

Interplay of Sketching & Prototyping in Early Stage Product Design

Qifang Bao¹, Daniela Faas², Maria Yang¹

*1. Department of Mechanical Engineering, Massachusetts Institute of Technology,
Cambridge, MA 02139, USA*

2. Olin College of Engineering, Needham, MA 02492, USA

Corresponding author:

Maria C. Yang

Department of Mechanical Engineering, Massachusetts Institute of Technology

Room 3-449B, 77 Massachusetts Avenue, Cambridge, MA 02139-4307, USA

E-mail: mcyang@mit.edu

Tel: 617-324-5592

Qifang Bao

Department of Mechanical Engineering, Massachusetts Institute of Technology

Room 3-446, 77 Massachusetts Avenue, Cambridge, MA 02139-4307, USA

E-mail: qfbao@mit.edu

Tel: 617-417-8604

Daniela Faas

Olin College of Engineering

Needham, MA 02492, USA

E-mail: Daniela.Faas@olin.edu

Tel: 781-292-2554

Interplay of Sketching & Prototyping in Early Stage Product Design

Research suggests that, for the design of simple mechanisms, sketching and prototyping are somewhat interchangeable in terms of their influence on idea quantity and quality. This study explores whether this interchangeability holds true for a consumer product design activity. Three conditions are compared: sketching only, prototyping only, and free prototyping & sketching. Eighteen novice designers participated in a one-hour individual design activity. Their resulting design ideas were evaluated by both design experts and potential users. A design evaluation metric, *idea distance*, is proposed to measure the breadth and depth of design space exploration. Results showed that, individuals who only sketched, on average, generated more ideas, explored broader design space, and had more novel final designs. However, participants who were allowed to both sketch and build prototypes explored the design space in more depth and tended to have final ideas that were perceived as more creative. Individuals who only prototyped generated designs that were perceived to be aesthetically more pleasing and performed better functionally. Exploring broader design space was found to correlate with more unique ideas. However, exploring too broad a design space reduced the depth of idea exploration, and was negatively linked to the functional performance of the final designs.

Keywords: Sketching, Prototyping, Idea Generation, Creativity, Design Idea Evaluation

1. Introduction

Sketching and prototyping are two tools that are frequently used to explore design concepts. A common strategy is for a designer to first use sketching to generate a large quantity of ideas, and then follow with a phase in which physical prototypes are created for deeper investigation of a subset of those ideas.

One reason that physical prototyping typically follows sketching is the relative speed and cost of sketching compared to prototyping. This reflects a strategy of minimizing resources spent when design uncertainty is the highest and progressively increasing resources as uncertainty is reduced (Eppinger & Ulrich, 1995). However, there are cases when this strategy may not hold. Recent work (Faas, Bao, & Yang, 2014)

suggests that, for a simple mechanical design task, sketching and prototyping are nearly interchangeable in terms of their influence on idea quantity and quality. In that study, the resulting designs were simple enough that prototypes were often direct, 3-D instantiations of sketches, such as a fulcrum or pulley.

This paper builds on that work and a follow-on study (Bao, Faas, & Yang, 2016) to explore the interplay of sketching and prototyping in a different context, that of product design. The expectation is that consumer products will be user-oriented and potentially more complex, rather than strictly function-oriented, thus the design process and outcome will be less obviously interchangeable between sketching and prototyping. A specific subset of prototypes was examined in this study. These are preliminary prototypes constructed with materials that are easy and fast to work with and thus facilitate explorative “conversation with materials” (Schön, 1992). These low-fidelity prototypes reduce the cost of failure, allowing practitioners to learn with lower risk, thus support creativity.

The following questions are asked:

- 1. In early stage design, how does sketching compare with prototyping in terms of the design idea exploration process and the quality of design outcome?*

It is expected that sketching will be linked to a higher quantity of designs, as well as more creative designs, because it is generally a faster, lower overhead tool as demonstrated in related studies (Neeley, Lim, Zhu, & Yang, 2013; Schütze, Sachse, & Römer, 2003). It is also anticipated that prototyping will be linked to better functional performance as it allows participants to have more realistic, physical testing of their designs (Viswanathan & Linsey, 2012).

2. *How do the breadth and depth of idea exploration influence design outcome?*

Does exploring narrower vs. broader, shallower vs. deeper design space relate to design creativity, functional performance and etc.?

The expectation is that both exploring broader and deeper design space will be linked to better design outcome.

To address these questions, an experiment was conducted with eighteen student designers who were presented with a design task and then asked to address it under one of three conditions: sketching only, prototyping only, or a combination of the two. The resulting designs were rated by five researchers, as well as evaluated via crowdsourcing.

2. Background

2.1 Design tools and idea generation

2.1.1 Sketching

Sketching is a flexible tool for designers to externalize their thinking (Ullman, Wood, & Craig, 1990). It enables visual reasoning and gives rise to new ideas in a process of interactive imagery (Goldschmidt, 1991, 1994). It also helps designers to gain deeper understanding of the design problem by visualizing their current solutions (Schön & Wiggins, 1992). In addition, sketching serves as an important medium of communication between designers (Ferguson, 1994) by allowing the quick capture and communication of ideas while preserving design freedom (Goel, 1995). Though it has been shown that sketching is not necessary for expert designers to develop ideas in the early phases of conceptual design (Bilda, Gero, & Purcell, 2006), it is generally believed that sketching facilitates designers in idea generation and development (Song & Agogino, 2004). Evidence has been found that design teams who were allowed to sketch performed better than teams that were not (Schütze et al., 2003). Importantly, it

is conjectured that sketching supports team creativity by providing integrated group process as well as by facilitating individual creativity (Van der Lugt, 2005).

2.1.2 Prototyping

Like sketching, the building of physical prototypes enables design exploration, but it can uncover issues that cannot be observed from a 2D visualization (Sass & Oxman, 2006; Schrage, 1993). Prototyping enables the evaluation of a design's quality (Houde & Hill, 1997), allows for proof-of-concept (Ullman, 2002), supports communication between stakeholders (Budde, Kautz, Kuhlenkamp, & Züllighoven, 1992), and supplements designers' mental models (Viswanathan & Linsey, 2012). The speed of constructing physical prototypes can vary widely. Simple prototypes, which are the focus of this study, can be built quickly, sometimes faster than sketching (Häggman, Tsai, Elsen, Honda, & Yang, 2015). These prototypes can still be informative and allow designers to learn from failures (Dijk, Vergeest, & Horváth, 1998; Gerber & Carroll, 2012). In fact, it has been observed in a design course that the creation of simpler prototypes was linked to better final designs (Yang, 2005).

2.1.3 Comparing Sketching and Prototyping

Substantial work exists examining sketching and prototyping separately as tools to facilitate design concept generation, design space exploration, and design team communication. However, limited research has compared them directly. Vidal, Mulet, and Gómez-Senent (2004) found that sketching allowed design teams to create more diverse ideas during brainstorming on functional design problems, however teams that prototyped created more valid ideas. Häggman et al. (2015) found that prototyping allowed individual designers to generate ideas more quickly than sketching or CAD in a product design task, and designs created by prototyping were perceived as more novel,

more aesthetically pleasing, and more comfortable to use. These two studies compared sketching and prototyping for idea generation, but didn't explore the interplay of the two tools. Viswanathan and Linsey (2013) compared sketching only with sketching and prototyping simultaneously and found that building physical models was linked to higher quality ideas. However, this work didn't investigate the role of sketching during idea generation or consider prototyping only. We believe more studies are needed to gain an understanding of the advantages and disadvantages of these two tools and the roles they play during design. This study attempts to fill this gap by comparing the performance of the two design tools in terms of both the efficacy of idea generation process and the quality of the resulting designs.

2.2 Design idea evaluation

2.2.1 Design creativity

Amabile, Conti, Coon, Lazenby, and Herron (1996) defined creativity as “the development of novel and useful concepts”, a bipartite definition that is broadly accepted by the research community (Runco & Jaeger, 2012). Even though most ideation methods encourage designers to think blue sky and suspend judgement based on feasibility (Pierce & Pausch, 2007), being “useful” is core to product creativity (Dean, Hender, Rodgers, & Santanen, 2006; Oman & Tumer, 2009), since engineering designs “serves purposes” and “perform tasks and solve problems” (Cropley & Cropley, 2005).

Consensual assessment technique (CAT), developed by Amabile (1982) and expanded by J. Baer, Kaufman, and Gentile (2004), has been a prevailing research method to assess creativity. In CAT, a panel of domain experts first evaluates the creativity of products individually, and then assessments are averaged across the

individuals, or they confer with one another to reach consensus (Kaufman, Baer, Cole, & Sexton, 2008). This method has been widely applied in assessing creative products including literary works, art and design.

From a user-centered design perspective, the user's evaluation of design ideas is as important as the expert assessment as their perceptions are directly related to their desire to own the product (Pérez Mata, Ahmed-Kristensen, & Yanagisawa, 2013). Evidence has been found that the user perception of design ideas can be significantly influenced by the mode and fidelity of the design representations (Häggman et al., 2015). Thus extra care needs to be taken to avoid evaluation bias.

2.2.2 Idea generation efficacy

Metrics commonly used to evaluate idea generation efficacy include: novelty, variety, quality and quantity (Shah, Vargas-Hernandez, & Smith, 2003). Research suggests that higher quantities of ideas is linked to better quality ideas (Linsey et al., 2011; Yang, 2009). In the early stage of design process, many ideation methods aim to generate large quantities of novel ideas, such as TRIZ (Altshuller, Shulyak, & Rodman, 1999), synectics (Gordon, 1961), and design-by-analogy (Linsey, Markman, & Wood, 2012).

According to Shah et al. (2003), novelty measures how unusual an idea is, while variety measures the amount of exploration of the solution space. The less similar a new idea is to existing ideas, the more novel it is; the less similar ideas are in a group of ideas, the higher variety of ideas there is. To measure the similarity between any pairs of two design ideas, McAdams and Wood (2002) created a quantitative metric by computing the inner product between vectors representing product functionality. In another study, latent semantic analysis was applied to a patent database to investigate the distance between existing solutions and new design problems (Fu et al., 2013). It

was found that analogical stimuli that are neither too ‘near’ nor too ‘far’ are most likely to inspire creative design ideas.

Clearly, not all ideas generated during ideation can be pursued, and the ability to discern the quality of ideas is important (Girotra, Terwiesch, & Ulrich, 2010). Though we do not discuss it here, idea generation is typically followed by a phase of concept selection to winnow down ideas for further development.

3. Methods

Eighteen participants were recruited from engineering departments at a US university to participate in an in-lab design experiment. Participants worked individually, and were randomly assigned to three groups. The groups were given the same design task but different tools: a sketching only group, a prototyping only group, and a prototyping & sketching group that was permitted to use both tools as desired. The experimental process is captured in Figure 1. Prior to the design task, participants provided demographic information such as design experience and familiarity with prototyping methods; if needed, brief training on prototyping techniques was given. Phases I and II were the active phases of the design task and are further discussed below. After the experiment, designs were tested for their reliability and all participants were asked to complete a debriefing survey.

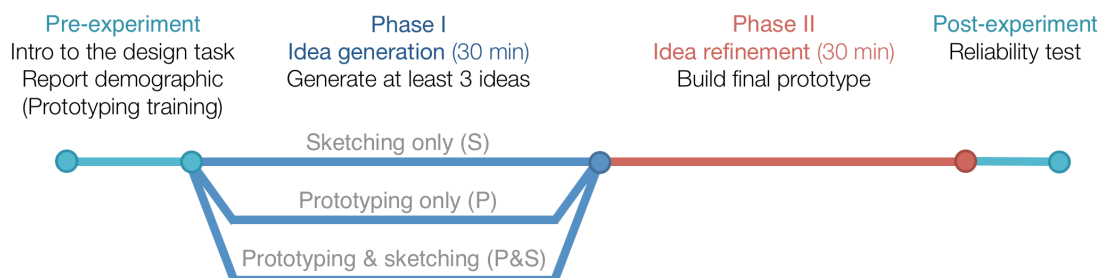


Figure 1 Flowchart of experimental process.

3.1 Participant Recruitment

In order to attract design-oriented students, recruitment emails were sent to both undergrad and graduate students in the Mechanical Engineering Department and the email list of an undergraduate dorm with a maker space. In addition, recruitment flyers were posted in the Mechanical Engineering Department. No previous design experience was required of participants. A \$5 gift card was provided as a token incentive. Those who responded to the recruitment emails or posters participated in the experiment, forming a self-selected sample.

3.2 Design Task

Participants were asked to design a package that could hold both a sandwich and a cup of coffee so that a user could carry both of them with one free hand. They were provided a sandwich (7"x3"x1") and a lidded coffee cup (10-oz, filled with water) at the beginning of the experiment, and were told that their designs would be tested with these items at the end of the experiment (Figure 2). This design task was developed because: 1) participants would likely be familiar with the experience of transporting a sandwich and a drink, 2) pilot testing suggested that participants would reasonably be able to sketch or prototype a design in the time allotted, and 3) the task was open enough that a wide variety of solutions would be possible.



Figure 2 Coffee cup and sandwich provided during the experiment. Both were filled with weights to simulate actual beverages and food items.

Pens, pencils and paper were provided for sketching. Prototyping materials included cardstock, foam-core, blue foam, wire, string, rubber bands, and Popsicle sticks. Tools including scissors, an X-acto knife, a ruler, a cutting mat, a glue gun, a hotwire cutter, and tapes were provided (Figure 3). The participants were supplied with unlimited amount of materials to reduce potential constraints on idea generation induced by limited resources.



Figure 3 Experimental setup including tools, materials and a timer.

The design task was split into two 30-minute phases. These time periods were chosen based on pilot testing that they were long enough for participants to finish the design task and short enough to avoid fatigue. Moderate time pressure has been found to foster creativity given a supportive environment (Amabile, Hadley, & Kramer, 2002; M. Baer & Oldham, 2006).

Phase I: Participants were randomly assigned to one of three conditions. In each condition, participants were asked to generate as many solutions to the design prompt as possible, with a minimum of three.

- Sketching only (S) - generate designs only by sketching. These participants were not allowed to build prototypes, but were permitted to touch and examine building material and tools.
- Prototyping only (P) - generate and explore designs only by prototyping, but not allowed to sketch.
- Prototyping and sketching (P&S) - allowed to sketch and prototype as desired to generate their ideas.

Phase II: All participants were instructed to build a prototype of their final design. The prototypes would undergo a reliability test in which the participants would use their prototypes to carry the sandwich and the cup for at least 30 seconds while walking at least 20 feet. A design “passed” the reliability test if the prototype did not break and no food/drink was spilled. This second phase of the experiment encouraged participants to iterate on their initial designs. It also allowed fair comparison between the design outcomes of the different conditions. If different design outcomes were observed, it could be concluded that the differences were introduced by sketching or prototyping during phase I.

3.3 Re-Sketching

The final designs from the second phase were all re-sketched by a professional industrial designer. Re-sketching ensured that the style and format of the designs would be consistent, and thus minimizing potential evaluation bias caused by different prototyping materials and craftsmanship levels of the participants. The re-sketched drawings showed the package itself, a sandwich, a cup, and a hand holding the package, indicating how the designs would be loaded with the sandwich and coffee, and how users would carry the package (Figure 4).

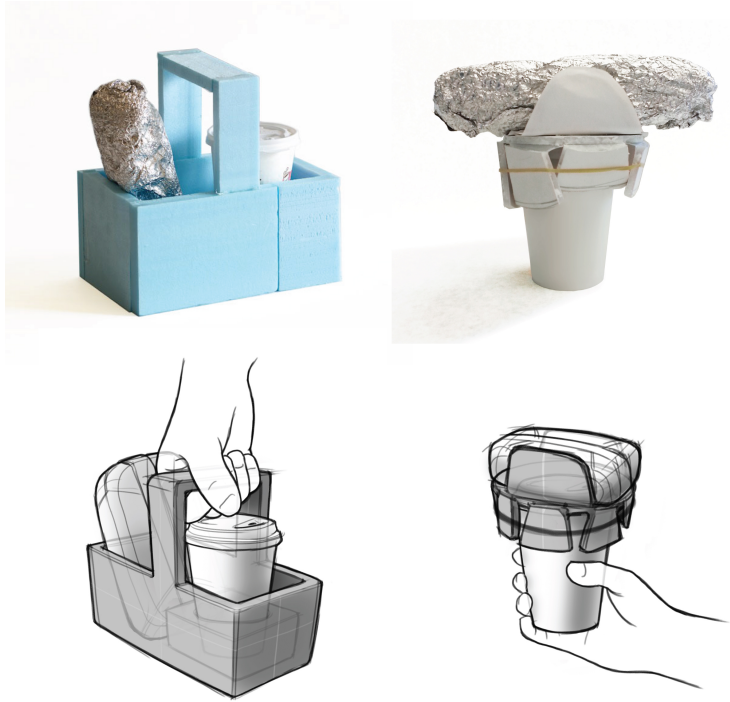


Figure 4 Top row shows examples of phase II prototypes. Bottom row shows re-sketched designs. Designs are shaded in grey, whereas hand and containers are in black/white.

3.4 Design Configuration Coding

Design ideas generated in both phases were carefully analyzed, and five key product attributes of differentiation were identified, each with two to four different configuration options. The descriptions and options of the attributes are summarized in **Table 1**.

Table 1 Package Configuration Coding Scheme

Attributes	Descriptions	Configuration Options
Package-Cup Relation	How does the package hold the cup?	<ul style="list-style-type: none"> • From the bottom • From the side • From the top • The package and the cup are detached
Cup-Sandwich Relation	What is the relative position between the cup and the sandwich?	<ul style="list-style-type: none"> • Side-by-side • The cup is on top of the sandwich • The cup is beneath the sandwich • The cup and the sandwich are detached

Sandwich Orientation	What is the orientation of the sandwich when in the package?	<ul style="list-style-type: none"> • The sandwich is being held upright • The sandwich is lying flat
Rigidity	Is the package rigid or flexible?	<ul style="list-style-type: none"> • Rigid (made of hard materials) • Flexible (made of soft materials) • Combination of rigid and flexible
User Interface	How should a user carry the package?	<ul style="list-style-type: none"> • The package has a handle or other features to hold • The package doesn't have significant features for holding but it can be grabbed easily • It's easier to carry the whole thing by holding the cup • The package interacts with parts of the body other than hands

Three evaluators (faculty, grad and undergrad student design researchers) independently coded all designs from both phase I and II. The designs were presented to the evaluators in random order. If the evaluators felt none of the predefined configurations matched the design, they were able to self-define the configuration with an “other” option. Krippendorff’s Alpha was used to evaluate the inter-rater reliability of the coding (Krippendorff, 2004). When there was disagreement among evaluators, the configuration chosen by the majority was used as the final.

3.5 Idea Distance

A new metric, the *design idea distance*, was formulated to measure the similarity between any pair of two ideas based on their configuration coding:

$$D_{i,j} = \sum_c d_{ij_c} \quad (1)$$

$D_{i,j}$ is the distance between a design idea i and another idea j , where d_{ij_c} is a dummy variable that denotes whether idea i and idea j are the same or not regarding configuration c . If they are the same, $d_{ij_c} = 1$; if not, $d_{ij_c} = 0$. Since there are five configuration attributes, the idea distance between any two designs can vary from 0 to 5. For example, if two ideas have the same *Package-cup relation*, *Cup-sandwich relation*

and *Sandwich orientation*, but have different *Rigidity* and *User interface*, then the distance between these two ideas is 2.

This idea distance measurement maps the ideas to a design space using the key attributes. It is similar to a metric developed by McAdams and Wood (2002) that compares the similarity of ideas in pairs. However, in this paper, the configurations of product features are compared directly with each other, instead of comparing the importance of product functions to user needs.

By computing different combinations of idea distance, different aspects of idea generation can be evaluated. In this paper, three measurements are calculated for each designer:

- Intrinsic idea distance:

$$D_{Intrinsic_p} = \frac{1}{n^2} \sum_{i_p=1}^n \sum_{j_p=1}^n D_{i_p,j_p} \quad (2)$$

where D_{i_p,j_p} is the distance between idea i and idea j of participant p , and n is the total number of ideas generated by participant p . This is a diversity metric given by the average distance among the design ideas of an individual participant, and applies to phase I. The Intrinsic idea distance measures the diversity of the pool of designs created by each individual designer in phase I. The larger this measurement is, the more diverse the pool of ideas is, and the broader the design space the designer had explored.

- Sequential idea distance:

$$D_{Sequential_p} = \frac{1}{n-1} \sum_{t=1}^{n-1} D_{t_p,(t+1)_p} \quad (3)$$

where $D_{t_p,(t+1)_p}$ is the distance between the t^{th} and the $t+1^{th}$ ideas generated by participant p . Again, n is the total number of ideas generated by participant p . This is an average of the idea distances between any consecutive pair of ideas of an

individual participant. Applies to phase I. The sequential idea distance measures how much a participant's ideas evolved. A high sequential idea distance indicates that the participant's ideas evolved at a fast pace; and a low sequential idea distance indicates that the participant was developing his/her ideas more steadily, making only small changes each time, or exploring designs of the same configuration with multiple trials before moving on to another configuration. Thus a low sequential idea distance could be regarded as a deeper exploration of the design space.

- Extrinsic idea distance:

$$D_{Extrinsic_p} = \frac{1}{N} \sum_{q=1}^N D_{p,q} \quad (4)$$

where $D_{p,q}$ is the distance between the phase II idea of participant p and the idea of participant q , and N is the total number of participants. This is a metric of uniqueness represented by the average distance of a design idea from all other designers' ideas. Relevant to phase II. The extrinsic idea distance is identical to Shah's Novelty measurement (Shah et al., 2003) when there is only one design stage plus all functions are weighted equally (derivation see Appendix). Thus it measures the uniqueness of any idea among a group of ideas. For the rest of the paper, *Novelty* will be used interchangeably when referring to this measurement.

3.6 Design evaluation

Two methods, the expert rating and the user evaluation, were used to assess the quality of the phase II design outcomes as well as to validate each other.

Method 1: Expert rating

Five design researchers (faculty, grad and undergrad) individually rated the phase II ideas according to the following criteria on a scale of 1-5.

- Creativity: which design looks more original/creative?
- Aesthetics: which design looks more aesthetically pleasing?
- Loading: which design looks easier to load the sandwich and cup into?
- Carrying: which design looks more comfortable to carry?
- Stability: which design looks more stable when holding the sandwich and cup?
- Storage: which design looks easier to store in large volumes at a restaurant or during shipping before it's given to a customer?

These criteria were formed based on a selection of Garvin's dimensions of product quality (Garvin, 1984): aesthetics, features, and performance, which could be evaluated based on sketches of preliminary design ideas. Loading, Carrying, Stability, and Storage criteria evaluate the features and performance of a design. The Creativity criterion emphasizes the originality of a design. The consistency of the ratings was measured by Cronbach's Alpha (Cronbach, 1951). The average rating for each criterion was calculated across the five researchers and considered as the expert evaluation result for the design ideas.

Method 2: User preference

To collect feedback on the designs from users, a survey was distributed on Amazon Mechanical Turk, a crowdsourcing platform. The same six criteria as the expert evaluation were included in the survey, as we believed they were important and reasonable for the users to assess. In addition, participants were asked to evaluate designs on:



- Likely-to-use: which design would you be more likely to use as a customer, if provided by a cafe or food truck?
- Better idea: which design is in general a better idea?

Detailed explanations of all criteria were introduced to the respondents at the beginning of the survey.

Rating individual designs would be hard for non-expert evaluators since they lacked a standard to assess preliminary design ideas, especially on design creativity, and thus ratings from different respondents would likely be inconsistent. On the other hand, ranking all 18 designs would be time consuming and impose cognitive burden on the respondents (Häggman et al., 2015). Therefore, pairwise comparison was chosen for respondents to evaluate design ideas. For each question, two re-sketched designs were presented side-by-side and the respondents would indicate the one that better addressed each criterion. Optional comment boxes were provided for respondents to explain the reasons for their choices. An example question is shown in Figure 5.

To ensure response quality, several control questions were interspersed throughout the survey (Kittur, Chi, & Suh, 2008; Mason & Suri, 2012). For example, respondents were randomly asked to “Click the left/right circle” when comparing two designs. After the first and last questions the respondents could be asked to describe one of the two designs they just saw to ensure that they were indeed looking at the designs. In addition, only respondents with a 99% approval rate on Amazon Mechanical Turk, which is an indication of past reliability, were allowed to take the survey.

Each respondent were presented with ten pairs of designs. The first and last pairs of design concepts were fixed in order to simplify quality checking. The remaining eight pairs of designs were randomly selected. Only the responses of the randomized eight pairs of designs were analyzed.

Which of the two designs

	Left	Right
looks more original/creative ?	<input type="radio"/>	<input type="radio"/>
looks more aesthetically pleasing ?	<input type="radio"/>	<input type="radio"/>
looks easier to load the sandwich and cup into?	<input type="radio"/>	<input type="radio"/>
looks more comfortable to carry ?	<input type="radio"/>	<input type="radio"/>
looks more stable when holding the sandwich and cup?	<input type="radio"/>	<input type="radio"/>
looks easier to store in large volume for shipping and storage?	<input type="radio"/>	<input type="radio"/>
would you be more likely to use as a customer?	<input type="radio"/>	<input type="radio"/>
is in general a better idea ?	<input type="radio"/>	<input type="radio"/>
click the Left circle	<input type="radio"/>	<input type="radio"/>

Comments (optional):
 Please tell us why you think one design is better than the other; what do you like or dislike about the designs, and etc.

Figure 5 Example pairwise comparison for design evaluation survey

3.7 Interpretation of the pairwise evaluation survey results

3.7.1 Pairwise comparison consistency

Equation (5) was used to check the consistency of the pairwise comparison results (Häggman et al., 2015):

$$\frac{\sum_{\text{all pairwise comparison}} \max(\text{count}(D_{a_c} > D_{b_c}), \text{count}(D_{a_c} < D_{b_c}))}{\text{count}(\text{all pairwise comparison of creterion } c)} \quad (5)$$

where D_{a_c} and D_{b_c} referred to *Design a* and *b* evaluated on *Criterion c*, respectively. If all pairwise choices made by different respondents are the same, then the consistency will be equal to 1. On the other hand, if all choices are random, then the consistency will be close to 0.5, since for any pair of designs, each one will be chosen half of the times.

3.7.2 Discrete choice model

A logistic regression model was used to convert the pairwise choices into continuous measurements for each criterion.

According to the model, the probability of *Design a* being chosen over a *Design b* when compared with each other is:

$$P(D_{a_c} > D_{b_c}) = \frac{\exp(u_{a_c})}{\exp(u_{a_c}) + \exp(u_{b_c})} \quad (6)$$

where u_{a_c} and u_{b_c} are the implicit measurements for the *Criterion c* of *Design a* and *b* respectively. For example, when measuring the creativity of the designs, if $u_{a_{creativity}}$ is larger than $u_{b_{creativity}}$, then *Design a* is considered to be more creative than *Design b*, and is more likely that *Design a* will be chosen over *Design b* regarding creativity.

To estimate the implicit measurements, the cross entropy loss (equation (7)) was minimized with L2 regularization:

$$Loss_{cross\ entropy_c} = \sum_{i=1}^K y_{i_c} \cdot \log(P_{i_c}) \quad (7)$$

where y_{i_c} is the actual choice made by respondents when evaluating designs on *Criterion c*; P_{i_c} is the corresponding probability of such choice being made according to equation (6), and K is the total number of pairwise comparisons.

3.7.2 Discrete choice model accuracy

To check how well the discrete choice models fit the data, the estimated measurements u_{x_c} were used to predict the pairwise choice results with Equation (6). If $P(D_{a_c} > D_{b_c}) > 0.5$, then it is predicted that *Design a* will be chosen over *Design b*; if $P(D_{a_c} > D_{b_c}) < 0.5$, then it is predicted that *Design b* will be chosen. The percentages of correct predictions are referred to as the model accuracies.

The higher the prediction accuracy, the better a model explains the choice results. A meaningless model would provide prediction accuracy of around 0.5, since it is not any different from randomly guessing the results.

4. Results

In total, 18 people (10 females and 8 males; 8 undergraduate students, 9 graduate students and a postdoc) participated in this study. Fourteen participants were from mechanical engineering, the rest were from architecture, chemical engineering, computer science or urban studies. All but one participant had self-reported experience in either mechanical design or product design for no more than 5 years (8 had 2 years or less design experience, 9 had 3-5 years of design experience), and one participant had design experience of 5-10 years. Participants with different design experience levels were evenly distributed among the group settings. Each experimental group had 6 participants. One-third of the participants had little experience with the tools and materials and were briefly trained on how to use them.

In total, 93 design ideas were generated in phase I. The sketching only group generated 38 ideas, the prototyping only group generated 22, and the prototyping & sketching group generated 33. Eighteen final designs were generated in phase II. All final designs passed the reliability test.

For the configuration coding, Krippendorff's Alpha among the three evaluators was 0.669 for the first phase ideas and 0.740 for the second, which is sufficient to draw tentative conclusions. For phase II idea evaluation, the expert ratings had Cronbach's Alpha of 0.940, which is considered excellent to draw conclusion.

The user evaluation survey collected 299 responses from Amazon Mechanical Turk. Screening on the quality control questions yielded 259 valid responses. Among the respondents who provided valid responses, 159 were female, 99 were male, and 1 preferred self-defined gender; 55 were aged 18-24, 113 aged 25-34, 51 aged 35-44, and 40 aged 45 or above; 35 had master's or doctorate degree, 102 had bachelor's degree, 92 had some college credit or associate degree, 17 had high school or equivalent education, and 13 had professional degree or vocational training. On average, each design was presented around 70 times, and each pair of designs was compared 13.5 times.

The software R was used to analyze the findings. All statistical analysis is given in the following format: mean \pm standard error of the mean.

4.1 Phase I: Efficacy of design space exploration

Examples of phase I design concepts for each of the three conditions are shown in Figure 6.

Prototyping only: The three designs in the top row were created by one participant, and are shown in the order they were built from left to right. The first and second ideas were very different from each other, in terms of the configuration coding described earlier; as do the second and third idea. The first and third ideas however, share more similarity. These resulted in a high idea change rate of 4, and a relatively low idea intrinsic distance of 2.22.

Prototyping & Sketching: The three ideas in the second row were generated by a single participant who chose to only sketch. All three ideas were very similar to each other. Therefore the intrinsic idea distance and the sequential idea distance were 0.89 and 1 respectively, both very low.

Sketching only: The nine ideas in the bottom two rows were created by one participant. These ideas were diverse, thus have a high intrinsic idea distance of 3.28. However, ideas close to each other in the sequence shared more similarity. For example, the first three ideas all had a container take the sandwich and the cup side-by-side, and the user would hold the packages with a part of the body other than hands. Thus the sequential idea distance of these ideas was 2.88, lower than that of the prototyping only examples provided above.



Figure 6 Phase I ideas of three participants from three experimental groups

Table 2 shows the distribution and the within group mean values of the number of ideas generated, the intrinsic idea distance and the sequential idea distance of all participants in phase I, categorized by the experimental conditions. Kruskal-Wallis H test (also called one-way ANOVA on ranks) (Kruskal & Wallis, 1952) was utilized to detect any differences among the groups. This method was chosen over the traditional ANOVA because of the relatively small sample size and possible non-Gaussian distribution of the data.

Table 2 Distribution and statistics of phase I number of ideas, intrinsic idea distance, and sequential idea distance of the three groups

(P - prototyping only, P&S - prototyping and sketching, S - sketching only)

	Number of ideas	Intrinsic idea distance	Sequential idea distance
Group Mean±SE			
P	3.7±0.9	1.76±0.25	2.66±0.40
P&S	5.5±0.8	1.79±0.33	1.92±0.33
S	6.3±0.8	2.31±0.34	2.59±0.26
Kruskal-Wallis H test			
Chi-squared	4.8078	2.2743	2.8806
p-value	0.0904	0.3207	0.2369

The majority of the respondents in the prototyping only group generated three (or sometimes fewer) ideas in phase I, barely satisfying the minimum requirement. However, one participant in this group generated eight ideas, indicating it was not impossible to generate ideas quickly with prototyping. The sketching only group, on average, generated the most ideas, and the prototyping & sketching group was in between. The differences among the groups are significant for $p < 0.1$. The sketching only group had a higher average intrinsic idea distance, while the average distances of the other groups were close to each other. The differences between the three groups were not statistically significant. In terms of the sequential idea distance, the prototyping only and the sketching only groups had similar average results, while the prototyping & sketching group was on average lower. Again, these differences were not statistically significant.

The Pearson correlation coefficient between the number of ideas and the intrinsic idea distance was 0.750 ($p\text{-value} = 0.0003$); and between the intrinsic idea distance and the sequential idea distance was 0.689 ($p\text{-value} = 0.002$). This suggests that the more ideas generated by a participant in phase I, the more likely that they were diverse; also, generating diverse ideas were positively linked to a fast evolution of ideas, meaning that on average two ideas generated in sequence shared less similarity. The correlation between the number of ideas and the sequential idea distance was not significant, with a coefficient of 0.124 ($p\text{-value} = 0.625$). This has been partially illustrated by the prototyping only examples in Figure 6, indicating that a small group of designs can still have large change intensity.

4.2 Phase II: Final design evaluation

4.2.1 Pairwise comparison consistency and model accuracy

The consistency of the pairwise choices and the accuracy of the discrete choice models are summarized in Table 3. If the choices were made in random, then the consistency and the model accuracy will be around 50%. The discrepancy between the consistency and model accuracy is likely because that the intransitivity of preference exists to some extent, and cannot be captured by logistic model. For example, design A was preferred over design B when compared together, and design B was preferred over design C, however, design C was preferred over design A when evaluated together. The Pearson correlation between the user evaluation and the expert rating results are also included in Table 3.

Table 3 Pairwise choice consistency, discrete choice model accuracy, and correlation between expert and user evaluations

	Consistency	Model Accuracy	Expert and User Evaluation Correlation (p-value)
Creativity	71.5%	66.1%	0.485 (0.041)
Aesthetics	79.4%	77.3%	0.499 (0.035)
Ease of loading	82.7%	81.6%	0.831 (<0.001)
Ease to carry	81.3%	77.5%	0.633 (0.005)
Stability	80.5%	77.6%	0.697 (0.001)
Storage	72.6%	69.4%	0.473 (0.048)
Likelihood to use	81.7%	80.8%	-
Better idea	81.4%	80.7%	-

The creativity and storage evaluations had relatively low choice consistency as well as model accuracy. Also, the assessment of these two criteria from the experts and users were less consistent. Discrepancy between expert and non-expert evaluation for creativity has been observed earlier by Kaufman et al. (2008). Different evaluation

standards of the two populations were attributed as the reason for this discrepancy. The lower consistency between the storage evaluations might be caused by the challenge of estimating volume via only sketches.

Aesthetics also had low consistency between the experts and users evaluation.

Considering the fact that the re-sketches presented to the experts and users were created by different sketchers thus different in style, this inconsistency was expected. Since the user evaluation consistency for aesthetics is high, it was used instead of the expert ratings for further analysis.

4.2.2 Overview of user evaluation results

Correlation analysis among the different criteria of user evaluation was conducted and the results are summarized in Table 4.

Table 4 Pearson correlation coefficients (and p-values) between criteria of users' evaluation for phase II design ideas

(Correlations significant on level of 0.05 are highlighted with a grey background)

	Creativity	Aesthetics	Loading	Carrying	Stability	Storage	Likely-to-use	Better idea
Creativity	-	0.242 (0.333)	0.221 (0.377)	0.335 (0.174)	0.317 (0.200)	-0.177 (0.482)	0.246 (0.326)	0.329 (0.183)
Aesthetics	0.242 (0.333)	-	0.853 (<0.001)	0.928 (<0.001)	0.795 (<0.001)	0.546 (0.019)	0.974 (<0.001)	0.961 (<0.001)
Loading	0.221 (0.377)	0.853 (<0.001)	-	0.830 (<0.001)	0.801 (<0.001)	0.418 (0.085)	0.903 (<0.001)	0.882 (<0.001)
Carrying	0.335 (0.174)	0.928 (<0.001)	0.830 (<0.001)	-	0.735 (0.001)	0.447 (0.063)	0.929 (<0.001)	0.944 (<0.001)
Stability	0.317 (0.200)	0.795 (<0.001)	0.801 (<0.001)	0.735 (0.001)	-	0.135 (0.594)	0.835 (<0.001)	0.870 (<0.001)
Storage	-0.177 (0.482)	0.546 (0.019)	0.418 (0.085)	0.447 (0.063)	0.135 (0.594)	-	0.561 (0.015)	0.460 (0.055)
Likely-to-use	0.246 (0.326)	0.974 (<0.001)	0.903 (<0.001)	0.929 (<0.001)	0.835 (<0.001)	0.561 (0.015)	-	0.977 (<0.001)
Better idea	0.329 (0.183)	0.961 (<0.001)	0.882 (<0.001)	0.944 (<0.001)	0.870 (<0.001)	0.460 (0.055)	0.977 (<0.001)	-

User perceived creativity was not significantly correlated with any other evaluations, indicating users were not referring to any other criteria to assess design creativity. Interestingly, perceived aesthetics was positively correlated with all criteria except creativity and all p-values were smaller than 0.05. A potential explanation is that, designs perceived to be more attractive are more likely to be considered better functionally, which is supported by of Yamamoto and Lambert (1994).

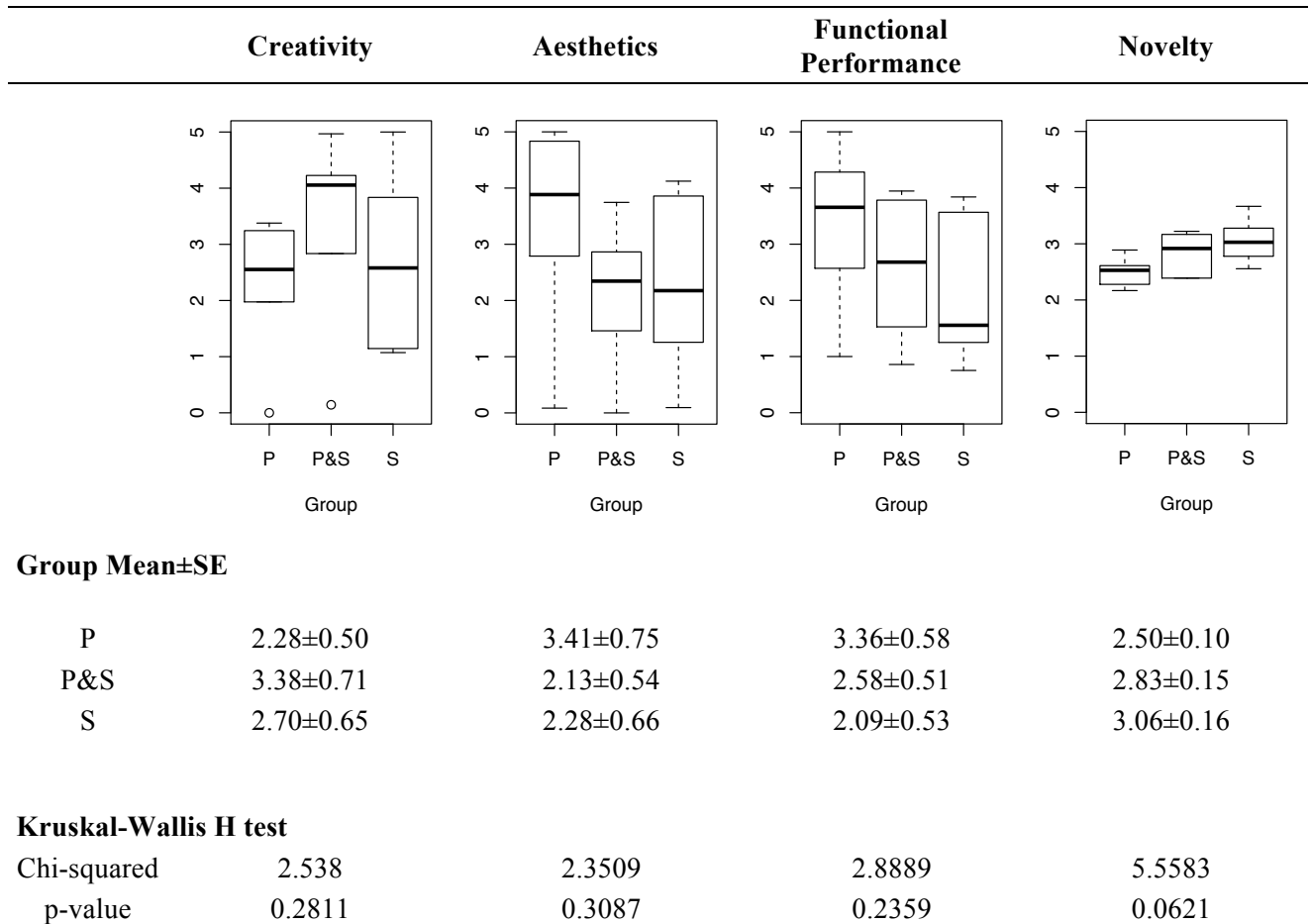
Evaluations for loading, carrying and stability were highly correlated with each other, while storage was less correlated with these three. Because storage also had lower evaluation consistency, it was excluded from further analysis. The average was calculated for the other three criteria to represent the functional performance of the designs.

The likely-to-use and better idea evaluations both provided an overall assessment for the designs. Thus it was not surprising that they were highly correlated with each other. Both of these evaluations were also highly correlated with the user perceived functional performance (loading, carrying and stability) and aesthetics of the designs.

The distribution of the creativity, aesthetics, and functional performance evaluations are summarized in Table 5. All evaluations were rescaled to the range of 0 to 5 to ease the comparison.

Table 5 Distribution and statistics of phase II idea evaluations (creativity, aesthetics, functional performance, and novelty) of the three groups

(P - prototyping only, P&S - prototyping and sketching, S - sketching only)



4.2.3 Idea creativity

The prototyping & sketching group had on average the highest user perceived creativity and the prototyping only group had on average the lowest. However, the Kruskal-Wallis H test showed these differences were not statistically significant (see Table 5).

Considering the moderate consistency of the creativity evaluation, the expert ratings were compared to the user evaluation. All designs were ranked according the two evaluations, and were divided into two groups: top-tier and bottom-tier designs. Designs that consistently were in the top or bottom tiers according to both the experts

and the users were investigated. Among these, three out of the five consistent top-tier designs were from the sketching & prototyping group, and four out of the five consistent bottom-tier designs were from the prototyping only group. This result matched the previous observation that P&S group had on average the highest creativity, and P group has on average the lowest creativity.

Table 6 Ranking comparison for the creativity evaluations

Group	Rank of User Evaluation	Rank of Expert Rating	Consistent Ranking
Prototyping Only	7	11	
	9	4	Top
	11	14	Bottom
	12	14	Bottom
	13	16.5	Bottom
	18	18	Bottom
Prototyping & Sketching	2	2	Top
	3	14	
	4	6	Top
	5	4	Top
	10	6	
	17	6	
Sketching Only	1	1	Top
	6	11	
	8	16.5	
	14	4	
	15	9	
	16	11	Bottom

Figure 7 shows an example design created by a participant in the free prototyping and sketching group. It ranked the second place according to both the expert rating and the user evaluation. However, its perceived aesthetics and functional

performance were lower, ranked 12th and 9th place respectively according to the user evaluation.



Figure 7 Design with high creativity evaluations and medium aesthetics and functional performance evaluations, from the prototyping & sketching group

4.2.4 Idea aesthetics and functional performance

Though the Kruskal-Wallis H tests didn't detect any significant differences on the average aesthetics and functional performance ratings between the three groups, the trend is clear that the prototyping only group had on average the highest evaluation for both (see Table 5).

It was observed that aesthetics and functional performance were highly correlated (with Pearson correlation coefficients of 0.929). However these two were not significantly correlated with creativity. Figure 8 presents an example design with high aesthetics and functional performance evaluation (both ranked the 2nd place according to the user evaluation), but low creativity evaluation (ranked the 12th and 14th place according to user and expert evaluations respectively).



Figure 8 Design with low creativity evaluation but high aesthetics and functional performance evaluations, from the prototyping only group.

4.2.5 Overall evaluations

Respondents tended to think designs that looked more aesthetically pleasing and had better functional performance were a better idea, and they would more prefer to use. For example, the design in Figure 8 had the highest and second highest evaluations for likely-to-use and better idea respectively. The design in Figure 7 ranked the 11th and 10th place according to the users' evaluation of likely-to-use and better idea.

Therefore, the prototyping only designs in general had the highest evaluation for likely-to-use and better idea, with mean values of 3.39 ± 0.75 and 3.47 ± 0.73 respectively; while the P&S designs had mean values of 2.28 ± 0.58 and 2.35 ± 0.65 respectively; and the S designs had mean values of 2.12 ± 0.65 and 2.38 ± 0.63 respectively.

4.2.6 Design idea novelty (extrinsic idea distance)

Phase II design ideas were measured by their extrinsic idea distance from each other, which was also Shah's novelty measurement. The distribution of the design idea novelty can also be found in Table 5. Design ideas from the sketching only condition had the highest novelty measurement on average, while those of the prototyping only group had

the lowest. These differences were significant on level of 0.1 according to the Kruskal-Wallis H test (see Table 5). Figure 9 shows an example of a design with the highest novelty evaluation.

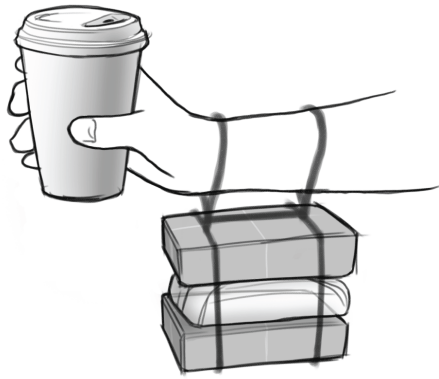


Figure 9 Design with high novelty, from the sketching only group

It's interesting to note that, while user perceived creativity of designs was not correlated with the novelty evaluation (Pearson correlation coefficient = 0.030, p-value = 0.908), the expert rating on creativity had higher consistency with novelty (Pearson correlation coefficient = 0.472, p-value = 0.048). Considering three out of the five researchers who rated design creativity had also encoded the design configurations based on which the novelty evaluation was calculated, it is likely that the expert evaluators constructed their creativity judgment more on the uniqueness of the design configurations, while the user evaluators did not.

4.3 Connecting the idea exploration process to the final idea outcome

To investigate the relationship between phase I design space exploration efficacy and the phase II final design outcomes, correlation analysis was conducted between the three criteria defined for phase I and the user evaluations for the phase II ideas.

Table 7 Pearson correlation coefficients (and p-values) between phase I design space exploration measurements and user evaluations of phase II design ideas

(Correlations significant on level of 0.05 are highlighted with a grey background)

		Phase II design evaluation					
		Creativity	Aesthetics	Functional Performance	Likely-to-use	Better idea	Novelty
Phase I design space exploration	Number of ideas	-0.042 (0.868)	-0.291 (0.241)	-0.294 (0.236)	-0.282 (0.258)	-0.269 (0.281)	0.484 (0.042)
	Intrinsic idea distance	-0.181 (0.281)	-0.495 (0.037)	-0.552 (0.017)	-0.502 (0.034)	-0.474 (0.047)	0.515 (0.029)
	Sequential idea distance	-0.178 (0.481)	-0.364 (0.137)	-0.475 (0.046)	-0.434 (0.072)	-0.400 (0.100)	0.136 (0.590)

User evaluations of aesthetics, functional performance, and overall qualities of the second phase designs were in general negatively correlated with the phase I number of ideas, the sequential idea distance, and especially the intrinsic idea distance. However, phase II idea novelty was positively correlated with these phase I idea exploration measurements, especially the number of ideas and the intrinsic idea distance. The relation between the user evaluated phase II idea creativity and phase I design exploration measurements were not clear.

These suggest that more ideas and more diverse ideas generated during the early phase linked to the high novelty and uniqueness of the final designs. However, these novel designs were not necessarily perceived as creative by the users, as mentioned previously. In addition, many and diverse ideas, as well as a fast evolving idea exploration process did not seem to be helpful for the perceived quality of the final designs. Given the limited time, generating design ideas that were too diverse and had little similarity with each other seemed to harm the designers' ability to delve deeply

into details of any single design idea. On the contrary, exploring ideas in depth, iterating on design ideas with fewer changes each time, seemed to be a better strategy to achieve favorable design outcomes.

4.4 Participants' design experience and experimental performance

To determine if the previous design experience of experimental participants had any impact on their performance, correlation analysis was run between the years of design experience (in four categories: 0 years, 1~2 years, 3~5 years and 6~10 years, treated as ordinal data) and all design evaluation criteria. Significant Spearman correlations were found between the design experience and the phase I number of ideas (Spearman's $\rho = 0.511$, $p\text{-value} = 0.030$), and between the design experience and the phase I intrinsic idea distance (Spearman's $\rho = 0.401$, $p\text{-value} = 0.099$).

However, none of phase II idea evaluation, including the expert and user evaluation and the design novelty, was significantly correlated with the design experience ($p\text{-value} > 0.1$). Specifically, the participant with the longest design experience (6~10 years) didn't appear to perform significantly better than the other participants: his phase II idea ranked the 2nd and the 4th places according to the expert and user evaluation on the creativity respectively; and ranked at best the 3rd place on other evaluation criteria, out of the six participants from the same experimental group.

These results suggest that, though the participants' design experience varied, they were ultimately novice designers without any professional experience. Their design experience did help them generate more and more diverse ideas in phase I; however, the impact was not carried over to phase II. Thus in this study, the design experience of the participants didn't significantly influence the experimental results, especially the quality of phase II ideas.

5. Discussion

This paper examined links between idea generation tools, design space exploration process, and design outcome. The experiment focused on a fast, provisional design-and-build activity. The results suggest the following answers to the proposed research questions:

1. *In early stage design, how does sketching compare with prototyping in terms of the design idea exploration process and the quality of design outcome?*

In this study, the use of sketching vs. prototyping in the early stages of a product design activity was found to make a difference in how designers generate ideas and explore the idea space.

As expected, in phase I of the experiment, the sketching only group generated the most ideas and the most diverse ideas, since sketching was a faster, easier tool to use and had fewer constraints. Limiting to prototyping only largely reduced the number and the diversity of the pool of ideas generated by each designer. The number of ideas generated by the sketching & prototyping group participants was on average in between of that of the other two groups. While the participants had the option of choosing sketch only to generate more ideas, most instead chose to use the two tools iteratively: draw some ideas on paper, play with the materials to test ideas, and then draw more. These iterations slowed down the pace of their idea generation process, but allowed them to explore the ideas more in depth. The low sequential idea distance of the prototyping & sketching group reflected this in-depth idea exploration pattern.

Almost all participants developed their final designs for phase II based on one or two ideas that were generated in the first 30 min of the experiment. Thus it was not surprising that the tools used in phase I were also linked with the final design outcomes, even though the sketching and prototyping limitation was only applied to the phase I. In

phase II, designs from the prototyping only group had on average highest perceived functional performance, and designs from the sketching only had on average the lowest. This indicates that prototyping early in the design process helped designers to explore what features would function better. The prototyping & sketching group appeared to have the most creative phase II ideas. Actually, it was observed that the sketching only group had more creative ideas in phase I. However, the most creative ideas might not be chosen for the second phase. Instead, less creative but more feasible ideas were selected. This phenomenon of abandoning creative ideas has been observed in previous research (Starkey, Toh, & Miller, 2016; Toh & Miller, 2016) and is certainly reasonable within the context of this study. Many of these ‘creative’ phase I ideas (e.g. a drone to carry the sandwich and drink) could not be feasibly built with the provided tools and material.

2. How do the breadth and depth of idea exploration influence design outcome?

Does exploring narrower vs. broader, shallower vs. deeper design space relate to design creativity, functional performance and etc.?

The phase I intrinsic idea distance was used to evaluate the breadth of design space exploration of each participant. The larger the intrinsic idea distance, the bigger design space explored. It was found that the intrinsic idea distance was significantly, positively correlated with the phase II idea novelty; however, it was significantly, negatively correlated with the perceived phase II idea aesthetics, idea functional performance, and the idea overall evaluations. This suggests that exploring broader design space early in the design process linked with the uniqueness of the final design. However, exploring too broad of a design space may have made it less likely for designs to delve more deeply into details and improve the designs’ functional performance, given the constraint of time. Time pressure is unavoidable in an industrial setting, where a designer is likely to have limited time to work on a design problem. Thus the

exploration of design space broadly or in-depth needs to be carefully balanced.

Sequential idea distance was used to measure the depth of idea exploration. A smaller sequential idea distance was considered to represent deeper idea exploration. The sequential idea distance significantly correlated with the perceived functional performance in a negative way, which provides evidence that exploring ideas in depth by iterating on design ideas with fewer changes each time was linked to the generation of better functioning designs. These findings are consistent with previous research that good designs are developed through series of highly interlinked ideas (Goldschmidt & Tassa, 2005), and more experienced designers pursue design thoughts more deeply than student designers (Suwa & Tversky, 1997).

One limitation of this study is that the number of experimental participants was low and varied from novices with no previous design experience to postgraduate designers. Even though designers with different experience were spread across the three groups and analysis ruled out the potential influence of this factor, this is a preliminary study and results and conclusions need to be further validated with a larger sample size.

6. Conclusion

In this study, a two-phase experiment was conducted to understand the role of two tools in the early stage of designing a product: sketching and prototyping. The breadth and depth of design space exploration were investigated using new measurements of intrinsic idea distance and sequential idea distance. The outcome designs were encoded by their configurations and evaluated by researchers as well as potential users.

Sketching-only was linked to the generation of more ideas and more diverse ideas in the early stage of the product design activity, which represents the exploration of a broader design space, and likely contributed to the higher novelty of their final

designs. Prototyping-only designers tended to test their ideas earlier through fabrication, which seemed to encourage the generation of more feasible ideas. Their final ideas tended to be perceived as more aesthetically pleasing and having better functional performance by potential users. Free sketching and prototyping allowed designers to iterate on design ideas with both tools, thus helped them to generate more ideas and explore the ideas more in depth. Consistent expert and user evaluations indicated that the free sketching and prototyping group generated the most creative ideas. These findings highlight the advantage and disadvantage of using either sketching or prototyping as the idea generation tools and cast light on how the interplay of the both can improve the idea generation efficacy, especially design creativity.

It was also found that exploring broader design space appeared to help designers generate more unique ideas. However, given the constraint of time, exploring too broad a design space reduced designers' chance of exploring the design ideas in depth. This might harm the functional performance of the final designs. This result suggests that, given the constraint of time, it is important to balance the idea quantity and diversity with the idea thoroughness as to achieve an optimal design outcome.

Acknowledgements:

The work described in this paper was supported in part by the National Science Foundation under Award CMMI-1334267. The opinions, findings, conclusions, and recommendations expressed are those of the authors and do not necessarily reflect the views of the sponsors. This work was also supported in part by Singapore University of Technology and Design (SUTD) - MIT graduate Fellowship. The authors would like to thank Julia Rue and Thomas Nelson for their expert assistance, and Ben Coble for his drawing expertise.

References:

- Altshuller, G. S., Shulyak, L., & Rodman, S. (1999). *The Innovation Algorithm: TRIZ, Systematic Innovation and Technical Creativity*: Technical Innovation Center, Inc.
- Amabile, T. M. (1982). The Social Psychology of Creativity: A Consensual Assessment Technique. *Journal of Personality and Social Psychology* 43(no. 5), 997–1013
- Amabile, T. M., Conti, R., Coon, H., Lazenby, J., & Herron, M. (1996). Assessing the Work Environment for Creativity. *The Academy of Management Journal*, 39(5), 1154-1184
- Amabile, T. M., Hadley, C. N., & Kramer, S. J. (2002). Creativity under the gun. *Harvard business review*, 80, 52-63
- Baer, J., Kaufman, J. C., & Gentile, C. A. (2004). Extension of the consensual assessment technique to nonparallel creative products. *Creativity research journal*, 16(1), 113-117
- Baer, M., & Oldham, G. R. (2006). The curvilinear relation between experienced creative time pressure and creativity: moderating effects of openness to experience and support for creativity. *Journal of Applied Psychology*, 91(4), 963
- Bao, Q., Faas, D., & Yang, M. (2016). Interplay of sketching & prototyping in early stage product design *The fourth international conference on design creativity*. Atlanta, GA.
- Bilda, Z., Gero, J. S., & Purcell, T. (2006). To sketch or not to sketch? That is the question. *Design studies* 27(5), 587-613
- Budde, R., Kautz, K., Kuhlenskamp, K., & Züllighoven, H. (1992). Prototyping *Prototyping* (pp. 33-46): Springer.
- Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *psychometrika*, 16(3), 297-334
- Cropley, D., & Cropley, A. (2005). Engineering creativity: A systems concept of functional creativity. *Creativity across domains: Faces of the muse*, 169-185
- Dean, D. L., Hender, J. M., Rodgers, T. L., & Santanen, E. (2006). Identifying good ideas: constructs and scales for idea evaluation
- Dijk, L., Vergeest, J. S., & Horváth, I. (1998). Testing shape manipulation tools using abstract prototypes. *Design Studies*, 19(2), 187-201
- Eppinger, S. D., & Ulrich, K. T. (1995). *Product design and development*.
- Faas, D., Bao, Q., & Yang, M. (2014). Preliminary Sketching and Prototyping: Comparisons in Exploratory Design-and-build Activities. *ASME International Design Engineering Technical Conferences, Buffalo, NY, 2014*
- Ferguson, E. S. (1994). *Engineering and the Mind's Eye*: MIT press.
- Fu, K., Chan, J., Cagan, J., Kotovsky, K., Schunn, C., & Wood, K. (2013). 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. doi: 10.1115/1.4023158
- Garvin, D. A. (1984). What Does “Product Quality” Really Mean? *Sloan management review*, 25
- Gerber, E., & Carroll, M. (2012). The psychological experience of prototyping. *Design Studies*, 33(1), 64-84. doi: 10.1016/j.destud.2011.06.005
- Girotra, K., Terwiesch, C., & Ulrich, K. T. (2010). Idea generation and the quality of the best idea. *Management science*, 56(4), 591-605
- Goel, V. (1995). *Sketches of thought*: MIT Press.
- Goldschmidt, G. (1991). The dialectics of sketching. *Creativity research journal*, 4(2), 123-143

- Goldschmidt, G. (1994). On visual design thinking: the vis kids of architecture. *Design studies*, 15(2), 158-174
- Goldschmidt, G., & Tatsa, D. (2005). How good are good ideas? Correlates of design creativity. *Design Studies*, 26(6), 593-611
- Gordon, W. J. J. (1961). *Synectics*: New York: Harper & Row.
- Häggman, A., Tsai, G., Elsen, C., Honda, T., & Yang, M. C. (2015). Connections Between the Design Tool, Design Attributes, and User Preferences in Early Stage Design. *Journal of Mechanical Design*, 137(7), 071101. doi: 10.1115/1.4030181
- Houde, S., & Hill, C. (1997). What do prototypes prototype. *Handbook of human-computer interaction*, 2, 367-381
- Kaufman, J. C., Baer, J., Cole, J. C., & Sexton, J. D. (2008). A Comparison of Expert and Nonexpert Raters Using the Consensual Assessment Technique. *Creativity Research Journal*, 20(2), 171-178. doi: 10.1080/10400410802059929
- Kittur, A., Chi, E. H., & Suh, B. (2008). Crowdsourcing user studies with Mechanical Turk *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 453-456): ACM.
- Krippendorff, K. (2004). *Content analysis: An introduction to its methodology*: Sage.
- Kruskal, W. H., & Wallis, W. A. (1952). Use of ranks in one-criterion variance analysis. *Journal of the American statistical Association*, 47(260), 583-621
- Linsey, J. S., Clauss, E., Kurtoglu, T., Murphy, J., Wood, K., & Markman, A. (2011). An Experimental Study of Group Idea Generation Techniques: Understanding the Roles of Idea Representation and Viewing Methods. *Journal of Mechanical Design*, 133(3)
- 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. doi: 10.1115/1.4006145
- Mason, W., & Suri, S. (2012). Conducting behavioral research on Amazon's Mechanical Turk. *Behavior research methods*, 44(1), 1-23
- McAdams, D. A., & Wood, K. L. (2002). A Quantitative Similarity Metric for Design-by-Analogy. *Journal of Mechanical Design*, 124(2), 173. doi: 10.1115/1.1475317
- Neeley, L., Lim, K., Zhu, A., & Yang, M. C. (2013). Building Fast to Think Faster: Exploiting Rapid Prototyping to Accelerate Ideation During Early Stage Design. *ASME International Design Engineering Technical Conferences, Portland, OR, 2013*
- Oman, S., & Tumer, I. Y. (2009). The Potential of Creativity Metrics for Mechanical Engineering Concept Design. *International Conference on Engineering Design (ICED'09), Stanford, CA, USA*
- Pérez Mata, M., Ahmed-Kristensen, S., & Yanagisawa, H. (2013). Perception of aesthetics in consumer products *19th International Conference on Engineering Design* (pp. 527-536).
- Pierce, J., & Pausch, R. (2007). Generating 3D Interaction Techniques by Identifying and Breaking Assumptions. *Virtual Reality*, 11(1), 15-21
- Runco, M. A., & Jaeger, G. J. (2012). The standard definition of creativity. *Creativity Research Journal*, 24(1), 92-96
- Sass, L., & Oxman, R. (2006). Materializing Design: the Implications of Rapid Prototyping in Digital Design. *Design Studies*, 27(3), 325-355.
- Schön, D. A. (1992). Designing as reflective conversation with the materials of a design situation. *Knowledge-based systems*, 5(1), 3-14

- Schön, D. A., & Wiggins, G. (1992). Kinds of seeing and their functions in designing. *Design studies*, 12(2), 22
- Schrage, M. (1993). The Culture (s) of Prototyping. *Design Management Review*, 4(1), 55-65
- Schütze, M., Sachse, P., & Römer, A. (2003). Support Value of Sketching in the Design Process. *Research in Engineering Design*, 14(2), 89-97
- Shah, J., Vargas-Hernandez, N., & Smith, S. (2003). Metrics for Measuring Ideation Effectiveness. *Design Studies*, 24(2), 111-134
- Song, S., & Agogino, A. M. (2004). Insights on Designers' Sketching Activities in New Product Design Teams. *ASME International Design Engineering Technical Conferences, Salt Lake City, UT, 2004*
- Starkey, E., Toh, C. A., & Miller, S. R. (2016). Abandoning creativity: The evolution of creative ideas in engineering design course projects. *Design Studies*, 47, 47-72
- Suwa, M., & Tversky, B. (1997). What do architects and students perceive in their design sketches? A protocol analysis. *Design studies*, 18(4), 385-403
- Toh, C. A., & Miller, S. R. (2016). Choosing creativity: the role of individual risk and ambiguity aversion on creative concept selection in engineering design. *Research in Engineering Design*, 27(3), 195-219
- Ullman, D. G. (2002). *The mechanical design process*: McGraw-Hill Science/Engineering/Math.
- Ullman, D. G., Wood, S., & Craig, D. (1990). The Importance of Drawing in the Mechanical Design Process. *Computers & Graphics*, 14(2)
- Van der Lugt, R. (2005). How sketching can affect the idea generation process in design group meetings. *Design studies*, 26(2), 101-122
- Vidal, R., Mulet, E., & Gómez-Senent, E. (2004). Effectiveness of the Means of Expression in Creative Problem-Solving in Design Groups. *Journal of Engineering Design*, 15(3), 285-298
- Viswanathan, V. K., & Linsey, J. S. (2012). Physical models and design thinking: A study of functionality, novelty and variety of ideas. *Journal of Mechanical Design*, 134(9), 091004
- Viswanathan, V. K., & Linsey, J. S. (2013). The Role of Sunk Cost in Engineering Idea Generation: An Experimental Investigation. *Journal of Mechanical Design*, 135(12), 121002. doi: 10.1115/1.4025290,
- Yamamoto, M., & Lambert, D. R. (1994). The impact of product aesthetics on the evaluation of industrial products. *Journal of Product Innovation Management*, 11(4), 309-324
- Yang, M. C. (2005). A Study of Prototypes, Design Activity, and Design Outcome. *Design Studies*, 26(6), 649-669. doi: 10.1016/j.destud.2005.04.005
- Yang, M. C. (2009). Observations on Concept Generation and Sketching in Engineering Design. *Research in Engineering Design*, 20(1), 1-11 doi: 10.1115/1.4003498

Appendix Derivative of Extrinsic idea distance to Shah's Novelty metrics

The extrinsic idea distance of idea i is:

$$D_{Extrinsic_i} = \frac{1}{T} \sum_{j=1}^T D_{ij} = \sum_c \left(\frac{1}{T} \sum_{j=1}^T d_{ij_c} \right) = \sum_c \frac{T - C_{i_c}}{T} = \sum_c S_{i_c}$$

Where T is the total number of ideas and C_{i_c} is the count of the same solution as idea i on configuration c . This is exactly Shah's Novelty measurement, with only one design stage ($m = 1$ and $f_j = 1$) and same weights to all functions ($p_k = 1$):

$$M = \sum_{j=1}^m f_j \sum_c S_{jc} p_c = \sum_c S_c$$

Table Content

Table 1 Package Configuration Coding Scheme

Table 2 Distribution and statistics of phase I number of ideas, intrinsic idea distance, and sequential idea distance of the three groups

Table 3 Pairwise choice consistency, discrete choice model accuracy, and correlation between expert and user evaluations

Table 4 Pearson correlation coefficients (and p-values) between criteria of users' evaluation for phase II design ideas

Table 5 Distribution and statistics of phase II idea evaluations (creativity, aesthetics, functional performance, and novelty) of the three groups

Table 6 Ranking comparison for the creativity evaluations

Table 7 Pearson correlation coefficients (and p-values) between phase I design space exploration measurements and user evaluations of phase II design ideas

Figure Content

Figure 1 Flowchart of experimental process.

Figure 2 Coffee cup and sandwich provided during the experiment. Both were filled with weights to simulate actual beverages and food items.

Figure 3 Experimental setup including tools, materials and a timer.

Figure 4 Top row shows examples of phase II prototypes. Bottom row shows re-sketched designs. Designs are shaded in grey, whereas hand and containers are in black/white.

Figure 5 Example pairwise comparison for design evaluation survey

Figure 6 Phase I ideas of three participants from three experimental groups

Figure 7 Design with high creativity evaluations and medium aesthetics and functional performance evaluations, from the prototyping & sketching group

Figure 8 Design with low creativity evaluation but high aesthetics and functional performance evaluations, from the prototyping only group.

Figure 9 Design with high novelty, from the sketching only group