Journal of Engineering Design
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/cjen20

Assessing the quality of ideas from prolific, early-stage product ideation

Barry Matthew Kudrowitz [a] & David Wallace [b]

[a] Department of Design, Housing and Apparel, University of Minnesota, 1985 Buford Avenue, St Paul, MN, 55108, USA

[b] Department of Mechanical Engineering, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, MA, 02139, USA

Available online: 05 Apr 2012

To cite this article: Barry Matthew Kudrowitz & David Wallace (2012): Assessing the quality of ideas from prolific, early-stage product ideation, Journal of Engineering Design, DOI:10.1080/09544828.2012.676633

To link to this article: http://dx.doi.org/10.1080/09544828.2012.676633

Journal of Engineering Design
iFirst, 2012, 1–20

Assessing the quality of ideas from prolific, early-stage product ideation

Barry Matthew Kudrowitz[a]* and David Wallace[b]

[a]*Department of Design, Housing and Apparel, University of Minnesota, 1985 Buford Avenue, St Paul, MN 55108, USA;* [b]*Department of Mechanical Engineering, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, MA 02139, USA*

(*Received 25 February 2011; final version received 9 March 2012*)

The goal of many blue-sky idea generation techniques is to generate a large quantity of ideas with the hope of obtaining a few outstanding, creative ideas that are worth pursuing. As such, a rapid means of screening the resulting sketches to select a manageable set of promising ideas is needed. This study explores a metric for evaluating large quantities of early-stage product sketches and tests the metric through an online service called Mechanical Turk. Reviewers' subjective ratings of idea creativity had a strong correlation with ratings of idea novelty ($r^2 = 0.80$), but negligible correlation with idea usefulness ($r^2 = 0.16$). The clarity of the sketch positively influenced ratings of idea creativity. Additionally, the quantity of ideas generated by an individual participant had a strong correlation with that participant's overall creativity scores ($r^2 = 0.82$). The authors suggest a metric of three attributes to be used as a first pass in narrowing a large pool of product ideas to the most innovative: novel, useful (or valuable), and feasible (as determined by experts).

Keywords: idea evaluation; idea generation; brainstorming; product design; creativity

1. Introduction

Product design idea generation (e.g. brainstorming) results in a large quantity of concept sketches which require a rapid screening process to down select to the most innovative for further exploration. In cases when there are many dozen ideas, participants may use a process called multi-voting where each participant chooses a few of their favourite ideas. This technique has no explicit criteria; participants' biases naturally influence the selection and ideas may get overlooked (Cronin 2001). There are more structured tools for comparing a *select* group of ideas such as the Pugh chart, quality function deployment matrix and the Pahl and Beitz' Utility Theory (King and Sivaloganathan 1999). This research deals with an earlier stage in the process before there is a smaller group of ideas to compare. As King and Sivaloganathan (1999) note, 'it is vital that the best initial concepts are selected as they determine the direction of the design embodiment'.

*Corresponding author. Email: barryk@umn.edu

ISSN 0954-4828 print/ISSN 1466-1837 online

http://dx.doi.org/10.1080/09544828.2012.676633
http://www.tandfonline.com

There are several studies that explore ways of evaluating ideas and they rarely suggest the same metrics for rating (Dean *et al.* 2006). The majority of these studies are specific to ideas for situational problems and not products. The few metrics in these studies that are intended for products are either specific to finished artefacts or are too detailed to evaluate many dozen ideas in a timely manner (Ward and Cox 1974, Besemer and Treffinger 1981, Amabile 1982, 1996, Besemer and O'Quin 1986, 1999, Besemer 1998, 2006, Shah *et al.* 2000, Christiaans 2002, Shah and Vargas-Hernandez 2003, Cropley and Cropley 2005, Horn and Salvendy 2006, 2009).

There is one metric, the NUF (novel, useful, and feasible) test, that is specific to the rapid evaluation of ideas from a brainstorming session, but it is not referenced in academic research and there are no formal studies on this test (Gray *et al.* 2010). Using this evaluation method (NUF), each idea should be quickly rated on these three criteria. The three criteria of NUF are common themes in other metrics.

When looking at prior methods of evaluating ideas, the main problem is that there are no standardised structured methods for down-selecting ideas from prolific product design brainstorming sessions. Pugh charts and similar techniques are structured, but are ideal for a smaller set of ideas. Methods like multi-voting are for large quantities of ideas, but are not systematic and good ideas may get overlooked. The NUF test is both structured and for large number of ideas, but there are no academic studies that have explicitly used this test.

Given that a prolific brainstorming session results in many ideas, it may be possible to crowd source the evaluation by having laymen rate the ideas in an online setting.

There are three main goals of this research: (1) Test a simple metric based on prior work (a variation of the NUF test) for quickly rating the innovativeness of a large quantity of ideas specific to product design. (2) Explore the interrelationships between the components of the metric. (3) Experiment with a method of online crowd sourcing to expedite the down selection, and also to allow the general public, being the consumers of the products, to influence the selection process.

The study focuses on evaluation of *sketches* of ideas, as it can be applied to the visual thinking methods widely used in many product design firms.

2. Idea generation

There are many methods that designers use to generate new ideas for products or creative solutions to problems. These idea generation (or *ideation*) methods can be grouped into categories in several ways.

2.1. *Free-form and structured idea generation methods*

One means of classifying methods is by the type of problems they are addressing. Problems can be ill structured (or free form) when they have poorly defined goals (Butler and Kline 1998), such as 'we need new types of water toys'. An example of an ill-structured task would be designing a house where there are no initial definite criterion to test a solution (Simon 1973). These types of problems often have a large number of solutions and there is no singular best answer. Problems that are structured, such as 'find the least expensive means of securely hanging this picture frame', have well defined goals and often there is an optimal solution. As ill-structured tasks begin to acquire structure and direction, they can be decomposed into many well-structured problems (Simon 1973).

Just as there are free-form and structured problems, there are also free-form and structured idea generation methods. Free-form idea generation methods include: brainstorming, free association and brain writing; and structured idea generation methods include theory of inventive problem solving and morphological analysis (Michalko 2006, Silverstein *et al*. 2008). Free-form methods tend to favour blue-sky or early-stage design (typically performed as a group), while the structured methods are more suitable for more advanced stages of design after a general concept has been selected (typically performed by individuals). Blue-sky idea generation is a term used in the design industry that means free from constraints and expected outcomes. A blue-sky idea generation session typically begins with an open-ended prompt such as 'the future of transportation' or 'gardening tools'. One study found that free-form methods tend to produce more creative ideas while structured methods tend to produce more useful ideas (Chulvi *et al.* 2012). This study relates to free-form ideation methods and, in particular, blue-sky brainstorming.

The contemporary design engineer is expected to be able to participate in the creative early stages of design in addition to the more technical elements (Nussbaum *et al.* 2005).

2.2. *Traditional and nominal brainstorming*

In a traditional brainstorming session, a group of people sit together to generate a large number of ideas in a short amount of time (Osborn 1963). The participants quickly sketch any idea that comes to mind and present it to the group. Each sketch should be drawn large and clearly on a separate sheet of paper and given an appropriate and simple title. As ideas are generated, a group facilitator calls on members to share their ideas. The group facilitator ensures everyone hears each idea and then records the concept title with a brief description.

A nominal brainstorm session follows the basic rules and sketching practices of a traditional brainstorm session, however, the participants do not talk to each other. They either share their ideas silently or do not share the ideas until after the session ends. In this study, the participants did *not* share their ideas. Nominal brainstorm sessions are used in studies, such as this one, when individual ideation ability is of interest. It has also been argued that nominal brainstorming is more effective than traditional brainstorming as it eliminates three theoretical factors of productivity loss: production blocking, evaluation apprehension and free riding (Diehl and Stroebe 1987). Production blocking occurs when a participant suppresses an idea because it seems less original when compared to other ideas, or when a participant forgets an idea because others were speaking. Evaluation apprehension is the term for the fear of negative evaluation from other group members. Free riding is the term for participants not contributing to the group as they feel that their contributions are dispensable. As the number of group members increases, the number of free riders, or non-participants, also increases (Bray *et al.* 1978).

IDEO is a leader in the business of product design consultancy and is known for helping to develop hundreds of cutting edge products. IDEO describes a successful brainstorming session as one that produces over a hundred ideas per hour and virtually every surface of the room should be covered with ideas (Kelly 2001). This practice of creating highly prolific brainstorming sessions is also used in many of the leading product design educational programmes.

3. Prior studies of idea rating

Between 1990 and 2005, there have been approximately 90 studies that have dealt with analysing ideas (Dean *et al.* 2006) and they rarely use the same metrics for rating. The majority of these studies involve prompts related to situational problems (e.g. increasing tourism in Tucson, improving the psychology department at a college, or coping with a drug-dealing roommate). Ideas for situational problems are quite different from ideas for products as the latter typically involve artefacts as well as drawings of these artefacts. Situational prompts are also highly structured where there can be 'more correct' responses and usefulness has a large impact on the perceived goodness.

The few studies that are specific to rating products typically involve existing, finished artefacts (not ideas) and/or are related to mechanisms where usefulness is a crucial component (Ward and

Cox 1974, Besemer and Treffinger 1981, Amabile 1982, 1996, Besemer and O'Quin 1986, 1999, Besemer 1998, Shah *et al.* 2000, Christiaans 2002, Shah and Vargas-Hernandez 2003, Cropley and Cropley 2005, Horn and Salvendy 2006, 2009). The work of Besemer (2006) is perhaps the only one that deals with *ideas for products*, but is better suited for reviewing and comparing a smaller set of ideas in detail. A condensed hybrid of these instruments is needed to evaluate product ideas from a high volume product ideation session.

3.1. *Metrics of evaluating ideas*

The 'criteria problem' is the term given to the issue of obtaining a valid assessment of the level of creativeness of a person (Besemer and Treffinger 1981). In studies that have dealt with product ratings, there are various sets of criteria that have been suggested to define a creative product.

There are two main categories of definitions for creative products: novelty-based and multi-attribute-based (Dean *et al.* 2006).

Multi-attribute-based definitions of creativity claim that products must be novel as well as possess other quality attributes. The most commonly used quality attribute to describe a creative product is *usefulness* (Dean *et al.* 2006). Mednick (1962) defines creative thinking as '… the forming of associative elements into new combinations which either meet specified requirements or are in some way useful'. Other quality attributes include relevance, appropriateness, clarity, workability, feasibility, etc. Almost every study in this field chooses different attributes to determine creativity. According to Shah and Vargas-Hernandez (2003), good engineering design should be both practical and novel; however, engineering design is not synonymous with product design. In the realm of product design, there are a lot of successful creative products that have no practical value outside of playfulness or entertainment.

Novelty-based definitions claim that products are creative based solely on originality and do not depend on any other quality attributes such as appropriateness, usefulness or applicability. Runco and Charles (1993) found that 'it is not necessary for an original idea to be appropriate (i.e. solve a problem) to be viewed as creative'. In a more recent study, Christiaans (2002) found that 'usefulness seemed not to be important in discriminating between designs with high and low creativity ratings'. Ideas that are novel, but not feasible or useful, are sometimes termed aesthetic innovation (Eisenman 2006), chindogu (Patton and Bannerot 2002), or novelty items. In either case, if usefulness does or does not contribute to the assessment of creativity, it is still an attribute that can be evaluated.

Amabile (1982) takes a stance that is divergent from these attribute-based assessments and claims that creativity cannot be determined objectively using metrics. She suggests a subjective Consensual Definition of Creativity: 'A product or response is creative to the extent that appropriate observers independently agree it is creative (1982)'. In other words, instead of determining if a product is creative by asking reviewers if the product is novel, useful (etc.), simply ask reviewers if the product is creative.

Table 1 presents relevant studies and the attributes (or dimensions) used to evaluate the idea or artefact.

The NUF test is specific to the rapid evaluation of ideas from a brainstorming session (Gray *et al.* 2010). Using this evaluation method (NUF), each idea is rated on these three criteria on a scale of 1–10. Summing the three values gives a score to each idea. The three criteria of NUF are common themes throughout the metrics presented in Table 1.

The reference for NUF (Gray *et al.* 2010) is a book called 'Gamestorming: A playbook for Innovators, Rulebreakers and Changemakers'. In the book, there is only a short paragraph describing the idea, and there are no references or studies cited for the technique. It is described as an adaptation on a process used to evaluate patents. This work is novel in the sense that it is the first academic study of Gray's proposed technique. NUF does not clearly take into account the realm

Table 1. Selection of prior studies involving a metric for the evaluation of ideas/artefacts.

Study	Summary	Products in review	Dimensions of creativity				
Dean *et al.* (2006)	Review of 90 constructs for idea evaluation	Ideas (e.g. increase tourism in Tucson)	Novelty (originality, paradigm relatedness)	Workability (acceptability, implementability)	Relevance (applicability, effectiveness)	Specificity (implicational explicitness, completeness and clarity)	
Besemer and O'Quin (1986, 1999)[a]	Objective metric for creative product evaluation	Artefacts (chairs)	Novelty (surprising, original)	Resolution (logical, useful, valuable, understandable)	Elaboration and synthesis (organic, well-crafted, elegant)		
Shah and Vargas-Hernandez (2000, 2003)	Evaluation of mechanical engineering designs	Artefacts and ideas (mechanical devices)	Novelty (unusual)	Quality (meets specifications)	Variety (explored solution space)	Quantity	
Horn and Salvendy (2006, 2009)[a]	Consumer-based assessment of product creativity	Artefacts (chairs and lamps)	Novelty (frequency, rarity)	Importance (relevance, significance)	Affect (appeal, desire, attraction, delight, stimulation, etc.)		
Amabile (1982)[b]	Subjective assessment method of creativity	Artefacts (artwork and poetry)	Creativity (as determined by appropriate judges)	Creativity Cluster (novel material use, novel idea, effort, detail, etc.)	Technical cluster (technical goodness, organisation, neatness, etc.)	Aesthetic judgment (liking, aesthetic appeal, would you display it?)	
Christiaans (2002)[b]	Creativity as one metric of design review	Artefacts (cabinets and telephone booths)	Creativity (as determined by appropriate judges)	Technical quality	Attractiveness	Interest	Goodness of example

[a]In these studies, each of the dimensions are made up of many sub-dimensions in the form of bipolar adjective scales.
[b]In these studies, the assessment dimensions in addition to 'creativity' do not determine creativity, but were used to find correlations with creativity.

of marketability (or desirability), which is an important consideration in product design (Horn and Salvendy 2006).

3.2. *Quantity of ideas*

The quantity of ideas produced in an ideation session (sometimes termed fluency or productivity) is a common dimension that is reviewed in addition to those that are used to measure creativity. For the most part, quantity is objective and easy to measure. In most studies, a count of non-redundant ideas constitutes the rating for this dimension.

Pauling (1977) said, '… you aren't going to have good ideas unless you have lots of ideas and some sort of principle of selection'. Research has shown a positive correlation between total number of ideas and total number of good ideas (Diehl and Stroebe 1987, Paulus *et al.* 2011). Another study found that quantity of ideas was positively correlated with original ideas and negatively correlated with feasible ideas (Rietzschel *et al.* 2006). Dean *et al.* (2006) found that 18 of the 90 idea review studies between 1990 and 2005 used quantity as the sole means of evaluating the ideation session.

Osborn (1963) claims that 'quantity breeds quality in ideation' and that 'early ideas are unlikely to be the best ideas generated during an ideation session'. The relationship between idea quality and idea quantity can be visualised with a curve called the ideation function (Reinig *et al*. 2007). The bounded ideation theory describes the ideation function as a positive s-curve where the majority of good ideas will come somewhere in the middle of an ideation session (Reinig *et al.* 2007). Similarly, van der Lugt (2003) and Goldschmidt and Tatsa (2005) suggest that the ideas that are perceived to be of higher quality are ones that have been built on earlier ideas and lead to further ideas.

Some studies evaluate only one idea per person or group, which makes it easier for a reviewer to evaluate those few ideas on several dimensions. When there are many ideas in need of review, the number of dimensions for which each idea is evaluated must be limited to prevent reviewer fatigue. Amabile (1982) found that levels of inter-judge agreement depend on the magnitude of effort required by the judges.

3.3. *Judging processes*

When evaluating creativity, the only requirement of judges is some basic level of familiarity with the subject matter (Amabile 1982). Expert judges are only required when the products in review are of a highly specialised nature. 'The more "cutting edge" a product is in a specific domain, the more likely an expert judge will be required (O'Quin and Besemer 1999)'. In the case of common household items (such as a toaster, umbrella or toothbrush), almost anyone could be considered an appropriate judge of creativity. A few studies have found, based on inter-rater agreement, that there is little difference between experts and non-experts in rating design creativity (Amabile 1982, Christiaans 2002). Although professionals or experts are capable of producing creative products, they may be unreliable in judging them as they rely on higher, esoteric, or idiosyncratic standards (O'Quin and Besemer 1999).

Perceptions of creativity and innovation are culture or society dependent. According to Boden (2004), two levels of creativity are said to exist: creative to the individual and creative in history. Others have recently added a third classification that lies between these two realms: creative to the situation/society/group (Shah *et al.* 2000, Sosa and Gero 2003). In product design, creative to the society is what matters most. Polling a section of the population online is a means of evaluating that type of creativity.

In some experiments, a panel rates attributes (such as novelty and practicality) on a scale (e.g. from 1 to 5) (Rietzschel *et al.* 2006, Fink and Neubauer in press). In some cases, the judges agree upon a standard for what a 1, 3, and 5 score would be or how to define each of the dimensions. When dealing with the dimension of creativity, Amabile (1982) suggests that the judges use their own definitions of creativity as opposed to attempting to define criteria.

If scores are to be combined for an individual's set of ideas, they will either be added or averaged. If added, a large number of bad ideas are helpful. For example, if participant A produced 20 ideas each with a low score of 1 and participant B produced 3 ideas each with a high score of 5, participant A would have a higher total score even though participant B had better ideas. Oppositely, if ideas are averaged, a large number of bad ideas will lower the overall score. For example, if participant A produced only two ideas, one with a score of 5 and one with a score of 1, this equates to 50% good ideas. If participant B produced 10 ideas, 2 with a score of 5 and 8 with score of 1, this would equate to 20% good ideas. Even though participant A had a higher percentage of good ideas, participant B had more good ideas. What matters most is the number of good ideas, not the number of bad ideas.

It is best to simply count the number of creative ideas as opposed to averaging or summing ratings to avoid the influence of a large number of bad ideas (Reinig and Briggs 2006). The count method is used in this study. An argument against this count method is that it does not take into account the degree of the creativity of the individual ideas.

4. Experiment

In short, participants were asked to generate ideas for blue-sky product design themes. Based on prior work, a metric was chosen to evaluate these ideas. An online crowd sourcing evaluation method was used to review the ideas with this metric.

4.1. *Selecting ideation themes*

Participants ideated around a product theme (ill structured) instead of developing solutions to a problem (structured). There are several benefits of an ill-structured or blue-sky idea generation over a structured theme. In structured problems, the solutions may not be related to products at all and it would then be hard to compare innovative designs. Designers and engineers may develop better solutions to structured problems simply because of their training in the scientific method. The structured problem prompt also suggests there is a solution or best solution and in product design idea generation there are many solutions (Butler and Kline 1998). Finally, in recent experiences, companies often ask for blue-sky innovation around a product theme such as peristaltic pumps, water guns, or cork. These are realistic scenarios. In the future, it may also be beneficial to look into a structured problem ideation as this may involve a different thought process and a different form of creative thinking.

The three themes used in the idea generation were umbrellas, toasters and toothbrushes. These themes are common enough for the general public to have a good understanding of the current state of the art. The themes are also common enough for the 'average' person to be considered an appropriate judge of creativity (Amabile 1982). A similar study evaluating the creativity of products chose two of these same themes (toasters and toothbrushes) because of their 'commonality' and 'moderate cost (Horn and Salvendy 2006)'. However, this similar study evaluated products that were currently on the market and not conceptual product ideas.

4.2. *Participants*

The test was administered to 84 participants (52 male, 32 female) of whom 24 were professional product designers, 21 were professional improvisational comedians, and 26 were mechanical engineering undergraduate students. The remaining 13 participants did not fit into these categories. The participants were chosen to represent a few disciplines that engage in regular idea generation sessions. The combined age range of the participants was between 18 and 63 years. The mean age was 28 and the median age was 23. The engineering students who volunteered to participate were all involved in one of three different product design-related courses. The majority of the students were underclassmen and more specifically, freshmen.

Participants were given the test in familiar locations, typically in their work surroundings in a group at a large conference table. Tests were performed individually but sitting together as a group to save time on the part of the administrator and to make the participants more comfortable.

4.3. *Procedure*

Nominal brainstorming was used in the test as opposed to a traditional interactive brainstorm to better evaluate the individual participants.

Participants were asked to develop as many innovative concepts they can around a given product theme. Each participant was given a stack of blank legal paper and a black fine tipped Sharpie® permanent marker. They were asked to sketch each idea with a title on a new piece of paper in the portrait orientation. They were asked that all sketches be made in the portrait orientation to simplify the scanning, processing and reviewing. Participants were asked to draw large, legibly and use as few words as possible to explain the concept. They were told that drawing ability does not count and that they should sketch all ideas that they feel are innovative in some way. Participants were told explicitly to not talk to each other or share ideas.

Each group of participants generated ideas around all three themes in a different order. Each brainstorm session lasted 12 min. A pre-test found that three 12-min sessions is an acceptable time to limit fatigue and still produce a good quantity of ideas. Time was called at 12 min and participants were allowed to finish any sketch they had started.

An identification number specific to each participant was placed on the back of each concept drawing and each drawing was scanned.

4.4. *Designing a metric*

In a smaller test experiment prior to the main study, there was confusion about the meanings of usefulness, practicality, feasibility, and novelty and whether these criteria actually determined innovativeness or creativity.

Appropriate evaluation attributes were chosen through referencing the prior studies of Table 1. The classification presented by Dean *et. al.* (2006), based on MacCrimmon and Wagner (1994), is thorough, but is also tailored to evaluate ideas for themes that are outside the realm of products. Dean suggests reviewing ideas by the following classification: Novelty (composed of Originality and Paradigm Relatedness), Workability (composed of Acceptability and Implementability), Relevance (composed of Applicability and Effectiveness), and Specificity (composed of Implicational Explicitness, Completeness, and Clarity).

Horn and Salvendy (2009) and Besemer and Treffinger (1981) are a few of the researchers that deal specifically with metrics for reviewing products. Their metrics are slightly long and detailed and are better suited to compare several ideas as opposed to hundreds of concepts.

In this study, ideas were evaluated on quantity as well as five qualitative attributes that are representative of prior metrics: creative, novel, useful, product-worthy, and clear. Descriptions of these attributes are as follows:

Creative – *Using the judge's own subjective definition of creativity*. Following the subjective assessment style of Amabile (1982) and Christiaans (2002), reviewers were asked to rate the ideas on several attributes, one of which is creativity as defined by the reviewer. It is then possible to assess which attributes correlate with the subjectively rated creative products.

Novel – *The concept is original* (Besemer and O'Quin 1986) *and uncommon* (Horn and Salvendy 2009). A gun that shoots out celery may not be considered useful or practical, but it should be considered novel. According to the multi-attribute definitions of creativity, novelty does not always imply creative. Take the celery gun example and imagine if the prompt was to ideate new toothbrush concepts. This concept would be considered original and uncommon, but would not make any sense in the realm of toothbrushes. As Mednick (1962) explains '7,363,474 is quite an original answer to the problem "how much is 12 + 12"' but it would not be considered creative.

Three types of novelty are said to exist: novel to the individual, novel to the society/group, novel to history (Shah and Vargas-Hernandez 2003). An idea for a product that is only novel to the individual would most likely not be considered innovative. In the realm of product design, societal and historical novelty are more important than individual novelty.

Useful – *The concept has practical applications in the product theme.* This is perhaps the most controversial attribute for rating concepts. MacCrimmon and Wagner (1994) and Dean *et al.* (2006) call this attribute 'Relevance' meaning applicable to the problem at hand. This wording is more appropriate for structured prompts as they have solutions. Horn and Salvendy (2009) call this attribute 'Importance' defined as relevant and significant. Besemer and Treffinger (1981) and Besemer and O'Quin (1999) would call this attribute 'Resolution' defined as the correctness of the solution; it is logical, useful and valuable.

Product-worthy – *The concept is both feasible and marketable*. Feasibility can be described as the ability of an idea to be physically made into a product. It takes into account technology, cost, safety, and manufacturability. Marketable takes into account the social and cultural variables and if people would want to purchase the product (is it too much work to use? is it too large to store? would it be too expensive for the intended users?). This attribute is termed 'Workability' by MacCrimmon and Wagner (1994) and Dean *et al.* (2006) and they define it as acceptability and implementability. This attribute is not addressed by Horn and Salvendy (2009) as their studies review products that are already on the market.

Product-worthy may be hard to determine for the layman, as there are many ways of manufacturing a product and many different technologies that are unknown to the general public. It is also difficult to evaluate a category when it is composed of subcategories, however, in this study feasible and marketable were grouped in the same manner as in MacCrimmon and Wagner (1994) to reduce the number of evaluation criteria.

Clear – *The concept is well communicated.* This attribute maps to what Dean *et al.* (2006) and MacCrimmon and Wagner (1994) term 'Specificity' meaning clear, complete and explicit. As this is early-stage brainstorming, it is not expected that ideas should be elaborate and complete, however ideas are expected to be clear and detailed enough that the judges are able to understand the concept. This attribute also relates to the Besemer and O'Quin (1999) attribute of 'Synthesis' which includes organic, well-crafted, elegant and, in an earlier study, understandable (1986). Unlike 'creative', 'novel', 'useful', and 'product-worthy', 'clarity' is independent of the concept and based solely on the idea presentation.

4.5. *Rating ideas*

To evaluate the product concepts, an online website approach was used as opposed to a physical review form. An online review has several benefits. The reviewers can be located in many different places, and this allows for a better general population sample. The reviewers can do the rating at their own convenience and in a comfortable setting. The data can be reviewed by hundreds of people as opposed to a select panel of expert individuals. The review data are also easy to collect and tabulate. As the product themes are common, it is acceptable to have a random laymen review panel (Amabile 1982). In a sense, these laymen are the experts when dealing with general commonplace products such as toasters. These reviewers are likely to own toasters (or toothbrushes or umbrellas), and given that they are early adaptors of contemporary Internet services, it is probable that these individuals are also early adapters (if not simply aware) of contemporary consumer goods.

Amazon Mechanical Turk is a website (http://www.mturk.com) that allows any user to post tasks for any other user to complete. The tasks are termed 'HITs' or 'human intelligence tasks', the person posting the task is termed a 'requester', and the people completing the tasks are termed 'workers'. Requesters post HITs and a monetary payment for completing each HIT. In this study, only workers could participate if they were located in the USA (based on IP address). Recent studies have shown that Mechanical Turk is a reliable source of experimental data and that the population of Mechanical Turk is at least as representative of the US Population as traditional subject pools (Paolacci *et al.* 2010). Other studies have found little to no difference in the magnitude of effects obtained using traditional methods and Mechanical Turk (Snow *et al.* 2008, Nowak and Ruger 2010).

Each worker was paid $0.15 to rate 20 randomly selected product ideas. Each product idea was presented as a scan of the original sketch alongside radio buttons for rating each attribute. An example of the online review page is shown in Figure 1.

Twelve different workers out of a pool of 397 rated each of 545 toaster ideas, 627 umbrellas ideas and 595 toothbrush ideas. The workers had an average age of 34.4. Approximately 38% were male.

Some prior studies that involved ratings of ideas used only two judges (Reinig and Briggs 2006, Rietzschel *et al.* 2006). As shown in the results, different people find different ideas creative and a review panel of two is not sufficient to accurately rate an idea. These prior studies also used Likert scales that range from binary to upwards of five points (Reinig and Briggs 2006, Rietzschel *et al.* 2006). A scale closer to binary is a more appropriate measure when dealing with hundreds of ideas.

Reviewers were asked to rate each idea on these five attributes: creative, clear, novel, useful, and product-worthy. To make these attributes easy to understand, each included a short explanation. Each attribute was rated on a three-point Likert scale (2 = yes, 1 = somewhat, 0 = no):

- *Creative*: The concept is creative.
- *Clear*: The concept is well communicated.
- *Novel*: The concept is original and uncommon.
- *Useful*: The concept has practical applications in the theme.

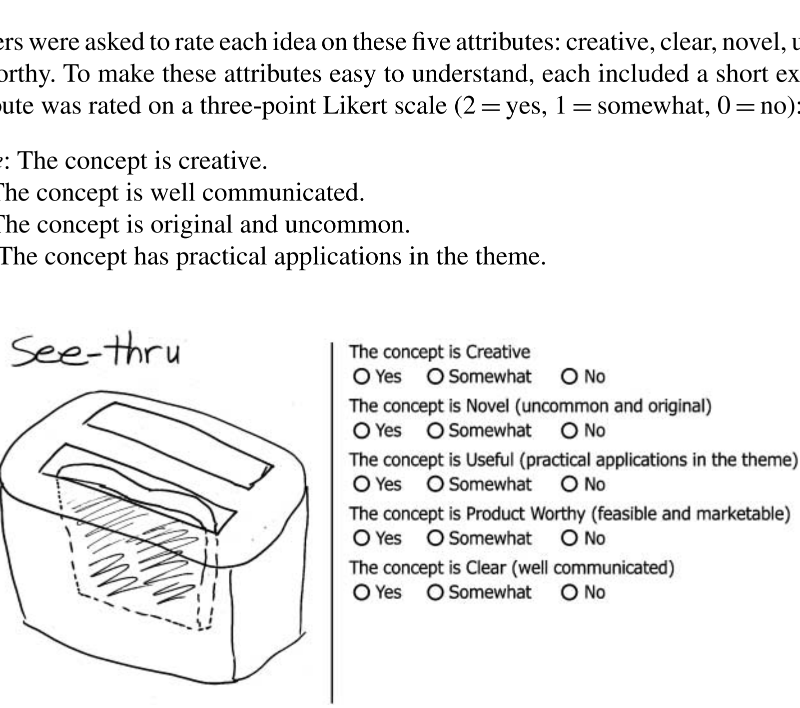

Figure 1. An example of an online product rating form.

Table 2. Interrelationships between attribute ratings (r^2).

	Clear	Creative	Novel	Product-worthy
Creative	0.32			
Novel	0.24	0.80		
Product-worthy	0.43	0.14	0.07	
Useful	0.42	0.16	0.08	0.86

- *Product-worthy*: The concept is feasible and marketable.

In evaluating individual ideas, an average of review scores for each attribute is acceptable. This would give each idea a set of average scores between 0 and 2. However, when evaluating a set of ideas (i.e. the creativity of a participant) a total count of 2s (yeses) is the most equitable measure to ensure that good ideas are not devalued by a large number of bad ideas (Reinig and Briggs 2006). A count of good ideas (e.g. number of ideas with a score over 1.0) is also an acceptable means of scoring the individual participants, however, the count of 2s gives a greater depth and finer resolution.

5. Results

5.1. *Interrelationships between attributes*

Comparing the average scores for each idea, Table 2 shows the interrelations between each attribute rating.

Useful and product-worthy ($r^2 = 0.86$) as well as creative and novel ($r^2 = 0.80$) are practically indistinguishable to reviewers. Clarity appears to be moderately correlated to product-worthy and useful ratings, and minimally correlated to creative and novel ratings. One can notice greater effects of clarity on creativity when comparing sets of similar ideas that are sketched with different levels of clarity.

Figures 2–4 show the most relevant inter-metric findings.

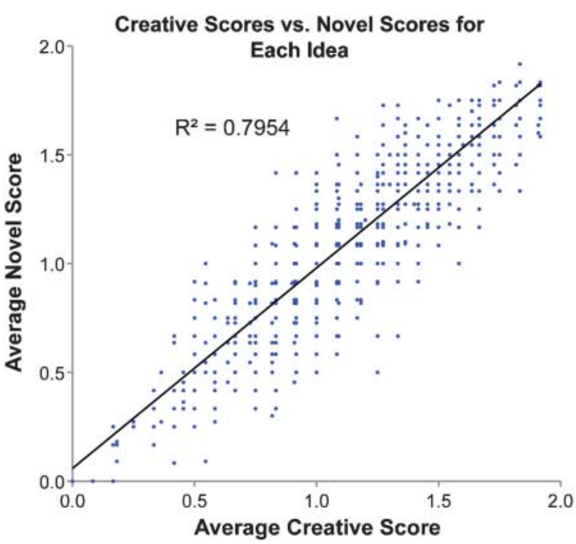

Figure 2. Creative *vs.* novel scores for each idea.

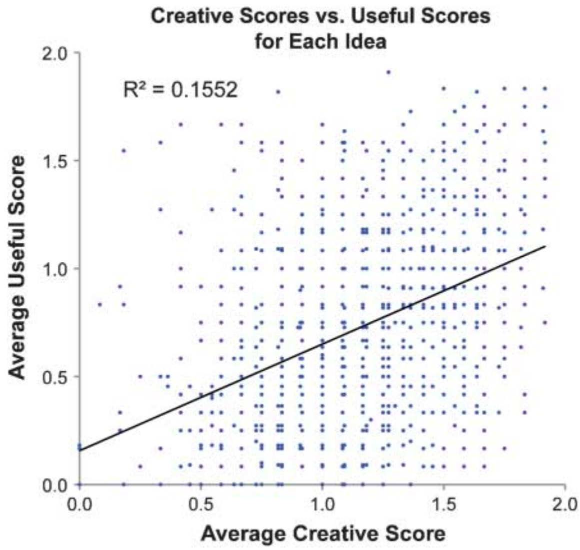

Figure 3. Creative *vs*. useful scores for each idea.

Figure 4. Product-worthy *vs*. useful scores for each idea.

5.2. *Quantity of ideas*

Quantity of ideas generated by an individual was highly correlated with overall individual creativity scores ($r^2 = 0.82$), as well as quantity of creative ideas (i.e. ideas with average creative scores over 1.0) ($r^2 = 0.64$). It may be argued that the high creative scores are an artefact of having a lot of ideas, however, the useful score is only minimally correlated with quantity of ideas ($r^2 = 0.38$). The quantity of ideas compared to individual creative and useful scores are shown in Figures 5 and 6, respectively.

The total number of product ideas per participant refers to a summation of the number of ideas that each individual produced in three product brainstorming sessions (toasters, umbrellas and toothbrushes). The most prolific idea generator in this study produced 52 ideas in the 36 min of

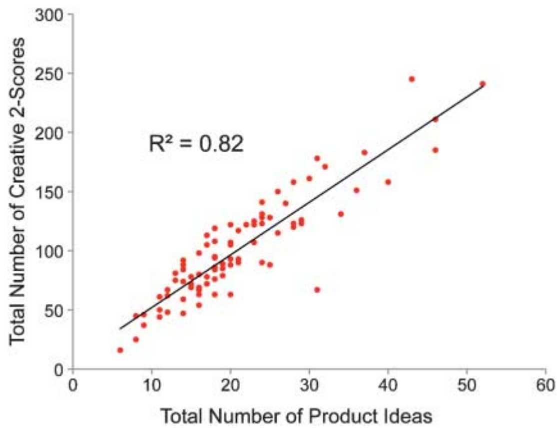

Figure 5. Quantity of ideas *vs.* overall creativity score per participant.

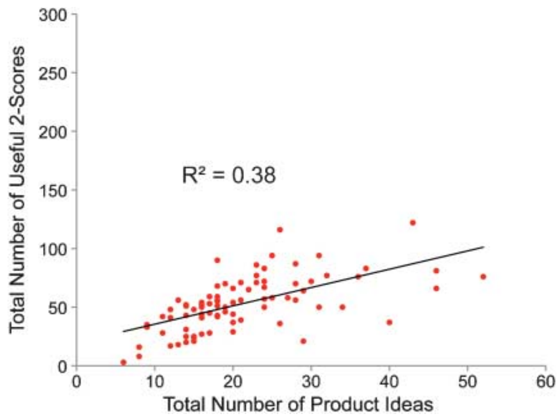

Figure 6. Quantity of ideas *vs.* overall useful score per participant.

brainstorming (12 min per product theme). This is approximately 4.3 ideas per minute. Even if these ideas are rough, any one of them could be a potential product and so each idea should be given the same evaluation treatment.

Plotting the quantity of ideas to the total average creativity score per participant shows there is a negligible negative/no correlation ($r^2 = 0.01$). This, however, is not an accurate measure as participants are penalised for having bad ideas mixed with good ideas. Having 10 good ideas out of 100 is better than one good idea out of two, even though the former has a much smaller percentage of good ideas. It is the number of good ideas that matter, not the number of bad ideas.

6. Discussion

6.1. *Quantity of ideas*

As quantity of ideas is highly correlated with creativity of ideas ($r^2 = 0.82$), it is reasonable to measure an individual's creativity simply on fluency. This, however, does not accurately represent that individual's ability to produce useful or feasible ideas. As found in this study, usefulness is not correlated with individuals' subjective definitions of creativity, and quantity is not correlated with usefulness. If usefulness (or feasibility or practicality) were grouped into the definition of creativity, this study would have come to the conclusion that creativity has a very low (if any) correlation with quantity. This is not to say that usefulness is unimportant, it is a separate attribute than the subjective laymen's perception of creativity. Not all things that are creative are practical or useful (e.g. Dip-n-Dots, Disneyworld, and Phillip Stark's Juicy Salif Juicer). These are all very successful and creative products that have no practical use aside from fun and enjoyment.

It could be argued that individuals that produce a lot of ideas are better at divergent or associative thinking, which is often related to creative thought process (Guilford 1956, Torrance 1972). As more associations are made, the probability of reaching a creative idea increases (Mednick 1962). One could also argue that individuals who are uninhibited will edit thoughts less, produce many ideas and their output should be less restrained and thus more creative. On the contrary, individuals who are logical thinkers may come up with more useful concepts, but will also be restrictive in their thought processes producing fewer concepts in total.

6.2. *Rating product worth*

In this study, product-worthy was defined as both feasible and marketable. These elements can be mutually exclusive, which made this attribute difficult for reviewers. In hindsight, it would have been more meaningful if the category of product-worthy were decomposed into its constituents (feasible and marketable) to gain further insight into these areas.

Laymen reviewers had difficulty rating marketability. Several ideas were rated as not marketable, when the idea was for a product on the market. An example of this is the Miracle Toaster shown in Figure 7, where 12/12 reviewers rated it as not product-worthy when it is currently a product on the market sold by Fred Inc. designed by Jason Amendolara (http://www.worldwidefred.com/holytoast.htm).

Laymen are also not appropriate reviewers of feasibility. A good percentage of reviewers rated a battery-powered toaster, a hand-cranked toaster and a cardboard box solar toaster as product-worthy.

As marketability depends on a variety of factors external from the concept itself, the authors suggest simplifying the attribute of product-worthy to feasible with the caveat that expert judges with

Figure 7. Toaster that produces holy images.

a technical background are required. It would not be possible to accurately rate this attribute using crowd-sourcing methods. Laymen may still be appropriate judges of creativity as per Amabile (1982), which we have defined as a separate category from usefulness, feasibility, and marketability.

6.3. *Concept clarity and creativity*

The general correlation between clarity and creativity scores was low-moderate ($r^2 = 0.32$). However, with the large number of ideas generated, there were many ideas that were repeated by different participants. Using these sets of similar ideas, one can better see the effect of clarity of sketch on perceived creativity.

Out of the 545 toaster ideas, four were ideas for toasters that optically detect if the toast is burning. Figure 8 shows the ratings of clarity and creativity for each of these toasters as measured by a count of 2s. In this example, the first sketch is practically illegible, the second sketch is annotated as 'burn detector' but the image is unclear, the third sketch does not feature a toaster but the annotated toast explains the concept, and the fourth sketch simply and clearly depicts a toaster with a sensor scanning for burned areas on the toast. As the clarity score increases, the creativity score increases even though these ideas are essentially the same. Another study found that quality of the sketch has a significant impact on the perceived creativity of the idea (Kudrowitz *et al.* 2012).

These findings suggest the importance of basic sketching ability for engineers. Creative ideas can be overlooked when they are poorly sketched.

6.4. *A map of product innovation*

Product design in general tends to favour innovation over invention. *Invention* is creating a new and useful technology, while *Innovation* is the combination of knowledge or technologies in original and non-obvious valued new products, processes or services (Luecke and Katz 2003). In essence,

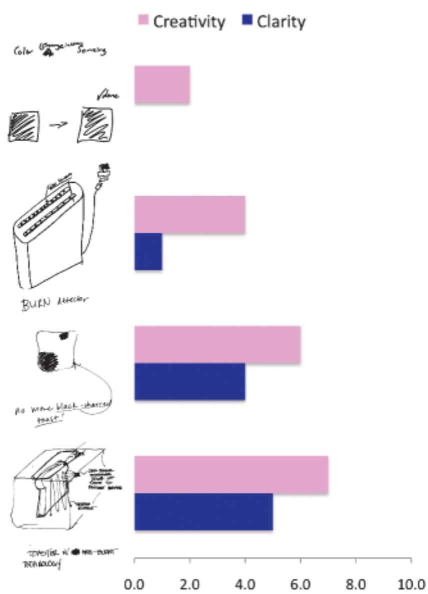

Figure 8. Creativity and clarity of optical burn-detecting toaster idea sketches.

what makes a product idea innovative is a combination of novelty, usefulness (which could be expanded to value), and feasibility. These equate to three of the four attributes of MacCrimmon and Wagner (1994) instrument: Novelty, Relevance, and Workability. MacCrimmon and Wagner's (1994) fourth attribute 'Specificity' which is similar to Besemer and O'Quin's (1986) attribute 'Elaboration' is not included in this instrument as it pertains to the detail, and in brainstorming sessions all of the sketches are quick and rough doodles. Although the clarity, specificity or elaboration may affect the perception of creativity, it is about the presentation of the idea and not the idea itself.

With strong correlations between creative and novelty scores ($r^2 = 0.80$), it seems appropriate to rate either creativity or novelty as both would be redundant. Another study had similar findings where expert judges were not able to distinguish between categories of creativity and novelty (Chulvi *et al.* 2012). These findings are consistent with the novelty-based definition of creativity and prior research (Runco and Charles 1993, Christiaans 2002).

One can visualise these three qualities (novel, useful, and feasible) as a spider plot as shown in Figure 9. Ideas can be mapped as a shape inside the triangle depending on their scores in each of these three areas. An idea that is innovative and that should be further explored would have high scores in all three areas as shown in Figure 10. This is similar to the aforementioned NUF test but with a rating of 0–2 for each attribute and in a visual format.

In engineering design or when dealing with solutions for a problem, the attribute of 'value' is most likely synonymous with 'usefulness'. When dealing with product design, 'value' can come from other dimensions. A product idea can be deemed innovative without having a practical application, just as an idea can be deemed creative without being feasible. Ideas that are novel and feasible, but not of practical use can include toys, entertainment, and novelty items. This is

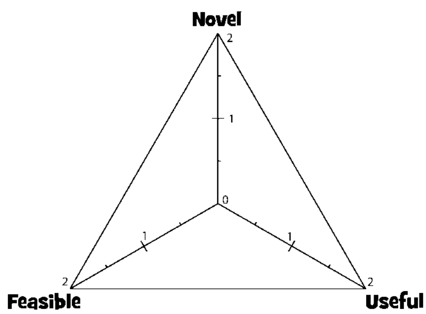

Figure 9. A map of product innovation as a spider plot.

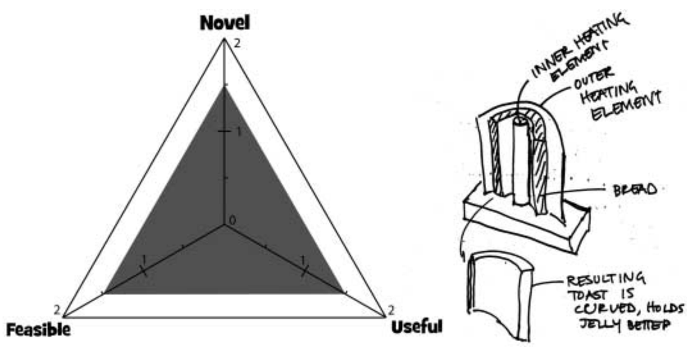

Figure 10. Example of innovative product idea score and innovative idea sketch: a toaster that curves the toast to hold toppings.

depicted in Figure 11. If it is important that these types of innovative product ideas do not get discarded or score lower, one can change the label of useful to something broader to reflect a potential market (e.g. valuable, appealing).

Ideas that are novel and useful but not feasible are desired but less attainable. These ideas are *chimera*, pipe dreams, and potential for invention. An example is depicted in Figure 12. Ideas that are useful and feasible, but not novel are likely to be similar to existing products. This is depicted in Figure 13. It is possible to have ideas that score high on only one of the three attributes. The most common are ideas that are novel, but not feasible or useful. These ideas are typically nonsense.

There are many methods to evaluate ideas and this article is perhaps a case study exploring one particular method (NUF). The attributes that make up this metric are relatively common as they

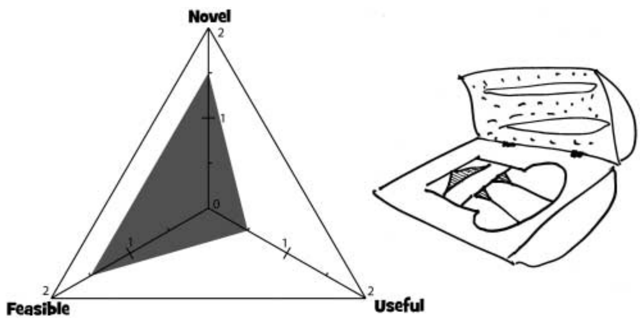

Figure 11. Example of novelty item idea score and novelty item idea sketch: a toaster that is shaped like a tanning bed and toasts a 'bikini tan' image onto the bread.

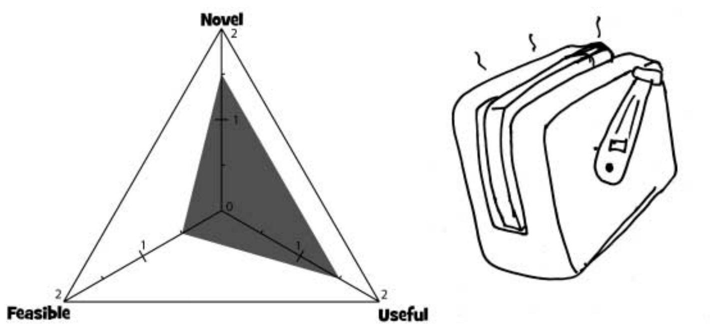

Figure 12. Example of chimera idea score and chimera idea sketch: a hand-crank-powered toaster.

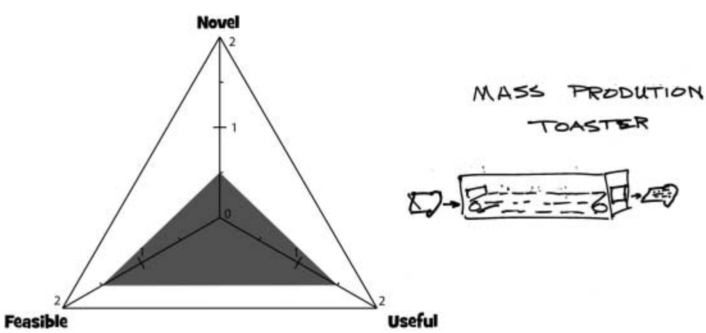

Figure 13. Example of existing product idea score and existing product idea sketch: a conveyor belt toaster for mass production.

are loosely based on the criteria to file a patent (Gray *et al.* 2010). The results of this study are based on laymen reviewers using an online rating system.

6.5. *Reviewing rate*

Using Mechanical Turk, it takes between 2 and 3 min for each reviewer to rate 20 product ideas on five attributes. As it is online, multiple reviewers can rate product sets simultaneously. It is always a good idea to document the output of a brainstorming session. Automatic document scanners allow for quick digitising and labelling of all ideas from a session.

In future studies, only three attributes are suggested (novel, useful or valuable, and feasible) to reasonably reduce hundreds of product concepts down to a more manageable set of promising ideas. Once a set of ideas is reduced (to say a few dozen), a more elaborate metric can be used to evaluate the ideas in detail. The metric presented by Besemer and O'Quin (1999) takes approximately 15 min to evaluate one product. This is over 100 times slower than the method presented in this study. To ask reviewers to rate hundreds of product ideas on 9+ attributes would be time intensive and would cause fatigue (Amabile 1982). Although the method presented in this article is less rigorous, it reasonably reflects the level of detail of the ideas that are being reviewed.

It is not suggested that an online crowd-sourcing method should always be used; this study was an experiment in crowd sourcing the initial filtering of ideas. Furthermore, the metric presented in this article does not need to be used in a digital setting. The participants of a brainstorming session could treat this metric as a first order check when sorting or discarding ideas as suggested by Gray *et al.* (2010).

7. Conclusions

Addressing the goals of this research, the authors reviewed the prior work in idea evaluation and suggest a three-attribute metric to quickly filter down a large selection of product ideas to the most innovative: novel (new to society), useful (or valuable), and feasible (as determined by experts with a technical background). There are many attributes that should be considered in choosing an idea to pursue and every prompt will have its own criteria; however, at a minimum an innovative product idea should be somewhat new, people should want it, and it should be able to be produced in a reasonable manner. This metric is not meant to replace a Pugh Chart or other idea evaluation techniques that are targeted at comparing a handful of select ideas, but rather as a tool to use as a first pass evaluation of a large pool of product ideas.

A secondary goal of this research is to explore relationships between the metric attributes tested in the study. This study has supported the notion that prolific idea generation *is* creative idea generation, as the quantity of ideas generated by the individual participants had a very strong correlation with each participant's overall creativity scores ($r^2 = 0.82$). This finding suggests that if we can encourage engineers and designers to come up with lots of product ideas then we are potentially increasing the number of *creative* product ideas.

The work also supports the novelty-based definition of creativity as reviewers' subjective ratings of idea creativity had a strong correlation with ratings of idea novelty ($r^2 = 0.80$), but negligible correlation with idea usefulness ($r^2 = 0.16$). As mentioned early, the novelty-based definition of creativity claims that products are creative based solely on originality and do not depend on any other quality attributes such as appropriateness, usefulness or applicability.

Sketch clarity had a moderate positive correlation with the perceived creativity of ideas. This supports an argument that engineers and anyone in a field related to innovation should have basic drawing abilities so their creative ideas are not overlooked.

A final goal of this research was to experiment with a method of online crowd sourcing to expedite the down selection, and also to allow the general public, being the consumers of the products, to influence the selection process. It was found that online reviewers using Mechanical Turk are able to rate between 7 and 10 ideas on five attributes per minute. This is a rapid means of taking a large collection of ideas and reducing them to a manageable set of the most promising ideas. It was found that laymen reviewers were not able to accurately rate ideas for feasibility or marketability. As a result of this, crowd-sourcing methods of evaluation, such as Mechanical Turk, are not appropriate for attributes such as feasibility for which reviewers require an understanding of current technology.

Further work is needed to explore sketch quality and creativity. Understanding what aspects of a sketch are most influential on perceived creativity could lead to better sketching and evaluation practices. As quantity of ideas was highly correlated with creativity of ideas, it is of interest to explore different methods of helping people become more prolific idea generators.

References

Amabile, T.M., 1982. Social psychology of creativity: a consensual assessment technique. *Journal of Personality and Social Psychology*, 43 (5), 997–1013.

Amabile, T.M., 1996. *Creativity in context.* Boulder, CO: Westview Press.

Besemer, S.P., 1998. Creative product analysis matrix: testing the model structure and a comparison among products-three novel chairs. *Creativity Reserach Journal*, 11 (4), 333–346.

Besemer, S.P., 2006. *Creating products in the age of design: how to improve your new product ideas!* Stillwater, OK: New Forums Press.

Besemer, S. and O'Quin, K., 1986. Analyzing creative products: refinement and test of a judging instrument. *The Journal of Creative Behavior*, 20 (2), 115–126.

Besemer, S.P. and O'Quin, K., 1999. Confirming the three-factor creative product analysis matrix model in an American sample. *Creativity Research Journal*, 12 (4), 287–296.

Besemer, S.P. and Treffinger, D.J., 1981. Analysis of creative products: review and synthesis. *The Journal of Creative Behavior*, 15 (3), 158–178.

Boden, M. 2004. *The creative mind: myths and mechanisms*. New York: Routledge.

Bray, R.M., Kerr, N.L., and Atkin, R.S., 1978. Effects of group size, problem difficulty, and sex on group performance and member reactions. *Journal of Personality and Social Psychology*, 36 (11), 1224–1240.

Butler, D.L. and Kline, M.A., 1998. Good vs. creative solutions: comparison on brainstorming, hierarchical, and changing perspectives heuristics. *Creativity Research Journal*, 11 (4), 325–331.

Christiaans, H.H., 2002. Creativity as a design criterion. *Creativity Research Journal*, 14 (1), 41–54.

Chulvi, V., *et al*., 2012. Comparison of the degree of creativity in the design outcomes using different design methods. *Journal of Engineering Design*, 23 (4), 241–269.

Cronin, J., 2001. *Process for facilitating the conception of inventions in a directed manner*. U.S. Patent Application 20010034629, filed 12 February 2001, and issued 25 October 2001.

Cropley, A. and Cropley, D., 2005. Engineering creativity: a systems concept of functional creativity. *In*: J. Kaufman and B. John, eds. *Creativty across domains: faces of the muse*. Mahway, NJ: Lawrence Erlbaum Associates, 169–185.

Dean, D.L., *et al*., 2006. Identifying quality, novel, and creative ideas: constructs and scales for idea evaluation. *Journal for the Association of Information Systems*, 7 (10), 646–649.

Diehl, M. and Stroebe, W., 1987. Productivity loss in brainstorming groups: toward the solution of a riddle. *Journal of Personality and Social Psychology*, 53 (3), 497–509.

Eisenman, M.L., 2006. *Essays on aesthetic innovation.* New York: Columbia University, Graduate School of Business.

Fink, A. and Neubauer, A.C., in press. Neuroscientific approaches to the study of creativity. *In*: J.S. Gero, ed. *Proceedings of the international workshop on studying design creativity.*

Goldschmidt, G. and Tatsa, D., 2005. How good are good ideas? Correlates of design creativity. *Design Studies*, 26 (6), 593–611.

Gray, D., Brown, S., and Macanufo, J., 2010. *Gamestorming: a playbook for innovators, rulebreakers, and changemakers.* Sebastopol, CA: O'Reilly Media.

Guilford, J.P., 1956. The structure of intellect. *Psychological Bulletin*, 52 (4), 267–293.

Horn, D. and Salvendy, G., 2006. Consumer-based assessment of product creativity: a review and reappraisal. *Human Factors and Ergonomics in Manufacturing*, 16 (2), 155–175.

Horn, D. and Salvendy, G., 2009. Measuring consumer perception of product creativity: impact on satisfaction and purchasability. *Human Factors and Ergonomics in Manufacturing*, 19 (3), 223–240.

Kelly, T., 2001. *The art of innovation: lessons in creativity from IDEO. America's leading design firm*. New York: Doubleday.

King, A.M. and Sivaloganathan, S., 1999. Development of a methodology for concept selection in flexible design strategies. *Journal of Engineering Design*, 10 (4), 329–349.
Kudrowitz, B., Te, P., and Wallace, D., 2012. The influence of sketch quality on perception of product-idea creativity. *Artificial Intelligence for Engineering Design, Analysis and Manufacturing*, 26 (3), in press.
Luecke, R. and Katz, R., 2003. *Managing creativity and innovation.* Boston, MA: Harvard Business School Press.
MacCrimmon, K.R. and Wagner, C., 1994. Stimulating Ideas Through Creative Software. *Management Science*, 40 (11), 1514–1532.
Mednick, S.A., 1962. The Associative Basis of the Creative Process. *Psychological Review*, 69 (3), 220–232.
Michalko, M., 2006. *Thinkertoys.* Berkeley, CA: Ten Speed Press.
Nowak, S. and Ruger, S., 2010. How reliable are annotations via crowdsourcing: a study about inter-annotator agreement for multi-label image annotation. *MIR '10 proceedings of the international conference on multimedia information retrieval*, 29–31 March, Philadelphia, PA. New York: ACM, 557–566.
Nussbaum, B., Berner, R., and Brady, D., 2005. Get creative! How to build innovative companies. *Bloomberg Businessweek*, 1 August, 60–68.
O'Quin, K. and Besemer, S., 1999. Creative products. *In*: M.A. Runco and S.R. Pritzker, *Encyclopedia of creativity.* San Diego, CA: Academic Press, 267–278.
Osborn, A., 1963. *Applied imagination: principles and procedures of creative problem solving.* New York: Charles Scribners Sons.
Paolacci, G., Chandler, J., and Ipeirotis, P., 2010. Running experiments on mechanical turk. *Judgment and Decision Making*, 5 (5), 411–419.
Patton, A. and Bannerot, R., 2002. Chindogu: a problem solving strategy for transforming uselessness into fearlessness. *Proceedings of the 2002 annual conference of the ASEE Gulf southwest section*. Lafayette, LA: American Society for Engineering Education.
Pauling, L., Aired June 1st, 1977. NOVA, Season 4, Episode 17. Interview. *In*: *Linus Pauling, crusading scientist.* Directed and Produced by Robert Richter. WGBH-Boston TV. 57 minutes.
Paulus, P., Kohn, N., and Arditti, L., 2011. Effects of the quantity and quality instructions on brainstorming. *The Journal of Creative Behaviour*, 45 (1), 38–46.
Reinig, B.A. and Briggs, R.O., 2006. Measuring the quality of ideation technology and techniques. *HICSS-39*, 4–7 January, Kauai, HI. Washington, DC: IEEE Computer Society, 20 pp.
Reinig, B.A., Briggs, R.O., and Nunamaker, J.F., 2007. On the measurement of ideation quality. *Journal of Management Information Systems*, 23 (4), 143–161.
Rietzschel, E.F., Nijstad, B.A., and Stroebe, W., 2006. Productivity is not enough: a comparison of interactive and nominal groups in idea generation and selection. *Journal of Experimental Social Psychology*, 42 (2), 244–251.
Runco, M.A. and Charles, R.E., 1993. Judgments of originality and appropriateness as predictors of creativity. *Personality and Individual Differences*, 15 (5), 537–546.
Shah, J.J. and Vargas-Hernandez, N., 2003. Metrics for measuring ideation effectiveness. *Design Studies*, 24 (2), 111–134.
Shah, J.J., Kulkarni, S.V., and Vargas-Hernandez, N., 2000. Evaluation of idea generation methods for conceptual design: effectiveness metrics and design of experiments. *Journal of Mechanical Design*, 122 (4), 377–384.
Silverstein, D., Samuel, P., and DeCarlo, N., 2008. *The innovator's toolkit: 50+ techniques for predictable and sustainable organic growth*. Hoboken, NJ: Wiley.
Simon, H., 1973. A. structure of ill structured problems. *Artificial Intelligence*, 4 (3–4), 181–201.
Snow, R., *et al.*, 2008. Cheap and fast – but is it good? Evaluating non-expert annotations for natural language tasks. *EMNLP 2008 proceedings of the conference on empirical methods in natural language processing*. Stroudsburg, PA: ACL, 254–263.
Sosa, R. and Gero, J., 2003. Design and change: a model of situated creativity. *In*: C. Bento, A. Cardoso and J.S. Gero, eds. *Proceedings of the IJCAI Creativity Workshop*, 9–10 August, Acapulco, Mexico. 25–34.
Torrance, P.E., 1972. Predictive validity of the torrance tests of creative thinking. *The Journal of Creative Behavior*, 6 (4), 236–252.
van der Lugt, R., 2003. Relating the quality of the idea generation process to the quality of the resulting design ideas. *In*: A. Folkeson, K. Gralen, M. Norell, and U. Sellgren, eds. *Proceedings of 14th international conference on engineering design*, 19–21 August, Stockholm. Glasgow, UK: Design Society, 19–21.
Ward, W.C. and Cox, P.W., 1974. A field study of nonverbal creativity. *Journal of Personality*, 42 (2), 202–219.