information Retrieval (e.g., [49]). In the so called confusion or error matrix, the prediction of a condition is compared to its designated condition [45]. Table 1 presents a confusion matrix for idea selection with four quadrants that exhibit the absolute values of classifications made by the crowd compared to the gold standard. In scientific research, the gold standard is usually established through multiple raters with domain knowledge (e.g., [3, 24]). When considering that an exclusion strategy strives for eliminating low quality ideas, positive predictions refer to the elimination of a low quality idea, while negative predictions describe ideas that were not eliminated and hence considered as high quality. The true positive (TP) quadrant, therefore includes ideas that have been classified by both, the crowd and the gold standard, as low quality ideas. The false positive (FP) quadrant, includes ideas that have been classified by the crowd as low quality, but as high quality by the gold standard. The false negative (FN) quadrant, includes ideas that have been classified by the crowd as high quality but as low quality by the gold standard. The true negatives (TN) quadrant, includes ideas that are classified as high quality by the crowd and the gold standard.

Table 1. Confusion Matrix and suggested measures for assessing evaluation performance

		Gold standard	
		Low quality	High quality
Prediction of the crowd	Low quality	*True positives (TP)*	*False positives (FP)*
	High quality	*False negatives (FN)*	*True negatives (TN)*

Measures to assess evaluation performance:

Accuracy: $ACC = \frac{\sum TP + \sum TN}{\sum TP + \sum FN + \sum TN + \sum FP}$

False negative rate: $FNR = \frac{\sum FN}{\sum TP + \sum FN}$

False positive rate: $FPR = \frac{\sum FP}{\sum FP + \sum TN}$

We argue for three measures of evaluation performance that are particularly relevant for idea selection. First, the crowd's evaluation of ideas should comply with the gold standards' rating and therefore, the elimination accuracy (ACC) should be high. Accuracy measures the proportion of all correct predictions and includes true positives as well as true negatives [30]. Thus, the crowd's elimination accuracy increases with the number of ideas that are correctly classified as low quality and correctly classified as high quality.

Second, the *false negative rate* represents the fraction of ideas that are incorrectly classified as high quality [30]. If the crowd wrongly classifies an idea as high quality even though it should be classified as low quality, more resources need to be deployed in the next phase. Contest organizers aim at avoiding allocating additional (financial and human) resources [36] in subsequent evaluation activities. Hence, the false negative rate (FNR) should be small.

Third, the *false positive rate* describes all incorrectly as low quality classified ideas [30]. Contest organizers might also be concerned with missing out on high quality ideas from the screening phase. Hence, the false positive rate (FPR) should be small to avoid eliminating high quality ideas.

Besides these idea related evaluation performance measures, also evaluator related measures should be considered. It has been established that idea evaluation is cognitively demanding (e.g., [3, 23]). When cognitive effort is perceived as high, people get tired more quickly and performance drops eventually. Some individuals might even experience information overload, because they feel overwhelmed by the amount of information.

Hence, we argue that evaluation performance (elimination accuracy, false negative rate, false positive rate) and cognitive demand (cognitive effort, information overload) are relevant measures for idea selection quality.

2.3. Idea Similarity

Many contest organizers pre-process ideas to support the identification of high quality ideas. For example, text mining was adopted in the IBM's Innovation Jam with more than 46,000 ideas to put similar postings in the same category for later assessment by experts [1]. Pre-processing ideas according to their similarity is useful in order to gain an overview of the opportunity space particularly when idea contests resulted in hundreds of ideas [24]. Identifying similar ideas during the evaluation process however rather distracts from evaluation task itself [11]. [42] found that categorization features in idea management system are positively associated to the effectiveness of and satisfaction with the idea assessment. Idea similarity has been investigated in the domain of crowdsourcing with different operationalization approaches as displayed in Table 2. There exist examples of human-based categorization efforts [24, 33, 52], of automated approaches [41, 49], or of hybrid approaches [43]. A handful of studies investigated the relationship between idea similarity and evaluation accuracy. Some studies found that dissimilar ideas are associated with higher selection probability [49] or higher creativity [47]. Other studies found that dissimilar ideas are not generally considered more valuable [24]. [18] found that either very similar or very dissimilar ideas are more likely to be implemented.

Table 2. Idea similarity in innovation contests

Author(s)	*Operationalization of similarity*	*Relevant contributions*
[49]	Text mining-based dissimilarity identification	Distinct ideas are associated with higher selection probability
[24]	Human categorization - grouping of similar ideas, indicating identical or essentially identical ideas	Ideas that are more distinct from other ideas are not generally considered more valuable
[47]	Text mining with statistical procedures	Ideas with semantic subnetworks that are more distinct (higher prototypical edge weight distribution) tend to be judged as more creative
[43]	Hybrid similarity comparison of 3 alternatives	Participants that see similar ideas generate ideas of higher creativity (not significant)
[18]	Text mining-based dissimilarity identification	Very similar or very dissimilar ideas are more likely to be implemented
[11]	Manual human identification of duplicate ideas	Duplicate identification detracts from identifying high quality ideas
[42]	Human categorization in idea management system	Idea management systems that have a categorisation feature are associated with higher idea assessment effectiveness and with higher satisfaction with idea categorization

This suggests that ideas, which potentially are of high quality, can be found at both ends of the idea similarity continuum. [24] suggested that high similarity among ideas can be an indicator of popularity. Hence, when there exist many similar or even identical ideas, evaluators might be persuaded to consider the type of idea to be in high demand and therefore valuable [27]. When subsets of ideas are randomly composed, chances decrease that evaluators choose from very similar ideas. They are more likely to end up with ideas from multiple categories, which are randomly positioned. This should make it harder to recognize frequent topics shared by multiple ideas. However, when subset creation depends on idea similarity, e.g., an idea belonging to the same category, evaluators are likely to recognize the common theme or topic among ideas. The truly good idea should stand out amongst its similar ones. Hence it should be easier to recognize their potential value, resulting in more true predictions (H1a), lower resource inefficiencies in subsequent evaluation activities (H1b) and lower fear of missing out on good ideas (H1c). Thus,

Hypothesis 1: Crowd workers that eliminate ideas from subsets of similar ideas will have higher evaluation performance in terms of elimination accuracy (H1a), FNR (H1b), and FPR (H1c) than crowd workers that eliminate ideas from subsets of random ideas.

In human-based categorization, people place the information they process into their mental schema. Depending on whether information is organized into macro or micro concepts greatly affects the number of alternatives they then need to consider [28]. Dealing with familiar concepts, e.g., ideas, people, or situations, induces cognitive ease for information processing [21]. Consequently, evaluators could use these freed up cognitive resources to investigate and compare ideas more in depth. We assume that the categorization or clustering of ideas according to similarity supports comparing alternatives with respect to their elaborateness. When an idea is described in detail (why the idea is relevant, how the idea can be implemented, for whom the idea is relevant, etc.), it is easier to estimate its potential benefit. The less elaborated an idea is, the more ambiguous it is, because potentially relevant information is missing from the idea description and selecting the idea into a consideration set becomes riskier. If evaluators need to choose from a random idea set, their chances to compare similar ideas are lower and therefore the choice is cognitively more demanding. Hence, we suggest:

Hypothesis 2: Crowd workers that eliminate ideas from subsets of similar ideas will perceive lower cognitive demand in terms of cognitive effort (H2a) and information overload (H2b) than crowd workers that eliminate ideas from subsets of random ideas.

3. Method

To test the hypotheses above, we conducted a between-subject web experiment manipulating idea similarity (categorized ideas vs. random). We presented participants with six idea subsets of five ideas each selected from a real innovation contest. We applied two distinct annotation processes with the goal to develop a) the category treatment (independent variable) and b) the gold standard (dependent variable).

3.1. Operationalization of Independent Variable

We drew a subset of 100 ideas from the ZEISS VR ONE App Contest[1]. The goal of the contest was to source ideas for apps or completed apps for the virtual reality (VR) and augmented reality (AR) headset. For each condition, we distributed ideas such that each idea was seen by multiple crowd workers who each received 30 distinct ideas in six subsets. All ideas were presented in random sequence to control for order bias. While the subsets in the random treatment contained randomly selected ideas, the crowd workers in the category treatment condition were presented with subsets of five similar ideas that had been pre-processed into categories beforehand.

To develop the category treatment, we applied a three-round categorization process that involved a team of two of the co-authors and four master and PhD students. In the first two hours, they built a shared understanding on the categories and their relationships using a subset of ideas. In order to identify and label categories they facilitated themselves using the Pin-The-Tail-On-The-Donkey (PD) ThinkLet, while the Theme Seeker (TS) [6] ThinkLet was used to categorize ideas. The result was a codebook of eight relevant category labels, their definitions and keywords. In the second round, the same process was repeated with the remaining ideas distributed among the team members.

We determined the final category for each idea by unanimity and majority (67%) rules: Entertainment (42 ideas), healthcare (14), travel (11), education (10), sports (6), shopping (6), design (5), work (4), safety (2). For the similarity treatment, we selected ideas for each

[1] The website of the contest is not publicly available anymore, information can be requested from the authors or found on the following websites: https://goo.gl/ZwnfWG and https://goo.gl/9wejm3

participant out of the entertainment category, because this category was the only one that included more than 30 ideas, which was necessary to create the six subsets of five similar ideas each. With respect to the random treatment, we presented a random selection of 30 ideas from all eight categories, which were split into six subsets of five ideas.

3.2. Subjects

Eighty-five subjects were recruited from Amazon Mechanical Turk, a platform that allows to outsource Human Intelligence Tasks (HIT) to crowd workers. The expected time to complete the HIT was determined during pre-tests and amounted to 30 minutes, which resulted in a reward of 3.00 US Dollar per HIT to comply with the minimum wage of the United States. We limited the pool of crowd workers to the United States to increase the participation probability of native English speakers and to those crowd workers that had at least 100 completed HITs and a HIT approval rate of 98%.

We rejected and denied compensation for 19 crowd workers that completed the HIT (see Table 3), because they had failed one or more predefined approval criteria: First, two crowd workers did not submit the correct survey code provided on the last page of the survey. Second, three crowd workers spent less than one minute on the platform compared to an average time spent of 7:12 minutes in the random treatment and 8:12 in the category treatment which raised doubts if they sufficiently paid attention to the task at hand. Third, sixteen respondents did not pass the attention check (adapted from [51]), which was "I felt there were too many cats in the idea. (Please select strongly disagree, as this is an attention check!)". Finally, we included 66 crowd workers in our data analysis out of which 41 were in the random and 25 were in the similarity treatment.

3.3. Procedure and instrumentation

Once crowd workers accepted the HIT, they were redirected to our experimental online platform (see Figure 1). On the welcome screen, crowd workers were informed about the task with the following prompt “Please reduce the ideas drastically and eliminate ‘bad’ ideas that you feel are insufficient for further consideration. You can eliminate zero, one or multiple ‘bad’ ideas from each set”. This binary assessment can be understood as a holistic rating scale, which means that only one score with a single trait is collected [15]. Hence, the meaning of “bad” was not further explained in order to avoid guiding the attention to multi-dimensional quality criteria and artificially inducing

Table 3. Crowd worker included in analysis

	Ran-dom	**Simi-larity**	**Total**
Completed HITs	**57 (67%)**	**28 (33%)**	**85 (100%)**
Rejected and unpaid	16	3	19
Wrong survey code	2	0	2
Less than 1 minute	3	0	3
Failed attention check	13	3	16
Included in analysis	**41 (62%)**	**25 (38%)**	**66 (100%)**

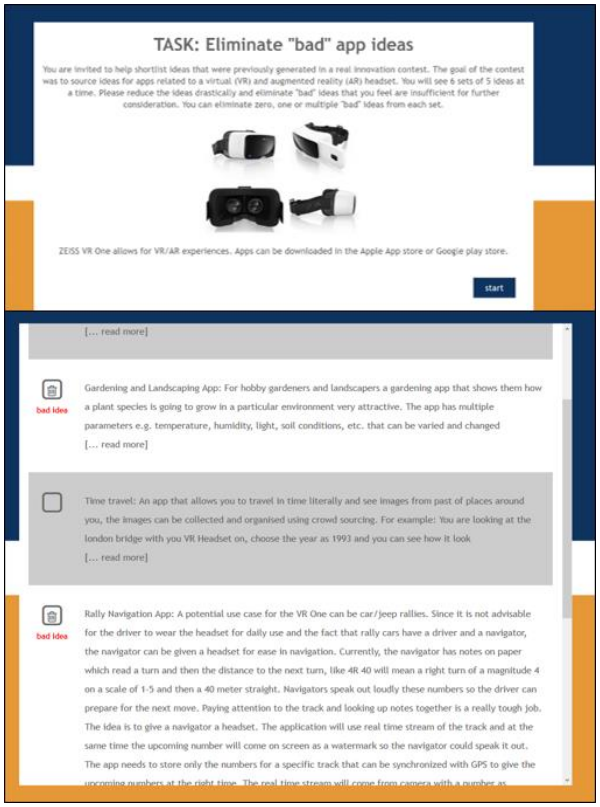

Figure 1. Screenshots of the elimination platform

higher cognitive effort during elimination. On each of the next six screens, five ideas were presented where crowd workers could check boxes to eliminate bad ideas indicated by a trash icon. As the ideas had different lengths and were described with up to 500 words, the screen showed the first 100 characters of the description. Crowd workers could click on “read more” to view the whole idea description. Once a crowd worker moved to the next idea screen, the selection of eliminated ideas was stored in the database including its start and end timestamps. The experiment ended with a survey that collected perception-based variables and demographic data.

3.4. Measures and Operationalization

The gold standard is necessary to *measure evaluation* and was set by two of the co-authors with necessary domain knowledge over a period of two weeks following a four-step approach. In the first step, the two raters checked the innovativeness of ideas by researching existing solutions on the market. They individually rated 30 ideas in terms of four criteria that indicate low quality: ideas are worn, uncreative, useless or not elaborated. In the second step, they discussed their individual assignments, built shared understanding of the evaluation criteria and agreed on a good/bad assessment for each of the 30 ideas in a two hour discussion. In the third step, both evaluated the remaining 70 ideas individually. We checked inter-rater reliability and achieved simple agreement of 69.5% and a Cohen's Kappa of .535 with $p < .001$, which is a fair to good value according to [10]. Finally, the two researchers discussed and resolved conflicts and therefore developed an agreed quality assessment of all ideas. To check for robustness, we correlated our assessment with the number of likes each idea had received from the online innovation community during idea generation. We found a positive correlation, $r(98) = 0.24$, $p = .014$, which further supports our assessment. Each crowd worker's idea assessments were then compared to the gold standard to determine the number of true positives, false positives, true negatives and false negatives. Afterwards, we calculated the metrics elimination accuracy, false positive rate and false negative rate.

The perception-based variables comprise cognitive effort [50] and information overload (adapted from product overload [16]) with a 7-point Likert scale (0=*strongly disagree*, 6=*strongly agree*). Cognitive effort was measured using three items: "The task of selecting ideas took too much time", "Selecting ideas required too much effort" and "Selecting ideas was too complex". Information overload was measured using five items: "There were so many ideas to choose from that I felt confused", "The more I learned about these ideas, the harder it seemed to choose between ideas", "It was difficult to obtain an overview over the ideas", "With that many ideas to choose between, I have had a hard time identifying distinguishing idea characteristics", and "With that many ideas to choose between, I found it difficult to compare competing ideas".

We performed reliability analysis with Cronbach's Alpha for perceived cognitive effort (alpha = .966) and information overload (alpha = .906). All perception-based constructs reached the recommended threshold of .7 [32]. To test convergent and discriminant validity, we performed exploratory factor analysis with Promax rotation. All items of our perception-based constructs loaded well on the resulting two factor solutions with factor loadings higher than .6. Cross-loadings were low and MSA-values higher than .5. All these values exceeded the recommended thresholds [32] and therefore convergent and discriminant validity are deemed satisfactory.

4. Results

This study investigates the effects of idea similarity on evaluation performance and cognitive demand. We first checked data against violation of statistical assumptions. For normal distribution, we inspected data visually with boxplots and histograms as well as skewness and kurtosis statistics. For the evaluation performance measures, boxplots and histograms indicated a close to bell curve; skewness was -0.561, -0.721 and 0.525 and kurtosis was -0.312, -0.796, and -0.398 for the three DVs, accuracy, FNR, and FPR. Hence, we deemed our data to be sufficiently normally distributed. We tested homogeneity of variance with Levene's statistics, which turned out to be non-significant (Accuracy: $F = 0.681$, $p = .412$; FNR: $F = 2.867$, $p = .095$; FPR: $F = 0.067$, $p = .796$) and hence satisfactory [13]. For the cognitive demand variables, the assumptions of normal distribution (skewness and kurtosis within the range of -1 and +1) and homogeneity of variance ($p < .05$) did not hold [14].

Table 4. Confusion matrix - random treatment

	Gold Standard	
Random	Low quality	High quality
Eliminated ideas	TP: 8.44 TPR: 36.04%	FP: 1.85 FPR: 28.47%
Remaining ideas	FN: 14.90 FNR: 63.96%	TN: 4. 80 TNR: 71.53%
Elimination Accuracy = 44.15%		

Table 5. Confusion matrix - similarity treatment

	Gold Standard	
Similarity	Low quality	High quality
Eliminated ideas	TP: 9.24 TPR: 43.13%	FP: 2.56 FPR: 30.02%
Remaining ideas	FN: 12.24 FNR: 56.87%%	TN: 5.96 TNR: 69.98%
Elimination Accuracy = 50.67%		

We performed one-way ANOVAs on elimination accuracy, FN rate and FP rate (Table 6). With respect to elimination accuracy, we found a significant treatment effect, which indicates that crowd workers presented with a set of similar ideas have higher elimination accuracy than crowd workers that were presented with random ideas, $F(1, 64) = 7.523$, $p = .008$, partial $\eta^2 = .105$. According to Cohen the effect of similarity on elimination accuracy can be referred to as medium to large [8]. The descriptive statistics are given in Table 4 for the random treatment and

Table 5 for the similarity treatment. The average elimination accuracy was 44.15% for crowd workers of the random treatment compared to 50.67% for the similarity treatment, a difference of 6.53 percentage points (see). We therefore accept Hypothesis H1a.

With respect to the *false negative rate (FNR),* the crowd in the random treatment had a FNR of 63.96%, which is significantly higher than the FNR of 56.87% in the similarity treatment ($F(1, 64) = 4.283$, $p = .043$, partial $\eta^2 = .063$). According to Cohen the effect of similarity on FNR can be referred to as medium [8]. We therefore accept Hypothesis H1b.[2]

The *false positive rates (FPR)* were similar for the random (28.47%) and similarity treatment (30.02%) with no significant differences between the groups, $F(1, 64) = 0.076$, $p > .784$, partial $\eta^2 = .001$. Hypothesis H1c was therefore rejected.

Furthermore, we tested differences between treatments for the two measures of cognitive demand, *cognitive effort* and *information overload*. As the data violated assumptions of ANOVA, we performed the non-parametric Mann-Whitney U test (see Table 7). Cognitive effort was significantly lower for crowd workers that eliminated ideas from a set of similar ideas (Mdn = 1.00) than for those that eliminated ideas from a set of random ideas (Mdn = 2.33), $U = 359.50$, $p = .033$, $r^2 = .069$. According to Cohen the effect size is medium [8] (see Table 8). We therefore accept Hypothesis H2a.

Information overload was not significantly lower for crowd workers that eliminated ideas from a set of similar ideas (Mdn = 3.60) than for those that eliminated ideas from a set of random ideas (Mdn = 2.60), $U = 461.50$, $p = .498$, $r^2 = .007$. We therefore reject Hypothesis H2b.

Table 6. ANOVA for elimination accuracy, FN rate, FP rate

Source	DF	Mean square	F	p-value	partial η^2
ANOVA Dependent variable: *Elimination accuracy*					
Treatment	1	0.066	7.523	**.008**	.105
Error	64	0.009			
ANOVA Dependent variable: *FN rate*					
Treatment	1	0.078	4.283	**.043**	.063
Error	64	0.018			
ANOVA Dependent variable: *FP rate*					
Treatment	1	0.004	0.076	.784	.001
Error	64	0.049			

Table 7. MANN-WHITNEY U-test for cognitive effort and information overload

Source	N	Mean rank	U	Z	p-value	r^2
MANN-WHITNEY U Dependent variable: *Cognitive effort*						
Random	41	37.23	359.500	-2.137	**.033**	.069
Similarity	25	27.38				
MANN-WHITNEY U Dependent variable: *Information overload*						
Random	41	32.26	461.500	-0.678	.498	.007
Similarity	25	35.54				

[2] We performed the ANOVA with less stringent criteria using (all completed HITs) and found the significant differences for H1a (Accuracy), but not for H1b (FN rate).

Table 8. Mean (M) and standard deviation (SD) of cognitive demand for the two treatments

		Cognitive Effort		Information Overload	
	N	M	SD	M	SD
Random	41	2.85	2.03	2.88	1.61
Similarity	25	1.79	1.20	3.18	1.40

5. Discussion

In this study, we used experimental methods to investigate the associations between idea similarity, evaluation performance and cognitive demand in a crowd setting, i.e. crowd workers were tasked to eliminate bad ideas from a set of ideas that were previously generated by a different crowd of ideators in an online innovation contest.

5.1. Theoretical implications

Our findings contribute to the literature on idea selection. We found that crowd workers who were prompted towards elimination and presented with subsets of similar ideas experienced lower cognitive effort and achieved higher elimination accuracy than crowd workers who were presented with subsets of random ideas. Thus, this paper provides first empirical evidence that just by presenting similar ideas in idea subsets, evaluation accuracy can be improved. This is in line with the finding in [23]. Yet, our study has some notable differences with respect to the setting and participants. First, [23] used an organization-internal crowd of 66 ideators that were also the raters. Hence, a potential endowment effect cannot be ruled out. In contrast to that, our study used real crowd-generated ideas from an open innovation contest and we tasked a different crowd to eliminate the bad ideas. [23] limited the number of potential exclusions to 10 and therefore had a fixed reduction rate (RR=10/48=0.21). Our study did not restrict the number of eliminations and the average reduction rates turned out to be higher in the random (RR=10.29/30=0.34) as well as in the similarity treatment (RR=11.80/30=0.39). In both treatments, the final sets also contained a higher ratio of high to low quality ideas than the original set. Therefore, we found that the crowd is capable to substantially reduce sets of ideas and to increase the proportion of high quality ideas to be further considered.

Moreover, we contribute to literature with the finding that presenting subsets with similar ideas was associated with lower cognitive effort. Raters that eliminated ideas from different categories experienced higher cognitive effort than raters that eliminated ideas from the same category. This implies that parts of the raters' cognitive demand can be reduced by allocating ideas into subsets of the same category before handing them over for elimination.

5.2. Implications for practitioners

Our findings also have implications for contest organizers: [40] found that processing one idea would cost approximately $500 and four hours of staff and management time in a Fortune 100 company. Compared with our experiment in which all 100 ideas had to be evaluated, our expenditures for 66 crowd workers amounted to $237.60, which worked in total 8 hours and 20 minutes. Hence, the challenge to keep costs and time for idea selection low is an important challenge. We are the first to have tested the evaluation performance variables FN and FP rate in the context of idea selection, which we argue provide insights to the challenges for contest organizers. The FN rate represents the objective to keep effort for subsequent evaluation activities low. It refers to the fraction of ideas that the crowd perceived as high quality even though they are of low quality according to the gold standard. In our elimination context, a lower FN rate indicates less evaluation effort. We found that idea similarity is associated with lower FN rates. Hence, the provision of similar ideas seemed to foster effective elimination. Yet, idea similarity was not associated with FP rates, which could give indication whether the elimination procedure could foster or decrease the fear of missing out on good ideas.

Moreover, we could show that ideas should be provided in subsets of similar ideas for improved idea selection. We provide empirical evidence that raters that eliminated ideas from subsets of similar ideas experienced lower cognitive effort and achieved higher evaluation accuracy.

5.3. Limitations and future research

There are also some limitations to our study that need to be considered and should be addressed by future research. First, the sample size is relatively small. Future work could repeat the experiment on a larger sample of crowd workers in order to increase statistical power.

Second, our gold standard assessment is correlated with the voting of the online innovation community and could therefore give indication on the popularity of ideas. We cannot rule out the possibility that community votes were distorted by manipulative tactics of community members [19]. Moreover, although the two raters had the required domain knowledge for idea evaluation, they are non-experts in the domain. By collecting background information and developing required knowledge in the domain of the contest, we

mitigated the risk of eliminating ideas that would not be in the interest of the contest sponsor's objectives. In addition, it was found that non-experts do not differ from experts when ranking ideas to determine high and low quality ideas [29]. Nonetheless, the gold standard assessment should be extended into a compound measure considering the opinion of domain experts in future research.

Third, the crowd workers in the similarity treatment were presented with similar ideas from the category entertainment, while we selected ideas from all categories in the random treatment. Even though the ideas were submitted to the same online contest, we cannot rule out that ideas differ with respect to the domain knowledge needed for their evaluation. Future research could explore the association between domain knowledge of crowd workers and the domains represented by the submitted ideas.

Fourth, the crowd workers in our experiment were asked to evaluate ideas in binary categories. For future work, non-binary categories such as "good/ mediocre/ bad" might provide a promising approach to further reduce cognitive effort by not enforcing decisions in situations, where crowd workers might not be able to make up their minds about an idea.

Finally, a considerable number of crowd members showed short task engagement as mentioned above. Future research could change incentives in such a way that a crowd worker gets additional rewards when the assessment is closer to the gold standard.

Acknowledgements

The research leading to the presented results was partially funded by the Austrian Science Fund (FWF): P 29765.

6. References

[1] Bjelland, O.M., and R.C. Wood, "An Inside View of IBM's 'Innovation Jam'", *MIT Sloan Management Review 50*(1), 2008, pp. 32–40.

[2] Blohm, I., J.M. Leimeister, and H. Krcmar, "Crowdsourcing: How to benefit from (too) many great ideas", *MIS Quarterly Executive 12*(4), 2013, pp. 199–211.

[3] Blohm, I., C. Riedl, J. Füller, and J.M. Leimeister, "Rate or Trade? Identifying Winning Ideas in Open Idea Sourcing", *Information Systems Research 27*(1), 2016, pp. 27–48.

[4] Bonabeau, E., "Decisions 2.0: The Power of Collective Intelligence", *MIT Sloan Management Review 50*(2), 2009, pp. 45–52.

[5] Boudreau, K.J., and K.R. Lakhani, "Using the Crowd as an Innovation Partner", *Harvard business review 91*(4), 2013, pp. 60–69.

[6] Briggs, R., and G.-J. de Vreede, *ThinkLets: Building Blocks for Concerted Collaboration by Robert Briggs (Paperback) - Lulu*, 2009.

[7] Briggs, R.O., and B.A. Reinig, "Bounded Ideation Theory", *Journal of Management Information Systems 27*(1), 2010, pp. 123–144.

[8] Cohen, J., *Statistical Power Analysis for the Behavioral Sciences*, 1988.

[9] Eppler, M.J., and J. Mengis, "The Concept of Information Overload: A Review of Literature from Organization Science, Accounting, Marketing, MIS, and Related Disciplines", *The Information Society 20*(5), 2004, pp. 325–344.

[10] Fleiss, J.L., B. Levin, and M.C. Paik, *Statistical methods for rates and proportions.*, J. Wiley, 2003.

[11] Di Gangi, P.M., M.M. Wasko, and R.E. Hooker, "Getting customers' ideas to work for you: Learning from Dell how to succeed with online user innovation communities", *MIS Quarterly Executive 9*(4), 2010, pp. 213–228.

[12] Girotra, K., C. Terwiesch, and K.T. Ulrich, "Idea Generation and the Quality of the Best Idea", *Management Science 56*(4), 2010, pp. 591–605.

[13] Hair, J.F., W.C. Black, B.J. Babin, and R.E. Anderson, *Multivariate data analysis; A global perspective*, 2010.

[14] Hair, J.F., W.C. Black, B.J. Babin, and R.E. Anderson, "Multivariate Data Analysis", *Vectors*, 2010, 816.

[15] Harsch, C., and G. Martin, "Comparing holistic and analytic scoring methods: Issues of validity and reliability", *Assessment in Education: Principles, Policy and Practice 20*(3), 2013, pp. 281–307.

[16] Heitmann, M., D.R. Lehmann, and A. Herrmann, "Choice Goal Attainment and Decision and Consumption Satisfaction", *Journal of Marketing Research XLIV*(May), 2007, pp. 234–250.

[17] Heller, D., I.P. Levin, and M. Goransson, "Selection of strategies for narrowing choice options: Antecedents and consequences", *Organizational Behavior and Human Decision Processes 89*, 2002, pp. 1194–1213.

[18] Hoornaert, S., M. Ballings, E.C. Malthouse, and D. Van den Poel, "Identifying New Product Ideas: Waiting for the Wisdom of the Crowd or Screening Ideas in Real Time", *Journal of Product Innovation Management 34*(5), 2017, pp. 580–597.

[19] Hutter, K., J. Füller, J. Hautz, V. Bilgram, and K. Matzler, "Machiavellianism or Morality: Which Behavior Pays off in Online Innovation Contests?", *Journal of Management Information Systems 32*(3), 2015, pp. 197–228.

[20] Jouret, G., "Inside Cisco ' s Search for the Next Big Idea", *Harvard Business Review*(September), 2009, pp. 43–46.

[21] Kahneman, D., "Thinking, Fast and Slow", 2011, pp. 499.

[22] Kahneman, D., J.L. Knetsch, and R.H. Thaler, "Anomalies: The Endowment Effect, Loss Aversion, and Status Quo Bias", *Journal of Economic Perspectives 5*(1), 1991, pp. 193–206.

[23] Klein, M., and A.C.B. Garcia, "High-Speed Idea Filtering With the Bag of Lemons", *Decision Support Systems 78*, 2015, pp. 1–25.

[24] Kornish, L.J., and K.T. Ulrich, "Opportunity Spaces in Innovation: Empirical Analysis of Large Samples of Ideas", *Management Science 57*(1), 2011, pp. 107–128.

[25] Kornish, L.J., and K.T. Ulrich, "Practical methods for assessing the quality of subjective selection processes", 2016.

[26] Levin, I.P., M.E. Huneke, and J.D. Jasper, "Information Processing at Successive Stages of Decision Making: Need for Cognition and Inclusion-Exclusion Effects.", *Organizational behavior and human decision processes 82*(2), 2000, pp. 171–193.

[27] Li, M., A. Kankanhalli, and S.H. Kim, "Which ideas are more likely to be implemented in online user innovation communities? An empirical analysis", *Decision Support Systems 84*, 2016, pp. 28–40.

[28] Maggitti, P.G., K.G. Smith, and R. Katila, "The complex search process of invention", *Research Policy 42*, 2013, pp. 90–100.

[29] Magnusson, P.R., E. Wästlund, and J. Netz, "Exploring Users' Appropriateness as a Proxy for Experts When Screening New Product/Service Ideas", *Journal of Product Innovation Management 33*(1), 2016, pp. 4–18.

[30] Metz, C.E., "Basic Principles of ROC Analysis", *Seminars in Nuclear Medicine 8*(4), 1978, pp. October.

[31] Nagar, Y., P. de Boer, and A.C.B. Garcia, "Accelerating the Review of Complex Intellectual Artifacts in Crowdsourced Innovation Challenges", *Thirty Seventh International Conference on Information Systems*, (2016), 1–17.

[32] Nunnally, J.C., *Psychometric Theory*, McGraw-Hill, New York, 1978.

[33] Onarheim, B., and B.T. Christensen, "Distributed idea screening in stage–gate development processes", *Journal of Engineering Design 23*(9), 2012, pp. 660–673.

[34] Paas, F., and J.G. van Merriënboer, "Variability of worked examples and transfer of geometric problem-solving skills: A cognitive load approach.", *Journal of Educational Psychology, 86*(1), 1994, pp. 122–133.

[35] Prpić, J., P.P. Shukla, J.H. Kietzmann, and I.P. McCarthy, "How to work a crowd: Developing crowd capital through crowdsourcing", *Business Horizons 58*(1), 2015, pp. 77–85.

[36] Riedl, C., I. Blohm, J.M. Leimeister, and H. Krcmar, "The Effect of Rating Scales on Decision Quality and User Attitudes in Online Innovation Communities", *International Journal of Electronic Commerce 17*(3), 2013, pp. 7–36.

[37] Riedl, C., N. May, J. Finzen, S. Stathel, V. Kaufman, and H. Krcmar, "An Idea Ontology for Innovation Management", *International Journal on Semantic Web and Information Systems 5*(4), 2009, pp. 1–18.

[38] Riedl, R., E. Brandstätter, and F. Roithmayr, "Identifying decision strategies: A process- and outcome-based classification method", *Behavior Research Methods 40*(3), 2008, pp. 795–807.

[39] Rietzschel, E.F., B.A. Nijstad, and W. Stroebe, "The selection of creative ideas after individual idea generation: Choosing between creativity and impact", *British Journal of Psychology 101*(1), 2010, pp. 47–68.

[40] Robinson, A.G., and D.M. Schroeder, *Ideas are free: How the idea revolution is liberating people and transforming organizations*, Berrett-Koehler Publishers, 2004.

[41] Schaffhausen, C.R., and T.M. Kowalewski, "Large scale needs-based open innovation via automated semantic tectual similarity analysis", *International Design Engineering Technical Conferences & Computers and Information in Engineering Conference*, ASME (2015), 1–11.

[42] Schulze, T., M. Indulska, D. Geiger, and A. Korthaus, "Idea assessment in open innovation: A state of practice", *European Conference on Information Systems, ECIS*, (2012), 1–13.

[43] Siangliulue, P., K.C. Arnold, K.Z. Gajos, and S.P. Dow, "Toward Collaborative Ideation at Scale - Leveraging Ideas from Others to Generate More Creative and Diverse Ideas Pao", *CSCW*, ACM (2015), 937–945.

[44] Simon, H.A., "A Behavioral Model of Rational Choice", *Source: The Quarterly Journal of Economics 69*(1), 1955, pp. 99–118.

[45] Stehman, S. V, "Selecting and interpreting measures of thematic classification accuracy", *Remote Sensing of Environment 62*(1), 1997, pp. 77–89.

[46] Sweller, J., J.J.G. Van Merrienboer, and F.G.W.C. Paas, "Cognitive Architecture and Instructional Design", *Educational Psychology Review 10*(3), 1998, pp. 251–296.

[47] Toubia, O., and O. Netzer, "Idea Generation, Creativity, and Prototypicality", *Marketing Science 36*(1), 2017, pp. 1–20.

[48] Velamuri, V.K., D. Schneckenberg, J.B.A. Haller, and K.M. Moeslein, "Open evaluation of new product concepts at the front end of innovation: objectives and contingency factors", *R&D Management 47*(4), 2015, pp. 501–521.

[49] Walter, T.P., and A. Back, "A Text Mining Approach to Evaluate Submissions to Crowdsourcing Contests", *46th Hawaii International Conference on System Sciences*, IEEE (2013), 1–10.

[50] Wang, W., and I. Benbasat, "Interactive Decision Aids for Consumer Decision Making in E-Commerce: The Influence of Perceived Strategy Restrictiveness", *Source: MIS Quarterly 33*(2), 2009, pp. 293–320.

[51] Wessling, K.S., J. Huber, and O. Netzer, "MTurk Character Misrepresentation: Assessment and Solutions", *Journal of Consumer Research 44*(1), 2017, pp. 211–230.

[52] Westerski, A., T. Dalamagas, and C.A. Iglesias, "Classifying and comparing community innovation in Idea Management Systems", *Decision Support Systems 54*(3), 2013, pp. 1316–1326.