Crowdsourcing inspiration: Using crowd generated inspirational stimuli to support designer ideation

Kosa Goucher-Lambert and *Jonathan Cagan*, Department of Mechanical Engineering, Carnegie Mellon University, Pittsburgh, PA 15213, USA

*Inspirational stimuli, such as analogies, are a prominent mechanism used to support designers. However, generating relevant inspirational stimuli remains challenging. This work explores the potential of using an untrained crowd workforce to generate stimuli for trained designers. Crowd workers developed solutions for twelve open-ended design problems from the literature. Solutions were text-mined to extract words along a frequency domain, which, along with computationally derived semantic distances, partitioned stimuli into closer or further distance categories for each problem. The utility of these stimuli was tested in a human subjects experiment (*N = 96*). Results indicate crowdsourcing holds potential to gather impactful inspirational stimuli for open-ended design problems. Near stimuli improve the feasibility and usefulness of designs solutions, while distant stimuli improved their uniqueness.*

Keywords: crowdsourcing, analogical reasoning, creativity, design cognition

Analogical reasoning, and more specifically, design-by-analogy, is a well-studied and active area of investigation within the design research community (Casakin & Goldschmidt, 1999; Chan et al., 2011; Linsey, Wood, & Markman, 2008; Moreno et al., 2014). As has often been observed, design practitioners can gain inspiration and insight from both the same or different domains as the problem, which serve to stimulate the formulation of new ideas during the product development process (Markman, Wood, Linsey, Murphy, & Laux, 2009; Vattam, Helms, & Goel, 2010). As a result, significant emphasis has been placed on trying to uncover the specific types of inspirational stimuli that are most beneficial for assisting productive design activity via analogy (Fu, Cagan, Kotovsky, & Wood, 2013).

Psychological theory posits that analogical reasoning hinges on the successful mapping of relations between a source and a target domain (Krawczyk, McClelland, Donovan, Tillman, & Maguire, 2010). Sometimes these domains are closely related to the problem domain. However, at times these domains

Corresponding author:
Corresponding author.
kosa@berkeley.edu

www.elsevier.com/locate/destud
0142-694X *Design Studies* **61** (2019) 1–29
https://doi.org/10.1016/j.destud.2019.01.001

may be more distant or unrelated to the problem. From a high level, the use of analogies in design has been studied in order to gain an understanding regarding how analogies affect the ideation process itself, as well as the impact analogies have on design outcomes. Analogical reasoning typically is thought to contain two disparate steps: retrieval and mapping (Forbus, Gentner, & Law, 1995). In this work, *inspirational stimuli* are gathered utilizing a crowd workforce and subsequently provided to participants (designers). As such, the pertinent relational mapping from the problem source to the target is left to the designer. While the provided inspirational stimuli are intended to seed analogical relations, which in turn facilitate the retrieval of useful concepts from memory, they are not themselves analogies (Goucher-Lambert, Moss, & Cagan, 2018).

One fundamental open problem within the design research community pertaining to inspirational stimuli and analogy is that it remains a challenge to find relevant stimuli (i.e., the source) in the first place. In this work, we first explore whether impactful inspirational stimuli can be obtained from an unlikely source: a crowd workforce. In controlled research studies it is typical for researchers to hand-select or create specific stimuli for the design problem(s) under investigation. While this practice is effective for a controlled human subject experiment, it is not practically useful for designers in the field. If design-by-analogy and related processes are to widely be used in practice, the appropriate stimuli have to be systematically available for new problems at the time they are being solved.

The creation of new supportive design tools hinges on the ability of appropriate stimuli being provided to the designer at the right moment. Such a tool would require the automated distribution of a wide variety of inspirational stimuli. Initial work in this area related to analogy included the use of the U.S. patent database to map and identify near and far analogies (Fu, Chan, Cagan, et al., 2013), as well as semantic verb mapping (Linsey, Markman, & Wood, 2012). The work presented in this paper introduces a different way to identify inspirational stimuli from individuals with no problem solving or domain expertise. Leveraging the vast power of a crowdsourcing workforce, it is possible to gain access to a high volume of workers for directed tasks. Here, this workforce is utilized in order to generate inspirational stimuli relevant to a wide variety of open-ended design problems for which this process of generating inspirational stimuli would otherwise be difficult.

This paper has two main aims. The first of these is to investigate whether it is feasible to obtain inspirational stimuli from an untrained workforce using crowdsourcing. Leveraging the crowd to quickly gather useful inspirational stimuli for design ideation could provide an opportunity to open up the ideation process in new ways. Sourcing inspirational stimuli from crowds enhances individual designer's natural capabilities by providing them with information at a scale that would likely be otherwise unobtainable. The second aim of this work is to test the effectiveness and impact of these crowdsourced inspirational stimuli during design concept generation using a human subject experiment of trained designers. This experiment will add to the design research literature by testing various distances of inspirational stimuli across multiple design problems. Incorporating multiple design problems into the cognitive study will allow for the determination of features of ideating with and without inspirational stimuli that are consistent and repeatable across multiple problem specifications and domains.

1 Background

This section briefly describes background literature relevant to this work at the intersection of crowdsourcing, analogical reasoning, and design research. First, crowdsourcing will be discussed, primarily by including an examination of how crowdsourcing techniques have been applied by design researchers. Next, analogical reasoning in design research will be examined as research related to targeted inspirational stimuli is typically referenced using this terminology in the design research literature. One feature of this work is the exploration of inspirational stimuli at varying distances. The distance of the inspirational stimuli refers to the proximity of each stimulus to the design problem being solved. As such, particular emphasis within this section is placed on prior research regarding what is termed analogical distance in the design research literature.

1.1 Crowdsourcing in design research

Crowdsourcing is a model in which a distributed network of individuals responds to an open call for proposals or work (Brabham, 2008; Howe, 2006). There are examples of specialized crowdsourcing design platforms, such as OpenIDEO (Lakhani, Fayard, Levina, & Pokrywa, 2012) and Local Motors (Norton & Dann, 2011), which offer domain experts the opportunity to create and collaborate with others. However, these platforms seek to use the crowd workforce for both the creation and validation of concepts. Using these services, crowd workers not only generate new solutions themselves, but also vote on designs or ideas that they think are best. Workers on these platforms tend to possess a relatively high level of experience and/or interest in the problems being presented. One reason for this is that challenges posted on OpenIDEO are typically sponsored either directly by IDEO or an external agency, and often offer a cash prize for winning submissions. Additionally, these problems are often very detailed; for example, Local Motors used crowdsourcing to help develop novel car designs. Due to the complexity and depth of the majority of posted problems, a high level of expertise is required to attempt a solution or make a contribution.

These platforms, while practical for rapid innovation in industry, have not been utilized widely in academic research. Primarily, the design research community has used crowdsourcing for evaluative and rating purposes (Fuge & Agogino, 2015; Kittur, Chi, & Suh, 2008; Kudrowitz & Wallace, 2013). For example, Kudrowitz and Wallace (2013) used a crowdsourced population to rate design concepts on a number of subjective measures, including creativity, novelty, clarity, and usefulness. However, one of the motivations for crowdsourcing inspirational stimuli in this work is to bring crowd workers into the innovation process itself. A prior study by Yu et al., asked crowd participants to search for analogies for design problems on an online repository of examples by providing them directly with and without search strategies (Yu, Kittur, & Kraut, 2014). The study by Yu et al. (2014) engaged crowd workers to participate in the process of finding inspirational stimuli by specifically guiding them through the search process and training participants regarding how to covert problem descriptions into a more abstract form useful for analogical transfer. More recently Yu, Kraut, and Kittur (2016) demonstrated that the ability of crowd workers to obtain useful far analogical stimuli can be further improved by abstracting the contextual information regarding the problem, but restricting the domain specific descriptors of key problem constraints (Yu et al., 2016).

Rather than have crowd workers directly search for analogical stimuli, in the work presented here, inspirational stimuli are instead gathered directly from work completed by crowdsourced participants. Perhaps another comparative approach would be prior work based purely on computational methods that search for analogies. Past approaches at solving this problem have been word-embedding models such as GloVe (Pennington, Socher, & Manning, 2014). In addition, work by Gilon et al. contributed a search engine for finding distant analogies for specific aspects of a product or design (Gilon et al., 2017). One exploratory aspect of the work in the current paper is to test whether a large-scale human effort can be more effective than purely computational approaches.

One of the reasons for the limited number of crowdsourcing applications in the design research literature is that it is difficult to identify individuals with expertise and domain specific knowledge within crowd-based communities. In situations where it has been possible to identify members of the crowd with domain specific knowledge, these members have generally been unable to provide consistent and accurate responses (Burnap et al., 2015; Ulu, Messersmith, Goucher-Lambert, Cagan, & Kara, 2019). Based on this limited sample size, it would appear that the reason for the sparing use of crowdsourcing in academic research (at least for creative tasks) is due to the difficulty in finding skilled workers. In this work, we attempt to leverage an open crowd-based workforce using an online crowdsourcing labor market (Amazon Mechanical Turk – MTurk) for a creative task in which each individual participating is asked to come up with a solution/idea to an open-ended design problem. Previous work has demonstrated that online crowdsourcing services, such as MTurk, provide a participant population pool that is representative of the United States population (Paolacci, Chandler, & Ipeirotis, 2010). For the purposes of this experiment, the MTurk populations are appropriate because it is assumed that crowd participants have no level of domain expertise. In fact, it does not particularly matter whether or not the crowdsourced population for this study is representative of a US population. In the case of this experiment, it is not a requirement that the concepts generated by the crowd participants be of high quality. This is due to the fact that the primary output from this analysis is to text-mine the responses from each crowd participant to extract inspirational stimuli (words) for designers.

For the purposes of testing different types of inspirational stimuli, the extracted stimuli are first separated based upon word frequency, with high frequency words approximating “near” stimuli and infrequently used words approximating “far” stimuli. However, while the frequency-based approach is easy to implement and highly scalable, it is possible that the approach is more closely representative of commonness rather than distance. As is discussed in Section 2.3, a more time-consuming computational approach was also used to re-categorize the stimuli based upon semantic distances and to test whether the or not the frequency-based approximation was an effective surrogate. Due to the inherent direction provided by a design problem (e.g., implicit and explicit problem constraints), it is believed that a high volume of crowd responses will lead to a diverse set of words, which holds enough contextual information be relevant to the design problem. A cognitive study is then used to test whether or not this information can be effective in inspiring designers during concept development.

1.2 Analogical reasoning and distance in design research

Analogical reasoning is the process by which information from a source is applied to a target through the connection of relationships or representations between the two (source and target) (Gentner, 1983; Moreno et al., 2014). From a broad perspective, design researchers examine the processes of analogical reasoning because prior examples (anecdotal, as well as academic research) have demonstrated that analogies can help promote the generation of additional solutions, or solutions with positive characteristics (e.g., novelty) (Bashir, 2001; Chan et al., 2011; Dorst & Royakkers, 2006; Fu, Chan, Cagan, et al., 2013; Fu, Moreno, Yang, & Wood, 2014; Goucher-Lambert et al., 2018; Linsey & Viswanathan, 2014; Linsey et al., 2008; Moreno et al., 2014; Murphy et al., 2014; Tseng, Cagan, & Kotovsky, 2012). A typical study on analogical reasoning in the design research community involves a participant set engaged in an open-ended design task. Participants work on the given design problem and, at some point, are presented with analogical stimuli. These stimuli can vary greatly in terms of similarity to the problem, domain, and modality. At the end of the study, participants' design output is scored or rated. Typically, this is done by domain experts, who evaluate the generated concepts on metrics such as quality, novelty, and fluency (Shah, Kulkarni, & Vargas-Hernandez, 2000; Shah, Smith, & Vargas-Hernandez, 2003).

One of the key factors influencing the process of retrieving relevant information from a source, and then applying useful connections to a target, is the analogical distance between the two. Primarily, research on analogical distance uses the terms "near" and "far" to describe the distance of the analogy from the problem being examined (Fu et al., 2013). The continuum of distance refers to the domain distance; a "near" analogy generally implies that the analogy comes from the same or closely related domain, whereas a "far" analogy comes from a distant domain. It has also been noted that near-field analogies share significant surface level (object) features with the target and far-field analogies share little or no surface features (Linsey et al., 2012). For example, when trying to design a device to reduce home energy use, a design team could take inspiration from smart thermostats, which learn and adjust heating and cooling schedules to match behavior and save energy (near analogy). Another approach could take inspiration from nature, where grazing animals sync their foraging cycles to plant growth cycles (far analogy). The authors agree with this perspective, in which inspirational stimuli (e.g., analogies) fall on a continuum of distance.

One open area of investigation relates to determining the analogical distance most likely to yield positive solution characteristics for a given problem. Several studies support the idea that more distant analogies positively impact ideation (Ward, 1998; Wilson, Rosen, Nelson, & Yen, 2010). A review regarding the use of analogies in industry found that far-field analogies are more beneficial in helping to create more novel solutions (Kalogerakis, Lüthje, & Herstatt, 2010). However, some empirical evidence disputes this (Chan, Dow, & Schunn, 2015). Fu et al. (2013) proposed that there exists a "sweet spot" of analogical distance that rests between an analogy being too near (where innovation is restricted, and fixation and copying are likely to occur) and too far (where the connections between the analogy and the problem are unable to be made). This work further contributes to this discussion by examining the differences in solution characteristics that are observed when the distance of the inspirational stimuli is varied. By classifying the crowd-sourced data into distance categories (e.g., near vs. far), any difference in impact based on the distance of the inspirational stimuli can be assessed. Using the crowd-generated inspirational stimuli, we examine their effect on several solution characteristics (e.g., novelty (aka. uniqueness), feasibility, and usefulness) for concepts developed for multiple design problems.

2 Methodology

The main aims of this paper are to test 1) whether it is feasible to obtain inspirational stimuli from an untrained workforce using crowdsourcing, and 2) the effectiveness and impact of varying distances of crowdsourced inspirational stimuli during design concept generation using a human subject experiment. To accomplish this, a four-step methodological approach was used (Figure 1). First, twelve open-ended design problems were identified from the design research literature. Next, these twelve problems were posted online on Amazon Mechanical Turk (MTurk) in an open call for crowd responses. With over 1300 responses obtained between the twelve problems, the textual data was examined using a natural language processing toolkit. Based upon word frequency, a variety of words were extracted as inspirational stimuli for a human subject study performed using a subset (four) of the original twelve design problems. Three experimental conditions were explored, each of which varied the distance of the inspirational stimuli from the problem statement. Results were analyzed to determine the impact of the inspirational stimuli on the feasibility, usefulness, quality, and novelty of solutions generated by the human subject participants.

2.1 Selecting design problems

Through a review of the design research literature, twelve design problems used in prior research investigations were chosen subjectively to include in the study. With the knowledge that these problems would be used within a crowdsourcing environment, some of them were modified such that design constraints were removed from the original problem statement. This was done primarily to limit the required time to provide a single idea for the problem to a few minutes, and to allow the crowd population (with no design domain expertise) to successfully provide a relevant idea. A diversity of problem domains was also sought in selecting the design problems. The adapted versions of each design problem used within the current study and relevant references are shown in Table 1. The modified forms of the problems were limited to a single sentence. Problem 13 (listed as “NA” in Table 1) was developed uniquely for use as training during the experiment. The results from this problem were not analyzed.

2.2 Crowdsourcing design solutions

The design problems shown in Table 1 were posted on MTurk, an online crowdsourcing labor market. Each problem was posted as a separate Human Intelligence Task (HIT), where the requesters (in this case, the authors) sought a minimum of 100 responses from workers for each design problem. In total, 1345 responses were made for the HITs. There were 45 rejected submissions due to workers not submitting fully completed assignments or exceeding the allotted time (20 min). The 97% acceptance rate for the HITs as a part of this work is in line with other MTurk submissions, as workers in the crowd-

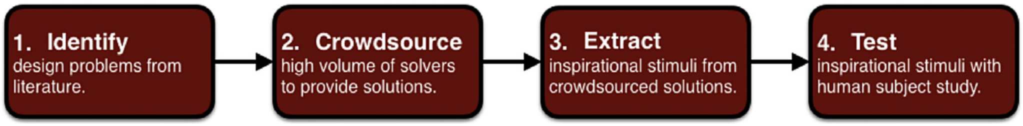

Figure 1 Methodological outline of experiment

Table 1 Design problems selected from literature for crowdsourcing experiment

Problem	*Reference*
1. A lightweight exercise device that can be used while traveling.	Linsey and Viswanathan (2014)
2. A device that can collect energy from human motion.	Fu, Chan, Cagan, et al. (2013)
3. A new way to measure the passage of time.	Tseng, Moss, Cagan, and Kotovsky (2008)
4. A device that disperses a light coating of a powdered substance over a surface.	Linsey et al. (2008)
5. A device that allows people to get a book that is out of reach.	Cardoso and Badke-Schaub (2011)
6. An innovative product to froth milk.	Toh and Miller (2014)
7. A way to minimize accidents from people walking and texting on a cell phone.	Miller, Bailey, and Kirlik (2014)
8. A device to fold washcloths, hand towels, and small bath towels.	Linsey et al. (2012)
9. A way to make drinking fountains accessible for all people.	Goldschmidt and Smolkov (2006)
10. A measuring cup for the blind.	(Jansson & Smith, 1991; Purcell, Williams, Gero, & Colbron, 1993)
11. A device to immobilize a human joint.	Wilson et al. (2010)
12. A device to remove the shell from a peanut in areas with no electricity.	Viswanathan and Linsey (2013)
NA. A device that can help a home conserve energy.	N/A

based community desire a high approval rating to garner more HIT opportunities. Workers responded to each HIT in return for $0.20 and no demographic information was sought through the collection of data. The only requirement placed through MTurk was that all workers were required to be U.S. citizens and at least 18 years of age.

For each HIT, workers were asked to provide an idea (solution) for a new product or device that addressed the given prompt. The instructions (Figure 2) for the HIT asked that the provided idea be something that workers believed did not currently exist. Workers were also instructed that they should not be concerned how, or if, what they were thinking of would be made. Once workers thought of an idea, they were asked to use as many words as necessary to describe it by writing into a free response text box. Next, participants were asked to provide up to six keywords (three nouns, three verbs) to serve as identifiers for the idea that they had entered into the free response box. Initial

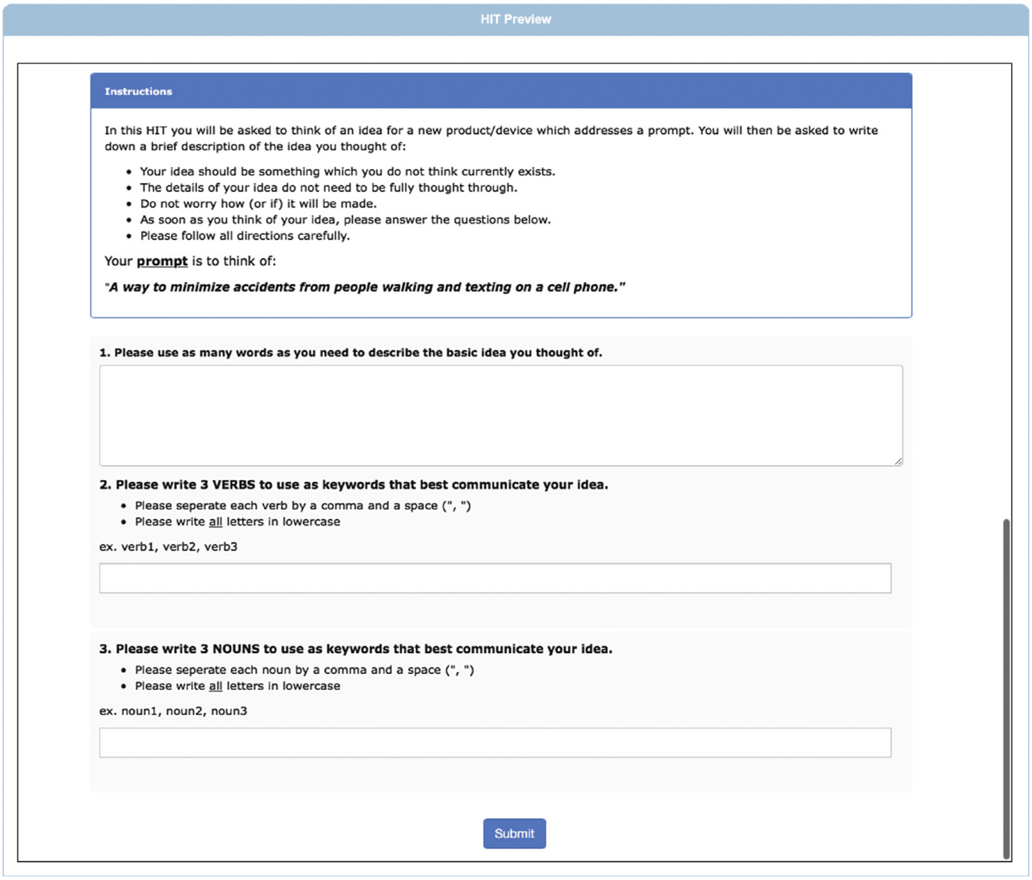

Figure 2 Amazon mechanical turk task example

analysis of pilot data indicated that participants were more likely to provide accurate keywords if they could be related to a specific design concept that the participant had already generated.

2.3 *Extracting and categorizing inspirational stimuli*

The three noun and three verb keywords provided with each HIT response from the MTurk task were used as the basis to obtain inspirational design stimuli at varying distances. In this work, the categorization of inspirational stimuli into different groupings (corresponding to distance from the problem space) was done in two ways: 1) based on word frequency and 2) using a computational approach based on path-length semantic distance.

The frequency approach simply used the word frequency within the crowd-sourced dataset to categorize the stimuli. This is based on the assumption that word frequency is a sufficient means to assess the relative distance of inspirational stimuli from the problem, while also providing a mechanism to gather stimuli that is straightforward to implement compared to computational approaches (also explored in this work and discussed below). Commonly used words within the response set were taken as near inspirational stimuli, and infrequently used words were taken as far inspirational stimuli. Due to the fact that word frequency provided a continuous distribution of words, a "medium" distance field set was also extracted from the crowdsourced responses. To accomplish this, the raw text from MTurk HIT responses was first collated together for each design problem. Using Python's Natural Language Processing Toolkit, individual word tokens were extracted from the raw text (Bird & Loper, 2004). The word token set was cleaned by removing stop words (e.g., "the", "is", "that", etc.), words that appeared in the problem statement (e.g., "reach" from Problem 5, "A device that allows people to get a book that is out of reach"), and by aggregating multiple tenses of words (e.g., "reach", "reaching", etc.). Following this, the new cleaned token sets were used to create a frequency distribution of words. Using the word frequency distribution, the crowd-generated word set was partitioned into three zones of distance: near, medium, and far. The top 25% most frequently used words became the "near" word set. Words that were only used once by the crowd respondents became the "far" word set. The "medium" word set were any entries that fell between these two ranges. Figure 3 gives an illustration of these three word set distance zones. Sample word extractions from each zone are shown in the Results section (Table 3).

The second categorization of the inspirational stimuli was done computationally, using a semantic measure of similarity defined using the scoring function in Equation (1) (Mihalcea, Corley, & Strapparava, 2006),

$$sim(T_1, T_2) = \frac{1}{(|T_1| + |T_2|)} \left(\sum_{w \in \{T_1\}} maxSim(w, T_2) + \sum_{w \in \{T_2\}} maxSim(w, T_1) \right). \tag{1}$$

This scoring function draws upon the WordNet library to define the word-to-word similarity between two different sets of words based on the maximum path similarity (Fellbaum, 1998). Here, the collection of words (T_1) defining each frequency-based category (near, medium, far, control) for a given problem was compared independently to the problem statement (T_2). For each word (w) in set T_1, the maximum similarity word was found in set T_2. Due to the fact that path similarity is not always symmetric, the inverse of this relationship was found and the mean of the two values was taken. Higher values (maximum of 1) indicate more similarity between the problem statement and the stimuli set.

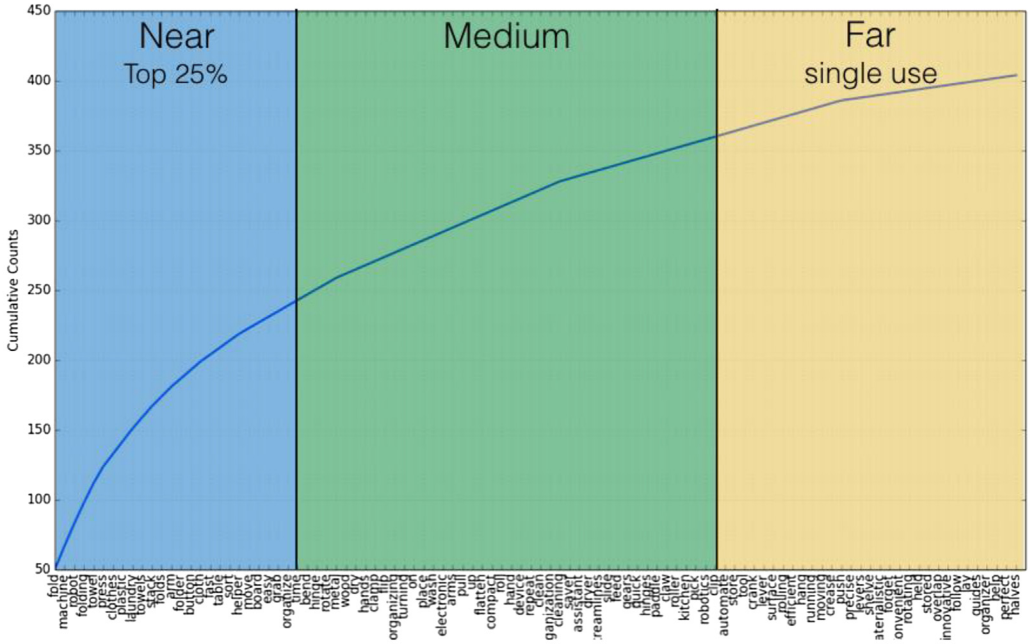

Figure 3 Illustrative frequency distribution from crowdsourced design problem showing near, medium, and far inspirational stimuli word pools

3 Exploring crowdsourced inspirational stimuli at varying distances using a human subject cognitive study

To test the impact of the crowdsourced inspirational stimuli across the three frequency-derived categorizations (near, medium, far) a human subject experiment was designed. Here, each of the three conditions was explored using a sampling of the crowd-generated inspirational stimuli for a subset of the original problems.

3.1 Participants

Participants for the cognitive study were recruited from junior, senior, and graduate level design and innovation courses at a major U.S university and offered course credit or $10 compensation for their participation. 95 participants were recruited from junior and senior level mechanical engineering design courses. An additional 16 participants were recruited from a

Table 2 Cognitive study group conditions

Problem	*Group A (N = 28)*	*Group B (N = 28)*	*Group C (N = 29)*	*Group D (N = 26)*
4	Near	Medium	Far	Control
7	Medium	Far	Control	Near
11	Far	Control	Near	Medium
12	Control	Near	Medium	Far

Table 3 Extracted inspirational stimuli, solution time, and lexical diversity of solutions from crowd-sourced concept generation experiment

Problem	*Avg. Time (s)*	*Lexical diversity*	*Near Words*	*Middle Words*	*Far Words*
1	239	0.537	pull, push, band, resist, bar	pedal, force, fill, fold, spring	roll, tie, sphere, exert, convert
2	216	0.543	store, charge, shoe, pedal, step	spin, transfer, spring, windmill, bracelet	beam, shake, attach, electrons, compress
3	262	0.649	light, sand, count, fill, decay	dial, grow, rotate, pass, slide	crystal, drip, pour, radioactive, gravity
4	199	0.539	spray, blow, fan, shake, squeeze	handle, puff, mesh, reservoir, hand	rotor, wave, cone, pressure, atomizer
5	201	0.460	extend, clamp, pole, hook, reel	magnet, fly, telescope, lasso, clip	pulley, hover, sticky, voice, angle
6	207	0.514	spin, whisk, heat, shake, chemical	inject, agitate, pump, beat, mix	surface, pulse, gas, gasket, churn
7	207	0.530	alert, flash, camera, sensor, motion	smart, beep, notify, background, recognize	emit, react, engage, lens, reflection
8	228	0.513	robot, press, stack, table, rotate	repeat, roll, turn, flip, clamp, flatten	deposit, cycle, rod, funnel, drain
9	213	0.469	adjust, lift, step, hose, nozzle	flexible, spigot, swivel, pull, extend	shrink, catch, attach, hydraulic, telescopic
10	220	0.417	braille, touch, beep, sound, sensor	weigh, heat, vibrate, lever, light	program, recognize, pressure, holes, cover
11	179	0.636	clamp, lock, cast, harden, apply	spray, strap, slide, magnet, inflate	shrink, inhale, fabric, condense, pressure
12	194	0.501	crack, crank, blade, squeeze, conveyor	pry, spin, mill, fall, drop	melt, circular, wedge, chute, wrap

multidisciplinary graduate course focusing on design innovation. There were 67 male and 44 female participants ranging in age from 19 to 26 (mean = 21.4). As discussed below, data from 15 participants was used for training the expert raters, and therefore was not included in the final analyses.

3.2 Experiment overview

Four conditions (three experimental, one control) were explored using the crowdsourced inspirational stimuli. These conditions varied the distance of the inspirational stimuli from the problem, defined in the cognitive study portion of this experiment, as the word frequency from the text-mined crowd-sourced data for that problem. Inspirational stimuli for the near, medium, and far conditions were extracted using the methods outlined in Section 2.3. Each of the four conditions was assigned four random words from within the available word set (approximately 600 words classified into three sets per problem). The words that were assigned to each problem−condition pair were fixed for the human-subject experiment. This was done in order to minimize possible sources of noise between subjects when analyzing the results. The control condition displayed four words verbatim from the problem statement. It was assumed that the control condition would provide no additional sources of inspiration for participants.

3.3 Experimental conditions

The cognitive study involved an approximately 1-hour session during which participants were asked to develop concept solutions to open ended design problems. Participants were told that they might receive a set of words during the problem that were intended to serve as inspiration for their concepts. Each participant saw the same four of the original twelve design problems (4, 7, 11, and 12 from Table 1) used in the crowdsourcing experiment. These four design problems were selected for the cognitive study due to high lexical diversity in their solution word set from the crowdsourced data and low completion time. A full factorial experimental design evenly split the conditions for each problem across study groups (Table 2), such that a given participant only saw one of the four conditions (near, medium, far, or control) for a given design problem.

At the start of the experiment, participants were provided with envelopes containing four separately marked problem packets, each containing a separate design problem. Participants were given ten minutes to work on each design problem, divided into two working blocks. Participants began each problem by first spending two minutes working to provide a single solution, along with up to six descriptors (three nouns, and three verbs) for the design problem. This initial procedure was meant to mirror the crowdsourced data. However, the descriptors generated by the cognitive study participants were not analyzed or used to generate inspirational stimuli of any kind. One reason for having participants initiate each brainstorming session before receiving the crowd-sourced inspirational stimuli, was that prior research on analogical reasoning in design has shown that inspirational stimuli (e.g., analogies) are more effectively applied if an open goal has been established (Tseng et al., 2008). After this initial period, participants were provided the crowd-sourced inspirational stimuli specific to their condition. These stimuli consisted of four words extracted from the text-mined MTurk dataset. Eight minutes of open idea generation was given following the presentation of the stimuli, where participants placed each generated idea into individual designated boxes provided within the problem packet. Each idea was time stamped at completion by the participant using a clock displayed at the front of the room during the study. Participants were allowed to use any combination of sketching and writing to express their ideas and were instructed to provide sufficient detail such that someone viewing their ideas later could understand the basic concept. Following each problem, a short questionnaire was provided to gauge participant's perceived usefulness and relevancy of the presented inspirational stimuli, as well as the overall quality and novelty of the generated solutions.

3.4 Analysis of design output from cognitive study

The design output from the participants was examined in order to determine the impact of crowdsourced inspirational stimuli at varying distances on solution characteristics. The following characteristics of the solution outputs were explored:

1. **Feasibility**: rated on an anchored scale from 0 (the technology does not exist to create the solution) to 2 (the solution can be implemented in the manner suggested).
2. **Novelty**: rated on an anchored scale from 0 (the concept is copied from a common and/or pre-existing solution) to 2 (the solution is new and unique). Of note "novelty" is considered to be the uniqueness of the solution with respect to the entire solution set.
3. **Usefulness**: rated on an anchored scale from 0 (the solution does not address the prompt and/or take into account implicit problem constraints) to 2 (the solution is helpful beyond status quo).
4. **Quality**: rated subjectively by each rater on a scale from 0 (low) to 2 (high).

One mechanical engineering PhD candidate and one mechanical engineering postdoctoral researcher, both specializing in design theory and methodology, were trained to perform all ratings for solution characteristics. Consistency was assessed over a subsample of the data using the intraclass correlation coefficient (ICC). In addition to the metrics noted above, perceived ratings for novelty and quality of the solutions were collected from the participants, along with the perceived usefulness and relevancy of the provided inspirational stimuli. At the conclusion of each problem, participants provided a rating for each self-rated metric between 1 and 5. This was done in order to test whether or not the perception of the extracted levels (near, medium, and far) for stimuli distance aligned with their predicted categories. Additionally, self-ratings allowed for the determination of how participants' impressions of their own ideas compared to the design expert evaluations.

4 Results

4.1 Crowdsourced inspirational stimuli

Using the methods outlined in Section 2.3, inspirational stimuli were extracted from the crowdsourced dataset provided by 1345 respondents, using word frequency as a measure of distance. Five inspirational stimuli were randomly extracted for each distance measure and are shown in Table 3. Table 3 also shows the mean response time for the HIT and the lexical diversity of the solution for each set of stimuli. The mean completion time provided insight into the difficulty of the problems for the crowd community. Lexical diversity measures the ratio of unique word entries within the submissions. Both of these measures were used to select four problems from the available twelve to use in the cognitive study, with a low response time and high lexical diversity seen as positive problem characteristics. The design problems selected for the cognitive study were Problems 4, 7, 11, and 12.

4.2 Cognitive study results

111 participants generated 1651 concepts across the four design problems. Each solution was rated using the methods outlined in Section 3.4. In addition to the rated solution characteristics, participants also provided perceived rating values for the relevancy and usefulness of the presented inspirational stimuli, as well as the quality and novelty of their own solutions.

Both raters evaluated a randomly selected subset of solutions from 15 participants (166 designs total) across the sub-dimensions of interest (Usefulness, Feasibility, Novelty, Quality) and consistency was assessed using the intraclass correlation coefficient (ICC). A strong level of correlation was obtained for three of the four metrics: Usefulness (ICC >0.65), Novelty (ICC > 0.71), and Feasibility (ICC > 0.77). ICC for the Quality metric was low to acceptable at ICC >0.50. Generally inter-rater reliability levels for this study are within the range of values typically found in behavioral studies with human raters (Cicchetti, 1994), and are consistent with past work in this area (Chan et al., 2011; Daly, Christian, Yilmaz, Seifert, & Gonzalez, 2012; Fu, Chan, Schunn, Cagan, & Kotovsky, 2013). The remaining 1485 concepts from 96 participants were rated by one of the two evaluators and included in the remaining analyses for this paper.

The correlation between the various outcome measures of interest was tested prior to completing the full analyses using Rstudio with base R and the corr function. A correlation matrix showing these relationships is shown in Figure 4. Novelty has no correlation between any of the other outcome measures. The feasibility and usefulness measures share a weak positive correlation with each other. However, the quality measure is strongly correlated with usefulness and moderately correlated with feasibility (both positive). Due to the strong correlation between the quality measure and other outcome measures of interest, little new information can be obtained from analyzing the quality measure in isolation from the other outcome measures. Additionally, the quality measure had the lowest ICC value of any outcome measure. For these two reasons, the quality measure was subsequently dropped from the analyses.

Figure 5 shows example solutions produced by two different participants during the human subject experiment. As participants were allowed to express their ideas using any combination of textual and pictorial information, most included some combination of the two. The time during the problem solving block that a solution was generated is noted in the top right quadrant of

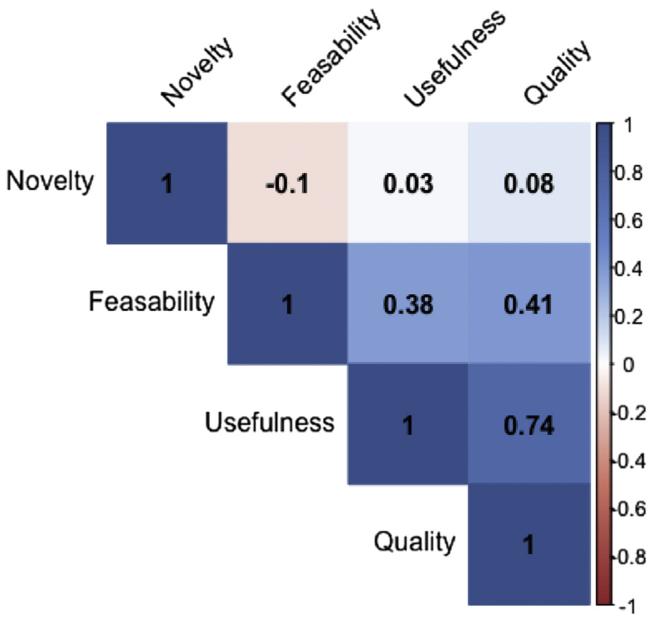

Figure 4 correlation matrix of relevant outcome measures

each solution. It should be pointed out that the solutions were not analyzed in order to understand the analogical transfer of concepts from the inspirational stimuli to the generated solutions.

4.2.1 Participant provided ratings and measures of stimuli distance

Participants provided four ratings following the presentation of each problem during the cognitive study. Two of these were gauged at assessing the inspirational stimuli that were presented for each design problem (usefulness and relevancy), and the other two sought to determine participants' subjective perception regarding the overall novelty and quality of the solutions they developed for that problem. Although quality has been removed from the expert rating analysis (Section 4.2.2) for reasons discussed previously (Section 4.2), the self-reported ratings are included here to test whether there was a perceptive difference in quality for participants. It should be noted that each participant only provided one rating for each of the metrics after each problem (even if they generated multiple solutions). Consequently, the provided ratings pertain to the entire set of solutions generated by each participant for each problem.

The results analyzing the participant self-rated data for the four questions previously discussed are shown in Figure 6. There was no significant difference between how participants rated the quality or novelty of their own solutions within the different conditions (Quality, $F(3,380) = 0.73$, $p = 0.53$; Novelty,

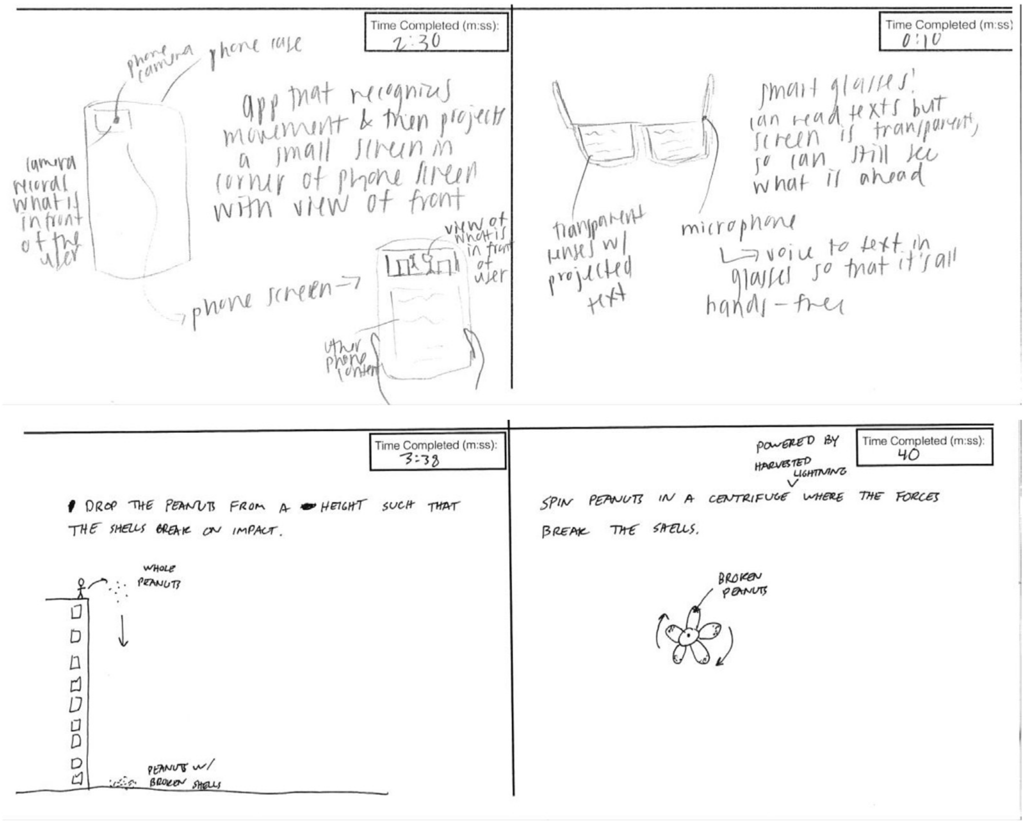

Figure 5 example solutions from cognitive study experiment

F(3,380) = 1.25, p = 0.29). While not statistically significant, the data suggests that participants may have perceived their solutions to be more novel as the distance of inspirational stimuli is increased. The largest (non-significant) difference was seen in the pairwise contrast between the control and far conditions, where the far condition led to more novel solutions (*F(2,285) = 3.04, p = 0.08)*. Additionally, there were no significant findings related to how participants perceived the quality of their solutions between the different conditions.

Study participants perceived less distant inspirational stimuli to be more useful than distant stimuli. A one-way ANOVA comparing the inspirational stimuli conditions (near, medium, far) was significant (*F(2,285) = 3.73, p = 0.03)*. As the control condition did not include inspirational stimuli, these questions were omitted from the rating form provided to participants. A post-hoc Tukey HSD (honest significant difference) test was used to conduct pairwise comparisons of individual conditions with significance values at a 95% confidence interval. These pairwise comparisons between the conditions revealed that only

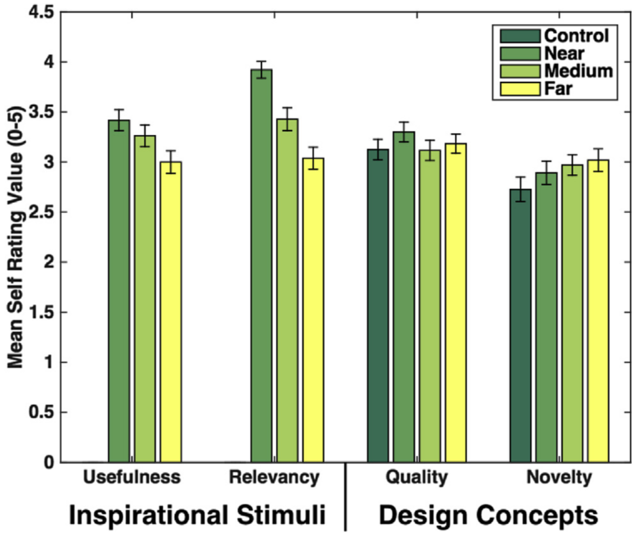

Figure 6 Participant provided ratings of inspirational stimuli and generated design concepts (+/−1 SE)

the contrast between the near and far conditions was significant (Near vs. Far: $p = 0.02$; Near vs. Medium: $p = 0.57$; Medium vs. Far: $p = 0.21$). As a result, it can be concluded that near inspirational stimuli are perceived as being more useful to designers than far stimuli.

One of the key assumptions of this work is that there exists a relationship between the frequency with which a word appears in a large set of (crowd-sourced) written solutions and the "distance" of this word when extracted and provided to a designer as inspirational stimuli. Words that appeared more frequently in the crowdsourced dataset were taken as near inspirational stimuli, and words that appeared less frequently were classed progressively further (i.e., medium and far). The validity of this hypothesis was tested using two methods: 1) explicit ratings of relatedness provided by human subject participants and 2) a computational approach based upon textual similarity (distance) using natural language processing.

Ratings of stimuli relevancy provided by participants in the cognitive study are also displayed in Figure 6. Participant ratings of the relevancy of the inspirational stimuli helps to provide further insight regarding whether or not the extracted inspirational stimuli appropriately aligned to the pre-determined categories (near, medium, far). Here, there was a clear trend in the perceived relevancy of the inspirational stimuli, where study participants perceived less distant inspirational stimuli to be more relevant to the design problem ($F(2,285) = 18.26, p < < 0.01$). Pairwise comparisons confirmed this was significant across all levels of the inspirational stimuli (Near vs. Medium: $p < 0.01$; Near vs. Far: $p < <0.01$; Medium vs. Far: $p = 0.02$).

Participants rating near inspirational stimuli as being more relevant to the design problem compared to far inspirational stimuli indicates that the condition groupings assigned based upon word frequency in the text-mined crowd responses were perceptible to the designers. However, it is also possible to test this relationship computationally without humans. As discussed in Section 2.3, the frequency-based categorizations were re-categorized based upon semantic distance.

The results from the computational distance analysis are shown in Table 4. As expected, similarity values for the control condition approach one, as the words for this set were extracted from the problem statement itself. Pairwise testing between the inspirational stimuli conditions revealed that there was a significant difference between the near and far experimental conditions ($p = 0.03$, $d = 0.86$). There was no significant difference between the medium condition and either the near ($p = 0.40$) or far ($p = 0.43$) conditions. In fact, the semantic distance for the medium condition only fell between the near and far conditions in 2 out of the 12 design problems, whereas the near and far conditions were correctly aligned in 8 out of 12 problems. This computational measure, based on semantic distance (WordNet path-length), supports the conclusion that word frequency is an acceptable mechanism to approximate the categorization of the stimuli into near and far distances. However, due to the highly variable and inconclusive results for the medium distance condition, it was decided that this condition would be excluded from analyses regarding the impact of the inspirational stimuli. Finally, when only considering the problems selected for the cognitive study (Problems 4, 7, 11, and 12), the mean computational distances were ordered correctly; however, no pairwise comparison between experimental conditions was significant due to the limited sample size (mean values: Near = 0.19, Medium = 0.16, Far = 0.13). This further highlights the benefit of comparing the results using both the frequency-based, as well as computationally-based categorizations.

4.2.2 The impact of inspirational stimuli on design solution outcome measures

In order to uncover the impact of inspirational stimuli of varying distances on measurable design outcome measures, cumulative link mixed models (CLMMs) were used. A CLMM is a type of ordinal regression model that allows for fixed and random effects. Here, CLMMs were used to examine the relationship between the expert evaluated outcome measures (Novelty, Usefulness, Feasibility) as a function of the Problem (levels: Problem 2 (Surface Coating), Problem 7 (Phone Accidents), Problem 11 (Joint Immobilization), Problem 12 (Peanut Sheller)) and Condition (Near and Far inspirational stimuli, and Control) being examined. A different model was constructed for each outcome measure and stimuli categorization method (word frequency and computational semantic distance) pair, creating a total of 6 separate models.

Table 4 Semantic distance of each condition word set from problem statement

Problem	*Near*	*Medium*	*Far*	*Control*
1	0.13	0.11	0.12	0.76
2	0.15	0.13	0.16	0.83
3	0.10	0.17	0.14	0.84
4	0.26	0.12	0.18	0.89
5	0.18	0.23	0.15	0.77
6	0.13	0.14	0.14	0.90
7	0.16	0.15	0.09	0.75
8	0.17	0.13	0.13	0.52
9	0.16	0.17	0.13	0.94
10	0.15	0.08	0.11	0.72
11	0.18	0.24	0.09	0.89
12	0.14	0.11	0.15	0.93
Mean	0.16	0.15	0.13	0.81
Std.	0.04	0.05	0.03	0.12

All analyses were conducted in Rstudio using base R and the CLMM (ordinal) package. For each model, participant was treated as a random variable. Parameter estimation was completed using a Gauss-Hermite approximation of the likelihood function with 10 quadrature points. The reference level selected for Condition was Control, and Problem 12 (Peanut Sheller) for the Problem variable. The ratings-based data provided by the experts for the various outcome measures were treated as ordered response categories, rather than continuous variables, due to the fact that prior research has identified that conventional linear regression and ANOVA models tend to produce over-confident results for variables with under approximately five values (Rhemtulla, Brosseau-Liard, & Savalei, 2012). Particularly, this is due to the fact that in the case of ratings-based data, the order of the categories is known, but the distance between them is not.

The results from the CLMMs are shown in Table 5 (Word Frequency) and Table 6 (Semantic Distance). Each table is broken into three sections pertaining to each separate model for each of the outcome measures of interest (Feasibility, Usefulness, and Novelty). The table includes values for each estimate (β_i, relative to Control), standard error of the estimate, and significance value, p. In addition, the odds ratio ($OR_i = e^{\beta_i}$) for each estimate value is included along with its 95% confidence interval. For brevity, the problem variable results are not included in each table due to the fact that the difficulty of each problem relative to the control problem was not of interest. However, it should be noted that in analyzing the data for all six models, there was significant variability in the problem variable. This implies that specific problems do lead to higher or lower estimates across the various outcome measures. Therefore, it was beneficial to capture this variation directly within each model. Since there were only four separate problems being analyzed it was feasible to add them

Table 5 Cumulative link mixed model results for cognitive study expert evaluations with data categorized based upon word frequency. Results are measured against reference of control condition

Outcome Measure	*Condition*	*Estimate*	*Std. Error*	p	*Odds Ratio*	*2.5% C.I.*	*97.5% C.I.*
Feasibility	Far	0.22	0.18	0.22	1.25	0.88	1.78
	Near	0.41	0.18	0.03*	1.51	1.05	2.16
Novelty	Far	0.31	0.14	0.03*	1.36	1.02	1.80
	Near	0.11	0.14	0.44	1.12	0.84	1.48
Usefulness	Far	−0.03	0.14	0.85	0.97	0.73	1.29
	Near	0.31	0.14	0.03*	1.36	1.03	1.80

into the CLMMs and obtain an estimate of this variability (rather than treat it as a random variable).

Maybe the most striking finding when examining the results from the CLMMs is that the conclusions produced by each categorization technique (word frequency vs. semantic distance) are the same. For the feasibility measure, near inspirational stimuli improve the feasibility of designs compared to the control condition (i.e., no stimuli) in both categorization techniques (word frequency: $p = 0.03$; semantic distance: $p < 0.01$). In addition to improving the feasibility of designs, near inspirational stimuli also improve the usefulness of designs compared to the control condition (word frequency: $p = 0.03$; semantic distance: $p = 0.01$). Finally, far inspirational stimuli improve the novelty (uniqueness) of design solutions compared to the control (word frequency: $p = 0.03$; semantic distance: $p = 0.03$). It should be noted that the semantic distance models led to more significant effects than the classification based solely on word frequency. Thus, while word frequency is a good approximation of distance within this dataset, semantic distance is a more accurate approach (however, at the expense of additional time and resources).

In terms of the total quantity of ideas, there was no significant difference found between the four conditions (Control: 369, Near: 375, Far: 362). Therefore, having inspirational stimuli did not increase the number of ideas that participants were likely to generate during the experiment. However, as discussed previously, results from the cumulative link mixed models demonstrate clear

Table 6 Cumulative link mixed model results for cognitive study expert evaluations with data categorized based upon computationally derived semantic distances. Results are measured against refernce of control condition

Outcome Measure	*Condition*	*Estimate*	*Std. Error*	p	*Odds Ratio*	*2.5% C.I.*	*97.5% C.I.*
Feasibility	Far	0.12	0.18	0.50	1.13	0.80	1.60
	Near	0.55	0.19	<0.01**	1.74	1.19	2.54
Novelty	Far	0.31	0.15	0.03*	1.37	1.03	1.82
	Near	0.10	0.15	0.52	1.10	0.82	1.47
Usefulness	Far	−0.10	0.15	0.51	0.91	0.68	1.21
	Near	0.39	0.15	0.01*	1.48	1.10	1.98

findings for each outcome measure across the different experimental conditions. Perhaps a more intuitive way to interpret these findings is through odds ratios (Figure 7). The odds ratio represents the likelihood of the estimate compared to the control, where odds = 1 demonstrates equal likelihood of the outcome measure being rated the same in that condition compared to the control. Figure 7 plots the odds ratio for each outcome measure using each categorization technique along with 95% confidence intervals around the estimate. Here, it is clearly visible that having a stimulus categorized as near has a significant overall impact on the likelihood that a design will be rated as more feasible and useful; in contrast, being provided with a far inspirational stimulus increases the likelihood that a design is rated as being more novel.

5 Discussion

This work uses crowdsourcing to obtain inspirational stimuli for future design problem solvers. By text-mining design solutions from crowd participants with no design expertise, commonly used words can be extracted and later serve as inspirational stimuli for new participants with design training. Here, more common words specify "near" inspirational and less commonly used words serve as "far" inspirational stimuli. A cognitive study tested these stimuli on participants with design domain expertise (all participants were students currently enrolled in undergraduate/graduate level engineering design courses), as they solved four open-ended design problems.

Results indicate that the methods employed in this work for crowdsourcing inspirational stimuli and using word frequency as a measure to approximate distance were successful. Extracted inspirational stimuli were categorized into separate bins representing varying levels of distance (near, medium, and far) from the problem space. Participants in the cognitive study were able to effectively judge these differences, as they rated more distant inspirational stimuli as having a lower level of relevancy for all of the design problems. More critically, a computational approach based on word path-length similarity demonstrated that the near and far word sets were significantly different from one another and are directionally appropriate (i.e., near word sets were more similar to the problem statement than far word sets). This is in line with previous research regarding analogical distance and participant perception of the relevance of analogies to the problem domain (Fu, Chan, Cagan, et al., 2013).

It is also important to note that participants rated more distant inspirational stimuli as being less useful than near stimuli. One possible theory is that this indicates participants were having difficulty connecting distant inspirational stimuli to the design problems. This was not apparent through the number of concept solutions generated in each condition. Based upon this measure, no experimental condition was significantly different. However, one

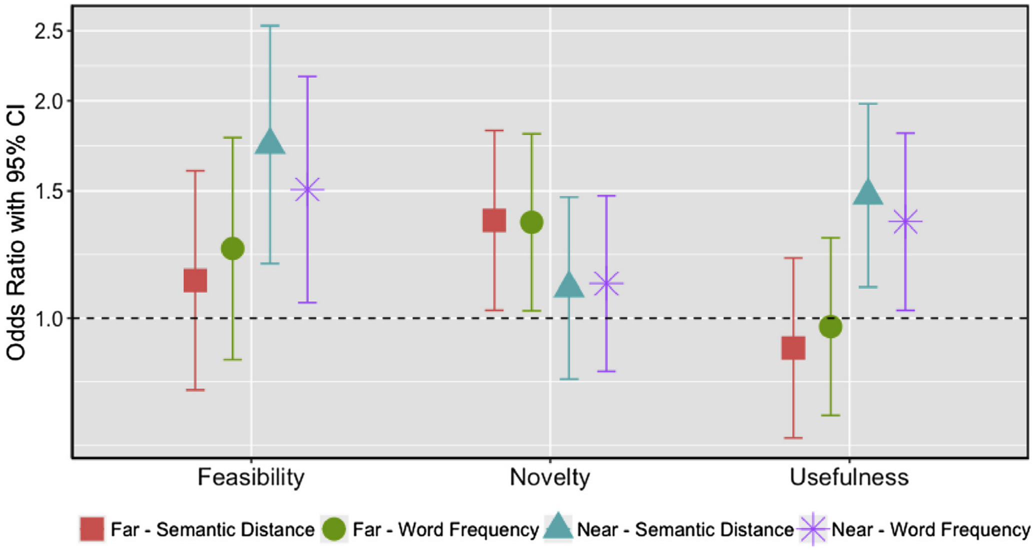

Figure 7 Odds ratio and 95% confidence intervals for each outcome measure on log scale. The dashed line represents even odds of outcome measure occurring compared to the control condition. Values above the line indicate that a higher value for the outcome measure is more likely compared to the control

potentially limiting factor in this work is the low amount of time that participants had to work on each design problem during the cognitive study (10 min). A short period was allotted for participants to develop an initial concept to the design problem before ideation. This time was intended to allow for open goals to be established for the problem (Tseng et al., 2008). Future work should consider in more depth the effect of allowing participants further time to develop open-goals prior to the presentation of inspirational stimuli. It is possible that, had the incubation time been longer, more distant stimuli may have become more impactful.

A separate goal of this work was to link the distance of inspirational stimuli to a variety of solution characteristics. The results, which analyze over 1000 solution concepts from 96 engineering design students, demonstrate that crowd-sourced inspirational stimuli significantly impact the novelty, feasibility, and usefulness of designs. Furthermore, the results were aligned regardless of whether the categorizations of the inspirational stimuli were based on word frequency or on semantic distance. Based on the results of this study, inspirational stimuli described as near increase the overall feasibility and usefulness of solutions compared to a control, while far inspirational stimuli increase their novelty. Quality was not included in the analysis (based on expert-ratings) due to a low inter-rater reliability evaluating this metric, as well as a high correlation between quality and the feasibility and usefulness metrics. While not based on any statistical testing, one can speculate that near field inspirational stimuli would have also produced higher quality designs based upon the correlation between quality, feasibility, and usefulness. Near field inspirational stimuli improving more design characteristics than far stimuli is in line with previous research that found less distant stimuli may actually be more beneficial for producing positive design outcomes (Chan et al., 2015; Gonçalves, Cardoso, & Badke-Schaub, 2013; Goucher-Lambert et al., 2018). While this work found far inspirational stimuli improved the novelty of design solutions compared to a control, it is difficult to accurately project whether increased novelty will necessarily lead to a more favorable design outcome. However, uniqueness (measured here as novelty) is generally considered a positive outcome measure due to the fact that a more diverse set of ideas increases the likelihood of a chosen solution being innovative (Terwiesch & Ulrich, 2009).

What type of inspirational stimuli is most impactful in assisting design ideation? Based on the results of this work, it would appear that near stimuli are better. The four design problems included in the cognitive study came from various domains. Yet, in order to maintain approximately the same level of complexity between the different selected design problems, many of the additional constraints associated with the problems' original versions were removed. While these results demonstrate there was variability in the difficulty across problems, the relative positive or negative impact of the inspirational stimuli across the outcome measures of interest were consistent. Even though near inspirational stimuli were more helpful, it is possible that the stimuli, although described as near, might have occupied a space closer to the "sweet spot" proposed by Fu et al. (2013). In other words, it is possible that even the inspirational stimuli described as "near" might not have been particularly near because they originated from a large database of crowdsourced solutions. Additional work is needed to develop and test theories on specific problem properties that are better suited for a specific stimuli distance, as well as how to accurately distinguish near vs. far stimuli. Doing so will allow for the determination of a specific sweet spot of inspirational stimuli distance that is required for a given design problem.

This work demonstrated that crowdsourcing could be an effective means to generate inspirational stimuli. One of the main benefits of this approach is that it allows for the collection of a large, diverse, and continuous set of inspirational for a given problem. Furthermore, utilizing a crowd workforce, this can be accomplished quickly and effectively, as demonstrated in this work. Future work should compare undirected methods of obtaining inspirational stimuli from crowdworkers (where workers are not explicitly guided through the process of searching for stimuli) to directed methods (e.g., Yu et al., 2014) and computational approaches (e.g., Pennington et al., 2014).

Inspirational stimuli were limited to text-based responses in order to improve the consistency of extracting inspirational stimuli at specific distances from the crowd. One area for future investigation could be to include diverse inspiration modalities (e.g., images, virtual models, etc.). Prior work has demonstrated that the modality of the stimulus can impact inspirational (analogical) transfer (Linsey et al., 2008). Additionally, future research should investigate the robustness of the results from the cognitive study. Here, only four inspirational stimuli were selected for each problem–condition pair. It is possible that these stimuli represent either poor or excellent stimuli from the available set. However, the consistency of the findings over a broad variety of problems and domains is promising for future investigations.

6 Conclusion

This work examined whether it is feasible to obtain inspirational stimuli using crowdsourcing techniques and how these sourced stimuli impact solution characteristics of design concepts generated by participants in a cognitive study. Results indicate that it is possible to obtain inspirational stimuli effectively using an untrained crowd workforce. Furthermore, the inspirational stimuli from crowdsourced design solutions are able to translate onto a continuous space of distance based on word frequency. Categorizations based upon computationally derived semantic distances significantly aligned with word frequency defined categorizations, further confirming that the frequency-based approximation was an effective surrogate. When testing the impact of distance of the crowdsourced inspirational stimuli (using both word frequency and semantic distance categorizations) on solution characteristics, results indicate a significant difference between multiple conditions. Using both categorization techniques, inspirational stimuli described as near improve the overall feasibility and usefulness of design concepts, while far inspirational stimuli improve the novelty (uniqueness) of designs. While additional work is needed to fully understand how designers will benefit from having specific types of inspirational stimuli, this paper demonstrates that the crowd can be the source of those stimuli.

Acknowledgments

The authors would like to thank Dr. Chris McComb and Josh Gyory for their assistance evaluating design concepts, as well as Jamie Amemiya, Leanne Elliot, and Naveen Shankar for their insightful discussions regarding analysis techniques. Additionally, the authors would like to thank the reviewers of this manuscript for their helpful comments and insight. This material is based upon work supported by the National Science Foundation Graduate Research Fellowship, and the Carnegie Mellon University Bradford and Diane Smith Fellowship. The authors would also like to thank the AFOSR for funding this research through grant FA9550-16-1-0049. A previous version of this paper was submitted to the International Conference on Engineering Design: Goucher-Lambert, K. and Cagan, J., (2017). Using crowdsourcing to provide

analogies for designer ideation in a cognitve study. *International Conference on Engineering Design 2017,* Vancouver, Canada.

References

Bashir, H. (2001). An analogy-based model for estimating design effort. *Design Studies, 22*(2), 157−167. https://doi.org/10.1016/S0142-694X(00)00015-6.

Bird, S., & Loper, E. (2004). NLTK: The natural language toolkit. In *Proceedings of the 42nd annual meeting of the association for computational linguistics (pp. 1−4)*. https://doi.org/10.3115/1118108.1118117.

Brabham, D. C. (2008). Crowdsourcing as a model for problem solving: An introduction and cases. *Convergence The International Journal of Research Into New Media Technologies, 14*(1), 75−90. https://doi.org/10.1177/1354856507084420.

Burnap, A., Ren, Y., Gerth, R., Papazoglou, G., Gonzalez, R., & Papalambros, P. Y. (2015). When crowdsourcing fails: A study of expertise on crowdsourced design evaluation. *Journal of Mechanical Design, 137*(3), 031101. https://doi.org/10.1115/1.4029065.

Cardoso, C., & Badke-Schaub, P. (2011). The influence of different pictorial representations during idea generation. *Journal of Creative Behavior, 45*(2), 130−146. https://doi.org/10.1002/j.2162-6057.2011.tb01092.x.

Casakin, H., & Goldschmidt, G. (1999). Expertise and the use of visual analogy: Implications for design education. *Design Studies, 20*(2), 153−175. https://doi.org/10.1016/S0142-694X(98)00032-5.

Chan, J., Dow, S. P., & Schunn, C. D. (2015). Do the best design ideas (really) come from conceptually distant sources of inspiration? *Design Studies, 36*(C), 31−58. https://doi.org/10.1016/j.destud.2014.08.001.

Chan, J., Fu, K., Schunn, C., Cagan, J., Wood, K., & Kotovsky, K. (2011). On the benefits and pitfalls of analogies for innovative design: Ideation performance based on analogical distance, commonness, and modality of examples. *Journal of Mechanical Design, 133*(8), 081004. https://doi.org/10.1115/1.4004396.

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment, 6*(4), 284−290. https://doi.org/10.1037/1040-3590.6.4.284.

Daly, S., Christian, J. L., Yilmaz, S., Seifert, C. M., & Gonzalez, R. (2012). Assessing design heuristics for idea generation in an introductory engineering course. *International Journal of Engineering Education, 28*(2), 1−11, Retrieved from. http://www.researchgate.net/publication/259104145_Assessing_design_-heuristics_for_idea_generation_in_an_introductory_engineering_course/file/3deec529fa6af1c6b4.pdf.

Dorst, K., & Royakkers, L. (2006). The design analogy: A model for moral problem solving. *Design Studies, 27*(6), 633−656. https://doi.org/10.1016/j.destud.2006.05.002.

Fellbaum, C. (1998)*WordNet: An electronic lexical database, Vol. 71*. Cambridge, London, England: MIT Press. https://doi.org/10.1139/h11-025.

Forbus, K. D., Gentner, D., & Law, K. (1995). MAC/FAC: A model of similarity-based retrieval. *Cognitive Science, 19*(2), 141−205. https://doi.org/10.1016/0364-0213(95)90016-0.

Fu, K., Cagan, J., Kotovsky, K., & Wood, K. (2013a). Discovering structure in design databases through functional and surface based mapping. *Journal of Mechanical Design, 135*(3), 31006−31013, Retrieved from. https://doi.org/10.1115/1.4023484.

Fu, K., Chan, J., Cagan, J., Kotovsky, K., Schunn, C., & Wood, K. (2013b). The meaning of "near" and "far": The impact of structuring design databases and

the effect of distance of analogy on design output. *Journal of Mechanical Design, 135*(2), 021007. https://doi.org/10.1115/1.4023158.
Fu, K., Chan, J., Schunn, C., Cagan, J., & Kotovsky, K. (2013c). Expert representation of design repository space: A comparison to and validation of algorithmic output. *Design Studies, 34*(6), 729−762. https://doi.org/10.1016/j.destud.2013.06.002.
Fu, K., Moreno, D., Yang, M., & Wood, K. L. (2014). Bio-inspired design: An overview investigating open questions from the broader field of design-by-analogy. *Journal of Mechanical Design, 136*(11), 111102. https://doi.org/10.1115/1.4028289.
Fuge, M., & Agogino, A. (2015). Pattern analysis of IDEO's human-centered design methods in developing regions. *Journal of Mechanical Design, 137*(7), 71405−71410, Retrieved from. https://doi.org/10.1115/1.4030047.
Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. *Cognitive Science, 7*(2), 155−170. https://doi.org/10.1016/S0364-0213(83)80009-3.
Gilon, K., Ng, F. Y., Chan, J., Assaf, H. L., Kittur, A., & Shahaf, D. (2017). *Analogy mining for specific design needs*. https://doi.org/10.1145/3173574.3173695.
Goldschmidt, G., & Smolkov, M. (2006). Variances in the impact of visual stimuli on design problem solving performance. *Design Studies, 27*(5), 549−569. https://doi.org/10.1016/j.destud.2006.01.002.
Gonçalves, M., Cardoso, C., & Badke-Schaub, P. (2013). Inspiration peak: Exploring the semantic distance between design problem and textual inspirational stimuli. *International Journal of Design Creativity and Innovation, 1*(4), 215−232. https://doi.org/10.1080/21650349.2013.799309.
Goucher-Lambert, K., Moss, J., & Cagan, J. (2018). A neuroimaging investigation of design ideation with and without inspirational stimuli—understanding the meaning of near and far stimuli. *Design Studies*. https://doi.org/10.1016/j.destud.2018.07.001.
Howe, J. (2006). *The rise of crowdsourcing by Jeff Howe | Byliner*.
Jansson, D. G., & Smith, S. M. (1991). Design fixation. *Design Studies, 12*(1), 3−11. https://doi.org/10.1016/0142-694X(91)90003-F.
Kalogerakis, K., Lüthje, C., & Herstatt, C. (2010). Developing innovations based on analogies: Experience from design and engineering consultants. *Journal of Product Innovation Management, 27*(3), 418−436. https://doi.org/10.1111/j.1540-5885.2010.00725.x.
Kittur, A., Chi, E. H., & Suh, B. (2008). Crowdsourcing user studies with Mechanical Turk. In *Proceeding of the twenty-sixth annual CHI conference on human factors in computing systems, Vol. 453*. https://doi.org/10.1145/1357054.1357127.
Krawczyk, D. C., McClelland, M. M., Donovan, C. M., Tillman, G. D., & Maguire, M. J. (2010). An fMRI investigation of cognitive stages in reasoning by analogy. *Brain Research, 1342*, 63−73. https://doi.org/10.1016/j.brainres.2010.04.039.
Kudrowitz, B. M., & Wallace, D. (2013). Assessing the quality of ideas from prolific, early-stage product ideation. *Journal of Engineering Design, 24*(2), 120−139. https://doi.org/10.1080/09544828.2012.676633.
Lakhani, K., Fayard, A.-L., Levina, N., & Pokrywa, S. H. (2012). *OpenIDEO, N1-612-066 New York*.
Linsey, J. S., Markman, a. B., & Wood, K. L. (2012). Design by analogy: A study of the wordtree method for problem re-representation. *Journal of Mechanical Design, 134*(4). 041009−041009. https://doi.org/10.1115/1.4006145.

Linsey, J. S., & Viswanathan, V. K. (2014). Overcoming cognitive challenges in bioinspired design and analogy. *Biologically Inspired Design* 221–244.
Linsey, J. S., Wood, K. L., & Markman, A. B. (2008). Modality and representation in analogy. *Artificial Intelligence for Engineering Design, Analysis and Manufacturing, 22*, 85–100. https://doi.org/10.1017/S0890060408000061.
Markman, A. B., Wood, K. L., Linsey, J. S., Murphy, J. T., & Laux, J. P. (2009). Supporting innovation by promoting analogical reasoning. *Tools for Innovation*. https://doi.org/10.1093/acprof:oso/9780195381634.003.0005.
Mihalcea, R., Corley, C., & Strapparava, C. (2006). Corpus-based and knowledge-based measures of text semantic similarity. In *Proceedings of the 21st National Conference on Artificial Intelligence, Vol. 1* (pp. 775–780). https://doi.org/10.1.1.65.3690.
Miller, S. R., Bailey, B. P., & Kirlik, A. (2014). Exploring the utility of Bayesian truth serum for assessing design knowledge. *Human Computer Interaction, 29*(5–6), 487–515. https://doi.org/10.1080/07370024.2013.870393.
Moreno, D. P., Hernández, A. A., Yang, M. C., Otto, K. N., Hölttä-Otto, K., & Linsey, J. S. (2014). Fundamental studies in design-by-analogy: A focus on domain-knowledge experts and applications to transactional design problems. *Design Studies, 35*(3), 232–272. https://doi.org/10.1016/j.destud.2013.11.002.
Murphy, J., Fu, K., Otto, K., Yang, M., Jensen, D., & Wood, K. (2014). Function based design-by-analogy: A functional vector approach to analogical search. *Journal of Mechanical Design, 136*(10), 1–16. https://doi.org/10.1115/1.4028093.
Norton, M., & Dann, J. (2011). Local motors : Designed by the crowd, Built by the customer. 9-510-062(September). *Harvard Business Review* 1–21.
Paolacci, G., Chandler, J., & Ipeirotis, P. G. (2010). Running experiments on amazon mechanical turk. *Judgment and Decision Making, 5*(5), 411–419. https://doi.org/10.2139/ssrn.1626226.
Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 1532–1543). https://doi.org/10.3115/v1/D14-1162.
Purcell, a. T., Williams, P., Gero, J. S., & Colbron, B. (1993). Fixation effects: Do they exist in design problem solving? *Environment and Planning B: Planning and Design, 20*(3), 333–345. https://doi.org/10.1068/b200333.
Rhemtulla, M., Brosseau-Liard, P.É., & Savalei, V. (2012). When can categorical variables be treated as continuous? A comparison of robust continuous and categorical SEM estimation methods under suboptimal conditions. *Psychological Methods, 17*(3), 354–373. https://doi.org/10.1037/a0029315.
Shah, J. J., Kulkarni, S. V., & Vargas-Hernandez, N. (2000). Evaluation of idea generation methods for conceptual design: Effectiveness metrics and de- sign of experiments. *Journal of Mechanical Design, 122*(4), 377–384.
Shah, J., Smith, S. M., & Vargas-Hernandez, N. (2003). Metrics for measuring ideation effectiveness. *Design Studies, 24*(2), 111–134. https://doi.org/10.1016/S0142-694X(02)00034-0.
Terwiesch, C., & Ulrich, K. (2009). *Innovation tournaments: Creating and selecting exceptional opportunities*. Harvard Business Press.
Toh, C. a., & Miller, S. R. (2014). The impact of example modality and physical interactions on design creativity. *Journal of Mechanical Design (Transactions of the ASME), 136*(9). [np]. https://doi.org/10.1115/1.4027639.
Tseng, I., Cagan, J., & Kotovsky, K. (2012). Concurrent optimization of computationally learned stylistic form and functional goals. *Journal of Mechanical Design, 134*(11), 111006. https://doi.org/10.1115/1.4007304.

Tseng, I., Moss, J., Cagan, J., & Kotovsky, K. (2008). The role of timing and analogical similarity in the stimulation of idea generation in design. *Design Studies, 29*(3), 203–221. https://doi.org/10.1016/j.destud.2008.01.003.

Ulu, N. G., Messersmith, M., Goucher-Lambert, K., Cagan, J., & Kara, L. B. (2019). Wisdom of micro-crowds in evaluating solutions to esoteric engineering problems. *Journal of Mechanical Design*. (submitted for publication).

Vattam, S. S., Helms, M. E., & Goel, A. K. (2010). A content account of creative analogies in biologically inspired design. *Artificial Intelligence for Engineering Design, Analysis and Manufacturing, 24*(04), 467–481. https://doi.org/10.1017/S089006041000034X.

Viswanathan, V. K., & Linsey, J. S. (2013). Design fixation and its mitigation: A study on the role of expertise. *Journal of Mechanical Design, 135*, 051008. https://doi.org/10.1115/1.4024123.

Ward, T. B. (1998). Analogical distance and purpose in creative thought: Mental leaps versus mental hops. In B. K. K. Holyoak, & D. Gentner (Eds.), *Advances in analogy research: Integration of theory and data from the cognitive, computational, and neural sciences* (pp. 221–230). New Bulgarian University Press. https://doi.org/10.1016/S0378-2166(98)80007-6.

Wilson, J. O., Rosen, D., Nelson, B. a., & Yen, J. (2010). The effects of biological examples in idea generation. *Design Studies, 31*(2), 169–186. https://doi.org/10.1016/j.destud.2009.10.003.

Yu, L., Kittur, A., & Kraut, R. E. (2014). Searching for analogical ideas with crowds. In *Proceedings of the 32nd annual ACM conference on human factors in computing systems CHI '14* (pp. 1225–1234). https://doi.org/10.1145/2556288.2557378.

Yu, L., Kraut, R. E., & Kittur, A. (2016). Distributed analogical idea generation with multiple constraints. In *Proceedings of the 19th ACM conference on computer-supported cooperative work & social computing – CSCW '16* (pp. 1234–1243). https://doi.org/10.1145/2818048.2835201.