TESTING TOURNAMENT SELECTION IN CREATIVE PROBLEM SOLVING USING CROWDS

Completed Research Paper

Yasuaki Sakamoto  
Stevens Institute of Technology  
Hoboken, NJ 07030 USA  
ysakamot@stevens.edu

Jin Bao  
Stevens Institute of Technology  
Hoboken, NJ 07030 USA  
jbao@stevens.edu

Abstract

We tested a technique for creative problem solving, which used crowd-based genetic algorithms; one crowd generated initial ideas, another crowd evaluated the quality of these ideas, and yet another crowd combined pairs of ideas selected by the computer. The pairs were selected through a tournament method, in which two ideas chosen were biased toward higher quality. To test the technique, we asked a crowd to evaluate a subset of 468 solutions for the 2010 oil spill in the Gulf of Mexico produced by another crowd of 1853 individuals, and 311 solutions by 311 experts. The crowd evaluated the most creative crowd solutions as creative as the most creative expert solutions. Moreover, tournament selection led to greater improvement in the creativity of combined solutions than random selection, in which two solutions were chosen randomly. Creative problem solving using crowd-based genetic algorithms can work with good design.

Keywords: Creative problem solving, crowd-based genetic algorithms, evaluation
Introduction

The ESP Game (von Ahn and Dabbish 2004), Peekaboom (von Ahn et al. 2006), Verbosity (von Ahn et al. 2006), Tag-A-Tune (Law et al. 2007), FACTory (game.cyc.com), and foldit (www.fold.it) are all examples of crowdsourcing, called Games with a Purpose. In these crowdsourcing activities, players perform some computation as part of a game, such as labeling an image by predicting how a partner may label the same image. They are ad hoc groups responding to an open call (Howe 2006) and coordinate across time and space (Wagner and Back 2008).

Instead of making games, other researchers have used Amazon's Mechanical Turk (AMT: www.mturk.com) to recruit workers for nominal fees. Workers complete short tasks, such as the labeling of data for machine vision (Sorokin et al. 2008) and natural language processing (Snoe et al. 2008; Bernstein et al. 2010). For example, Soylent is a word processing interface that enables writers to call on AMT workers to edit their documents, a small part at a time.

In the above examples, the crowd collectively process a large amount of information that is unreasonable for a few professionals to process, but each individual in the crowd performs a simple task that is easy for humans to complete (except foldit); the interesting observation is not the ability of humans to complete the task, but the way in which the task is broken down and the responses are combined (Adar, 2011), which contribute to the research of coordination in information systems (Boh et al. 2007; Chan 2008; Durcikova 2009; Heckman et al. 2005; Hsiao 2008; Kane et al. 2007; Okoli et al. 2004).

By contrast, other crowdsourcing systems try to solve ill-defined, open-ended problems that are not easy for humans. For example, the articles on Wikipedia and the software on SourceForge are user-generated (Benkler 2002) and can be regarded as crowdsourced content. For these tasks, it is interesting that the crowd is successful in the sense that the collective solutions of the crowd can be as good as those of experts (Surowiecki 2004). However, the task decomposition is unclear in these systems, and it is hard to study which aspects of design contribute to the success of these systems.

Previously, we have developed a crowdsourcing system, in which the crowd produces creative output for ill-defined problems through the process of generating, evaluating, and combining ideas (Nickerson and Sakamoto 2010). The mechanisms of the system were testable; the system was experimentally tested using AMT workers who designed sketches for children’s chair (Nickerson et al. 2011; Yu and Nickerson 2011a; Yu and Sakamoto 2011) and alarm clocks (Yu and Nickerson 2011b; Yu et al. 2011) as well as using undergraduates who generated text ideas for reducing the negative influences of misinformation on Internet users (Tanaka et al. 2011).

The purpose of the current work is to further test this crowdsourcing system by evaluating ideas it produces. The system is detailed in the next section. The current work differs from the past work in the following three ways:

First, we focus here on a creative problem solving by the crowd, in which the output is free-form text ideas for solving a social problem. In contrast, design sketches have been the focus of most of the past work in this research program (e.g., Yu and Nickerson 2011a). The social problem we used was how to deal with the 2010 oil spill in the Gulf of Mexico. The one study (Tanaka et al. 2011) that examined generation of text ideas involved undergraduates, not the crowd, and used a different social problem. The current work thus tests the generality of the system.

Second, we examine two different ways of combining ideas. Other studies have examined whether or not combining ideas resulted in more creative output than simply generating ideas (Yu and Nickerson 2011b) or criticizing and modifying ideas (Tanaka et al. 2011). Past work has also studied how evaluation process changes the distribution of the creativity of ideas that are produced by the system (Bao et al. 2011). We add to this knowledge a comparison of two mechanisms of selecting pairs of ideas the crowd combines.

Third, we compare the creativity of ideas produced by the crowd against that by experts. We asked a crowd of 166 individuals to evaluate a subset of 468 solutions produced by 1853 individuals who participated in the crowdsourcing system, and 311 solutions by 311 experts. These ideas were collected while the crisis was happening. If none of the crowdsourced ideas is as creative as the most creative ideas by experts, then the crowdsourcing system will not be useful even if it is time and cost effective.
Through the evaluation of these ideas, we try to answer the following interrelated research questions:

**Q1 – Can the crowdsourcing system produce creative ideas?** To answer this question, we compare ideas produced by the crowd to ideas generated by experts. The expert ideas were generated by professors, scientists, and engineers, posted on the Principal Investigators Association Website (http://www.principalinvestigators.org/).

**Q2 – Does combining ideas improve creativity?** To address this question, we compare the creativity of non-combined ideas with that of combined ones. We describe the detail in the next section.

**Q3 – Does it matter which two ideas are combined with respect to improving creativity?** We address this question by comparing tournament selection, in which the selection of two ideas to be combined is biased toward better ideas, with random selection, in which two ideas are randomly selected for combination. We explain why we use combination and these selection mechanisms in the next section.

In the remainder of the paper, we first describe the design of the crowdsourcing system in relation to theoretical foundations and hypotheses. We then present the method for collecting ideas and evaluating the system, followed by the evaluation results. Finally, we discuss the future directions for this line of research and conclude that creative problem solving by the crowd can work.

![Diagram of Amazon Mechanical Turk](Image)

**Figure 1.** In the crowdsourcing system, crowds perform different steps through many generations.

### Theoretical Foundations and Hypotheses

In this section, we describe the crowdsourcing system, and present our hypotheses. We are interested in whether the system, not each individual, can produce creative ideas.

#### The Crowdsourcing System

Figure 1 visualizes the design of the crowdsourcing system, in which different crowds generate and combine ideas, and the computer selects ideas based on the crowd’s evaluation of the ideas. In the present work, five crowds perform the generation, evaluation, and combination tasks. As overviewed in Figure 1, Crowd 1 generates a set of ideas, and Crowd 2 evaluates these ideas. Then computer algorithms select
pairs of ideas to be combined based on the evaluation by Crowd 2. Crowd 3 generates ideas by combining the selected pairs of ideas. Crowd 4 evaluates the combined ideas from Crowd 3. Computer algorithms select pairs of ideas for combination based on the evaluation of Crowd 4. Finally, Crowd 5 combines selected pairs of ideas.

The design of the crowdsourcing system is based on the ideas of genetic algorithms, which have been successful in optimizing many tasks (e.g., Fogel 2006; Goldberg 1989). In genetic algorithms, combining part of one genome and part of another results in a new genome. In relation to our work, combining parts of Parents (ideas) produces a child (idea) who benefits from the good qualities of its parents. Combination can filter out poor features of the parents and preserve the best of the parents. We ask the crowds, not the computer, to carry out the combination process of genetic algorithms (Kosorukoff 2001), because genetic algorithms need clear objective functions and solution representations, neither of which can be provided for ill-defined social problems as the one used here. The crowdsourcing system thus is a crowd-based genetic algorithms system in that the crowds combine ideas selected by the computer.

By applying genetic algorithms to the crowdsourcing system, we are essentially creating specialized crowds that collectively complete processes akin to divergent thinking and convergent thinking, which are thought to underlie creative production (Guilford 1967; Osborn 1957). Divergent thinking involves producing a variety of ideas. Convergent thinking involves filtering and synthesizing these ideas to find the “correct” solution. Generating a set of ideas in our system is like divergent thinking, and evaluating ideas is like convergent thinking. Combining ideas involves both divergent and convergent thinking as it is about generating new ideas by integrating existing ideas in meaningful ways.

The Geneplor model (Finke et al. 1992) further suggests that generating a diverse set of candidate ideas (akin to divergent thinking) can be useful for producing something original, and subsequent explorations and refinements of these ideas (akin to convergent thinking) can produce something practical. Originality and practicality are regarded as two important components of creativity (Mumford 2003; Runco & Pritzker 1999) and are often used as measures of creativity (Ward et al. 1997). Consequently, we measure creativity with respect to originality and practicality in the present work. We are interested in whether the crowds can produce original and practical ideas through generating, evaluating, and combining ideas (Q1). Although we think that crowds will produce many ideas that are not creative, given the relationship between our system and past work on creative production mentioned here, we believe that some of the crowd ideas will be refined through the processes of evaluation and combination:

**H1 – Some of the crowd ideas will be as original as the most original ideas by experts.**

**H2 – Some of the crowd ideas will be as practical as the most practical ideas by experts.**

The crowdsourcing system proceeds to request AMT workers to generate a population of solutions, and then improve it through repetitive applications of evaluation and combination. The system generated solutions for dealing with the 2010 oil spill in the Gulf of Mexico. Figure 2 shows an example of how the system asks the crowds to generate, evaluate, and combine ideas. Next we detail the generation, evaluation, and combination steps.

**Generation**

Initially a crowd generates many individual solutions, which form an initial population of ideas. For example, a typical starting population size is 100 in the genetic algorithms research (cf. Deb 1999). One advantage of crowdsourcing is that members of a crowd tend to have diverse backgrounds, and thus the initial population can results in a range of ideas. In related work that combines humans and computer algorithms, the machine typically generates the initial generation of solutions (and combine existing solutions), and humans are used only for evaluating the solutions (e.g., Takagi 1998).
Generation

Instruction:

Please read the material:

On April 20, 2010, an oil well explosion caused a sea-floor oil spill in the Gulf of Mexico. More than a month later, the oil continues to flow into the gulf, towards land, and out to the ocean. The latest solution by experts may work, but better, faster, and more innovative solutions for the current and future oil spill crises are needed.

Please use the space below to share your creative and novel idea for stopping or cleaning an oil spill like the one in the Gulf of Mexico. Other workers on Mturk will vote for all ideas. We will pay a $2 bonus to players whose idea ranks in top three!

Two participants’ response:

Idea 1 (Parent 1): Spilled oil should be skimmed as fast as possible. It’s hard to do this using large ships as it is done today. One good option would be to create robotic unmanned floating (or underwater) drones equipped with sensors that detect oil presence that could collect it and deposit in other, larger, floating autonomous storage. Such robot swarms that work in large teams could be deployed to skim the spilled oil as fast as possible.

Idea 2 (Parent 2): I think that using a kind of absorbent fibers will help to stop an oil spill.

Evaluation

Instruction:

Select one out of seven points representing how good each idea is.

1. I think that using a kind of absorbent fibers will help to stop an oil spill.

Very poor ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ 6 ○ 7 Very good

Combination

Instruction:

Please read the material:

On April 20, 2010, an oil well explosion caused a sea-floor oil spill in the Gulf of Mexico. More than a month later, the oil continues to flow into the gulf, towards land, and out to the ocean. The latest solution by experts may work, but better, faster, and more innovative solutions for the current and future oil spill crises are needed.

Please use the space below to combine two ideas for stopping or cleaning an oil spill like the one in the Gulf of Mexico. Other workers on mturk will vote on all combined ideas. We will pay a $2 bonus to players whose combined idea ranks in top three!

Idea 1. Spilled oil should be skimmed as fast as possible. It's hard to do this using large ships as it is done today. One good option would be to create robotic unmanned floating (or underwater) drones equipped with sensors that detect oil presence that could collect it and deposit in other, larger, floating autonomous storage. Such robot swarms that work in large teams could be deployed to skim the spilled oil as fast as possible.

Idea 2. I think that using a kind of absorbent fibers will help to stop an oil spill.

One participant’s response:

Combined idea (Child): Using absorbent fibers wrap around the robotic unmanned floating (or underwater) drones equipped with sensors that detect oil presence that could collect it and deposit in other, larger, floating autonomous storage. Will make the work faster.

Figure 2. Crowds generated, evaluated, and combined ideas. Three responses are also shown. Idea 1 was ranked 78 out of 500 and Idea 2 was ranked 248 out of 500. Combined idea was ranked 5 out of 500.
Selection and Evaluation

Like genetic algorithms, machine selects a portion of the existing population to breed a new generation of ideas through combination. Unlike standard genetic algorithms, this selection is based on a crowd’s evaluation of the ideas: a new crowd evaluates the ideas.

Following genetic algorithms, tournament selection (Goldberg 1989) is used in the system, which is found to be effective in machine learning (Fogel 2006). In tournament selection, the machine randomly selects two ideas. Then, the machine selects the better idea based on the crowd’s evaluation of the ideas. The machine randomly selects two more ideas, and selects the better idea based on crowd’s evaluation. The two selected ideas will be presented to the next crowd who combines the two ideas. The machine selects pairs of ideas in this way. In tournament solution, the selection of pairs of ideas is biased toward better ideas.

Despite its usefulness in machine learning, whether or not tournament selection will work for humans is unclear. Thus, we compare tournament selection with random selection through an experiment. In random selection, the machine selects pairs of ideas randomly. Whereas one set of crowds combined ideas chosen by tournament selection, the other set of crowds combined ideas chosen by random selection:

**H3** – Tournament selection will result in more original ideas than random selection.

**H4** – Tournament selection will result in more practical ideas than random selection.

Combination

The next step is to generate a second-generation of ideas through combination: a new crowd generates a new idea by combining the two ideas selected by the computer.

In addition to the use of genetic algorithms in the system, combination has been thought to underlie creative acts. For example, creativity is proposed to be due to novel combinations of mental representations (Thagard & Stewart 2011). Research in cognitive science has shown a link between creativity and conceptual combinations, in which separate ideas or concepts are merged (e.g., Thagard 1984; Ward 2004), and many original arts and scientific discoveries are said to be the results of synthesizing two opposing ideas (Rothenberg 1979). Furthermore, work in brainstorming has shown that shared ideas can stimulate the subsequent idea generators to think of other ideas or categories of ideas (Dugosh et al. 2000; Rietzschel et al. 2007), if they do not result in production blocking (Gallupe et al. 1992) and cognitive fixation (Kohn & Smith 2010). In particular, groups of people can generate more creative combinations when given appropriate ideas to combine (Kohn et al. 2011).

For these reasons, coupled with the success of genetic algorithms machine learning (e.g., Fogel 2006; Goldberg 1989), we think combining ideas will improve creativity:

**H5** – Combined ideas will be more original than the non-combined ideas.

**H6** – Combined ideas will be more practical than the non-combined ideas.

Of course, combination interacts with selection. For example, it could be that combined ideas become more creative than non-combined ideas when tournament selection is used. **H3-6** together answer **Q2** and **Q3**, and test if the crowd-based genetic algorithms system is a viable method for generating creative ideas through combination.

Will the Crowdsourcing System Work?

We can recruit many people from diverse background using crowds and collect numerous ideas in a short time. These advantages make the crowdsourcing of creative problem solving attractive and promising. Recently, other researchers have successfully conducted large scale experiments using different crowdsourcing systems, in which humans perform tasks akin to computers performing processing (e.g. Kittur et al. 2008; Little et al. 2010; Raykar et al. 2010).

We think that the crowdsourcing system can produce creative solutions. One reason is the power of the wisdom of crowds. In a variety of tasks, the median of the individual guesses collected independently from
a large group of people becomes extremely close to the actual value, and is often better than any individual’s guess (e.g., Galton 1907; Lorge et al. 1958). Of course, we do not take the median of the crowdsourced ideas, and our task is not to guess some values. Nevertheless, the repeated combinations of ideas may overcome problems associated with each idea, just as the aggregate of guesses can cancel out noises resulting from individual errors.

Another reason why we think the system can produce creative solutions is the lack of direct interaction among participants. Direct interaction can actually hurt the creativity processes. For example, the participants in a brainstorming session often see one another and discuss all the proposed ideas. This presence of others can prevent self-conscious individuals to contribute ideas, and vocal members can sway the opinion of the group (Asch 1951; Lorge et al. 1958; Mullen et al. 1991). Production blocking (Gallupe et al. 1992) and cognitive fixation (Kohn & Smith 2010) are other potential disadvantages of direct interaction.

Although there is no direct interaction in our crowdsourcing system, the design of the system allows for some creative collaboration. Participants collaborate through their responses across time. For example, Crowd 3 modifies Crowd 1’s ideas, and Crowd 2’s responses influence which two ideas Crowd 3 modifies. Nevertheless, participants in our crowds are not co-present and do not communicate. Such indirect collaboration may be the key to successful crowdsourcing systems.

To foreshadow our results, we found that experts overall produced more creative ideas than crowds, judged by the crowds on originality and practicality. However, the most creative ideas from the crowds were as creative as the most creative ideas from the experts (Q1: H1 and H2). Thus, the crowdsourcing system is promising for generating creative solutions. We also found that combining ideas using tournament selection leads to greater improvement in creativity of ideas than combining ideas using random selection (Q2: H5 and H6). This suggests that whether or not combining ideas improves creativity depends on which two ideas are combined. Tournament selection, in which pairs are biased toward better ideas, can be more effective in producing creative ideas than random selection, in which pairs are randomly chosen (Q3: H3 and H4).

**Method**

To test our hypotheses, we conducted an experiment using the crowdsourcing system. Participants were 1853 AMT workers who collectively produced 468 ideas for dealing with the 2010 oil spill in the Gulf of Mexico, while the crisis was happening. Our crowds were 57% male with the mean age of 30 (SD = 10). They participated online for nominal fees. Each worker participated in only one task. The data collection lasted from July to November 2010. The ideas of 311 experts, including professors, scientists, and engineers, were collected from Principal Investigators Association’s Website.

**Design and Procedure for Crowdsourcing**

**Crowd 1 (Parents)** Crowd 1 generated 100 ideas, which we call Parents. Examples of parent ideas, Parent 1 and Parent 2, are shown in the generation part of the Figure 2. Each of 100 participants was asked to generate one idea. There were 7 empty ideas. One of our hypotheses was whether these parents would be judged less or more creative than the combined ideas (i.e., their offspring).

**Crowd 2 (Evaluation Scores)** Crowd 2 produced evaluation scores. Each of 465 participants was presented with one parent idea and asked to evaluate its quality using a 7-point scale (1= very poor; 7 = very good). The evaluation scores were used to order the parents by rank so that the computer could determine the better idea during tournament selection. In tournament selection, the computer randomly selected two ideas. The one receiving a higher evaluation score by Crowd 2 won. This procedure was repeated to find another winner. These two winners were chosen as a pair, and presented to Crowd 3.

**Crowd 3 (Children)** Crowd 3 combined pairs of parents into new ideas, which we call Children. One example of combined idea is shown in the combination part of Figure 2. Each of 184 participants was given two parents and asked to combine them. Participants were randomly assigned to either the tournament or random condition. In the tournament condition, each of 92 participants received two parents based on tournament selection. In the random condition, each of 92 participants received two
ideas that were randomly selected. The purpose of the random condition was to test if tournament selection is a good candidate for selecting ideas in our human-based genetic algorithms. Crowd 3 generated 184 combined ideas, 92 in the tournament condition, and 92 in the random condition. Following genetic algorithms, we used elitism, in which the best 8 parents persisted through the end; elitism makes sure that the system does not devolve through combination.

**Crowd 4 (Evaluation Scores)** Crowd 4 produced evaluation scores in the same manner as Crowd 2 did; 920 participants evaluated the 184 Children ideas from Crowd 3. The machine selected pairs of ideas to be combined by Crowd 5.

**Crowd 5 (Grandchildren)** Each of 184 participants in Crowd 5 generated a Grandchildren idea by combining two Children ideas from Crowd 3. The design and procedure of Crowd 5 were identical to those of Crowd 3.

**Procedure for Final Evaluation: Prediction Voting**

Our main interests were comparing (1) crowds' ideas with experts', (2) parents with Children and Grandchildren, and (3) ideas from the tournament condition with ideas from the random condition, in the originality and practicality dimensions. We randomly sampled 30 ideas from each of the following six groups: Parents, Children tournament, Children random, Grandchildren tournament, Grandchildren random, and Experts. The random condition and tournament condition shared the same Parents.

Additional 166 workers, who were not in Crowds 1-5, participated in the final evaluation tasks for nominal fees. One set of participants evaluated the originality of these 180 ideas (30 ideas x 6 groups). A different set of participants evaluated the practicality of the same 180 ideas.

We used prediction voting as our evaluation method. We settled on this method for evaluation after testing prediction voting, five-point Likert scale rating, and other methods several times. We found that although prediction voting and Likert scale rating highly correlate, they have different goals: whereas prediction voting focuses evaluators on identifying the very best solutions, the rating focuses evaluators on the entire range of solutions (Bao et al. 2011). Prediction has several advantages over rating. One is that the crowd is more motivated to participate in prediction voting than Likert scale rating; more workers participated in prediction voting than rating. Second, in the crowdsourcing system, there are many poor quality solutions that need to be filtered out. Prediction voting is more efficient than rating for the purpose of filtering (Bao et al. 2011). Third, for the final evaluation, we are interested in comparison of the best quality ideas in different groups. Predicting the winner focuses participants on the best ideas. We are biasing their responses toward non-winner, but that is exactly what we want: pick the very best and filter out everything short of being the best.

In prediction voting, for the originality evaluation participants were presented with the instruction:

“Recently we collected 180 ideas for solving oil spill problems. One idea that was most novel and surprising received an originality award.”

There was an idea below the instruction. Participants predicted whether the idea is a good candidate for an originality award as shown in Figure 3. The procedure for the practicality award was identical except that the instruction referred to practicality.

---

<table>
<thead>
<tr>
<th>This idea is</th>
<th>☐ the winner</th>
<th>☐ NOT the winner.</th>
</tr>
</thead>
</table>

**Figure 3.** Our final evaluation method was prediction voting, in which participants predicted whether the given idea is a winner of an originality or practicality award.
The participants saw one idea at a time. Ten different participants rated each idea. The value each solution gets from one evaluation is a binary value, where 0 means “not the winner” and 1 means “the winner”. The evaluation score of each idea is defined as the sum of values for this specific idea, which is also the number of votes as the winner each idea received. In other words, the maximum evaluation score one idea can have is 10 and the minimum is 0.

**Results**

Figure 4 shows the distribution of evaluation scores for ideas in each group. It is a combination of a box plot and a kernel density plot; wider means many ideas in that area, and taller means there are ideas with high scores. For example, the distribution of originality scores in Parents group shows that there are many solutions that received 1–3 votes, and fewer solutions that received 4–6 votes. For the Children Random group, 17 of 30 solutions received only 1 vote for the practicality dimension, and thus the box and whisker are missing. Combining ideas through tournament selection results in a different evolution of ideas from combining ideas through random selection, with respect to creativity.

We found that the evaluation scores for practicality are in general lower than that at originality. For originality, the highest evaluation scores came from Grandchildren tournament and expert; ideas from these two groups resulted in similar distribution of evaluation scores, indicating the two groups produced
ideas that are similar in originality. For practicality, the idea with the highest evaluation score came from Grandchildren tournament.

At the right side of the figure the Grandchildren tournament group (the purple plot with red border) and the expert condition (the dark yellow plot) shows visually that at least some ideas from the crowds are as creative as the best ideas from the experts, supporting $H_1$ and $H_2$. Furthermore, both Children and Grandchildren in the tournament condition seem to have higher evaluation scores than those two groups in random selection condition, especially for practicality, supporting $H_3$ and $H_4$. Grandchildren tournament appears to have higher evaluation scores than Parents for both originality and practicality, consistent with $H_5$ and $H_6$.

Next we test our hypotheses through inferential statistics.

<table>
<thead>
<tr>
<th>Table 1. Group Evaluation Scores</th>
<th>Parents</th>
<th>Children</th>
<th>Grandchildren</th>
<th>Experts</th>
</tr>
</thead>
<tbody>
<tr>
<td>Originality</td>
<td>Random</td>
<td>73</td>
<td>78</td>
<td>75</td>
</tr>
<tr>
<td></td>
<td>Tournament</td>
<td>82</td>
<td>93</td>
<td></td>
</tr>
<tr>
<td>Practicality</td>
<td>Random</td>
<td>27</td>
<td>32</td>
<td>25</td>
</tr>
<tr>
<td></td>
<td>Tournament</td>
<td>41</td>
<td>50</td>
<td></td>
</tr>
</tbody>
</table>

**Hypotheses Testing**

Table 1 shows the evaluation scores of 6 conditions for the originality and practicality dimensions. The score of each group is defined as the sum of all 30 ideas’ scores in that group. Table 2 summarizes our hypotheses and statistical results.

**Crowds vs. Experts**

Chi-square tests supported $H_1$ and $H_2$; examination of the top 5 ideas in each group shows that the difference between experts and crowds in the creativity of the best ideas they produced were small. For originality, the total score of top 5 ideas at the Expert group is 30, and the total score of the top 5 ideas at the Grandchildren tournament group is 28 [$\chi^2 < 1$]. For practicality, the top 5 score in Experts is 22, and the top 5 score in Grandchildren tournament group is 22 [$\chi^2 < 1$]. Most creative ideas by crowds are as creative as most creative ideas by experts.

Furthermore, Grandchildren tournament did not differ statistically from Experts on both originality dimension [$\chi^2(1, N=200) = 0.98$, $p = .32$] and practicality dimension [$\chi^2(1, N=114) = 1.72$, $p = .19$]. However, Experts dominated crowds other than Grandchildren tournament on both originality and practicality dimensions, as can be seen in Table 1. For the originality dimension, Experts resulted in a significantly higher score than Parents [$\chi^2(1, N=180) = 6.42$, $p = .01$], Children random [$\chi^2(1, N=185) = 4.55$, $p = .03$], and Grandchildren random [$\chi^2(1, N=182) = 5.62$, $p = .02$]. The difference between Experts and Children tournament approached significance [$\chi^2(1, N=189) = 3.31$, $p = .06$].

The pattern of results was similar for practicality. In general, the experts produced ideas that are much more practical than the crowds. Except for Grandchildren tournament, Experts resulted in a significantly higher score than Parents [$\chi^2(1, N=91) = 15.04$, $p < .001$], Children random [$\chi^2(1, N=96) = 10.67$, $p = .001$], Grandchildren random [$\chi^2(1, N=99) = 17.09$, $p < .001$] and Children tournament [$\chi^2(1, N=105) = 5.04$, $p = .02$].

Coupled together, the results indicate that although experts produced more creative ideas than the crowds in general, the best ideas from the crowds were judged as creative as the best ideas from the experts. This suggests to us that the crowdsourcing system is promising.
Table 2. Summary of Hypotheses and Results

<table>
<thead>
<tr>
<th>Hypotheses</th>
<th>Results</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>H1</strong> – Some of the crowd ideas will be as original as the most original ideas by experts.</td>
<td>Yes, the total originality score of top 5 ideas from the experts was not significantly different from that from the crowds.</td>
</tr>
<tr>
<td><strong>H2</strong> – Some of the crowd ideas will be as practical as the most practical ideas by experts.</td>
<td>Yes, the total practicality score of top 5 ideas from the experts was not significantly different from that from the crowds.</td>
</tr>
<tr>
<td><strong>H3</strong> – Tournament selection will result in more original ideas than random selection.</td>
<td>No, the tournament selection group and random selection group did not differ significantly in originality scores. But, the improvement in originality scores from Parents to Grandchildren in the tournament condition was significantly greater than that in the random condition.</td>
</tr>
<tr>
<td><strong>H4</strong> – Tournament selection will result in more practical ideas than random selection.</td>
<td>Yes, the tournament condition and the random condition differed significantly in practicality scores. Moreover, the improvement in practicality scores from Parents to Grandchildren in the tournament condition was significantly greater than the improvement in the random condition.</td>
</tr>
<tr>
<td><strong>H5</strong> – Combined ideas will be more original than the non-combined ideas.</td>
<td>No, the difference in originality scores between Grandchildren tournament and Parents did not reach significance, but there was a trend in the right direction.</td>
</tr>
<tr>
<td><strong>H6</strong> – Combined ideas will be more practical than the non-combined ideas.</td>
<td>The practicality score of ideas in Grandchildren tournament was significantly higher than that in Parents.</td>
</tr>
</tbody>
</table>

**Tournament Selection vs. Random Selection**

To compare tournament selection and random selection, we defined the tournament condition score as adding the score of the Children and Grandchildren groups in the tournament selection condition together, and did the same for the random selection condition.

For originality, the difference between the tournament condition and random condition did not differ significantly $(\chi^2(1, N=328) = 1.48, p = .22)$. However, whereas Children and Grandchildren in the tournament condition were becoming more creative than Parents, the pattern was flat in the random condition. The score of Grandchildren tournament was 20 higher than that of Parents. In contrast, the score of Children random, which was the highest in this condition, was only 5 higher than that of Parents. The improvement from the Parents to the Grandchildren group in tournament condition...
was significantly higher than the improvement in the random selection [$\chi^2(1, N=25) = 9.0, p = .003$]. This indicates that originality scores improved more through tournament selection than random selection, partially supporting $H_3$.

We found significant differences of scores between the tournament condition and the random condition (91 vs. 57) [$\chi^2(1, N=148) = 7.81, p = .005$]. Further, the significant difference of the improvement from the Parents to the Grandchildren group between the tournament condition and the random condition supported $H_4$ [$\chi^2(1, N=28) = 11.57, p < .001$].

These results suggest that tournament selection is more effective in improving creativity of ideas than random selection. Combining ideas through tournament selection especially improved the practicality of ideas.

**Parents vs. Children vs. Grandchildren**

As can be seen in Figure 4 and Table 1, whereas the Children and Grandchildren in the tournament condition were becoming more original than Parents, the pattern was flat in the random condition. Although the difference in originality between Grandchildren tournament and Parents did not reach significance [$\chi^2(1, N=166) = 2.41, p = .12$], the trend was in the predicted direction. There was no significant difference in originality between Parents and Children, and between Parents and Grandchildren in the random condition [$\chi^2 < 1$ for both]. Combining ideas using random selection did not help improve originality. Thus $H_5$ was not fully supported, but there was a trend in the tournament condition.

The pattern was similar but stronger for practicality. The score of Grandchildren tournament group was significantly higher than the score of the Parents [$\chi^2(1, N=77) = 6.87, p = .009$]. In the random condition, the scores of both Children and Grandchildren did not differ from the score of Parents [$\chi^2 < 1$ for both]. $H_6$ was supported in the tournament condition, indicating that combining ideas could improve creativity through tournament selection. Whether or not combining ideas improves creativity depends on which two ideas are combined.

**Discussion**

We examined the ability of crowds to perform creative problem solving by testing a crowdsourcing technique, in which crowds and machine sequentially engaged in the process of generating, evaluating, selecting, and combining ideas. The technique involved crowd-based genetic algorithms, in which crowds processed the generation and combination of ideas, and the computer processed the selection of pairs for combination.

**Findings from the Present Work**

There were three main findings, which answered the three research questions in the current work.

**Q1 – Can the crowdsourcing system produce creative ideas?** The answer is yes. Although experts produced more creative ideas than crowds in general, the best ideas from the crowds were as original and practical as the best ideas from the experts.

**Q2 – Does combining ideas improve creativity?** Yes, with good design. Consistent with the past work on conceptual combination in creativity (e.g., Ward et al. 1997), combining ideas increased creativity, but only when the ideas selected for combination were biased toward better ones in tournament selection, not when two ideas were randomly selected for combination in random selection.

**Q3 – Does it matter which two ideas are combined with respect to improving creativity?** Yes, combining ideas using tournament selection led to more creative ideas than combining ideas using random selection, especially on the practicality dimension.

Our results show that $H_1$, $H_2$, $H_3$, $H_4$, and $H_6$ were supported. Although $H_5$ was not supported, there was a trend that combining ideas in tournament condition improved originality.
Here is the process by which one of the most original ideas from the crowds was generated in tournament selection:

Child 1 was: “Using Chemical dispersant’s Materials that break down the oil into its chemical constituents. This helps disperse the oil and make it less harmful to wildlife and shorelines”. (Originality score: 3)

Child 2 was: “Insert a large estopple after clamping it to keep it in place and inflate the stopple to shut off the flow. With the flow shut off remove any damaged parts and install a new valve on top with the remote cutoff that will work. Deflate the stopple and remove it through the new valve. Now close the valve and install a new line. A stopple is just a large very heavy walled balloon that can be inflated to very high pressure and stop the flow of steam or hydrocarbon”. (Originality score: 3)

The Grandchild idea that combined the two Children above was: “first Using Chemical dispersant’s Materials that break down the oil into its chemical constituents. This helps disperse the oil and make it less harmful to wildlife and shorelines then treat this like a large high-pressure steam line with a broken valve. Insert a large estopple after clamping it to keep it in place and inflate the stopple to shut off the flow. With the flow shut off remove any damaged parts and install a new valve on top with the remote cutoff that will work. Deflate the stopple and remove it through the new valve. Now close the valve and install a new line. A stopple is just a large very heavy walled balloon that can be inflated to very high pressure and stop the flow of steam or hydrocarbon”. (Originality score: 6)

In this example, the originality scores of the two Children were 3, but Grandchild’s score was 6, the second highest in Grandchildren tournament. Its originality is improved by the addition of the new features during the process of combination. In this way, combining ideas can result in highly original ideas.

From these observations, we think that the idea of crowdsourcing creativity has great potential. Nevertheless, there is plenty of room for improvement.

**Limitation**

One limitation of our study is that the final evaluators of the ideas were a crowd. It could be that crowds judge ideas by other crowds, and thus biased in their evaluation. We are currently working collecting evaluation from experts; initial results seem to suggest that, like crowds, experts judge ideas by experts to be generally more creative, but there are some ideas by the crowds that are highly creative. We plan to carefully compare the ratings by crowds and experts, and see whether they show different kinds of biases in their judgments. For some responses, such as customer sentiment and product liking, crowds may be the more appropriate judges. For social problems, evaluations from both crowds and experts may be needed to address both social desirability and practicality.

Another limitation is the generality of the results. Although we have tested the system in a variety of domains and examined a range of mechanisms used in the system, the space is large given the complexity of the system. More work is needed to better understand when the crowdsourcing system works and when they do not.

It is possible that some workers paste a solution they find online. We tried to catch such behavior by searching crowd-generated ideas online. We found only a few instance of this gaming behavior using our search method, but we could not know for sure that all the ideas by the crowds are their own. At the same time, the spirit of crowdsourcing is for each individual to complete a simple task quickly; it may be OK for some members of the crowd to search solution online.
Finally, in the current work, we saw quite a few cases of the crowd simply aggregating two ideas without integrating the features of the ideas. Nevertheless, combination through tournament selection improved the creativity of the crowd’s ideas, and some of these ideas were judged as creative as the most creative ideas by the experts.

**Future Directions**

**Combination Process** Combining ideas through tournament selection led to large improvement in practicality. It could be that during combination, features of ideas that are impractical are refined or filtered out (cf. Yu & Sakamoto 2011).

The effect of combination was not as strong for originality. To further improve originality through combination, we need to examine the combination method more closely. Studies on conceptual combination in creativity show that combining dissimilar ideas can lead to the discovery of more original features that did not exist in the initial ideas (Estes & Ward 2002; Kunda et al. 1990; Wilkenfeld & Ward 2001). Combining two opposing ideas may be even more effective in increasing originality (Rothenberg 1979). But two ideas that are irrelevant may be too hard to combine, because finding distant analogies is not an easy task (e.g., Duncker 1945; Solomon 1994). Applying these findings to crowds, we could have a crowd that evaluates the similarity of ideas, and control the selection of two ideas based on the similarity evaluation.

Moreover, different combination processes may be primed depending on the similarity of the paired ideas. Whereas priming abstract representations can be effective in combining highly dissimilar ideas in a creative way, priming shared features can be effective in finding a creative solution for combining related idea (Mumford et al. 1997). Thus the combination process should be tailored to match the nature of the pairs provided to the crowds.

**Other Tasks** There are tasks other than combination that may improve the creativity of ideas. For example, one crowd may criticize the ideas by identifying problems associated with the ideas. The next crowd may address the problems (cf. Tanaka et al. 2011). Critical crowds can facilitate the refinement process, and perhaps especially contribute to the practicality of ideas. Another task may be to replace features that are not surprising to something surprising, which tends to be selected by the crowd during combination (Yu and Sakamoto 2011). This crowd can contribute to the originality of ideas.

These different tasks can take place in different phases of crowdsourcing. For example, critical crowds may be most beneficial later in crowdsourcing so that a diverse set of ideas is available early on (cf. Osborn 1957). Ideas with new elements added on may be most useful early on, and combining such ideas may result in surprising ideas. Critical crowds can later make these surprising ideas practical.

**Interface Design** Some items are easier to combine than others. When features of items are readily identifiable, the items can be integrated nicely by combining the features. Ideas in free texts may not be easy to combine and evaluate. This is because whereas it is relatively easy to find features of a two-dimensional design for a chair to combine, finding features of a solution for a social problem seems hard. Consequently, combining two text ideas is harder than combining two graphic designs. Moreover, in the present work, crowds wrote ideas freely, and thus ideas varied in their structures. For instance, some ideas may have a summary before examples. Other ideas may contain examples throughout the ideas. The differences in structures likely influence people’s evaluation of creativity even if the basic concepts are the same (cf. Schwarz and Clore 1983; 2007).

Future work should explore different ways of generating and combining ideas without using free texts. For example, ideas can be broken down into a meaningful set of words, not each word, and crowds can be constrained on how they can combine two ideas. Such a constraint may actually help improve the combination of ideas, by highlighting the features of the ideas. Furthermore, the interface can constrain the crowd to structure the ideas in the same way. Alternatively we can ask the crowd to find features in ideas, and structure the ideas in a certain way.
Conclusions

We tested a technique for creative problem solving, which used crowd-based genetic algorithms; one crowd generated initial ideas, another crowd evaluated the quality of these ideas, and yet another crowd combined pairs of ideas selected by the computer. The best ideas produced by this crowdsourcing system were evaluated to be as original and practical as the best ideas from the experts. Tournament selection, in which two ideas chosen for combination were biased toward higher quality, was critical in improving the creativity of combined ideas in this system. There are many other mechanisms that need to be tested, and there are also many other tasks the system should be tested on. While the idea of crowdsourcing creative problem solving has great potential, more work is needed to better understand crowd behavior and improve the design of crowdsourcing systems.

Acknowledgements

We thank the members of the Center for Decision Technologies for intellectual support. This work was financially supported by the National Science Foundation under grant IIS-0968561.

References


Nickerson, J.V. and Sakamoto, Y. “Crowdsourcing Creativity: Combining Ideas in Networks,” Workshops on Information in Networks, 2010
Surowiecki, J. 2004. The wisdom of crowds: Why the many are smarter than the few and how collective wisdom shapes business, economies, societies, and nations, New York: Random House.


