CrowdConcept
assessing the crowd’s creative capacity

by

Jeffrey S. Mekler

Submitted to the Engineering Systems Division and Department of Mechanical Engineering on February 3, 2014, in partial fulfillment of the requirements for the degrees of

Master of Science in Technology and Policy
Master of Science in Mechanical Engineering

Abstract

Crowdsourcing creative work – the process of outsourcing a creative task to a large online community or crowd – has grown in popularity over the last decade and is now considered a serious design strategy. However, there has been little research examining how the composition of the crowd and the nature of the crowd’s interactions affect the crowd’s creative capacity. Building upon related findings in the group creativity literature, this thesis examines the effects social diversity, cognitive diversity, and communication capability on the crowd’s ability to generate high quality design concepts. A design challenge to “reimagine the public restroom experience” was developed, and challenge participants were placed into crowds with varying levels of diversity and communication capability. Findings from the study show that social diversity and cognitive diversity had positive effects on the quality of participants’ concepts. Furthermore, diverse, communication-capable crowds generated higher quality concepts than participants in crowds with no communication capability, but low diversity, communication-capable crowds performed worse than crowds with no communication capability. These findings suggest that designers of CCW platforms should consider the diversity of the crowd when deciding how much communication capability to build into crowdsourcing platforms. Future work is needed to investigate the generalizability of these findings for different problem types.
Acknowledgements

To my advisor, David Wallace: thank you for helping me rediscover my creative confidence. Your passion for design teaching and your selfless dedication to MIT’s students inspires me to make an impact through design and to help others find their own creative confidence so that they too can make a positive impact on this world.

To Lucienne Blessing, Warren Seering, and Maria Yang: thank you for introducing me to the world of design research. Your advice and guidance have had an impact throughout the pages of this thesis.

To my fellow students and friends in the CADLAB: your camaraderie, support, and (too) frequent Nerf battles, Frisbee sessions, and Flour trips have made the late nights and long hours in the dark dungeon that our office infinitely more enjoyable. I feel honored to work among such a creative, competent, and inspired group of people.

To Alissa: you’ve transformed how I look at the world and think about design in ways that I cannot begin to describe. Thank you for always being there, listening, and asking questions.

Finally, to my parents: thank you for your love and constant support. From my earliest plans to travel into space, you’ve always believed in my ability to succeed in whatever I set my mind to, even when I doubted myself.
Table of Contents

Chapter 1
  Introduction 6

Chapter 2
  Examples of Crowdsourced Creative Work 9
    Threadless: crowdsourced graphic design 10
    InnoCentive: solutions to complex, technical problems 10
    OpenIDEO: design for social impact 11
    Quirky: social product development 12
    DARPA FANG: open design of an infantry fighting vehicle 13
    Summary of CCW Examples 14

Chapter 3
  Literature Review 16
    The Nature of Creative Work 16
    Is Crowdsourcing Creative Work a Viable Strategy? 21
    The Creative Capacity of the Crowds 29

Chapter 4
  Experimental Design 39
    Reimagining the Public Restroom Experience 39
    Experimental Design 44
    Measuring Diversity 48
    Study Procedures 59
Chapter 1
Introduction

No matter who you are, most of the smartest people work for someone else.

- Bill Joy, co-founder of Sun Microsystems

It has long been recognized that the firm’s capability to innovate is critical to its long-run existence. Competition, evolving market demands, and a desire to increase financial returns all encourage the firm to continually improve upon its current products, services, and processes.

To sustain innovation, the classic private investment model of innovation argues the firm should invest financial and human capital into research and development with the hope that these investments generate financial returns in the form of new products, services, and processes (Chesbrough 2003; Ulrich and Eppinger 2012). When the firm does not have enough internal innovation capacity, this process may be outsourced to innovation consultants. In either case, innovation occurs within the closed boundaries of the firm and its partners.

In contrast to these closed innovation strategies, researchers have begun to suggest that open innovation models, which allow the firm to capture knowledge and expertise that exist outside the firm’s boundaries, can, under certain conditions, be more effective than traditional models (Chesbrough 2003). Under the umbrella of open innovation, a number of related strategies exist, such as lead-user innovation (Von Hippel 2005), open-source development (Balka 2011), and crowdsourcing (Surowiecki 2005; Howe 2006). With the introduction of these new innovation models, two important questions arise:

Selection When should the firm use an open innovation strategy?
Design How can the selected open innovation strategy be utilized to maximize its innovation returns?

This thesis aims to contribute to the limited but growing body of literature about the design of crowdsourcing models. Crowdsourcing is method where the firm outsources a task, in the format of an open call for submissions, to a large collection of individuals or firms who collectively constitute the “crowd” (Howe 2006). The term is a portmanteau of “crowd” and “outsourcing.”
Today, crowdsourcing is one of the more popular models of open innovation, and there are numerous examples of successful crowdsourcing platforms. Quirky, a consumer products manufacturer, has successfully leveraged its online community of nearly 400,000 people to generate and develop concepts for over 300 consumer products (Quirky 2013). The growth of crowdsourcing’s popularity is due in part to the prevalence of Internet access and the growth online communities. These two factors have made it very easy for firm to broadcast problems to a large, global audience.

In the design community, there has been increasing interest in using crowdsourcing in the concept generation phase of the design process. In this phase, many design concepts are generated to address latent user needs (Ulrich and Eppinger 2012). However, understanding user needs and translating those needs into design concepts is a “wicked” process (Buchanan 1992). Empirical evidence suggests that crowdsourcing can help manage this wickedness by engaging individuals with diverse perspectives in the concept-generation process. For example, recent study found that a firm’s customers generated more novel concepts for a new infant feeding product than the firm’s own internal product development team (Poetz and Schreier 2012).

Studying the Crowds

While limited empirical evidence suggests that crowdsourcing may be an effective concept-generation strategy, the underlying mechanisms that govern its success or failure are still poorly understood. Popular examples of crowdsourcing reveal that there are few best practices geared towards improving the crowd’s creative output. For example, some crowdsourcing models encourage collaboration among members of the crowd, while others force the crowd’s members to work in isolation. Furthermore, the crowd has been regarded as a largely anonymous collection of individuals, and there has been little effort to understand how the composition of the crowd affects ideation performance.

This thesis aims to clarify how the composition of the crowd and the nature of collaboration affect the crowd’s ability to generate creative concepts. These issues are examined from perspective of group creativity research, which studies the creative performance of teams in organizational settings. Group creativity research has highlighted the countervailing effects of social diversity and cognitive diversity on team performance (Dahlin 2005). Cognitive diversity is cited for having a generally positive effect on team performance as it allows the team to understand the problem from many perspectives and prevents the team from becoming fixated on one perspective. Social diversity is thought to have a more deleterious effect, particularly when diversity is high, because it erodes perceived team cohesion and leads to social categorization. Since crowds are ostensibly similar to teams, these findings suggest that social and cognitive diversity may affect the crowd’s ideation performance in a similar fashion.
However, there are important differences between crowds and teams that need to be considered. Unlike a team, members of the crowd may not innately have any capacity to communicate with each other. In these situations, the crowd is analogous to many individuals working independently to generate concepts. The creativity literature has examined the ideation performance of these so-called isolated or nominal groups, where individuals work in isolation. The findings from the creativity literature indicate that nominal groups often generate better concepts than team because social diversity and other group failures limit the team’s ideation effectiveness (Kavadias and Sommer 2009). This suggests that communication capability may affect the crowd’s performance and mediate the effects of cognitive and social diversity.

This thesis presents findings from a study developed to examine the effects of communication capability and diversity on the crowd ideation performance. A design challenge was created to “reimagine the public restroom experience.” Participants in the design challenge were placed into crowds with varying levels of diversity and communication capability to see how these factors affected the quality of participants’ concepts. Findings from the study suggest that both social and cognitive diversity have positive effects on ideation performance. Crowds with greater communication capability did not perform significantly better than nominal crowds. However, when the communication capable crowds were highly diverse, they performed significantly better than the nominal crowd.

Chapter 2 presents several examples of crowdsourced creative work (CCW). The examples serve to familiarize the reader with current crowdsourcing practices and examine key differences in the design of CCW models.

Chapter 3 discusses the nature of creative work and reviews research from the crowdsourcing literature discussing the conditions where CCW is a viable innovation strategy. Then, findings from group creativity research regarding the effects of collaboration and diversity are discussed. Based on these findings and a comparison between teams and crowds, three hypotheses are developed regarding the effects of social diversity, cognitive diversity, and communication capability on the crowd’s ideation performance.

Chapter 4 explains the development of the public restroom design challenge, experimental methods, and a metric used to measure social diversity and cognitive diversity.

Chapter 5 presents findings from the research study.

Chapter 6 discusses implications for the design of CCW platforms and future group creativity research. Limitations of the study and opportunities for future research are also explored.
Chapter 2
Examples of Crowdsourced Creative Work

Following the publication of Howe's (2006) Wired article on the rise of crowdsourcing, there have been numerous examples of crowdsourcing in popular media. From the development of Phylo (McGill University 2012) and Foldit (Moore 2011) – computer games that harnesses normal people's pattern-recognition abilities to help scientists solve problems in genetic biology – to the rise of Kickstarter (2014) and Indiegogo (2014) – crowdfunding sites that help entrepreneurs and artists fund their projects through small donations and advanced purchases – it is clear that crowdsourcing has wide reach across different domains and problems. While these examples all share some similarities characteristic to crowdsourcing (an open call for participation, freely revealed information, etc.), they differ substantially when one examines the nature of the task performed by the crowd and the incentives provided to participate.

This is no less true in the realm of crowdsourced creative work (CCW), where the crowd is asked to generate novel and useful solutions to a problem. This chapter highlights several notable examples of crowdsourced creative work. The examples reveal key differences in the design of the crowdsourcing models for creative work. These differences demonstrate that models of CCW can be adapted to work across industry settings for a variety of design problems.
Threadless: crowdsourced graphic design

Threadless.com is an online retailer of community-designed apparel, which runs design challenges, such as the one depicted in Figure 1, soliciting graphic designs for t-shirts, backpacks, and other apparel. Anyone can submit their own design, which is then reviewed by the Threadless community of over 2.5 million users. Following a one-week review period, the most popular designs are printed and sold on Threadless’ online marketplace. The creators of winning designs are typically awarded cash, credit towards future Threadless.com purchases, and prizes in-kind. Since its founding in 2000, Threadless users have submitted more than 500,000 designs, approximately 1% of which have been selected for production (Threadless 2013).

InnoCentive: solutions to complex, technical problems

In 2007, the Oil Spill Recovery Institute (OSRI) posted a challenge on InnoCentive, which asked solvers to develop a method to separate oil and water that had become frozen together (InnoCentive 2007). The solution came from an unusual person: an expert from the concrete industry named John Davis. Davis suggested that OSRI use vibration – a method widely used in the concrete industry to prevent concrete from hardening – to prevent the oil from freezing (InnoCentive 2007). Using pre-existing tools from the concrete industry, OSRI researchers were able to develop a method to separate oil and water that had become frozen together.
industry, OSRI confirmed that Davis’ solution would work and awarded him $20,000 (InnoCentive 2007). Since 2001, InnoCentive has administered more than 1,650 challenges focused on solving complex research and development problems. Nearly 85% of challenges have been solved, and solvers have earned over $40 million (InnoCentive 2013).

An InnoCentive challenge closely follows the format of a blind contest. Solvers work independently and are not able to see other solvers’ submissions. The challenge poster reviews each submission and awards prizes if they determine that a solution meets the necessary criteria provided in the challenge. The deliverables for each challenge vary. In some cases, challenge posters are simply looking for creative ideas and in other cases posters may require detailed documentation and/or reduction to practice. The challenge poster is given discretion, and may award one, multiple, or no prizes. Prizes for winning solutions range from $5,000 to over $1 million depending on the complexity of the problem and the challenge poster’s desire to find a solution (InnoCentive 2013).

A notable characteristic of InnoCentive’s platform is the protections it provides for challenge posters and solvers. Large companies, with a strong interest in creating intellectual property and protecting their competitive advantage, post a number of challenges on InnoCentive. To protect these companies’ interests, challenge solvers are typically required to assign all intellectual property rights to the company if the solver submits a winning solution (Dean 2008). If the solution is not accepted, however, the solver retains all intellectual property rights to their ideas. InnoCentive also only reveals the challenge poster’s identity to solvers who submit a winning solution. This ensures that companies posting challenges on InnoCentive do not indirectly reveal details about their research and development to competitors.

OpenIDEO: design for social impact

OpenIDEO is an open innovation platform developed by IDEO, a well-known global design consultancy. OpenIDEO partners with non-profit organizations to tackle a wide range of social problems, such as: inspiring young people to develop creative confidence, finding ways to improve wellbeing for the elderly, and improving maternal health in low-income countries (IDEO 2013a).

---

**FIGURE 2: OPENIDEO’S STRUCTURED BRAINSTORMING PROCESS**

source: Zhang (2010)
OpenIDEO challenges follow a structured process, shown in Figure 2, to generate and evaluate concepts. During the “inspiration” phase, OpenIDEO members submit examples of existing design solutions that are insightful or relevant to the challenge. The inspiration phase serves to educate the community on prior art, help identify possible design strategies, and inspire possible solutions. After several weeks of inspiration, members of the community are allowed to submit their own concept solutions. Typically these concepts include several images and a few paragraphs of text to describe the concept. After a period of concepting, the challenge moves into the evaluation phase, where members of the community rate and review the concepts that have been submitted. The OpenIDEO team and challenge sponsors then select a short-list of the top-rated concepts for further review and select one or more of these concepts as the winning solutions.

Collaboration is at the core of OpenIDEO. When a user submits an inspiration or concept on OpenIDEO, it is immediately visible to the rest of the community. Commenting and applauding features allow the community to express their opinions about a submission and provide feedback to the user about possible improvements. Users are encouraged to build upon other users’ submissions. When a user submits a concept, they can reference other submissions that influenced their concept. The resulting network of influence can view on OpenIDEO’s website.

Quirky: social product development

Quirky is a consumer goods company that leverages the crowd to generate ideas for new consumer products, refine the product vision for these ideas, and evaluate the market demand for these products. Compared to other examples of CCW, Quirky is notable because it engages the crowd throughout a product’s development process. Quirky users pay a small fee to submit an idea to Quirky’s team for consideration. Each week, Quirky’s in-house design team chooses several ideas for further development. In some cases, Quirky’s design team generates design variations and asks the community to vote and comment on their favorite designs. In other instances, Quirky asks the community to get creative and generate possible names, taglines, and branding for the product. In the later phases of development, Quirky leverages the community to do pricing studies in order to identify a suitable sale price for the product.

Quirky also has a very interesting shared revenue model that incentives the Quirky community to collaborate and contribute throughout the development of the product. Ten percent of gross revenue from direct and indirect sales of Quirky products is shared with members of the community. This revenue is paid out to members based on their “influence” in the project. Quirky determines how much influence a member has by
awarding specific percentages of influence when a member contributes to the project in a specific way. For example, a user who submits an idea for consideration is awarded 40% influence; while a user who comes up with the name for the product is awarded 3% influence.

For some inventors, their participation in the Quirky community has been immensely profitable. Jake Zien, a student at RISD, had an idea for a new power strip that would work better with bulky power adapters. He submitted his idea on Quirky, and after a year of development and contributions from over 800 community members, Quirky released Pivot Power. Through his influence in the project and subsequent variations of Pivot Power, Jake has earned more that $500,000, and Quirky has paid out more than $1,000,000 to the Quirky community for their contributions to Pivot Power alone (Quirky 2014a; Quirky 2014b).

For some inventors, their participation in the Quirky community has been immensely profitable. Jake Zien, a student at RISD, had an idea for a new power strip that would work better with bulky power adapters. He submitted his idea on Quirky, and after a year of development and contributions from over 800 community members, Quirky released Pivot Power. Through his influence in the project and subsequent variations of Pivot Power, Jake has earned more that $500,000, and Quirky has paid out more than $1,000,000 to the Quirky community for their contributions to Pivot Power alone (Quirky 2014a; Quirky 2014b).

DARPA FANG: open design of an infantry fighting vehicle

The Fast Adaptable Next-Generation Ground Vehicle (FANG) program is a DARPA initiative that seeks to accelerate the design and development of advanced infantry fighting vehicles (IFVs) through open source design challenges (DARPA Tactical Technology Office 2011). The first FANG challenge – a $1 million competition to design the drivetrain of an IFV – engaged more than 1,000 participants and more than 200 teams (DARPA 2013). The winning team of three was geographically dispersed.

Compared to the other examples above, the DARPA FANG challenges are especially ambitious in terms of their scale and engineering complexity. The IFV and its subsystems will require the design, simulation, and assembly of thousands of parts, electronic
components, and software code. To obviate the need for physical prototyping and testing, participants in the FANG challenges are given free access to a number of advanced simulation and verification tools (DARPA Tactical Technology Office 2011). These tools, collectively called the META toolkit, make it possible to conduct trade studies, simulate the performance of components, and evaluate the manufacturability of designs in one environment. Just as software code is compiled and tested, the performance of each FANG design is evaluated through simulation in the META environment to determine whether it meets requirements and specifications (DARPA Tactical Technology Office 2011).

Summary of CCW Examples

The examples of CCW presented above suggest that crowdsourcing can be used in many different design domains. All of the examples share some common elements, such as an open call for submissions. However, the differences between the examples are potentially more interesting from a research perspective. The examples show that the design of crowdsourcing platforms should be closely tailored to the characteristics of the problem, the capabilities of the crowd, and the broader organizational and industry context. Some key characteristics from each example are summarized in Table 1.

<table>
<thead>
<tr>
<th>Domain</th>
<th>Innov. Model</th>
<th>Review Format</th>
<th>Incentives</th>
<th>Community</th>
</tr>
</thead>
<tbody>
<tr>
<td>Threadless</td>
<td>graphic design</td>
<td>competition</td>
<td>community</td>
<td>$1k - $3k</td>
</tr>
<tr>
<td>InnoCentive</td>
<td>science &amp; tech</td>
<td>blind contest</td>
<td>private</td>
<td>$10k +</td>
</tr>
<tr>
<td>OpenIDEO</td>
<td>social needs</td>
<td>collaboration</td>
<td>mixed</td>
<td>recognition</td>
</tr>
<tr>
<td>Quirky</td>
<td>products</td>
<td>collaboration</td>
<td>mixed</td>
<td>$0 - $500k</td>
</tr>
<tr>
<td>DARPA FANG</td>
<td>engineering</td>
<td>competition</td>
<td>private</td>
<td>$1,000,000</td>
</tr>
</tbody>
</table>

The examples suggest that models of CCW often use one of two innovation models: competitions and collaborations. In practice, innovation models span this continuum. At the competitive end of the spectrum is the InnoCentive model, where participants work independently and have no knowledge of other participants' progress. OpenIDEO and Quirky are at the collaborative end of the spectrum, and both strongly encourage the sharing and co-development of ideas.

The format for reviewing the crowd's ideas also differs across CCW models. On InnoCentive, only organizations posting a challenge can see and review submissions, while Threadless users are both the creators and reviewers of ideas.
The size, type, and distribution of incentives offered to the crowd also vary across the models. Monetary awards are common because they encourage users to generate their best ideas. While many competition-based crowdsourcing models only give awards to the top ideators, Quirky distributes royalties to each member of the crowd that contributes. Unlike the other platforms, which offer some form of monetary incentives, OpenIDEO's members are driven solely by their interest in OpenIDEO's challenges, their passion for IDEO, and their desire to work with and be recognized by other designers.
Chapter 3
Literature Review

Everyone designs who devises courses of action aimed at changing existing situations into preferred ones.
– Herb Simon, The Sciences of the Artificial

This chapter takes a critical look at the merits of crowdsourced creative work. Through a review of descriptive models and empirical research, the opportunities, limitations, and risks of CCW are examined. Open research questions are outlined, and from these open questions, the central hypotheses of the thesis are developed.

The Nature of Creative Work

While the examples of CCW discussed in Chapter 2 reveal that the crowd can generate creative concepts in a number of design domains, they provide no indication CCW is better than more traditional design processes used by firms. To assess whether crowdsourcing is a better innovation strategy, one must first understand the nature of design problems and the processes that designers use to find solutions to these problems.

Design Problems

The problems that practicing industrial designers, architects, and engineers encounter on a daily basis are often very different from the problems posed in university classrooms and examined in the science and math domains.

The Diving Board Example

Suppose an engineering student were asked to compute the deflection of a fixed-free steel beam with an applied tip load. How would the student’s teacher evaluate the student’s response? The student would be likely expected to recall the beam bending equations, compute bending moments and inertias, look up the elastic
properties of steel, and plug terms into the beam bending equations to compute the tip deflection. The correct solution to this problem is known at the outset.

What if the student were instead asked to design a diving board? One might model a diving board as a fixed-free beam with a geometry that produces large deflections and consequently large spring forces. Conceptualized this way, the student must simply rearrange terms in the bending equations to solve for the length, width, and thickness of the diving board given a reasonable load-deflection ratio. By convention, the length and width may be constrained to values typical of diving boards, yielding a deterministic relationship between the beam deflection and thickness. Designing a diving board sounds very similar to the beam deflection problem, but is it?

Despite the similarities between this problem and the preceding problem, designing a good diving board requires more thought than a simple beam deflection computation. The problem is not adequately constrained, so the student is left to his or her own devices to select a suitable beam-deflection ratio, identify geometric parameters that should be held constant, and choose a material for the diving board. All of these decisions necessitate that the student to seek out and process new information that was not provided in the problem as presented. For example, steel may have favorable elastic properties, but the student might decide that steel is not a suitable material choice because a steel diving board will rust in the presence of water.

Evaluating the student’s response is an equally difficult task. There are an infinite number of designs varying in shape and material that satisfy the beam bending equation. A single, “correct” solution does not exist, so the teacher must evaluate the student’s response on the basis of the design issues the student thought were salient and the degree to which the student’s response is consistent with those issues.

All design problems share some common characteristics observed in the diving board example. In contrast to well-defined problems, such as the beam bending problem, design problems are characterized by significant indeterminacy. Design problems lack a “correct” solution, but solutions can still be good or bad.

While the nature of design problems varies across disciplines, they share some common characteristics. Design theorists have often emphasized that design problems are all characterized by some degree of “wickedness” (Buchanan 1992). Rittel and Webber (1973) first introduced the notion of wicked problems in the context of social policy and planning work, and its general applicability to design problems was recognized later by Buchanan (1992). Rittel and Webber suggested that wicked problems are characterized by several properties, reproduced below from Buchanan (1992):
1. Wicked problems have no definitive formulation, but every formulation of a wicked problem corresponds to the formulation of a solution.
2. Wicked problems have no stopping rules.
3. Solutions to wicked problems cannot be true or false, only good or bad.
4. In solving wicked problems there is no exhaustive list of admissible operations.
5. For every wicked problem there is always more than one possible explanation, with explanations depending on the Weltanschauung of the designer.
6. Every wicked problem is a symptom of another, "higher level," problem.
7. No formulation and solution of a wicked problem has a definitive test.
8. Solving a wicked problem is a "oneshot" operation, with no room for trial and error.
9. Every wicked problem is unique.
10. The wicked problem solver has no right to be wrong – they are fully responsible for their actions.

(Buchanan 1992)

This is an impressive list that speaks to the nature of design problems, designers, and the design process. For someone with a background in the science or math domains, these properties can be counterintuitive. If a designer’s explanation of a wicked problem depends upon his or her Weltanschauung (5) and each formulation of a wicked problem corresponds to the formulation of a solution (1), it follows that designers with different backgrounds will arrive at different solutions to the same problem. This suggests that a critical part of the designer’s task is framing the problem. By deciding what design issues are important to consider, the designer can reduce an ill-defined, wicked problem, such as “design a diving board,” to a more tame, well-defined problem, such as “determine the appropriate thickness of a fixed-free beam to produce a desired deflection.” However, there is no guarantee that the designer’s solution to the tamed problem is a good solution to the wicked problem. The quality of the solution will depend upon how adequately the designer’s framing of the problem captures important issues.

The Design Process – Managing Wickedness

Designers use formal and informal methods to manage the wickedness encountered in design problems. Divergent thinking techniques, such as brainstorming, help the designer consider many problem framings and prototype solutions before selecting a solution to move forward with. Designers also utilize a pattern of “reflective thinking,” where they continually reframe the problem as they discover more information relevant to the problem.
New Product Development: An Process-Centric Model of Design

There are a number of formalized, process-centric models that prescribe the steps a designer or design firm should follow. Just in the realm of engineering design, researchers and practitioners have developed a number of these process models, including the Pahl and Beitz model (Pahl and Beitz 1996), the German VDI 2221 guidelines (Jänisch and Birkhofer 2006), and Ulrich and Eppinger's NPD model (Ulrich and Eppinger 2012).

![FIGURE 4: NEW PRODUCT DEVELOPMENT PROCESS](source: Ulrich and Eppinger (2012))

These models all share some common characteristics. There is a divergent phase of the development process, where the problem is defined and many possible design solutions are proposed and considered. These concepts are usually not highly detailed and can be described by a few words or quick sketch on a post-it note or napkin. In Ulrich and Eppinger's NPD model, shown in Figure 4, the divergent phase includes the "Planning" and "Concept Development" steps. Once one or more favorable concepts are selected, the design process enters a convergent phase, where design concepts are detailed, prototyped, tested, and ultimately produced. The convergent phase of Ulrich and Eppinger's NPD model includes the "System Design" and later phases of the NPD model in Figure 4.

Reflective Thinking: An Epistemological Model of Design

Designers also use less formal methods to manage wickedness. In his well-known book *The Reflective Practitioner*, Donald Schön studied the tacit thought processes of seasoned architects, psychoanalysts, engineers, business managers, and town planners (Schön 1983). Schön observed that these practitioners all utilized a common, "reflective" pattern of thinking to solve problems they encountered in their work. Through this "reflective conversation" with their work, the designers that Schön observed engaged in an iterative process of defining and redefining the problem. Each problem frame and prototype
solution revealed new information about the problem, which could then be internalized and incorporated into an updated problem definition.

While Rittel and Webber argued that solving wicked problems is a “oneshot” operation, the reflective thinking process allows designers to test out many problem definitions and prototype solutions artificially. Before breaking ground on a new building, an architect will consider thousands of design variations through sketches and scale models. An engineer will use computation methods and bench top tests to ensure that the diving board they are developing can withstand the repeated stresses of use. An industrial designer will create foam “looks-like” models of a new medical device to better understand how doctors want to interact with the device. Observations from these tests help the designer better understand the issues of the wicked problem and adapt their problem definition to capture these issues.

Managing Wickedness

<table>
<thead>
<tr>
<th>Uncertainty / patterns / insights</th>
<th>Clarity / Focus</th>
</tr>
</thead>
<tbody>
<tr>
<td>Research</td>
<td>Concept</td>
</tr>
<tr>
<td>Design</td>
<td></td>
</tr>
</tbody>
</table>

FIGURE 5: DAMIEN NEWMAN'S DEPICTION OF THE DESIGN PROCESS
source: Newman (2014)

The informal and formal methods employed by designers help them manage the wickedness of design problems. For the purposes of this thesis, the design process is conceptualized in two broad phases: a divergent phase followed by a convergent phase. In the divergent phase, many possible solutions are considered, which frame the problem in different ways. This phase serves to help the designer reduce, or tame, the wickedness of the problem. Damien Newman’s graphical depiction of the design process, shown in Figure 5, captures both the uncertainty inherent in this phase of the process as well as the end goal of reducing uncertainty before selecting a concept for further development. Once the problem is understood and one or more favorable design concepts are identified, the
design process enters the convergent phase, where design concepts can be detailed, prototyped, and produced.

Is Crowdsourcing Creative Work a Viable Strategy?

Determining whether crowdsourcing offers advantages to traditional innovation models is a firm- and problem-specific question. Characteristics of the firm, problem, crowd, and industry all affect the firm’s ability to crowdsource solutions. This section highlights a number of factors that should be considered when deciding whether a crowdsourced innovation model is suitable.

The Locus of Knowledge

The crowd’s capabilities are often best utilized in the divergent stages of the design process to generate and evaluate many ideas. In later phases of the design process, the relative capabilities of the firm often outweigh the benefits of the crowd.

Finding Solutions

Crowdsourcing is a good innovation strategy when the firm encounters a problem that they do not have the knowledge and expertise to understand or solve (Afuah and Tucci 2012; Aitamurto, Leiponen, and Tee 2011; Malone, Laubacher, and Dellarocas 2009). The firm has several options in such a situation: develop or acquire the required capabilities internally, outsource the work to an independent contractor, or ask the crowd for solutions. Afuah and Tucci (2012) argue that crowdsourcing is often an economical, expedient, and reliable strategy to identify possible solutions.

When the firm first encounters a problem, it is often unclear what knowledge and expertise are needed to find solutions. For example, the Oil Spill Recovery Institute had no idea it should be asking people from the concrete industry for methods to separate frozen water and oil. This makes it difficult for the firm to hire the employees or find an appropriate independent contractor. Since crowdsourcing engages a wide audience in the problem solving process, crowdsourcing increases the likelihood of finding one or more individuals who understand the problem.

In other instances, the problem may be more wicked or indeterminate. For example, the problem may be: “How might we reduce water usage in homes?” There is no best solution for such a question, but it is often useful to look at the problem from different perspectives.
to identify multiple solution strategies. Because crowdsourcing engages a large and diverse audience, it is often a very effective method to generate a wide variety of solutions to these problems.

The widespread growth of the Internet and the rise of social communities has greatly reduced the cost of broadcasting problems to a wide audience. Furthermore, the rise of crowdsourcing platforms, such as InnoCentive and OpenIDEO, make it even easier for the firm to crowdsource solutions. These platforms already have large competent crowds and experience facilitating crowdsourced challenges.

Developing Solutions

While the crowd may be good at generating ideas to solve a problem, firms often have the capabilities to carry a proposed solution through the convergent phases of the development process. In general, the firm should take over the development of solutions when the firm’s own capabilities exceed the crowd’s capabilities. In some cases, the crowd may develop detailed solutions before turning the work over to the firm. For example, participants in the DARPA FANG challenges were expected to devise, simulate, and verify the performance of their IFV designs before submitting them for consideration. In other cases, the firm may assume development much sooner. Ideas submitted on OpenIDEO often require significantly more development work before they can be implemented.

The Problem

Not all problems are amenable to crowdsourcing. There are a number of factors that may make it difficult for the firm to broadcast the problem or limit the ability of the crowd to solve the problem.

Intellectual Property

The creation of intellectual property – novel ideas and processes protected through patents and other legal structures – is often an important driver of innovation for firms in competitive industries. However, crowdsourcing may be difficult in patent-intensive industries. Crowdsourcing platforms where ideas are freely revealed are generally incompatible with intellectual property laws since public disclosure of an idea can make it difficult to patent. Matters are further complicated if an idea is developed collaboratively by the crowd because it can be difficult to ascertain who should be listed as an inventor on the patent. Firms must also ensure that ideas generated by the crowd do not infringe upon existing patents as this can lead to costly licensing fees and litigation.
While intellectual property challenges may limit the applicability of CCW, thoughtfully designed crowdsourcing platforms have successfully navigated around these challenges. When a solver submits ideas on InnoCentive’s platform, those ideas are not revealed to other solvers in order to protect any intellectual property generated by the solver. In addition, agreements stipulating that solvers must transfer intellectual property rights to challenge posters if their solutions are chosen as winners provide guarantees that the firm will capture intellectual property generated on InnoCentive.

A recent partnership between GE and Quirky demonstrates that creative business practices can enable crowdsourcing in patent-protected industries. In 2013, GE granted Quirky members exclusive rights to use over 200 GE patents in their idea submissions, and GE plans to eventually give Quirky members access to thousands of its patents (Kim 2013). The financial arrangement between GE and Quirky has not been disclosed, but the efforts have already led to a line of smart, internet-connected electronic devices that can be controlled with a mobile phone.

**Decomposability**

Problems are more readily crowdsourced when they can be decomposed into smaller problems that require more specific types of knowledge (Lakhani, Lifshitz-Assaf, and Tushman 2012; Afuah and Tucci 2012). Several types of problem decomposition are important to consider: decomposition around knowledge domains and decomposition of the design process.

**Decomposition around Knowledge Domains**

When problems are complex and require knowledge from multiple domains, they can be difficult to solve. However, if problems can be decomposed into smaller sub-problems which require knowledge from a specific domain, it will be easier for members of the crowd to contribute solutions (Afuah and Tucci 2012). This is a strategy commonly used on OpenIDEO. For example, a recent challenge asked: “How might we inspire young people to cultivate their creative confidence?” (IDEO 2013b). While the problem is posed in a broad manner that does not limit possible solutions strategies, the challenge brief also suggests several sub-themes, such as “taking the first step,” “hands-on making,” and “creative catalysts.” These sub-themes help the problem solver decompose the problem into something manageable and actionable.

**Process Decomposability**

Different stages in the design process require different types of knowledge and expertise. Creativity, benchmarking, and ethnographic research are necessary skills in the planning and concept development phases of NPD (Ulrich and Eppinger 2012). However, as design
concepts reach the detailed design phase, engineering, industrial design, and manufacturing skills become more important. When design processes are structured into phases that require different types of knowledge and expertise, it is easier for the firm to efficiently utilize the capabilities of the crowd and the firm’s own employees. This is especially true when the crowd’s capabilities are complementary to the knowledge and expertise of the firm. Quirky provides an excellent example of process decoupling. Quirky leverages the crowd’s creativity and customer insight to generate product concepts and evaluate market demand for new products, while Quirky’s experienced in-house design team translates these concepts into fully-realized sketches, CAD models, and prototypes.

When process decoupling is difficult or impossible, it may be more challenging to crowdsource designs solutions because work cannot be efficiently distributed between firm and crowd. This is especially true in design domains where the design process is characterized by quick iteration cycles. In these cases, it is often more efficient for one entity to carry a design through an entire design cycle. If the crowd does not have the requisite knowledge and experience to complete an entire design cycle, it is unlikely that the crowd will be able to produce high quality work. However, it may be possible to develop tools that help linearize the design process and enable individuals in the crowd to incorporate insights from domains that are unfamiliar. For example, simulations provided in the META toolkit allow a FANG challenge participant to evaluate the manufacturability of a design with minimal knowledge of manufacturing processes. This helps the challenge participant capture and incorporate design insights that would typically be identified later in the design process.

The Crowd

The composition of the crowd also influences its ability to generate high quality work. The size and diversity of the crowd have been shown to influence the crowd’s creative performance.

The Size of the Crowd

Crowds are, by definition, large. However, it is not clear how large the crowd should be to maximize its creative performance. Research findings suggest that increasing participation has advantages and disadvantages. The magnitude of each effect depends upon the nature of the problem and the motivations of the crowd.

The innovation literature has often described the problem solving process as a search for solutions on an unknown solution landscape (Simon 1978; Afuah and Tucci 2012; Kavadias and Sommer 2009). Because problem solvers are have limited knowledge, resources, and time, their understanding of the solution space is incomplete. Without full knowledge of
the solution space, it is impossible for the problem solver to know whether better solutions that exist beyond the problem solver’s search boundaries (Afuah and Tucci 2012). One strategy to improve the search performance is to add additional problem solvers. Because problem solvers have different experiences and knowledge, they search different regions of the solution space. When the number of problem solvers is large, the likelihood of finding a good solution is high. This parallel search process creates a parallel path effect, which improves the performance of the solution search process.

The parallel path effect is predicted to play a significant role in the crowd’s ability to identifying good solutions that are distant from the region of the solution space where the firm resides. As long as the firm is capable of implementing distant solutions, the crowd will generally outperform both the firm and any independent contractor that the firm could hire.

While the parallel path effect increases the likelihood that a good solution will be found, this is only possible if a large number of individuals participate. Classic economic theory suggests that this may be difficult because competitions can create disincentives that limit participation (Boudreau, Lacetera, and Lakhani 2011). Without any a priori knowledge about the competency of participants, one can only assume that each participant has an equal probability of submitting the winning solution \( p = 1/N \). As more participants join the competition (as \( N \) increases), the likelihood that any one participant’s submissions will be selected as the winning solution falls. It may be difficult to recruit participants when the expected returns from participation are low.

Even if an individual chooses to participate, the participation disincentive may reduce the quality of work submitted by that individual. In a study of over 9,000 software design contests, Boudreau, Lacetera, and Lakhani (2011) found that increasing the number of competition participants negatively impacted individual participant’s performance and average performance in the competition. The participation disincentive had a greater impact on individuals scoring in the upper percentiles of the coding competitions, as can be seen in Figure 6. While the quality of work among the upper percentiles performers decreased substantially as competition participants increased, the number of participants had no significant impact on the max score from each competition. This discontinuity from the downward trend suggests that the parallel path effect may counteract the participation disincentive.
The relative impacts of the parallel path effect and participation disincentive may depend upon the type of problem being solved. Upon a closer analysis of the data, Boudreau, Lacetera, and Lakhani found that the participation disincentive had a less negative impact on the quality of solutions when problems required knowledge from multiple domains (Figure 6, right). Multi-domain problems are considered to be more complex because they have more interacting factors and problem solvers need to integrate knowledge from multiple domains to identify solutions. As a consequence, the solution space of multi-domain problems is expected to be more rugged than the solution space of single-domain problems. The authors found that the maximum score increased slightly with each additional competitor for multi-domain problems and decreased for single-domain problems. This indicates that the participants were less discouraged by the participation disincentive when the problem exhibited greater uncertainty of identifying a solution.

The rationale behind the participation disincentive is founded upon the assumption that participants gain nothing unless they win. However, research has cast doubt on this assumption and shown that participants are often intrinsically motivated by the joy of solving problems (Jeppesen and Lakhani 2010) or recognition by others (Malone, Laubacher, and Dellarocas 2009).

The Composition of the Crowd

When crowds are more diverse, they may be more likely to find high quality solutions. In a study of 166 InnoCentive challenges, Jeppesen and Lakhani (2010) found that a significant number of challenges were solved by “technically marginal” solvers – individuals from
scientific disciplines than were significantly different from the scientific domain of the problem. These findings are corroborated by the earlier discussion of the parallel path effect. When it is likely that innovations lie in a distant part of the solution space, and it is difficult to predict where the innovative solutions exist, the crowd's diversity accentuates the parallel path effect by broadening the search for solutions.

Open Questions

Much of the discussion about the characteristics of the crowd has assumed that members of the crowd work independently. For example, the metaphor of solution space search assumes that each individual searches a limited region of the solution space constrained by his or her own cognitive limitations and time or resource constraints. However, the examples of CCW discussed earlier demonstrate that collaborative modes of participation also exist. There has been little research regarding the differences between collaborative crowds and competitive crowds or the effect that crowd composition has on the relative performance of each crowd type. These open questions are discussed later.

Empirical Evidence

Despite numerous examples of successful CCW and indications that crowdsourcing may be useful in the divergent phases of the design process, there have been few empirical studies evaluating these claims.

Case Study: MAM Baby Products

A recent study by Poetz and Schreier (2012) reveals strong evidence that the crowd can generate higher quality concepts than internal innovation teams. Their study focused on an innovation cycle at MAM, a high-end European baby products company. Market research had revealed a strong need to develop easier methods to feed babies supplemental mash and solid foods, so MAM initiated a new product development cycle to develop concepts addressing this need. At the same time, MAM asked users to submit their own solutions and concepts through MAM's website. Users were offered a €500 prize for the best concept and other merchandise to be raffled off at random.

1 Jeppesen and Lakhani also report that women were more likely to identify a solution than men. Citing the gross gender imbalance in many scientific fields, the authors argue that this is evidence that social marginality also plays a role in the problem solving process.
Users generated 70 concepts and MAM's internal team generated 51 concepts. Of the 70 user-generated concepts, 18 were discarded because they were not of sufficient detail or clarity to evaluate. MAM's CEO and head of R&D scored each of the remaining concepts on its novelty, potential customer benefit, and feasibility. An overall quality score was also computed for each concept by taking the product of the average novelty, customer benefit, and feasibility (novelty x customer benefit x feasibility). Average results for professional- and user-generated concepts are summarized in Figure 7.

### Table 1: Idea Quality Comparison

<table>
<thead>
<tr>
<th>Idea Quality</th>
<th>Professional Ideas (n=51)</th>
<th>User Ideas (n=52)</th>
<th>Mann-Whitney-U Test</th>
<th>Z-value (p value)*</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean (SD)</td>
<td>Mean (SD)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Novelty</td>
<td>2.12 (1.14)</td>
<td>2.60 (1.27)</td>
<td>-1.956 (.050)</td>
<td></td>
</tr>
<tr>
<td>Customer benefit</td>
<td>1.86 (.66)</td>
<td>2.44 (1.01)</td>
<td>-3.010 (.003)</td>
<td></td>
</tr>
<tr>
<td>Feasibility</td>
<td>4.33 (.91)</td>
<td>3.91 (1.21)</td>
<td>-1.856 (.063)</td>
<td></td>
</tr>
<tr>
<td>Three-way interaction</td>
<td>16.75 (12.15)</td>
<td>24.93 (19.24)</td>
<td>-1.973 (.048)</td>
<td></td>
</tr>
</tbody>
</table>

* We use Mann-Whitney-U tests instead of simple t-tests because the dependent variables are not normally distributed.

The authors find that, on average, user-generated concepts are more novel and generate greater customer benefit than professional concepts, but professional concepts may be easier to realize. While these mean trends provide some insights into the differences between user- and professional-generated concepts, it is equally important to assess trends among the top-ranked concepts, since these will be carried forward into later stages of development (Ulrich and Eppinger 2012). An assessment of the top-ranked concepts revealed that users were more likely to generate top-rated concepts than MAM's design team.

### Table 2: Top-Ranked Concepts

<table>
<thead>
<tr>
<th>Idea Quality</th>
<th>Company Users (n=51)</th>
<th>Users (n=52)</th>
<th>Expected Frequency</th>
<th>Obs. (Exp.)</th>
<th>Obs. (Exp.)</th>
<th>Chi-square (p value)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Top ideas</td>
<td>6 (10.9)</td>
<td>16 (11.1)</td>
<td>2 (5.0)</td>
<td>8 (5.0)</td>
<td>42 (39.1)</td>
<td>5.536 (.019)</td>
</tr>
<tr>
<td>Other ideas</td>
<td>45 (40.1)</td>
<td>36 (40.9)</td>
<td>49 (46.0)</td>
<td>44 (47.0)</td>
<td>9 (11.9)</td>
<td>3.859 (.049)</td>
</tr>
<tr>
<td>Chi-square (p value)</td>
<td>2.318 (.128)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* Top ideas are defined as those that score higher than three in the respective quality dimension.

The study by Poetz and Schreier (2012) reveals strong evidence that the crowds can generate more novel and useful concepts that a firm's own employees. Unfortunately, no other studies exist comparing the ideation performance of the crowd and firm, and the
authors of the study caution that their findings are not generalizable without additional research.

**Conclusions**

The descriptive models and empirical research reviewed above suggest that the crowd can, under the right conditions, be a valuable asset in the design process. In general, crowdsourcing works best when the capabilities of the crowd are complementary to the capabilities of the firm. For example, the crowd could be used to generate creative product concepts when the firm does not have experience or expertise in the divergent phases of the design process. However, if the firm has deep experience and expertise in the convergent phases of the design process, where concepts are detailed and produced, it is wise for the firm to keep these phases of the design process in-house.

There are some important limitations that may make it difficult for the firm to leverage the crowd. For example, the firm’s design process must be decomposed so that specific design tasks can be outsourced to the crowd. Furthermore, in industries that are highly protected by intellectual property, it may be difficult for the crowd to generate usable concepts because those concepts may infringe upon patents. While these factors may make it more difficult to crowdsource creative work, the CCW models reviewed earlier demonstrate that with some thought the crowdsourcing model can often adapt to work within these limitations.

**The Creative Capacity of the Crowds**

Recent research published in *Nature* indicates that group performance in a number of tasks, including creative thinking, can be explained by a collective intelligence factor, which is similar to personal intelligence factors, such as IQ, that affect individual performance (Woolley et al. 2010). The authors of the study found that this collective intelligence factor depends primarily upon “the average social sensitivity of group members, the equality in distribution of conversational turn-taking, and the proportion of females in the group” and shows no clear relationship to the individual intelligence of group members (Woolley et al. 2010, 686). The existence of a collective intelligence factor suggests that individuals working together in a group may be more (or less) intelligent than they would otherwise be working independently.

While collective intelligence has only recently been studied in general terms, researchers have been assessing the creative performance of groups for some time. There has been
extensive debate in the creativity literature about whether teams, where individuals work collectively to brainstorm ideas, are more creative than so-called "nominal" groups, where individuals work independently and collect all ideas at the end of the idea generation process (Hennessey and Amabile 2010; Kavadias and Sommer 2009). While the crowd differs in some respects from the co-located groups studied in the creativity literature, the CCW platforms discussed earlier provide examples of both collaborative and nominal crowds. Currently, there has been no attempt to apply and test the findings from group creativity research to the crowd. Understanding how the insights from group creativity research apply to the crowds would influence the design of CCW platforms may improve the quality of ideas generated by the crowd.

Research into the effects of collaboration on crowd performance may also benefit group creativity research. While it is difficult to regulate individual behavior in a co-located group, platform features, such as commenting, largely regulate the interactions available to the crowd. Because the features implemented by crowdsourcing platforms constrain interactions between the crowd's actors, the distinction between collaborative crowds and nominal crowds becomes a question of degree rather than kind. Through careful experimental design, creativity researchers may be able to create crowdsourcing platforms that enable or limit certain types of interactions between individuals, which are thought to have either positive or negative effects on group performance. These platforms can be used as research tools to isolate and study the factors that affect group performance.

Research findings on the performance of teams and nominal groups are summarized in the following sections. Based on these findings and a brief discussion about the differences between co-located groups and crowds, the implications of these findings for the crowd are examined.

**Group Creativity Insights**

Today, most of the creative work in industry occurs in teams, and the collaborative brainstorming techniques introduced by Osborn (1953) are widely used and celebrated (Kavadias and Sommer 2009). However, theoretical and empirical research by organizational scientists and social psychologists has cast doubt on the effectiveness of these collaborative ideation techniques and suggested that nominal group techniques are more effective (Kavadias and Sommer 2009; Hennessey and Amabile 2010).

**Advantages of Teams**

In *Applied Imagination*, Osborn (1953) claimed that brainstorming teams generate twice as many ideas as individuals working alone. While this claim was never substantiated, the collaborative ideation process may provide some advantages over nominal groups. The
wicked formulation of design problems posits that an idea is derived from a specific framing of the problem. These problem framings in turn emerge from an individual’s perspectives, education, and experience. If this is true, individual generating ideas in isolation may only consider problem framings that are representative of their own perspectives. In a team setting, however, individual are exposed to many more ideas and framings, which are representative of other team members’ Weltanschauungen. Just as Schön’s reflective practitioners deepen their understanding of the problem by assessing current solutions, individuals in the team may learn from the alternative ideas and framings presented by others.

Anyone who has engaged in a group brainstorming process has likely experienced such a phenomenon when one person’s idea stimulates a wave of new ideas from other group members. These new ideas emerge from a recombination of existing ideas and problem framings. In a controlled experiment where participants were asked to generate ideas by recombining two previous ideas generated by others, Nickerson, Sakamoto, and Yu (2011) found that the recombined ideas were more novel and practical than the original ideas. This suggests that reflective thought processes may improve the ideation performance of teams.

While the role of reflective thought is not examined here in detail, Woolley et al. (2010)’s finding that the average social sensitivity of teams is strongly correlated with collective intelligence further indicates that reflective thought processes play a role in team performance.

Failures in Team Settings

Despite the prevalence of teams in industry and the arguments that collaboration can improve the ideation performance, research findings show that teams seldom realize their performance potential because of a number of social factors that lead to group failures (Hennessey and Amabile 2010; Dahlin 2005; Kavadias and Sommer 2009). In fact, there is general consensus in the creativity research community that nominal groups generate more ideas than brainstorming teams and also that nominal group’s ideas are more creative (Hennessey and Amabile 2010).

Social Dynamics

There are a number of social dynamics within the team that may reduce the team member effort or limit team members from contributing. For example, outspoken members of the team may dominate the idea generation discussion, thus limiting the ability for less vocal members to share their ideas. There are a number of other social factors, such as cultural differences, that may lead to social categorization and limit the ability of team members to
work together (Dahlin 2005). In a nominal group setting, these social dynamics do not exist, so every team member can contribute their full effort the idea generation process.

Production Blocking

Another factor that leads to reduced team performance is production blocking. When one team member is speaking about his or her idea, other members are “blocked” in the sense that they cannot share their ideas until there is a break in the discussion (Hennessey and Amabile 2010). Due to the effects of production blocking, a nominal group will likely generate more concepts than a team in a fixed amount of time (Kavadias and Sommer 2009).

Social Loafing

In the nominal group setting, each team member is expected to generate as many ideas as possible. However, in the team setting, less motivated team members might be inclined to contribute less to the team than they would have otherwise contributed if they were working alone (Hamilton, Nickerson, and Owan 2004). This form of social loafing leads to an underutilization of the team’s creative capacities.

Idea Fixation

When team members generate concepts, they can sometimes become fixated on one type of solution. Teams that become fixated on a class of ideas cannot take advantage of differences in team members cognitive and culture backgrounds to search the solution space (Kavadias and Sommer 2009).

The Effect of Problem Structure

In an attempt to reconcile experimental research findings with practical experience, researchers have suggested that the choice between team and nominal ideation methods may depend on the structure of the problem (Kavadias and Sommer 2009). Proponents of team brainstorming argue that the problems considered in the research literature are simpler and require less domain knowledge than the problems encountered in practice (Kavadias and Sommer 2009).

Using simulations of the solution search process described by Simon (1978), Kavadias and Sommer (2009) attempted to explain how the effects of problem structure and team diversity impact the performance of teams and nominal groups. Their results suggest that the choice between teams and nominal groups depends upon the structure of the problem, how cognitively diverse the group is, and the degree of production blocking and evaluation apprehension present in the team. In general, the authors’ findings are consistent with previous research results: nominal groups generally outperform teams. The benefits of
nominal groups are most significant when problems require specialized knowledge from one domain. However, when problems require knowledge from multiple domains and problems cannot be readily decomposed into independent sub-problems, teams identify better solutions than nominal groups. The simulations also show that the relative advantage of nominal groups diminishes as production blocking and evaluation apprehension decrease. Figure 9 summarizes these findings by denoting the regions of the problem-structure space where nominal group and team (denoted brainstorming) ideation methods are preferred.

These findings are consistent with the examples of CCW presented earlier. Nominal groups are preferred when problems require deep domain knowledge, such as those considered on InnoCentive. However, collaborative approaches, such as those used on OpenIDEO, are favorable when problems are multi-disciplinary and require a synthesis of information from different viewpoints.

![Figure 9: Performance Frontiers for Teams and Nominal Groups](source: Kavadias and Sommer (2009))

The Role of Team Diversity on Idea Generation Performance

Research also suggests that diversity mediates the positive and negative factors affecting team performance. Diversity in cognitive attributes, such as education background, is thought to help broaden the search for solutions by encouraging the group to considering alternative framings of the problem. Diversity is social characteristics, such as nationality and gender, is thought to have a more deleterious effect on team performance because it can increase social dynamics that make it difficult for the team to work effectively together (Hennessey and Amabile 2010).

In a study examining how educational and national diversity impacted information usage in small teams, Dahlin (2005) found that the relationship between diversity and information
use is not necessarily straightforward. The study analyzed three aspects of information use: the range or breadth of topics considered by the team, the amount of depth or detail present in the teams analysis of each topic, and the degree to which information from different topics was integrated by the team to form a cohesive understanding of the problem. The results indicate an inverse curvilinear relationship between educational diversity and range and depth of information. This suggests that cognitive diversity helps the team broaden its solution search up to a point, but when diversity is very high, communication barriers emerge that prevent the team from effectively utilizing its diverse perspectives.

Downselecting

The creativity literature reviewed above has focused primarily upon a group’s ability to generate ideas. However, in an organizational setting, it is equally imperative that the group selects the best ideas to progress into the next phase of the design process. The process of downselecting is complicated because the ideas generated by the group often lack sufficient detail to evaluate their technical feasibility, customer appeal, and business viability. Social factors, such as a group member’s attachment to his or her own ideas, further complicate the downselection process. To make an informed decision, the group must consider a number of important issues and organize these insights so that the group can choose ideas with the greatest probability of success.

Regardless of whether nominal or collaborative methods are used to generate ideas, the downselection process is collaborative. Collaboration and group characteristics likely affect the group’s ability to make informed decisions in the downselection process. For example, nominal groups may be less effective than teams in the downselection process because members of the nominal group may exhibit greater attachment to their own ideas. Since diversity affects team cohesiveness, it will also likely impact the downselection process. Dahlin (2005) found that teams with diverse education backgrounds were less effective at organizing and synthesizing information, while teams with moderate amounts of national diversity were better at organizing and synthesizing information.

Collaboration and group characteristics, such as group diversity and group size, are important factors that may affect the group’s ability to identify good solutions in the downselection process. Descriptive reasoning indicates that teams may be better than nominal groups in the downselection process. Additional research is needed to examine these findings.

Summary of Group Creativity Literature

The group creativity literature suggests that collaborative teams are, in general, less creative than nominal groups. A number of failures in team settings, such as production blocking
and social loafing, lead to poor ideation performance. However, when problems require multidisciplinary knowledge, the benefits of collaboration may outweigh these failures, and teams may identify better solutions than nominal groups. Team may also be more effective in the downselection process, where the most promising ideas are selected for further development.

Group diversity also has varying effects on nominal group and team performance. Cognitive diversity is thought to have benefits in the divergent phases of the design process because it helps broaden the search for solutions. However, cognitive diversity can also limit the group’s ability to build integrate information in the downselection and convergent phases of the design process. Social diversity tends to promote social categorization and other failures in team settings, so it is thought to degrade team performance in both the divergent and convergent phases of design. However, research findings have also shown that moderate amounts of social diversity can be beneficial.

**From Groups to Crowds**

While debate over the superiority of collaborative and nominal ideation methods continues, the creativity literature suggests that problem structure and group diversity are important factors affecting group performance. There is currently little understanding of how these findings apply to the large, uncoordinated groups found on crowdsourcing platforms. The limited findings in the crowdsourcing literature indicate that individuals from distant domains often identify the best solutions. This suggests that the cognitive diversity of the crowd influences the crowd’s performance. Better understanding of how collaboration, problem structure, and crowd composition affect the crowd’s creative capacity will inform the design of CCW models and hopefully improve design outcomes. This thesis takes a first step in assessing how the insights from the creativity literature apply to the crowds.

**Differences between Groups and Crowds**

Before hypothesizing how findings in the creativity literature extend to the crowd, it is important to consider how groups and crowds differ. These differences can be broken into three categories: size and involvement, communication capability, and coordination.

**Crowd Size**

Crowds are larger than the groups found at firms. InnoCentive challenges often attract nearly 200 problem solvers, though only 10% typically submit solutions, and Quirky product development cycles may engage over 1,000 members. Because crowds are larger than groups, participation disincentives may be more deleterious to ideation performance.
At the same time, the large size of the crowd may lead to parallel path effects that improve ideation performance.

Communication Capability

The capability for members of the crowd to communicate with each other is more limited than in team settings. While a team usually works in a co-located fashion to generate and share ideas, the crowd’s members are often geographically dispersed and can only communicate through features available on the crowdsourcing platform. In general, these features include the ability to: share ideas in the form of pictures, videos, and/or text; comment or rate ideas that have been shared; and evaluate or express approval of ideas.

The communication features available on a crowdsourcing platform regulate crowd’s ability to share and discuss ideas. When features facilitating communication are eliminated, as is the case on InnoCentive’s platform, the crowd cannot collaborate and resembles a nominal group. As communication capability increases, the type of interactions available to the crowd more closely resembles those found in a team setting. Even on the most communication-capable crowdsourcing platforms, the crowd will typically have less communication capability than a co-located team.

![COMMUNICATION CAPABILITY](image)

FIGURE 10: COMMUNICATION CAPABILITY OF THE CROWD, NOMINAL GROUP, AND COLLABORATIVE TEAM

While limited amounts of communication capability may make it more difficult for the crowd to collaborate, it may also reduce some adverse effects of collaboration observed in the creativity literature. In a team setting, differences in nationality, gender, age, physical fitness, and other social factors are readily apparent. However, because the crowd’s communication capability is more limited, social differences in the crowd are not as easily observed. By reducing the observability of these social factors, the negative effects of social diversity on ideation performance should be reduced. This leads to the first hypothesis:

Hypothesis 1 Social diversity will have a negligible effect on the crowd’s ideation performance regardless of the crowd’s communication capability.

Coordination

Teams are highly coordinated entities. During the ideation process, the team’s members need to meet at a common location and time. In contrast to teams, the crowd’s members can operate in a much less coordinated fashion. It is common for a member of the crowd to
submit an idea, leave, and return hours or days later to see what other ideas have been submitted. In this sense, the crowd’s members act in a relatively asynchronous manner, while members of a team are more synchronous. In fact, the asynchronous behavior of the crowd closely resembles the parallel search processes found in nominal groups. Because members of the crowd act asynchronously, the crowd is not subject to the same production blocking effects observed in synchronous teams.

Conclusions

Key differences in size, communication capability, and coordination of groups and crowds suggest that some of the findings in the group creativity literature may not be applicable to the crowds. While social diversity has been shown to reduce collaborative group performance, crowds are less likely to exhibit these failures because social differences in the crowd are less visible than in a group setting. In addition, the crowd’s asynchronous nature should limit the effects of production blocking on collaborative crowds. These findings collectively suggest that the crowd is less susceptible to collaborative group failures.

Cognitive Diversity

While there are differences between groups and crowds, some of the findings in the group creativity literature are likely relevant. For example, greater cognitive diversity should broaden the search for solutions and improve the crowd’s ideation performance. Limited evidence from the crowdsourcing literature reviewed earlier supports this proposition. Jeppesen and Lakhani (2010) found that technically marginal individuals were more likely to identify good solutions to problems posted on InnoCentive. These findings lead to the second hypothesis:

Hypothesis 2  Cognitive diversity will have a positive effect on the crowd’s ideation performance.

The Nominal Crowd and the Collaborative Crowd

The group creativity literature suggests that nominal groups typically outperform collaborative teams (Hennessey and Amabile 2010). This conclusion is based on the reasoning that collaborative teams exhibit a number of group failures that reduce ideation performance. However, differences between groups and crowds suggest that crowds may be less susceptible to these failures (Hypothesis 1). If this is confirmed, the relationship between collaboration and the crowd’s creative capacity warrants further study.

Assuming that Hypothesis 1 is confirmed, what other factors might be important to consider? Problem structure has also been shown to influence the tradeoff between nominal and collaborative ideation methods (Kavadias and Sommer 2009).
are multidisciplinary, simulations of the solution search process indicate that collaborative teams outperform nominal groups despite the presence of group failures. If the crowd is less susceptible to these failures, collaborative crowds may outperform nominal crowds for a wide range of problem types.

While it may be difficult to identify specifically when collaborative ideation methods are better than nominal methods, design problems, as defined earlier, are multidisciplinary. Therefore, collaborative crowdsourcing models may be best suited to these types of problems. The earlier discussions of creative work demonstrate that most problems encountered in the design domains are wicked. These problems elude absolute definition and are only understood when they are examined from multiple domain perspectives. Collaboration is valuable in these situations because it helps organize insights from different perspectives to develop a cohesive problem understanding. This suggests that collaborative crowds may be more creative than nominal crowds.

The degree of collaboration in the crowd depends on the communication capabilities of the crowdsourcing platform and the extent to which the crowd utilizes these capabilities. Designers of crowdsourcing platforms have little control over the crowd and their utilization of the platform, so it is difficult to control for these factors. In contrast, designers have full control over the communication capabilities built into the crowdsourcing platform. These capabilities also limit the extent of collaboration on crowdsourcing platforms. Because communication capability is readily controlled in the design of crowdsourcing platforms and is expected to have a significant effect on the amount of collaboration observed on the crowdsourcing platform, communication capability is used as an indicator for collaboration capability. This leads to a third and final hypothesis:

\[ \text{Hypothesis 3} \quad \text{Communication capability will have a positive effect on the crowd's ideation performance.} \]

---

2 Assuming other factors, such as concerns about protecting intellectual property, do not preclude the use of collaborative crowdsourcing.

38
Chapter 4
Experimental Design

A design challenge and observational study were devised in order to assess the effects of communication capability, social diversity, and cognitive diversity on the crowd's creative performance. Participants were asked to "reimagine the public restroom experience" and submitted concepts online using CrowdConcept, a website developed to support this study that is similar to existing crowdsourcing platforms. CrowdConcept can be found online at http://crowdconcept.herokuapp.com, and screenshots of the site are included in the appendices.

Reimagining the Public Restroom Experience

Participants in the study were asked to develop concepts to "reimagine the public restroom experience." This section outlines the design and motivation behind the challenge.

Motivation Behind the Design Challenge

The design challenge took inspiration from similar challenges found on OpenIDEO and suggestions from the Stanford d.school's K12 Lab (Crandall 2012). OpenIDEO challenges address broad social issues, and participants' proposed solutions span the range of design artifacts, including products, graphic campaigns, mobile applications, and services. The accessible and multidisciplinary nature of OpenIDEO challenges is consistent with the types of wicked design problems discussed earlier. These types of problems were hypothesized to be most amenable to crowdsourcing methods, so a challenge problem that followed the OpenIDEO format was desired.

Brainstorming sessions were held with design students at MIT to identify possible problems that the design challenge could address. Participants in these brainstorming sessions were encouraged to think of problems that were:

    Relevant: the problem should be a compelling and participants should be familiar with it
Open-Ended  the problem should lend itself to many different approaches, interpretations, and solutions

Human-Centered  the problem should require an empathy and understanding of users needs

Accessible  participants in the challenge should not need specialized technical skills to understand and contribute to the challenge

More than 70 potential problems were suggested, and the group selected several problems to investigate further. The public restroom was selected as the basis for the design challenge because the design issues are universally understood and the problem affords for many different types of design solutions. Emphasis was placed on improving the experience of public restroom users rather than improving the public restroom itself to broaden the scope of issues considered by participants.

Design Brief

A short design brief was created to educate participants about the challenge, explain the concept review process, and provide some general suggestions to guide them through the brainstorming process. Another goal of the brief was to make sure that participants had a common understanding of what constituted a public restroom and the many uses of public restrooms. The brief included images of different types of public restrooms, including: normal public restrooms found in stores, aircraft lavatories, porta-potties, and restroom facilities at parks. Images of less common bathroom users and uses were also included, such as: janitors cleaning bathrooms, parents changing a baby, people using bathroom mirrors to put on makeup, and handicapped individuals using an accessible bathroom. These images were intended to open up the design space and encourage participants to focus on specific users and their needs rather than general public restroom issues, such as sanitation.

Before participants could submit concepts, they were asked to review the design brief, and at any point during the study, participants could return to the design brief, which appeared in a pop-up window. The total amount of time each participant spent reviewing the brief was recorded to see if there were any correlations between time spent reviewing the brief and ideation performance.

---

3 A print version of the design brief is included in the appendices.
Concept Submission

Study participants could submit up to three concepts via a web form on the CrowdConcept site. Participants were asked to provide a descriptive title for their concept, a brief elevator pitch, and at least one representative sketch or image. They were also asked to provide 1-2 paragraph responses to three questions about their concept:

- What problem or need does your concept address?
- What is your concept and how does it work?
- What barriers might prevent your concept from being realized?

These three questions were included to encourage participants to think critically about their concept and also to ensure that concepts had enough information for reviewers to adequately evaluate.

Reviewing Concepts

Amazon Mechanical Turk (Amazon 2014) was used to review concepts. Amazon Mechanical Turk (AMT) is a micro-task marketplace developed by Amazon where individuals (requesters) post short human intelligence tasks (HITs) for others individuals (workers) to complete. Common HITs on AMT include surveys, transcription tasks, image categorization tasks, and user studies (Ross, Irani, and Silberman 2010). Requesters pay workers a small reward for each completed HIT and Amazon takes an additional 10% transaction fee, which is paid by the requester.

While an expert review panel may be preferred to a novice review, it would have been difficult to organize an expert panel large enough to provide multiple reviews for each concept. In contrast to an expert panel, AMT’s large worker population made it easy to obtain many reviews for each concept in an inexpensive and expedient manner. In addition, workers are probably just as representative of a typical public restroom user as an expert, so they should be able to reasonably assess the concepts.

Review HITs were limited to workers from the United States with a 90% or greater approval rating because previous experiences with AMT indicated that US workers generally provide higher quality and more thoughtful reviews than international workers. Limiting the worker sample to the US also ensured that reviewers would have a fairly consistent experience with public restrooms. Workers were paid $0.10 for each review and could review as many concepts as they pleased. Participants from the ideation phase were prohibited from participating in the review process to limit possible review bias.
Workers scored concepts on five criteria:

- **Clarity**: is the concept well articulated and easy to understand?
- **Need**: does the solution address a compelling problem or need?
- **Novelty**: do similar concepts exist or is the concept highly unique?
- **Usefulness**: would users find the concept useful or desirable?
- **Feasibility**: could the concept be reasonably implemented from a technical and business perspective?

**Novelty** and **usefulness** stem from patent criteria and are used to evaluate concepts in the NPD process. **Need** was included to assess whether the concept addresses a real and compelling problem in public restrooms. Assessing whether a concept is clear and easy to understand was added to control for possible differences representation style and descriptive quality of concepts. Finally, **feasibility** was included to capture technical and business risks that might prevent the concept from being developed into a mature product or service.

The review criteria are similar to those used in past studies (Poetz and Schreier 2012; Shah, Smith, and Vargas-Hernandez 2003; Nickerson, Sakamoto, and Yu 2011) and by design practitioners in the downselection process (Ulrich and Eppinger 2012). Reviewers were presented with a normative statement, such as “The concept is a novel solution the problem or need,” and were asked to assess their agreement with the statement on a 5-point Likert scale: (1) Strongly Disagree, (2) Disagree (3), Neutral, (4) Agree, (5) Strongly Agree. Reviewers could also provide optional comments about their ratings.

**Treating Categorical Data as Continuous Data**

Categorical data from a Likert scale cannot generally be treated as continuous data. This is because it is impossible to say whether the distance between “(1) Strongly Disagree” and “(2) Disagree” is equal to the distance between “(2) Disagree” and “(3) Neutral”. When the distance between categories cannot be determined, one cannot compute parametric statistics, such as the mean or standard deviation. Instead, non-parametric statistics, such as the median, must be used. These non-parametric statistics are based on differences in rank rather than degree.

While treating Likert data like continuous data is not statistically sound, it is common practice. In general, the approach relies on the assumption that the Likert categories are equally spaced. When this assumption is true, the Likert point values can be directly converted to a continuous scale. Other more advanced methods, such as (Snell 1964), attempt to rescale the categorical data so that the ratings are representative of a normal
distribution. This study followed common conventions defining the five Likert categories, so ratings are directly converted to a continuous on the range of 1 to 5.

Overall Concept Quality Metrics

While each concept was reviewed on several criteria, it is useful to define an overall quality metric to facilitate general comparisons between concepts. This quality metric should capture the key factors a firm considers when deciding which concepts to move into the next stage of development. Ulrich and Eppinger (2012) suggest that the firms developing new products should evaluate product concepts by asking a few screening questions:

- **Is it real?** The product must address a compelling problem or need, and there needs to be a sufficient number of customers who find the product useful or desirable enough to purchase it.

- **Can we win?** The product must be able to compete or outperform existing products on the market. In other words, the product must be sufficiently novel to distinguish itself from other competitors.

- **Is it worth it?** The product needs to provide financial returns at an acceptable level of risk. This suggests that the product must be both financially and technically feasible.

The Real-Win-Worth it screening method suggest that need, novelty, usefulness, and feasibility are all important criteria to consider when evaluating a concept. If a concept scores poorly on any of these criteria, it may not be viable to move forward into later stages of development.

To capture the combined effects of these four criteria, a four-way interaction term was defined: \( \text{need} \times \text{novelty} \times \text{usefulness} \times \text{feasibility} \) (NNUF). The use of such an interaction term has some basis in the literature. Poetz and Schreier (2012) use a three-way interaction term, which includes novelty, usefulness, and feasibility, to assess and compare the overall quality of concepts. The addition of need is useful in the context of this study because there are many sub-problems that participants could develop concepts for, but not all problems are equally prevalent or important to people using public restrooms.

A secondary interaction term was defined that includes need, novelty, and usefulness (NNU) but does not include feasibility. This was included because reviewers may not have the knowledge or expertise to evaluate the technical and/or business feasibility of concepts.
Experimental Design

An experiment was designed to examine how communication capability and diversity affect the quality of concepts generated during the design challenge.

Assessing Communication Capability

The nominal groups and ideation teams studied in the creativity literature have significantly different amounts of communication capability. Members of nominal groups have no capacity to communicate with each other, while members of collaborative ideation teams have full use of their verbal, visual, and graphic communication skills. In this sense, the creativity literature examines differences in groups at the extremes of the communication capability spectrum.

To assess whether the findings from the creativity literature were applicable to the crowds, an experimental design was needed to create crowds with varying degrees of communication capability. Study participants were assigned to one of two groups:

- **Isolated (Co)** participants completed the study in isolation and could not see concepts generated by other participants.
- **Collaborative (Ci)** participants could view concepts submitted by other participants while generating their own concepts.

While it would have been valuable to see how crowds with greater communication capability performed relative to the Co and Ci crowds, replicating such an environment in a research setting is difficult. When crowds are collaborative, members of the crowd frequently visit the crowdsourcing platform to see new concepts, provide comments and feedback, and submit their own concepts. These collaborative crowdsourcing sites rely on an active community of users. However, these communities do not immediately coalesce when a crowdsourcing platform is created. Developing the community that has its own social norms and expectations is a slow process and often requires deliberate intervention from administrators of the crowdsourcing platform. Creating a research environment with such a community would have been difficult and could have introduced complications from a research perspective. While some participants might be extremely engaged and returned to the site frequently to contribute, others might submit a concept and never return. The analysis of collaborative crowds would need to account for differences in additional factors, such as engagement, that might affect the quality of concepts generated by participants. While these are important factors to consider, they are left to future studies.
Although the C1 crowd is similar to an ideation team, it differs in several key respects. These differences and their potential effect on ideation performance are discussed below.

Information Structure and Flows

The nature of information sharing for C1 participants is more structured and limited than normal co-located teams. While information flows bidirectionally between members of a team, information flows in the study are necessarily unidirectional. Once a participant submits concepts and completes the study, they are unable to submit additional concepts. This is analogous to preventing members of a team from sharing ideas if they already suggested an idea. The active C1 participant can see and reflect upon previous participants’ concepts, but these previous participants have no way of seeing the active participant’s concept because they have already completed the study. This limits the magnitude of impacts from cognitive diversity because the active participant is the only member of the crowd who can take advantage that diversity.

Information sharing is also limited in the study because concepts are the only mechanism to share information. Informal communication and other mechanisms that allow a member of an ideation team to articulate their framing of the problem and prototype solution do not exist on CrowdConcept. The absence of these communication modes may reduce the positive effects of cognitive diversity by reducing the likelihood that the participant will notice and absorb different problem framings. At the same time, limiting communication prevents dominance and free-riding behavior, which should reduce the negative effects of social diversity on crowd’s performance.

Anonymized Concepts

The effect of observable social factors shown to degrade ideation team performance was likely diminished in this study because no information about participants accompanied the crowd concepts. Even though information was not specifically disclosed, gender- and age-specific concepts reveal information to the active participant about the social characteristics of other members of the crowd. This suggests that social diversity should have little effect on ideation team performance.

Upon examining the differences between the C1 crowd and an ideation team, the effects of cognitive diversity and social diversity will likely be more muted in the C1 crowd than in an ideation team. Although the magnitude of both effects is expected to decrease, it is unclear how the relative magnitude of the social and cognitive diversity effects will compare as communication capability is reduced. The negative effects of social diversity should be
significantly reduced since crowd concepts are anonymized. The positive effects of cognitive diversity will also be reduced, but the magnitude of the reduction may be less significant because differences in cognitive styles should be apparent when participants view their crowd concepts.

Assessing Social and Cognitive Diversity

Developing an experiment to assess differences between diverse and homogenous crowds is challenging because diversity is not well suited to traditional treatment-control methods, like those used to study the effects of communication capability. The issues are most readily seen by way of example. Suppose you wanted to see how occupational diversity affects the performance of crowds composed of three types of individuals: engineers, industrial designers, and architects. A typical treatment-control formulation might randomly assign participants to a low occupational diversity group (control) or high occupational diversity group (treatment). Participants assigned to the low diversity group would be placed in a crowd with participants who have similar occupational backgrounds, so an industrial designer would be placed into a crowd with other industrial designers. Participants in the high diversity group would be placed in a crowd that includes a mix of all three occupations. The effect of diversity would be evaluated by comparing the performance of the treatment and control crowds. The difficulty arises because a crowd needs to be created for every distinct occupation in the low diversity group. As more participants enroll in the study with different occupations, more low diversity crowds need to be created. This approach quickly becomes intractable without a very large participant sample because there are simply not enough participants to adequately populate each crowd. A more nuanced approach was needed to overcome the problems associated with the treatment-control method.

Rather than assigning $C_1$ participants to a specific diversity group, the "crowd concepts" shown to each participant were varied to achieve one of four diversity objectives: minimize social diversity, maximize social diversity, minimize cognitive diversity, or maximize cognitive diversity. In other words, a custom set of crowd concepts was created for each participant. For example, an engineer in the low cognitive diversity group would be shown concepts generated by other engineering participants, while a 20-year-old woman in the high social diversity group would be shown concepts from participants of varying ages and genders. The selection algorithm used to generate crowd concept sets and the measurement of social and cognitive diversity is discussed in a later section of this chapter.

To maintain consistency throughout the study, crowd concept sets were limited to twelve concepts. While actual crowds typically generate many more concepts, there was concern that there might not be enough concept submissions to generate larger crowd concept sets. Therefore, limiting the crowd concept sets to twelve was a fair compromise. To further
ensure that there would be enough concepts to generate crowd concept sets, the experiment was run in two phases. In the first phase of the study, all participants were assigned to the C₀ (isolation) group. This ensured that there were enough concepts to create C₁ participants’ crowd concept in the second phase of the study.

The crowd concepts approach circumvents the challenges of a treatment-control study by creating a curating set of crowd concepts for each C₁ participant. One limitation of this approach, however, is that the composition of the crowd concepts is less controlled than the composition of fixed-population crowds. For example, two industrial designers in the low cognitive diversity group may have crowd concept sets that differ substantially in terms of average concept quality. To control for these effects, the average and maximum scores for each crowd concept set were recorded to assess whether the quality of the crowd concepts had any effect on the quality participants’ concepts.

Summary of Experimental Design

Five unique design groups were created to study the effects of collaboration, social diversity, and cognitive diversity. These groups are depicted in Figure 11. Participants in the isolation group completed the design task independently and were not shown any other concepts. Participants in the four collaborative groups were shown a collection of 12 concepts submitted by other participants. These concepts were selected from concepts previously submitted to fulfill one of four diversity objectives: minimize/maximize social diversity or minimize/maximize cognitive diversity. For example, a mechanical engineer in the high cognitive diversity group would be shown concepts generated by participants with different occupational backgrounds.

<table>
<thead>
<tr>
<th>Participant Group</th>
<th>Participant Concepts</th>
<th>Crowd Concepts</th>
</tr>
</thead>
<tbody>
<tr>
<td>Isolation</td>
<td><img src="image" alt="Isolation Concepts" /></td>
<td></td>
</tr>
<tr>
<td>Collaborative Max Social Diversity</td>
<td><img src="image" alt="Collab Max Social Diversity Concepts" /></td>
<td></td>
</tr>
<tr>
<td>Collaborative Min Social Diversity</td>
<td><img src="image" alt="Collab Min Social Diversity Concepts" /></td>
<td></td>
</tr>
<tr>
<td>Collaborative Max Cognitive Diversity</td>
<td><img src="image" alt="Collab Max Cognitive Diversity Concepts" /></td>
<td></td>
</tr>
<tr>
<td>Collaborative Min Cognitive Diversity</td>
<td><img src="image" alt="Collab Min Cognitive Diversity Concepts" /></td>
<td></td>
</tr>
</tbody>
</table>

FIGURE 11: SUMMARY OF DESIGN GROUPS
To ensure that there was a large pool of concepts to select from, the study was broken into two phases. In the first phase, all participants were assigned to the isolation group, and in the second phase, all participants were assigned to one of the four collaboration groups. Participants from each group were allowed to submit up to three concepts.

Measuring Diversity

Diversity is a familiar social concept. Stories in popular media highlight the importance of cultural diversity in university classrooms, the role of gender diversity in industry, and the effect of diminished species diversity on ecosystem health. Despite our familiarity with the concept of diversity, our understanding of diversity is largely phenomenological. In a research setting, a quantitative measure of diversity is needed so that crowds accurately and reliably compared.

Throughout this section, we distinguish between two types of diversity: observable diversity and total diversity. Observable diversity is a measure of diversity based on some observable attribute of the crowd, such as: age, gender, or occupation. Past research literature examining the effects of diversity on team performance has used measures of observable diversity as indicators for social and cognitive diversity. For example, Dahlin (2005) used national diversity and educational diversity as representative measures of social diversity and cognitive diversity in her assessment of team information usage. While these metrics are useful from a research standpoint, our phenomenological notions of social and cognitive diversity encapsulate more than differences in a single attribute. A team that is nationally diverse but contains individuals that are the same age and gender might be considered as less socially diverse than a team that is moderately diverse in all age, gender, and nationality. This suggests that a more comprehensive measure of “total diversity,” captures differences within and across group attributes, is needed to generate truly diverse crowd concepts for each Ci participant. Both social diversity and cognitive diversity are examples of total diversity, since they depend on a number of more readily observed factors.

For the purpose of this study, operational definitions of social diversity and cognitive diversity were required to permit quantitative comparison between crowds. Developing these operational definitions presented two measurement challenges:

1. How should diversity be measured for observable attributes of crowds, such as age and gender?

2. How can observable diversity measures be used to quantify the total diversity of crowds?
This study adopted Shannon entropy as a measure of observable diversity. Building upon the work of (Balch 2000), a method to calculate total diversity was developed and used to estimate the social diversity and cognitive diversity of C1 participants’ crowd concept sets.

**Observable Diversity**

In 1948, Claude Shannon developed a mathematical measure to quantify the unpredictability of information contained in a communication signal (Shannon 1948). Shannon’s work developing this measure, which is today known as Shannon entropy, had major implications for the efficient transcription, transmission, and reconstruction of messages.

The unpredictability of information is closely related to the diversity of a group. Suppose a message is created by randomly generating a sequence of A’s, B’s, and C’s. The message might look like: ABACBCACB. Shannon entropy asks: how difficult is it to correctly predict the next letter in the message? Calculating the diversity of a group is analogous to calculating Shannon entropy of the message. For example, suppose individuals in a class were assigned to one of three groups – A, B, or C – based on their scores in a test. If individuals were randomly selected and their test scores were recorded, the sequence of scores would look very similar to the sequence of characters in Shannon’s message. When the class has a diverse range of scores, it is more difficult to correctly guess whether a randomly selected individual scored in the A, B, or C range. However, if everyone scores in the A range, the class has no score diversity because a randomly selected individual’s score can be guessed with certainty. Thus Shannon entropy is a good measure of group diversity.

**Definition of Shannon Entropy**

Shannon entropy provides a quantitative, ordinal measure of diversity. Groups with higher Shannon entropy are, by definition, more diverse than groups with lower Shannon entropy. However, care must be taken when comparing groups with different entropy values because the measure is ordinal. A doubling in Shannon entropy is not equivalent to a doubling in diversity.

Shannon entropy can be readily calculated using the following method:

Given a population \( R \) with \( N \) members \( \{r_1 .. r_N\} \) and a classification criterion \( (C) \), which organizes the members of \( R \) into \( M \) groups \( \{c_1 .. c_M\} \), the Shannon entropy \( (H) \) is calculated using Equation 1.
EQUATION 1: SHANNON ENTROPY DEFINITION

\[ H(R) = - \sum_{i=1}^{M} p_i \ln p_i \text{ where } p_i = \frac{|c_i|}{N} \]

For example, the gender diversity – or Shannon entropy in gender – of a crowd with \( N \) individuals can be calculated with Equation 2.

EQUATION 2: SAMPLE CALCULATION OF GENDER DIVERSITY

\[ H(R) = - \left( \frac{N_{\text{female}}}{N} \ln \left( \frac{N_{\text{female}}}{N} \right) + \frac{N_{\text{male}}}{N} \ln \left( \frac{N_{\text{male}}}{N} \right) \right) \]

Group Distance and the Choice of Classification Criterion

Shannon entropy makes no distinction about the degree of difference between groups, so it is important to use care when defining the classification criterion. For example, if members of the crowd were grouped by age, a crowd with groups of 21 and 22 year olds would be treated in the same manner as a crowd with groups of 21 and 51 year olds. If the intent was to identify significant differences in age, it may make more sense to group participants by age range (e.g. 21-30, 31-40, 41-50, etc.) rather than specific age. Ultimately, the choice of classification criterion depends on the type of distinctions that want to be identified.

While it may seem inconvenient that Shannon entropy makes no distinction about the distance between groups, it is also extremely helpful in situations where it is difficult or impossible to define such a distance. For example, while one might be willing to say that a mechanical engineer and an industrial engineer are more similar than a mechanical engineer and architect, it is difficult say whether mechanical and industrial engineers are more similar than landscape architects and urban planners. Even when distance metrics cannot be defined, it is important to identify what level of classification is appropriate to answer the questions of interest.

Dependence on Number of Groups

Once a classification criterion is selected, the members of the crowd can be assigned into groups. For a given classification criterion, crowds with a greater number of groups are generally considered to be more diverse. For example, a crowd with an equal number of Americans, Canadians, and Australians is considered to be more diverse than a crowd of with only Americans and Canadians. The calculation of Shannon entropy captures this effect.
When the proportion of individuals in each of the $M$ groups is constant ($p_i = constant$), the calculation of Shannon entropy reduces to $H(R) = \ln(M)$, and entropy grows logarithmically with the number of groups ($M$). In this situation, Shannon entropy simply measures the number of bits required to represent the groups in base $e$. For example, a fair coin toss has two, equally probable outcomes, heads (H) and tails (T). Representing these outcomes in binary would require 1 bit of information ($H = 0b0$, $T = 0b1$). Similarly, an eight-sided die has eight possible outcomes, which can be represented by 3 binary bits ($1 = 0b000$, $2 = 0b001..8 = 0b111$). The mathematical relationship between the number of groups $M$ and the number of bits required to represent those groups in a base $b$ is equal to $\log_b(M)$. Shannon entropy adopts the natural base $e$ for simplicity.

Shannon entropy’s dependence on the number of groups is consistent with the common notion of diversity. Crowds with more distinct groups are generally considered to be more diverse than crowds with fewer distinct groups. When the proportion of individuals in each group is constant, Shannon entropy grows logarithmically with the number of unique group sizes.

Dependence on Relative Proportion of Groups

For a given number of classification groups, Shannon entropy is maximized when each group has an equal proportion of individuals ($p_i = constant$). While the mathematical proof of this is beyond the scope of this thesis, this property is consistent with the common notion of diversity. Crowds with an equal number of men and women are considered to be more gender diverse than crowds with an unequal proportion of men and women. As the distribution of individuals becomes less uniform, Shannon entropy falls because it becomes easier to correctly guess the classification group of a randomly selected individual. In the limit that all participants are placed in one group ($M = 1$), Shannon entropy approaches zero because an individual’s classification group can be predicted with certainty. This is also consistent with the common view of diversity. A crowd composed of only men or only women is the least diverse configuration possible. Therefore, crowds with Shannon entropy of zero are the least diverse configuration possible.

Total Diversity

Shannon entropy provides a useful measure of observable diversity for specific attributes of the crowd, such as age or gender. However, its application is limited because our notion of diversity captures differences in many attributes simultaneously. This suggests that a more comprehensive measure of “total diversity” is needed.
Motivating Example

Consider the three populations depicted in Figure 12. Each population has four individuals who are assigned to groups based on their age and gender. Which of these populations is the most diverse?

Intuitively, one would say that population A is the most diverse and population C is the least diverse. However, this conclusion is not so easily reached by computing Shannon entropies. There are several possible classification criteria that can be used to compute the Shannon entropy of each population. Age and gender are obvious attributes, and there is one additional classification criterion, which is slightly less obvious: what if the classification criteria captured differences in age and gender?

Independent, Joint, and Diversity Attributes

The third classification criterion is an example of a joint attribute. A joint attribute represents the combination of two or more independent attributes, such as age or gender. In general terms, for a population that is characterized by \(k\) independent attributes, there are a total of \(2^k - 1\) independent and joint attributes that may be used as classification criteria. The full set of independent and joint attributes are hereafter referred to as diversity attributes because each attribute may be used as a classification criterion to compute the Shannon entropy. The degree of a diversity attribute is defined as the number of independent attributes it joins. Following from this definition, one would expect to find \(\binom{k}{d}\) diversity attributes of degree \(d\). The relationship between independent attributes and diversity attributes is demonstrated in Table 2 for independent attributes \(X_1, X_2, X_3\) (\(k = 3\)).

**TABLE 2: INDEPENDENT ATTRIBUTES VS DIVERSITY ATTRIBUTES WHEN \(k=3\)**

<table>
<thead>
<tr>
<th>Degree</th>
<th>Independent Attributes</th>
<th>Diversity Attributes</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>(X_1, X_2, X_3)</td>
<td>(X_1, X_2, X_3)</td>
</tr>
<tr>
<td>2</td>
<td>(X_1X_2, X_1X_3, X_2X_3)</td>
<td>(X_1X_2X_3)</td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
The Relationship between Observable Diversity and Total Diversity

Returning now to the example in Figure 12, Shannon entropy can be computed for three diversity attributes—age, gender, and age x gender. The Shannon entropy for each population and diversity attribute is shown in Table 3.

<table>
<thead>
<tr>
<th>Diversity Attribute</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>Total Diversity (D)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intuition</td>
<td>Most</td>
<td>Middle</td>
<td>Least</td>
<td>D(A) &gt; D(B) &gt; D(C)</td>
</tr>
<tr>
<td>Age</td>
<td>ln(2)</td>
<td>ln(2)</td>
<td>ln(2)</td>
<td></td>
</tr>
<tr>
<td>Gender</td>
<td>ln(2)</td>
<td>ln(2)</td>
<td>0</td>
<td>D(A), D(B) &gt; D(C)</td>
</tr>
<tr>
<td>Age x Gender</td>
<td>ln(4)</td>
<td>ln(2)</td>
<td>ln(2)</td>
<td>D(A) &gt; D(B), D(C)</td>
</tr>
</tbody>
</table>

No single attribute reveals the intuitive conclusion that population A is the most diverse and population C is the least diverse. Classifying by the age attribute indicates that populations A and B are more diverse than C. Similarly, the joint attribute indicates that population A is more diverse than B and C. By combining the conclusions from each diversity attribute, the intuitive conclusion is logically reached.

Hierarchic Social Entropy

Observable diversity and total diversity are related but different concepts. The above example demonstrates that no single measure of observable diversity is sufficient to determine which crowds are more diverse from a total diversity perspective. However, by considering multiple observable diversity measures, a picture of total diversity emerges. This suggests that total diversity can be computed by combining observable diversity measures. A method developed by Balch (2000) called hierarchic social entropy (HSE) provides a robust method to aggregate multiple Shannon entropy measures. This method is adapted to compute total diversity.

Hierarchic social entropy stems from the insight that the Shannon categorization criteria used to assign individuals to groups do not work well when the items being classified exhibit differences at multiple scales. For example, a classification criterion that grouped individuals of the crowd based on their occupation cannot simultaneously distinguish between large occupational differences and more minute differences. A detailed criterion might distinguish between mechanical engineers, industrial engineers, and architects, while a broader criterion would group engineers into one group and architects into another. Neither of these criteria is completely consistent with our notion that mechanical engineers and industrial engineers are more similar to each other than they are to architects.
This suggests that a better representation of groups might use a taxonomic scheme, such as the one shown in Figure 13.

Hierarchic social entropy provides a method to calculate diversity in the presence of such a taxonomic scheme. The method is most readily explained graphically. A sample crowd is reproduced from (Balch 2000) in Figure 14. Individuals are distinguished by their occupations, which are denoted in abbreviated form (i.e. ME indicates mechanical engineer).

Using the occupational taxonomy from Figure 13, individuals can be placed into groups, denoted in red, at different classification levels \( h \). The classification level is a quantitative measure used in the field of numerical taxonomy, which defines the largest permissible distance between two members of the same group. When \( h = 0 \), every individual is assigned to a unique group, and when \( h = 1 \), all individuals are grouped together in one group. For the purposes of this study, the classification level is defined to be \( h = \frac{1}{M} \), where \( M \) is the maximum number of unique groups that can exist at that location in the taxonomy. For example, in the lowest branch of the occupation taxonomy, there are four possible groups, so \( h = \frac{1}{4} \). In general, not all groups must be populated. For example, a crowd may
not have any architects. Despite this, $h = \frac{1}{4}$ because there can be, at most, four unique groups (mechanical engineer, industrial engineer, architect, business consultant).

Hierarchic social entropy is defined as the area under the curve found by plotting the Shannon entropy ($H$) against classification level ($h$). This is demonstrated in Figure 15.

Hierarchic social entropy provides a useful measure of diversity when a classification taxonomy can be defined, and Balch (2000) provides several additional examples of the HSE calculation that demonstrate its utility over the simpler Shannon entropy calculation. Since HSE is computed by combining individual Shannon entropy measures, this suggests that HSE is closely aligned with the notion of total diversity. While HSE and total diversity are similar, there is one important distinction. The example of HSE presented here is motivated by a desire to capture differences within a single attribute, such as occupation, while total diversity must capture differences across attributes, such as age and gender.

From Hierarchic Social Entropy to Total Diversity

To bring the notion of total diversity into alignment with the Balch's HSE calculation method, a taxonomy is required that captures differences across attributes. A method to generate such a taxonomy is presented here and is based upon the combinatorial nature of diversity attributes.

Recall the example in Figure 12 that examined the diversity of three populations. How might one generate a taxonomy based on the three diversity attributes that characterize
these populations? The diversity attributes for the populations can be represented in a tree structure, as shown in Figure 16. Each branch of the diversity attribute tree represents a possible multi-attribute taxonomy. Since there are two branches in this example, there are two possible taxonomies that can be used to group the populations. These two taxonomies are shown in Figure 16 below the tree.

![Diversity Attribute Tree](image)

**FIGURE 16: THE DIVERISITY ATTRIBUTE TREE AND MULTI-ATTRIBUTE TAXONOMIES**

At the finest level of detail, both taxonomies group by the joint age x gender attribute. However, as the classification level increases, one taxonomy groups by age while the other taxonomy groups by gender.

This method is readily generalized to any number of attributes. For \( k \) independent attributes, there will be \( 2^k-1 \) unique diversity attributes in the diversity attribute tree, and the tree will have \( k! \) branches, which each represent a possible multi-attribute taxonomy. Since none of the taxonomies is strictly preferred to the others, total diversity is defined as the average HSE across these taxonomies. For example, Figure 17 shows the HSE graphs for the three populations presented earlier and their relation to total diversity.

**Conclusion**

In this section, a method to compute total diversity was presented. Given several attributes that should be incorporated into a total diversity metric, a set of multi-attribute taxonomies is generated. For each taxonomy, diversity is calculated using the formulation of HSE, introduced by Balch (2000). Total diversity is defined as the average HSE across all multi-attribute taxonomies. This method is used to compute social diversity and cognitive diversity.

56
Measuring Social Diversity

The metric used to measure the social diversity of crowd concept sets incorporated two attributes: age and gender. Participants were asked to report their age and gender in a background survey before they could submit concepts. Based on the participants' responses, they were assigned to one of five age groups: (1) 18-24, (2) 25-34, (3) 35-44, (4) 45-54, (5) 55+; and one of two gender groups: (1) female, (2) male. Participants were also asked to report their nationality in the background survey. While it would have been interesting to incorporate nationality into the social diversity metric, the recruitment procedures, which are discussed later, effectively limited participation to individuals in the US. Since nationality was expected to be predominantly limited to the US, it would not meaningfully contribute to overall social diversity and was therefore excluded from the social diversity metric.
Measuring Cognitive Diversity

The metric used to measure the cognitive diversity of crowd concept sets was based on the US Bureau of Labor Statistic's Standard Occupational Classification (SOC) system (US Department of Labor 2010). The SOC is an occupational taxonomy, which differentiates between occupations at 4 classification levels. The broadest classification level, the system distinguishes between 23 occupations, and finest classification level distinguishes between 840 occupations. A description of the system with several examples is provided in Table 4.

<table>
<thead>
<tr>
<th>Class. Level</th>
<th># of Groups</th>
<th>Sample Group 1</th>
<th>Sample Group 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Major</td>
<td>23</td>
<td>Architecture &amp; Engineering Occupations</td>
<td>Art, Design, Entertainment, Sports &amp; Media Occupations</td>
</tr>
<tr>
<td>Minor</td>
<td>97</td>
<td>Engineers</td>
<td>Art &amp; Design Workers</td>
</tr>
<tr>
<td>Broad</td>
<td>461</td>
<td>Mechanical Engineer</td>
<td>Designers</td>
</tr>
<tr>
<td>Detail</td>
<td>840</td>
<td>Mechanical Engineer</td>
<td>Commercial and Industrial Designers</td>
</tr>
</tbody>
</table>

Since the SOC system is already constructed as a 4-level taxonomy, normal HSE calculation methods can be used to calculate occupational diversity. While factors beyond occupation may contribute to cognitive diversity, occupation is expected to play a dominant prominent role. Because of this, the 4-level occupational taxonomy was taken as a representative of cognitive diversity.

Other Background Survey Factors

Participants were asked several additional questions in the background survey about other factors that might contribute to cognitive diversity. While these factors were not incorporated into the cognitive diversity metric.

Highest Education Level Attained

Participants were asked to report their highest attained level of education. While differences in occupation are probably a better indicator of cognitive differences, the highest education level gives a sense of how specialized a participant's knowledge is in their respective field.

Years of Professional Experience

Participants were asked to report how many years of work experience they have in their profession. This may be a useful factor in situations where more senior
participants from a given occupation approach problems differently than more junior participants.

Importance of Creative Thinking in Work Tasks

Participants were asked to assess how important creative thinking is in their day-to-day work activities on a 5-point Likert scale. While participants' responses are not necessarily an indicator of their creativity, it is reasonable to expect that participants who use creative thinking skills frequently at work may generate more creative concepts.

Study Procedures

Recruitment

Many participants were needed to simulate the large crowds present on crowdsourcing sites and generate enough concepts to reveal any statistically significant differences between the crowds. Recruitment was conducted through a number of different channels, including: email, print flyers, Facebook, and Amazon Mechanical Turk.4

Recruitment emails were sent to all MIT academic departments as well as design-related departments at Yale, Stanford, Georgia Tech, University of Washington, Northwestern, University of Michigan, and the Art Center College of Design. Emails were also sent out to contacts at Boston-area design firms and other acquaintances. Print flyers describing the design challenge and providing a link to CrowdConcept were placed in bathrooms around MIT’s campus, and web links were also placed on Facebook to reach a wider audience.

Use of Amazon Mechanical Turk

Amazon Mechanical Turk (AMT) was also used to recruit participants for the study. AMT has many favorable qualities as a tool for human research. There are over 400,000 workers registered on AMT, and requesters have the ability to limit access to by geographic region and worker approval rates. AMT worker (turkers) were included in the study for several reasons. First, turkers have been used exclusively as subjects in previous creativity studies (Nickerson, Sakamoto, and Yu 2011; Morris, Dontcheva, and Gerber 2012), but it is not well understood how turker performance compares to other common subject populations, such

4 Sample recruitment materials are included in the appendices.
as students or design practitioners. By including both turkers and other populations, it was possible to investigate differences in turker and non-turker performance. Second, assessing whether diversity impacts ideation performance requires a diverse population. Because other participants were expected to come from a more narrow range of design-related backgrounds, including turkers helped diversify the participant pool. Finally, recruiting participants is a slow and tedious process. Using turkers made it possible to increase the sample size by a factor of two.

Allowing turkers to participate in the study also presents several challenges. Turkers work on AMT to earn money, so they need to be compensated reasonably for their participation in the study. Assuming participants would take approximately 30 minutes to complete the study, each turker was paid $2.00 for successful completion of the study. This works out to an expected hourly salary of $4/hr, which is consistent with the compensation in previous studies (Kress and Schar 2012). Due to limited financial resources, however, non-turk participants were not paid to participate in the study. This introduces some inconsistencies in the incentives offered to turk and non-turk participants. While these inconsistencies could potentially create incentive biases, the turkers' compensation was sufficiently low that only turkers who also found the challenge interesting were expected to participate.

Incentives

To increase participation and incentive high quality work, study participants were offered a $50 Amazon gift card if their concept was selected as one of the five best concepts. This is a fairly large monetary incentive for research work, and it was offered for several reasons. First, a large number of concepts were required to study the combined effects of communication capability and diversity. However, without adequate incentives, it was unlikely that non-turk individuals would participate in the study. The size of the monetary incentive was also chosen to minimize potential incentive bias stemming from the fact that turk participants were compensating for participating while non-turk participants received nothing.

Participant Workflow

Participants completed the study in a structured manner, depicted in Figure 18. After signing up and consenting to the terms of the study, participants completed a brief background survey. Responses from this survey were used to generate crowd concepts for each C1 participant. Upon completing the background survey, participants began the design task discussed earlier. Participants were allowed to submit one to three design concepts and could move on to the feedback stage at any point after submitting one design concept. The feedback stage provided an opportunity for the participants to give the
researchers general feedback about the study and the features on CrowdConcept. Finally, participants were directed to a page where they could view all concepts submitted to date.

FIGURE 18: PARTICIPANT WORKFLOW
Chapter 5
Results and Discussion

Overview of Participant Sample

Over a four-week period, 244 individuals enrolled in the study and 66 of those submitted one or more concepts. Table 5 provides a summary the number of participants and concepts submitted for each of the five test groups.

<table>
<thead>
<tr>
<th>Group</th>
<th># of Participants</th>
<th># of Concepts</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Total</td>
<td>Non-Turk</td>
</tr>
<tr>
<td>Total</td>
<td>66</td>
<td>39</td>
</tr>
<tr>
<td>Isolation</td>
<td>19</td>
<td>9</td>
</tr>
<tr>
<td>Low Social Div.</td>
<td>10</td>
<td>6</td>
</tr>
<tr>
<td>High Social Div.</td>
<td>12</td>
<td>7</td>
</tr>
<tr>
<td>Low Cognitive Div.</td>
<td>12</td>
<td>8</td>
</tr>
<tr>
<td>High Cognitive Div.</td>
<td>13</td>
<td>9</td>
</tr>
</tbody>
</table>

Participant Survey Responses

Participants completed a brief survey before submitting their concepts. The survey responses reveal some differences between the turk and non-turk populations. For example, participants in the non-turk population are generally from technical or scientific backgrounds, while the turk population has greater representation from the art and design domains. Responses for the all participants as well as the turk and non-turk populations are summarized for below.

Gender

The population showed a high degree of gender balance. Of the 66 participants, 35 were male and 31 were female. There were no significant differences in gender representation in the turk and non-turk populations.
Age

Participants were moderately young compared to the teams found at many organizations. Eighty percent of participants reported being under the age of 35. The turk population was older (median age of 33) than the non-turk population (median age of 23).

Highest Level of Education

Participants were well educated. More than half of participants reported completing at least a Master’s degree, and an additional 27 percent report completing at least some college. Sixty percent of participants who had completed some college were still under the age of 21, which suggest that they are actively pursuing an Associate’s or Bachelor’s degree. Turkers were less educated than other participants. Only 37 percent of turkers reported having at least a Bachelor’s degree compared to 63% from non-turk participants.

Importance of Creative Thinking at Work

Participants were asked to assess how important creative thinking was in their day-to-day activities at work. While participants’ responses to this question are not necessarily representative of their actual creative capacities, the responses give a sense of how frequently participants use creative thinking patterns at work. The distribution of responses is depicted in Figure 19. The responses suggest that non-turk participants may be more familiar with creative thinking than the turkers.

![Figure 19: Importance of Creative Thinking at Work](image)

Occupation

Participants were asked to report their occupation using the US Bureau of Labor Statistics’ Standard Occupational Classification system (US Department of Labor
Responses summarized in Figure 20 indicate that turk and non-turk participants came from significantly different occupational backgrounds. Non-turk participants came predominately from technical and scientific backgrounds. While these occupations were much less common in the turk population, it had a greater fraction of participants from art and design disciplines.

![Figure 20: Participant Occupations](image)

### Completion Time

The amount of time participants spent completing the ideation phase of the design study was recorded to see if completion time was correlated with other factors in the study, such as the number of concepts submitted. The data shows a strong, nearly linear relationship between the completion time and number of concepts submitted. The median completion time for participants that submitted, one, two, and three concepts was 13.1, 27.4, and 43.2 minutes, respectively.

Turk participants completed their concept submissions more quickly than non-turk participants ($p < 0.01$). The median completion time for a turk participant was 12.7 minutes, while the median completion time for a non-turk participant was 31.2 minutes.

One possible reason that turk participants completed the study more quickly is that they submitted fewer concepts on average than non-turk participants. Only 33% of turk participants submitted two concepts and none submitted three concepts. In contrast, 45% of non-turk participants submitted two or more concepts and 23% submitted three concepts. These differences are not particularly surprising given that turk participants were paid $2.00 regardless of how many concepts they submitted.

### Concepts

All 101 concepts submitted during the study can be found online at [http://crowdconcept.herokuapp.com](http://crowdconcept.herokuapp.com). Several examples of concept submission are also included in the appendices.
Reviewing Submissions

Concepts were reviewed by workers on Amazon Mechanical Turk. Reviews were screened to remove reviews completed in haste or without care, and the remaining reviews were used to calculate average need, novelty, usefulness, and feasibility ratings for each concept. The review data was characterized by low inter-rater reliability. While the review data was still used for this study, future research should seek to understand what factors led to low inter-rater reliability.

Review Statistics

Review HITs were posted for each concept and could be reviewed by up to 50 workers. In total, over 700 AMT workers participated in the review process and completed 4,692 reviews. The median number of reviews per worker was 2, but several workers completed over 50 reviews. The median time to review a concept was just over 70 seconds, and 68% of reviews included at least one comment. In nearly 60% of the reviews, workers indicated that they were target users of the proposed concept, which validates the earlier assumption that AMT workers are representative of public restroom users.

Screening Reviews

Past studies using AMT workers highlight the need to screen for reviews completed without due care and consideration (Kress and Schar 2012). For example, some workers may randomly select ratings so that they can complete as many reviews as quickly as possible to increase their compensation. Many techniques have been used to screen reviews. Questions can be reverse-scored to identify possible biases, submissions completed too quickly can be discarded, or reviews that deviate significantly from the mean or median values can be flagged. All of these methods seek to identify suspect workers so their contributions can be removed from the dataset. In the context of this study, three indicators were used to flag suspect workers.

Review Time

The most obvious indicator of effort is the amount of time a worker spends completing a review. Workers who spend more time reviewing a concept are likely taking time to read and understand how the concept works, while workers who complete it quickly likely do not have enough time to understand how the concept works. The median review time was 72 seconds. An average review time was computed for each worker. Workers were flagged if their average review was below
the 25th percentile review time (29s). Approximately 5% of workers fell below this threshold and were flagged.

Comment Word Count

The frequency and amount of comments a worker provides is another good indicator of their engagement in the review task. For each review, the total number of words provided in the comment regions of the review was tabulated. Using this data, the 75th percentile comment word count was computed for each worker. Any worker whose 75th percentile comment count was zero was flagged. Approximately 20% of workers were flagged based on the comment criteria.

Rating Error

Reviews that deviate significantly from the average are also suspect. However, the rating criteria are subjective in nature, so care must be made to avoid eliminating workers who simply have different opinion from the mainstream workers. For example, one worker may consider a concept to be very novel, while another use may be aware of prior art that is similar to the concept.

To identify workers whose reviews deviate significantly from the typical ratings, the median ratings for each concept were computed. Differences between a concept’s median ratings and review’s ratings were computed, and the root mean square (RMS) rating error was calculated from differences in the five rating criteria. Finally, for each worker, an average RMS rating error was calculated. Any worker with an average RMS rating error greater than the 75% RMS rating was flagged. This resulted in approximately 25% of workers being flagged.

A total of 60 workers received at least two flags. Each flagged worker’s submissions were reviewed to assess whether they were truly suspect. Of the 60 flagged workers, 34 workers completed only one review. These workers’ reviews were not excluded because there were not sufficient data to determine whether their reviews were suspect. Furthermore, these reviews represented less than 1 percent of all reviews, so the impact of the reviews on the dataset would be small. Of the remaining 26 workers, 24 were identified as highly suspect, and their reviews were removed from the dataset. In total, 717 reviews, or about 15% of all reviews, were excluded from the dataset.

Summary of Review Trends and Reliability

After filtering out suspect reviews, 3,975 reviews remained. Next, the reviews were analyzed to determine whether there was general agreement between the 700+ reviewers that participated in the review process. Figure 21 depicts the distribution of review ratings
for each criterion. The first four criteria - clarity, need, novelty, and usefulness - exhibit a significant amount of upward bias. Feasibility is more uniformly distributed.

![Graph showing distribution of ratings](image)

**FIGURE 21: DISTRIBUTION OF RATINGS**

**Inter-Rater Agreement**

In general, the degree of agreement between reviewers was low. Average standard errors were on the order of 1-point for all five rating criteria. Since the ratings are on a 5-point scale, this represents an error of approximately 20%. Standard errors were slightly lower for more highly rated concepts, which suggests that there may have been greater agreement about the quality of the best concepts.

Inter-rater reliability (IRR), which provides a more robust measure of the degree of agreement between reviewers, was also found to be low. Inter-rater reliability metrics compare the differences in reviewer ratings for each concept to the differences that might arise by chance if reviewers randomly assigned ratings. While there are a number of common IRR metrics, Krippendorff's alpha was chosen to evaluate IRR because all other metrics require that each reviewer evaluate each concept. Since most reviewers only reviewed a subset of concepts, the other IRR metrics could not be used. Krippendorff's alpha compares the average difference between pairs of ratings for a given concept (in-group differences) to the average difference between all possible pairs of ratings (between-group differences). When the in-group differences are much lower than the between-group differences, Krippendorff's alpha approaches 1 and agreement between reviewers is high. Typically, an alpha of 0.65 or greater is indicative of strong agreement between reviewers (Poetz and Schreier 2012).

Krippendorff's alpha was computed for each of the five rating criterion. Alpha was found to be: 0.10 for clarity, 0.18 for need, 0.06 for novelty, 0.25 for usefulness, and 0.17 for feasibility. All of these values are well below the acceptable threshold of 0.65. Alpha was consistently low despite attempts to remove additional reviews from the dataset. For example, limiting the reviews to users who reported that they were a target user of the concept had no significant effect on alpha. Imposing stricter screening rules also had marginally positive effects on the alpha.

There are several possible reasons why IRR is so low. First, there was little overlap between reviewers because there were a large number of reviewers and the average number of reviews per reviewer was low. This means that reviewers may not have seen enough
common concepts to calibrate their ratings. It is also possible that some suspect reviews were not removed from the dataset. Since suspect reviews are effectively random, they would significantly reduce the IRR calculation. Finally, the narrow distribution of mean ratings, shown in Figure 21, also likely contributed to low IRR values. Krippendorff’s alpha uses the overall distribution of ratings to estimate the expected differences between reviews, so situations where the ratings are tightly clustered lead to a low alpha. This effect is clearly seen in the calculation of clarity and novelty, which both have very tight rating distributions and therefore low IRR. It is unclear whether the ratings are tightly clustered because reviewers were overly kind or if the concepts were actually comparable. Ideally, the ratings would make full use of the 5-point Likert scoring range.

While inter-rater reliability was low, the workers’ reviews were still used to compute the mean scores for each concept. Future work should attempt to better understand if differences in the opinions of AMT workers are systematic, which would suggest that workers are not suitable reviewers, or if they arise due to other factors, such as too little overlap between reviewers.

Overview of Concept Ratings

Concept reviews were averaged to generate mean need, novelty, usefulness, and feasibility ratings for each concept. The need x novelty x usefulness (NNU) and need x novelty x usefulness x feasibility (NNUF) overall quality metrics were calculated for each concept by multiplying the mean ratings.

The distributions of mean ratings are displayed in Figure 22. Ideally, one would hope that the mean rating distributions were more uniformly or normally distributed, so that a small percentage of concepts were among the highest rated. The data, however, shows that mean ratings are upward biased for the need, novelty, and useful criteria. Therefore, a large percentage of concepts are among the highest rated. This makes it difficult to assess differences between the best-rated concepts and other concepts because rating differences between concepts in the upper quartile are small. For example, the difference in need rating between the 75th and 95th percentile concept is only 0.19 while the difference between the 5th and 25th percentile concept is 0.77.
Assessing Ideation Performance

All of the hypotheses in this thesis claim that relationships exist between specific configurations of the crowd and the crowd's creative performance. However, there is no widely accepted method to evaluate creative performance. In the research community, ideation performance is often evaluated in terms of the average quantity, variety, novelty, or quality of ideas generated (Shah, Smith, and Vargas-Hernandez 2003; Nelson et al. 2009). While these metrics give a sense of how effectively the design space was searched, they are not necessarily relevant in a design context, where it is more important to identify which configurations of the crowd produce the best ideas. There are two general ways to formulate this question:

1) Does one crowd configuration generate more top-ranked ideas than other configurations?

2) Do the best ideas from one crowd configuration score significantly higher than the best ideas from other configurations?

The first question looks at the quantity of top ideas while the second question assesses the quality of top ideas. This section reviews various data analysis techniques used to evaluate ideation performance. The rationale and methods for using each technique are outlined in this section. Since these methods are used to test all of the hypotheses, results are presented in later sections for each hypothesis.

Null Hypothesis Testing

Null hypothesis testing is an important first step in the statistical assessment of a hypothesis. Typically, a hypothesis asserts that there is a specific relationship between two
observed phenomena. The null hypothesis represents the default assertion that the two phenomena have no relationship. Demonstrating that the null hypothesis can be rejected with a high level of confidence suggests that some relationship between the phenomena exists. Generally, confidence estimates are based on the probability that differences between the treatment and control groups could have been observed by chance. While, rejecting the null hypothesis is important, it does not demonstrate that the phenomena necessarily exhibit the predicted relationship. Additional tests are needed to confirm or reject this relationship.

The hypotheses in this thesis are formulated as both conventional and null hypotheses. Hypotheses 2 and 3 assert that cognitive diversity and collaborative capability are positively correlated with the quality of the concepts generated by the crowd. The null assertion for these hypotheses is that cognitive diversity and collaborative capability have no relationship to concept quality. While the H2 and H3 are formulated as typical hypotheses, H1 is formulated as a null hypothesis: social diversity has no effect on concept quality. To confirm this hypothesis, it is sufficient to show that the H2 cannot be rejected with a significant level of confidence.

There are a number of common tests used to assess null hypotheses. The most prevalent are the Student t-test, Z-test, F-test, and analysis of variance (ANOVA). These four tests are parametric tests that require continuous data and assume that the data is normally distributed. The Student t-test and Z-test compare differences between the mean value for a specific group and the whole population mean. When the variance of the population is known and the sample size is large, the Z-test is preferred. The F-test and ANOVA compare differences in variance between independent groups within a population. When more than two groups need to be compared, ANOVA is used.

Similar test exist for non-normal and non-parametric datasets. These tests are preferred in this study because the mean concept ratings were not normally distributed for most of the rating criteria. The Kruskal-Wallis (KW) test was selected to assess null hypotheses. The KW test is an extension of the ANOVA test and was preferred because it can be used to compare three or more groups simultaneously. The KW test relies on one key assumption: the distribution for each group must have a similar shape.

Best-Versus-Rest Comparison

While null hypothesis tests assess whether the distribution of concept quality is statistically different between the groups, comparing the overall distributions is of little significance in a design setting. The more important question to ask is how do the groups compare in ability to generate the top-ranked concepts. For example, is the collaborative group more likely to generate highly novel concepts than the isolated group?
The first step in operationalizing this question is determining what constitutes a top concept. One approach is to identify concepts that do well in one or more rating criteria. Poetz and Schreier (2012) take this approach and define thresholds on the novelty, usefulness, and feasibility ratings. Concepts that exceed the rating threshold are put into the top concept group for that particular criterion. The best overall concepts are those that are in the top concept group for multiple criteria. This definition is consistent with the earlier discussion of the Real-Win-Worth it screening method. The most promising ideas are those that are rated highly on all of the relevant performance criteria.

A best-versus-rest comparison method was used to evaluate whether differences in the number of top concepts generated by each group was significantly different from what would otherwise be expected in the absence the treatment criterion. For each of the four rating criteria (need, novelty, usefulness, and feasibility) and the two overall quality criteria (NNU and NNUF), concepts were flagged as a top concept if they had a mean rating in the upper quartile of the respective criterion.

To test whether the treatment criterion had any effect on the likelihood of generating a top concept, the KW test was again used. In this test, concepts were assigned a value of 1 if they were in the top concept group and a value of 0 otherwise. When statistically significant differences are observed between the actual and expected number of top concepts from each group, there is a strong indication that the treatment criterion, which differentiates those groups, has a significant effect on the likelihood of generating a top concept.

**Bootstrap Percentile Performance**

The best-versus-rest comparison method assesses whether different groups are more or less likely to generate top concepts. While this method is useful, it does not necessarily capture differences in degree between the top concepts. For example, while one group may be more likely to generate a top concept, this does not necessarily suggest that the best concepts from other groups are worse than the best concepts from the first group. In fact, other groups may produce fewer top concepts, but those concepts may receive significantly higher ratings than the first group’s concepts. This suggests that another useful way to look at the data is to see how the concept ratings from different groups compare in the upper percentiles. For example, does the 95% concept from one group score higher than the 95% concept from another group?

Bootstrap resampling techniques can be helpful when comparisons, such as percentile performance, need to be made between groups. The general idea behind bootstrapping is to create many new populations that are sampled with replacement from the original population. An estimator, such as best concept score, is computed for each resampled
bootstrapped population, and the distribution of the estimator across the bootstrap populations gives an indication of both the mean and confidence interval for that metric.

Bootstrapping can be used to compare two groups to each other as well. For example, to compare the best concepts between two groups, many bootstrapped group populations could be generated. The maximum scores from each pair of bootstrapped groups can be compared to compute the probability that the best concept from one group scores better than the best concept from another group. When the scores are close to each other, the bootstrap resampling technique will report a probability near 50%, which indicates that the differences are not significant. As the probability approaches 100%, the differences between the groups become very significant, and one can conclude that one group’s best concept scores statistically higher than other group’s best concept.

If the two groups are not the same size, it is important to be cautious how the bootstrap groups are created. Consider a simple example where group A has N concepts and the group B has 2N concepts. Also suppose that the quality of the concepts is distributed normally, and, unbeknownst to the researcher, the grouping has no effect on the quality of concepts (i.e. concept quality is independent of the grouping criteria). If the researcher compares the means of the two groups, he or she will find that the mean values are similar, assuming N is large enough to reasonably sample the underlying concept quality distribution. However, if the researcher compares the best concept scores from groups A and B, he or she will likely conclude that the best concepts from group B score better than the best concepts from group A. However, since the grouping has no effect on the quality distribution, this conclusion would be erroneous. Because group B has twice as many concepts as group A, it is statistically more likely that one of group B’s concepts is drawn from the tail of the concept quality distribution. This effect of group size on the sample estimator may be important when there is reason to believe that differences in group size are meaningful. For example, the crowd is generally larger than a team, so the crowd may be statistically more likely to identify an extreme solution even if there are no underlying differences between the crowd and group. However, when differences in group size are not meaningful and arise because of conditions of the research study, the bootstrap comparison technique becomes biased.

To control for this bias, it is useful to limit the size of the bootstrap populations to the size of the smallest group being compared. While this will not totally eliminate the bias, it significantly reduces it. Consider groups A and B again, and assume that 1,000 bootstrap populations of size N were generated for each group. Because the bootstrap population size for group B is cut in half, the probability that extreme values appear in the bootstrap populations is lower. Therefore, the comparison between groups A and B will be less biased by differences in sampling size. Based on these examples, a simple rule is used to make comparisons between groups of different sizes: limit the size of bootstrap populations to the size of the smallest group being compared.
To compare differences in the performance of groups, a bootstrapping method similar to the one described above is used. Bootstrap population size is limited to the smallest group size, and 1,000 bootstrap populations are generated for each group. To compare the performance across groups, the score is computed for each bootstrap population at every 5th percentile from 0% to 100%. The percentile scores of one group's populations are compared to the other group's populations to calculate a p-value at each percentile performance level. When differences between the groups are significant at percentiles above the 75%, it is reasonable to conclude that the best ideas from one group score better than the best ideas from another group.

**Summary of Data Analysis Techniques**

Three analysis techniques were introduced in this section that will be used to assess the validity of the three hypotheses described earlier. The null hypothesis test assesses whether differences between groups arise by chance. A best-versus-rest analysis compares the likelihood of different groups to top-rated concepts. Finally, using bootstrap analysis methods, the magnitude and significance of differences in the quality of concepts from different groups can be evaluated. Each of these analysis techniques lends a different type of insight into the performance of the crowd.

**Effect of Social Diversity**

_Hypothesis 1_ Social diversity will have a negligible effect on the crowd's ideation performance.

Social diversity was predicted to have a negligible effect on the crowd's ideation performance because the social characteristics of the participants could not readily observed in the study. The data suggest that social diversity has a positive effect on ideation performance. This finding contradicts predictions from the creativity literature, and suggests that further research is needed to understand the causal link between social diversity and creative performance.

Figure 23 depicts the social diversity distribution of the 47 C1 participants' crowd concept sets. While it is possible to evaluate differences between concepts generated by participants assigned to the low and high social diversity objective groups, a better method is to define new social diversity groupings based on the overall distribution of crowd concept sets. This method is preferred for several reasons. First, there are some instances where the crowd concept selection algorithm did not perform well. For example, two crowd concept
sets from the low social diversity group appear with high social diversity scores. Furthermore, the sample size is effectively doubled when participants assigned to the low and high cognitive diversity objective groups are included in the analysis.

To assess the first hypothesis, concepts generated by the collaborative group (C1) were split into two social diversity groups:

$$SD_{\text{low}}$$  ideas from participants whose curated crowds had a social diversity of less than 0.6.

$$SD_{\text{high}}$$  ideas from participants whose curated crowds had a social diversity greater than 0.6.

The threshold value was chosen because it corresponded to a significant discontinuity in the social diversity distribution. Of the 72 total concepts generated by the collaborative group, 31 were assigned to $$SD_{\text{low}}$$ and 41 were assigned to $$SD_{\text{high}}$$.

**Null Hypothesis Tests**

The null hypothesis tests indicate that a relationship between social diversity and ideation performance is highly probable. If H1 were valid, the Kruskal-Wallis tests should indicate that the null hypothesis cannot be rejected. However, the tests summarized in Table 6 indicate that the differences between the ratings for low and high social diversity groups are statistically different. The novelty criterion is the one significant exception to this finding ($p = 0.200$).
TABLE 6: NULL HYPOTHESIS TESTS FOR SOCIAL DIVERSITY

<table>
<thead>
<tr>
<th>Rejected?</th>
<th>Need</th>
<th>Novelty</th>
<th>Useful</th>
<th>Feasible</th>
<th>NNU</th>
<th>NNUF</th>
</tr>
</thead>
<tbody>
<tr>
<td>p-value</td>
<td>0.012</td>
<td>0.200</td>
<td>0.001</td>
<td>0.024</td>
<td>0.004</td>
<td>0.001</td>
</tr>
</tbody>
</table>

An example output from the KW test is shown in Figure 24. The graph depicts the distribution of concept scores for each social diversity group. The test is primarily geared towards analyzing differences in median values, depicted as a red line, and the interquartile range, shown by the extents of the blue box. The KW test shows fairly conclusively that the median concepts from the high social diversity group score better than the 75th percentile concepts from the low social diversity group. This finding suggests that significant differences will also be observed between in the BVR and bootstrap comparisons.

![FIGURE 24: KRUSKAL-WALLIS TEST COMPARING SOCIAL DIVERSITY GROUPS FOR NNUF](image)

**Best-Versus-Rest Comparison**

The BVR comparison indicates that the high social diversity group is more likely to generate top-rated concepts in the usefulness ($p = 0.003$), NNU ($p = 0.036$), and NNUF ($p = 0.065$) rating criteria. Similar trends were observed for need, but the differences between groups are not significant enough to draw conclusions ($p > 0.10$). These results suggest that socially diverse crowds produce a greater number of highly useful and high quality concepts.

The results from the BVR comparison are depicted in Table 7. The table compares the observed number of top concepts from each diversity group to the number that would be expected if the social diversity groupings had no effect on the number of top concepts.
Expected numbers are displayed in parentheses below each observation. A KW test is used to assess whether the number of top concepts from each group is significantly different from the expected numbers, and a p-value is reported for each rating criteria.

**TABLE 7: BVR COMPARISONS FOR SOCIAL DIVERSITY GROUPS**

<table>
<thead>
<tr>
<th>Social Diversity</th>
<th><strong>Need</strong></th>
<th></th>
<th><strong>Novelty</strong></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SD\text{low}</td>
<td>SD\text{high}</td>
<td>SD\text{low}</td>
<td>SD\text{high}</td>
</tr>
<tr>
<td>Top Concepts</td>
<td>(n=31)</td>
<td>(n=41)</td>
<td>(n=31)</td>
<td>(n=41)</td>
</tr>
<tr>
<td>Obs (Exp)</td>
<td>4</td>
<td>13</td>
<td>8</td>
<td>14</td>
</tr>
<tr>
<td></td>
<td>(7.7)</td>
<td>(10.2)</td>
<td>(9.5)</td>
<td>(12.5)</td>
</tr>
<tr>
<td>Other Concepts</td>
<td>(n=31)</td>
<td>(n=41)</td>
<td>(n=31)</td>
<td>(n=41)</td>
</tr>
<tr>
<td>Obs (Exp)</td>
<td>26</td>
<td>28</td>
<td>23</td>
<td>27</td>
</tr>
<tr>
<td></td>
<td>(23.3)</td>
<td>(30.8)</td>
<td>(21.5)</td>
<td>(28.5)</td>
</tr>
<tr>
<td>p-value</td>
<td>0.113</td>
<td>0.450</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Social Diversity</th>
<th><strong>Useful</strong></th>
<th></th>
<th><strong>Feasible</strong></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SD\text{low}</td>
<td>SD\text{high}</td>
<td>SD\text{low}</td>
<td>SD\text{high}</td>
</tr>
<tr>
<td>Top Concepts</td>
<td>(n=31)</td>
<td>(n=41)</td>
<td>(n=31)</td>
<td>(n=41)</td>
</tr>
<tr>
<td>Obs (Exp)</td>
<td>3</td>
<td>17</td>
<td>5</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td>(8.6)</td>
<td>(11.4)</td>
<td>(7.3)</td>
<td>(9.7)</td>
</tr>
<tr>
<td>Other Concepts</td>
<td>(n=31)</td>
<td>(n=41)</td>
<td>(n=31)</td>
<td>(n=41)</td>
</tr>
<tr>
<td>Obs (Exp)</td>
<td>28</td>
<td>24</td>
<td>26</td>
<td>29</td>
</tr>
<tr>
<td></td>
<td>(22.4)</td>
<td>(29.6)</td>
<td>(23.7)</td>
<td>(31.1)</td>
</tr>
<tr>
<td>p-value</td>
<td>0.003</td>
<td>0.197</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Social Diversity</th>
<th><strong>NNU</strong></th>
<th></th>
<th><strong>NNUF</strong></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SD\text{low}</td>
<td>SD\text{high}</td>
<td>SD\text{low}</td>
<td>SD\text{high}</td>
</tr>
<tr>
<td>Top Concepts</td>
<td>(n=31)</td>
<td>(n=41)</td>
<td>(n=31)</td>
<td>(n=41)</td>
</tr>
<tr>
<td>Obs (Exp)</td>
<td>5</td>
<td>16</td>
<td>4</td>
<td>13</td>
</tr>
<tr>
<td></td>
<td>(9.0)</td>
<td>(12.0)</td>
<td>(7.3)</td>
<td>(9.7)</td>
</tr>
<tr>
<td>Other Concepts</td>
<td>(n=31)</td>
<td>(n=41)</td>
<td>(n=31)</td>
<td>(n=41)</td>
</tr>
<tr>
<td>Obs (Exp)</td>
<td>26</td>
<td>25</td>
<td>27</td>
<td>28</td>
</tr>
<tr>
<td></td>
<td>(22.0)</td>
<td>(29.0)</td>
<td>(23.7)</td>
<td>(31.3)</td>
</tr>
<tr>
<td>p-value</td>
<td>0.036</td>
<td>0.065</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Bootstrap Comparisons

Bootstrap comparisons were made between the low and high social diversity groups to investigate the relationship between social diversity and the quality of the best concepts from each diversity group. The results indicate that the best concepts from SD\textsubscript{high} score better in the need, useful, feasible, and NNUF criteria. These findings provide generally strong support for H1 as well.

Two bootstrap comparisons are shown in Figure 25. Each graph compares the distribution of scores for SD\textsubscript{low} and SD\textsubscript{high} using the bootstrap comparison method discussed earlier. The graphs depict the mean scores with one standard error range at each percentile. This permits comparisons such as: how does the 50\textsuperscript{th} percentile concept from SD\textsubscript{low} compare to the 50\textsuperscript{th} percentile concept from SD\textsubscript{high}? When standard error regions of the groups overlap, it indicates that the differences in mean values are not statistically significant. A bootstrapped p-value is plotted at each percentile to help determine whether the differences between the two groups are statistically significant.

Using these graphs, one can assess whether best concepts from the high social diversity group outscore the best concept the low social diversity group by comparing the mean values in the upper percentiles. To operationalize these comparisons, the percentile difference between the groups is computed. A summary of these comparisons and the statistical significance of the findings are shown in Table 8.
TABLE 8: SUMMARY OF BOOTSTRAP COMPARISONS FOR SOCIAL DIVERSITY GROUPS

<table>
<thead>
<tr>
<th>Difference</th>
<th>Need</th>
<th>Novelty</th>
<th>Useful</th>
<th>Feasible</th>
<th>NNU</th>
<th>NNUF</th>
</tr>
</thead>
<tbody>
<tr>
<td>$SD_{\text{high}} - SD_{\text{low}}$</td>
<td>%</td>
<td>-0.3%</td>
<td>4.7%</td>
<td>16.3%</td>
<td>11.0%</td>
<td>35.9%</td>
</tr>
<tr>
<td>$SD_{\text{low}}$</td>
<td>(p)</td>
<td>(0.040)</td>
<td>(0.456)</td>
<td>(0.031)</td>
<td>(0.026)</td>
<td>(0.110)</td>
</tr>
</tbody>
</table>

The findings from the bootstrap comparison are generally consistent with the BVR analysis. The results indicate that the best concepts from $SD_{\text{high}}$ score better than the best concepts from $SD_{\text{low}}$ in the need, useful, feasible, and NNUF criteria. The differences between groups for the need and useful criteria are relatively small. The best concepts from $SD_{\text{high}}$ score 3-5% better than the best concepts from $SD_{\text{low}}$. More significant differences are observed for feasibility ratings and the NNUF ratings. In the BVR analysis, statistically significant differences between the low and high diversity groups were only observed for the useful criteria. In contrast, the bootstrap comparison revealed statistically significant differences for all criteria except novelty and NNU.

Discussion

An assessment of the C1 participants' concepts suggests that low social diversity crowds produce lower quality concepts than higher social diversity crowds. This finding is inconsistent with H1, which predicted that no relationship would exist between social diversity and ideation performance. These results also deviate from the general position in the creativity literature, which suggests that greater social diversity reduces team cohesiveness and degrades concept quality (Hennessey and Amabile 2010).

TABLE 9: WERE DIFFERENCES OBSERVED BETWEEN $SD_{\text{LOW}}$ AND $SD_{\text{HIGH}}$?

<table>
<thead>
<tr>
<th>Comparison Method</th>
<th>Need</th>
<th>Novelty</th>
<th>Useful</th>
<th>Feasible</th>
<th>NNU</th>
<th>NNUF</th>
</tr>
</thead>
<tbody>
<tr>
<td>BVR</td>
<td>N</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td>Y</td>
</tr>
<tr>
<td>Bootstrap</td>
<td>Y</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td>Y</td>
</tr>
<tr>
<td>Conclusion</td>
<td>N</td>
<td>N</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>Y</td>
</tr>
</tbody>
</table>

findings show support when $p < 0.10$

One reason that social diversity may have improved ideation performance is that social diversity can expose study participants to problems encountered by different age and gender groups. For example, a concept about accessibility for the elderly might reveal a niche of restroom users with acute needs that a participant was not aware of or had not considered. This may help the participant identify problems and develop solutions that are catered to specific needs rather than more general issues, such as sanitation. When
concepts address a more concrete problem, one would expect that the ratings related to the severity of the need and the usefulness of the concept should increase.

While the findings differ from the conclusions in the creativity literature, they should not suggest that the crowd is generally immune to the negative effects of social diversity observed in team settings. Communication capability in the study’s collaborative group was extremely limited, even compared to most CCW platforms. Participants were not informed about the backgrounds of other participants; they were simply shown concepts generated by others. Information about other participants could only be inferred through their concepts. For example, it is probably fair to assume that a woman generated a concept that dealt with issues in women’s restrooms.

While the small amount of communication capability limits the generalizability of results from this study, it also highlights an important difference between crowds and ideation teams. In a team setting, communication capability cannot be controlled, but in a crowd setting, the designers of CCW platforms have much greater control over the interactions between crowd members. Since communication capability is an important factor that affects the observability of social differences in the crowd, designers can control the amount of communication capability to mediate any negative effects of social diversity.

Further research is needed to better understand how social diversity affects ideation performance as communication capability increases. It seems reasonable to hypothesize that increasing communication capability will increase the observability of social differences, which may lead to reduced ideation performance. Regardless of whether this hypothesis is confirmed or denied, understanding how communication capability mediates the effects of social diversity will help the designers of crowdsourcing platforms make informed decisions about the amount of communication capability needed to maximize the crowd’s creative output. Experimental designs to test for these complex effects will be complex, so it may be best to analyze existing datasets from CCW platforms.

Effect of Cognitive Diversity

_Hypothesis 2_ Cognitive diversity will have a positive effect on the crowd’s ideation performance.

Cognitive diversity was predicted to have a positive effect on the crowd’s ideation performance. This prediction was based on findings in the creativity literature, which showed that cognitively diverse teams were more likely to identify good solutions when problems were complex and multidisciplinary (Kavadias and Sommer 2009). The data reveals some support for this hypothesis. C1 participants with more cognitively diverse
crowd concept sets were more likely to generate highly novel and highly feasible concepts than other participants. These best concepts from these participants also scored significantly higher in the need, useful, feasible, and NNUF scoring criteria. These findings collectively confirm H2 and suggest that cognitive diversity has a positive effect on ideation performance.

Overview of Sample

The methods used to assess H2 were similar to those used to assess H1. The cognitive diversity distribution for the 47 crowd concept sets is shown in Figure 26. Compared to the social diversity distribution, cognitive diversity is more evenly distributed across the full diversity range, so there is no clear distinction between the high diversity crowd concept sets and the others. Instead, crowd concept sets were split into three diversity groups corresponding to the lower (CDlow), middle (CDmid), and upper (CDhigh) quartile ranges. The grouping criteria resulted in 16, 38, and 18 concepts in the low, mid, and high diversity groups, respectively.

Figure 26 also shows that the crowd concept selection algorithm worked well for cognitive diversity. The total cognitive diversity for crowd concept sets in the low and high cognitive diversity design groups are tightly distributed at the extremes of the diversity spectrum.
Null Hypothesis Tests

The Kruskal-Wallis tests reject the null hypothesis for all criteria except novelty with at least 95% confidence. Some criteria are rejected with even high confidence. For example, the null hypotheses for the feasibility and NNUF criteria are rejected at the 99% confidence. While the null hypothesis for novelty is not rejected at 95% confidence level, the differences between cognitive diversity groups appear to have a much more significant effect on novelty \((p=0.06)\) than the differences between the social diversity groups \((p=0.20)\).

<table>
<thead>
<tr>
<th></th>
<th>Need</th>
<th>Novelty</th>
<th>Useful</th>
<th>Feasible</th>
<th>NNU</th>
<th>NNUF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rejected?</td>
<td>Y</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>p-value</td>
<td>0.047</td>
<td>0.060</td>
<td>0.027</td>
<td>0.005</td>
<td>0.017</td>
<td>0.002</td>
</tr>
</tbody>
</table>

Best-Versus-Rest Comparison

The BVR comparison, which summarized in Table 11, reveals some evidence supporting H2. Participants with more cognitively diverse crowd concepts sets were significantly more likely to generate top-rated concepts for the feasible \((p = 0.030)\), NNU \((p = 0.053)\), and NNUF \((p = 0.009)\) criteria. Similar trends were observed for other rating criteria, but the results are not statistically significant.

Statistically significant differences between the three groups were also found for novelty; however the findings indicate that that both the CD_{low} and CD_{high} groups produced a greater number of highly novel concepts than expected. The fraction of top-rated novel concepts was 7% higher than expected for the low diversity group and 25% higher than expected for the high diversity group. Since these results were different than expected, potential bias terms were investigated. Closer analysis revealed that participants in the low and high cognitive diversity groups rated the importance of creative thinking more highly than participants in the mid diversity group. This might suggest that participants in the low and high diversity groups had more experience with creative thinking, and therefore might be expected to generate more novel concepts. However, no correlation was found between novelty and the importance of creative thinking.

Since CD_{low} only generated one more top-rated novel concept than expected, it is also possible that the difference between expected and observed outcomes is a statistical artifact due to an outlier. To see whether differences between CD_{low} concepts and other C1 concepts were systematic, a KW test was performed. The statistical significance of this test was low \((p = 0.50)\), which indicates that CD_{low} concepts were no more likely to be top-rated concepts than other C1 concepts. A similar test was performed for CD_{high}, which indicated that CD_{high} participants were more likely to generate top-rated novel concepts than other C1...
participants ($p < 0.01$). These two tests suggest that the more than expected number of top-rated novel concepts for $\text{CD}_{\text{low}}$ was likely a statistical artifact, while the increase for $\text{CD}_{\text{high}}$ was systematic.

**TABLE 11: BVR COMPARISONS FOR COGNITIVE DIVERISTY GROUPS**

<table>
<thead>
<tr>
<th>Cognitive Diversity</th>
<th>Need</th>
<th>Novelty</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\text{CD}_{\text{low}}$</td>
<td>$\text{CD}_{\text{mid}}$</td>
</tr>
<tr>
<td>Top Concepts</td>
<td>4</td>
<td>7</td>
</tr>
<tr>
<td>Obs (Exp)</td>
<td>(4.0)</td>
<td>(9.5)</td>
</tr>
<tr>
<td>Other Concepts</td>
<td>12</td>
<td>31</td>
</tr>
<tr>
<td>Obs (Exp)</td>
<td>(12.0)</td>
<td>(28.5)</td>
</tr>
<tr>
<td>p-value</td>
<td>0.260</td>
<td>0.009</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Usefulness</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
<tr>
<td>Top Concepts</td>
</tr>
<tr>
<td>Obs (Exp)</td>
</tr>
<tr>
<td>Other Concepts</td>
</tr>
<tr>
<td>Obs (Exp)</td>
</tr>
<tr>
<td>p-value</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Feasibility</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
<tr>
<td>Top Concepts</td>
</tr>
<tr>
<td>Obs (Exp)</td>
</tr>
<tr>
<td>Other Concepts</td>
</tr>
<tr>
<td>Obs (Exp)</td>
</tr>
<tr>
<td>p-value</td>
</tr>
</tbody>
</table>
Bootstrap Comparisons

The bootstrap comparisons, summarized in Table 12, are generally consistent with the BVR findings and show strong support for H2. Because there were 3 cognitive diversity groups, three bootstrap comparisons were needed to make pairwise comparisons between the groups. The findings indicate that CD\textsubscript{high} participants generated higher rated concepts than CD\textsubscript{low} participants for all but the novel and NNU criteria. CD\textsubscript{high} participants also outperform CD\textsubscript{mid} participants for most of these criteria, but the confidence in these findings is lower.

**TABLE 12: BOOTSTRAP COMPARISONS OF COGNITIVE DIVERSITY GROUPS**

<table>
<thead>
<tr>
<th>Difference</th>
<th>Need</th>
<th>Novelty</th>
<th>Useful</th>
<th>Feasible</th>
<th>NNU</th>
<th>NNUF</th>
</tr>
</thead>
<tbody>
<tr>
<td>CD\textsubscript{mid} - CD\textsubscript{low}</td>
<td>% 1.3%</td>
<td>- 4.1%</td>
<td>0.4%</td>
<td>11.0%</td>
<td>-3.7%</td>
<td>8.2%</td>
</tr>
<tr>
<td></td>
<td>(p)</td>
<td>(0.261)</td>
<td>(0.105)</td>
<td>(0.472)</td>
<td>(0.148)</td>
<td>(0.313)</td>
</tr>
<tr>
<td>CD\textsubscript{high} - CD\textsubscript{mid}</td>
<td>% 3.4%</td>
<td>3.6%</td>
<td>2.6%</td>
<td>16.1%</td>
<td>13.8%</td>
<td>27.5%</td>
</tr>
<tr>
<td></td>
<td>(p)</td>
<td>(0.082)</td>
<td>(0.132)</td>
<td>(0.298)</td>
<td>(0.011)</td>
<td>(0.058)</td>
</tr>
<tr>
<td>CD\textsubscript{high} - CD\textsubscript{low}</td>
<td>% 4.8%</td>
<td>-0.5%</td>
<td>3.3%</td>
<td>28.7%</td>
<td>9.6%</td>
<td>37.5%</td>
</tr>
<tr>
<td></td>
<td>(p)</td>
<td>(0.013)</td>
<td>(0.519)</td>
<td>(0.037)</td>
<td>(0.006)</td>
<td>(0.128)</td>
</tr>
</tbody>
</table>

While cognitive diversity was found to affect nearly all of the rating criteria, the bootstrap comparisons did not reveal any correlations between novelty and cognitive diversity. This is surprising since the BVR analysis did reveal a statistically significant difference in the number of top concepts generated by the three groups. The bootstrap comparison of novelty score for CD\textsubscript{high} and CD\textsubscript{low} are shown in the left graph of Figure 27. The comparison indicates that statistically significant differences were observed in the lower percentile ranges, but the groups performed similarly above the 50\textsuperscript{th} percentile. This finding contradicts the conclusions of the BVR analysis and suggests that it may be difficult to make conclusions about the relationship between cognitive diversity and concept novelty.
Discussion

Both the bootstrap and BVR analyses show some support for H2. The conclusions from the BVR and bootstrap comparisons are summarized in Table 13. Overall, the results indicate that participants with more cognitively diverse crowd concepts outperformed participants with less cognitively diverse crowd concept sets. These participants also generated concepts with higher ratings in the need, usefulness, feasibility, and NNUF criteria. No conclusive differences were observed between the three cognitive diversity groups for the novelty or NNUF criteria.

TABLE 13: WERE DIFFERENCES OBSERVED BETWEEN CDLOW, CD MID, AND CDHIGH?

<table>
<thead>
<tr>
<th>Comparison Method</th>
<th>Need</th>
<th>Novelty</th>
<th>Useful</th>
<th>Feasible</th>
<th>NNU</th>
<th>NNUF</th>
</tr>
</thead>
<tbody>
<tr>
<td>BVR</td>
<td>N</td>
<td>Y</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>Bootstrap</td>
<td>Y</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td>Y</td>
</tr>
<tr>
<td>Overall</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>Y</td>
<td>N</td>
<td>Y</td>
</tr>
</tbody>
</table>

findings show support when $p < 0.10$
Contradictory Novelty Findings

It is unclear why the BVR and bootstrap analysis arrived at different conclusions regarding the relationship between cognitive diversity and novelty. One possible explanation is that the novelty scores were tightly clustered compared to other quality criteria. Therefore, the effective difference between a 25th percentile concept and a 75th percentile concept was small. Since the BVR analysis only analyzed differences in rank and not degree, the test may have indicated that more diverse crowds produce more novel concepts when the magnitude differences were small. In contrast, the statistical noise introduced by the bootstrap comparison method helps ensure that small rating differences are not treated with statistical significance.

It is also possible that reviewers had difficulty assessing the novelty of concepts.

Feasibility Findings

It is unclear why participants with more cognitively diverse crowd concept produced concepts with higher feasibility ratings than other participants. Given the strong degree of statistical significance ($p < 0.01$), it is unlikely that this was a chance finding, but it is also unclear how cognitive diversity affects feasibility. Differences in the composition of the groups were studied to see if there was any sample bias, but no significant differences were found. For example, further analysis showed that a greater percentage of $CD_{\text{high}}$ participants had technical backgrounds compared to the participant population overall. While one might expect participants with more technical backgrounds to generate more feasible concepts, no correlations were found to support this conclusion.

Participants in a cognitively diverse crowd might be more likely to consider different dimensions of feasibility when generating concepts. For example while a participant with a business background may tend to focus on the financial viability of a concept, showing that participant concepts generated by engineers, which discuss the technical feasibility of their concepts, may encourage the business participant to think about the technical feasibility of their concept. When the participant caters their concept to the many dimensions that affect feasibility, it is more likely to receive high feasibility ratings. Further contextual analysis of the concepts was beyond the scope of this study, but future work could see if participants in more cognitively diverse groups were more likely to discuss multiple dimensions of feasibility in the section of their submission where they were asked to assess implementation risks.
Isolating the Effects of Social and Cognitive Diversity

While the preceding analysis looked at the effects of cognitive diversity and social diversity independent, each C1 participant’s crowd concepts only controlled for one type of diversity, so there was no guarantee that social diversity and cognitive diversity were independently distributed. The combined distribution of cognitive and social diversity is shown in Table 14 and reveals that social and cognitive diversity were not independent factors in the study. This likely occurred because many non-turk participants were students from technical universities, so they were expected to have very similar age and occupational profiles.

<table>
<thead>
<tr>
<th>TABLE 14: COMBINED DISTRIBUTION OF SOCIAL DIVERSITY AND COGNITIVE DIVERSITY</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
<tr>
<td>Social Diversity</td>
</tr>
<tr>
<td>Cognitive Diversity</td>
</tr>
<tr>
<td>Low</td>
</tr>
<tr>
<td>Mid</td>
</tr>
<tr>
<td>High</td>
</tr>
</tbody>
</table>

Because social and cognitive diversity were not independently distributed in the study, one must be careful when drawing conclusions about the effects of a single diversity factor. For example, since all concepts in CD_{high} are also in SD_{high} and all concepts in CD_{low} are also in SD_{low}, one cannot draw conclusions about differences in the high and low cognitive diversity groups without demonstrating that social diversity does not affect the quality metric under consideration. Therefore, it is important to attempt to isolate the effects of cognitive diversity and social diversity to ensure that the findings discussed above are reliable.

The combined distribution reveals three potential group comparisons that isolate the effects of social diversity or cognitive diversity:

- Comparing concepts in SD_{low} and SD_{high} for the CD_{mid} group
- Comparing concepts in CD_{low} and CD_{mid} for the SD_{low} group
- Comparing concepts in CD_{mid} and CD_{high} for the SD_{high} group

BVR and bootstrap comparisons were performed for these three groups to see whether the findings substantially deviated from the uncontrolled comparisons above. The isolated comparisons revealed no major differences in the major findings, but some differences were observed for less significant findings.
Isolated BVR Comparisons

BVR analyses were performed on the three subsets where it was possible to isolate the effects of social diversity or cognitive diversity. The p-values from these isolated BVR comparisons are summarized in Table 15. The p-values from the earlier BVR comparisons are also included for comparison.

<table>
<thead>
<tr>
<th>Hypothesis</th>
<th>Need</th>
<th>Novelty</th>
<th>Useful</th>
<th>Feasible</th>
<th>NNU</th>
<th>NNUF</th>
</tr>
</thead>
<tbody>
<tr>
<td>(SD_{high} &gt; SD_{low}) \cap CD_{mid}</td>
<td>0.136</td>
<td>0.741</td>
<td>0.003</td>
<td>0.499</td>
<td>0.020</td>
<td>0.346</td>
</tr>
<tr>
<td>Social Diversity BVR</td>
<td>0.133</td>
<td>0.450</td>
<td>0.003</td>
<td>0.197</td>
<td>0.036</td>
<td>0.065</td>
</tr>
<tr>
<td>(CD_{mid} &gt; CD_{low}) \cap SD_{low}</td>
<td>0.172*</td>
<td>0.131*</td>
<td>0.083*</td>
<td>0.129</td>
<td>0.020*</td>
<td>0.324*</td>
</tr>
<tr>
<td>(CD_{high} &gt; CD_{mid}) \cap SD_{high}</td>
<td>0.388</td>
<td>0.012</td>
<td>0.770*</td>
<td>0.062</td>
<td>0.208</td>
<td>0.028</td>
</tr>
<tr>
<td>Cognitive Diversity BVR</td>
<td>0.260</td>
<td>0.009</td>
<td>0.412</td>
<td>0.030</td>
<td>0.053</td>
<td>0.009</td>
</tr>
</tbody>
</table>

* opposite hypothesis was found to be true at specified p-level

In general, the isolated BVR comparisons are consistent with the earlier BVR analysis, which suggests that findings discussed earlier are reliable.

Social Diversity Effects

The findings from the isolated BVR indicate that statistically significant effects observed in the earlier social diversity BVR are generally reliable. Group differences for the useful and NNU rating criteria increased in significance, which suggests that social diversity has a strong positive impact on usefulness and overall quality, as measured by the NNU criterion.

Differences between the high and low social diversity groups are less statistically significant for the feasibility and NNUF criteria. This is not surprising. Cognitive diversity was found to have a strong effect on feasibility. Since SD_{high} and CD_{high} are coupled, one would expect the earlier analysis to suggest that social diversity has some effect on feasibility and NNUF.

Cognitive Diversity Effects

The isolated comparisons reveal some interesting dynamics for the cognitive diversity groups. Comparisons between CD_{high} and CD_{mid} reproduce the earlier findings that more cognitively diverse crowds are more likely to generate novel and useful concepts. The
results also indicate that more cognitively diverse crowds produce a greater number of concepts that score highly in the NNUF criteria.

While the results comparing $\text{CD}_{\text{high}}$ and $\text{CD}_{\text{mid}}$ are consistent with the earlier BVR analysis, the isolated BVR indicates that $\text{CD}_{\text{low}}$ outperforms $\text{CD}_{\text{mid}}$ in many of the criteria. These findings are statistically significant for the useful ($p = 0.083$) and NNU ($p = 0.020$) criteria. This may suggest that the relationship between cognitive diversity and concept quality is curvilinear. It is not clear, however, why crowds that are moderately diverse in cognitive attributes would be expected to generate lower quality concepts than crowds that are more cognitively homogenous. Further research is needed to see whether the nonlinear relationship between cognitive diversity and concept quality was systematic or simply a statistical artifact.

The isolated BVR comparisons are generally consistent with, and in some cases strengthen, the earlier BVR findings. Social diversity was found to have a significantly positive effect on the usefulness of concepts, and cognitive diversity was found to have a significantly positive effect on concept novelty and feasibility. The statistical significance of complementary findings (i.e. the effect social diversity on novelty and feasibility) decreased for the isolated BVR comparisons, which suggests that some the dependence of social diversity and cognitive diversity was present in the earlier BVR analyses. However, this coupling did not affect any of the significant findings discussed earlier.

**Isolated Bootstrap Comparisons**

Bootstrap comparisons were also performed for the three isolated groups. The results are generally consistent with the isolated BVR analysis and earlier bootstrap comparisons. The results from the isolated bootstrap comparisons are summarized in Table 16.

**Summary of Isolated Effects**

The isolated effects analysis shows that the earlier conclusions regarding the effects of social diversity and cognitive diversity on ideation performance were not significantly affected by the codependence of social diversity and cognitive diversity. A summary of the isolated comparisons is shown in Table 17. The summary indicates that social diversity has a positive effect on the useful and NNU rating criteria, while cognitive diversity has positive effects on the feasibility and NNUF criteria.
TABLE 16: SUMMARY OF ISOLATED BOOTSTRAP COMPARISONS

<table>
<thead>
<tr>
<th>Difference</th>
<th>Need</th>
<th>Novelty</th>
<th>Useful</th>
<th>Feasible</th>
<th>NNU</th>
<th>NNUF</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \frac{SD_{high} - SD_{low}}{SD_{low}} ) &lt;sub&gt;isolated&lt;/sub&gt;</td>
<td>% 4.0%</td>
<td>2.5%</td>
<td>13.5%</td>
<td>3.2%</td>
<td>25.5%</td>
<td>30.8%</td>
</tr>
<tr>
<td>(p)</td>
<td>(0.245)</td>
<td>(0.145)</td>
<td>(0.000)</td>
<td>(0.309)</td>
<td>(0.000)</td>
<td>(0.109)</td>
</tr>
<tr>
<td>Social Diversity Bootstrap</td>
<td>% 3.3%</td>
<td>-0.3%</td>
<td>4.7%</td>
<td>16.3%</td>
<td>11.0%</td>
<td>35.9%</td>
</tr>
<tr>
<td>(p)</td>
<td>(0.040)</td>
<td>(0.456)</td>
<td>(0.031)</td>
<td>(0.026)</td>
<td>(0.110)</td>
<td>(0.025)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Difference</th>
<th>Need</th>
<th>Novelty</th>
<th>Useful</th>
<th>Feasible</th>
<th>NNU</th>
<th>NNUF</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \frac{CD_{mid} - CD_{low}}{CD_{low}} ) &lt;sub&gt;isolated&lt;/sub&gt;</td>
<td>% -1.9%</td>
<td>-5.7%</td>
<td>-10.2%</td>
<td>8.6%</td>
<td>-20.8%</td>
<td>-12.0%</td>
</tr>
<tr>
<td>(p)</td>
<td>(0.517)</td>
<td>(0.011)</td>
<td>(0.002)</td>
<td>(0.177)</td>
<td>(0.001)</td>
<td>(0.125)</td>
</tr>
<tr>
<td>Cognitive Diversity Bootstrap</td>
<td>% 1.3%</td>
<td>-4.1%</td>
<td>0.4%</td>
<td>11.0%</td>
<td>-3.7%</td>
<td>8.2%</td>
</tr>
<tr>
<td>(p)</td>
<td>(0.261)</td>
<td>(0.105)</td>
<td>(0.472)</td>
<td>(0.148)</td>
<td>(0.313)</td>
<td>(0.463)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Difference</th>
<th>Need</th>
<th>Novelty</th>
<th>Useful</th>
<th>Feasible</th>
<th>NNU</th>
<th>NNUF</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \frac{CD_{high} - CD_{mid}}{CD_{mid}} ) &lt;sub&gt;isolated&lt;/sub&gt;</td>
<td>% 2.8%</td>
<td>3.0%</td>
<td>1.1%</td>
<td>15.7%</td>
<td>10.3%</td>
<td>17.5%</td>
</tr>
<tr>
<td>(p)</td>
<td>(0.105)</td>
<td>(0.188)</td>
<td>(0.456)</td>
<td>(0.014)</td>
<td>(0.101)</td>
<td>(0.084)</td>
</tr>
<tr>
<td>Cognitive Diversity Bootstrap</td>
<td>% 3.4%</td>
<td>3.6%</td>
<td>2.6%</td>
<td>16.1%</td>
<td>13.8%</td>
<td>27.5%</td>
</tr>
<tr>
<td>(p)</td>
<td>(0.082)</td>
<td>(0.132)</td>
<td>(0.298)</td>
<td>(0.011)</td>
<td>(0.058)</td>
<td>(0.048)</td>
</tr>
</tbody>
</table>

* opposite hypothesis was found to be true at specified p-level

TABLE 17: DO ISOLATED FACTORS AFFECT RATING CRITERIA?

<table>
<thead>
<tr>
<th>Social Diversity</th>
<th>Need</th>
<th>Novelty</th>
<th>Useful</th>
<th>Feasible</th>
<th>NNU</th>
<th>NNUF</th>
</tr>
</thead>
<tbody>
<tr>
<td>BVR</td>
<td>N</td>
<td>N</td>
<td>Y</td>
<td>N</td>
<td>Y</td>
<td>N</td>
</tr>
<tr>
<td>Bootstrap</td>
<td>N</td>
<td>N</td>
<td>Y</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>Overall</td>
<td>N</td>
<td>N</td>
<td>Y</td>
<td>N</td>
<td>Y</td>
<td>N</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Cognitive Diversity</th>
<th>Need</th>
<th>Novelty</th>
<th>Useful</th>
<th>Feasible</th>
<th>NNU</th>
<th>NNUF</th>
</tr>
</thead>
<tbody>
<tr>
<td>BVR</td>
<td>N</td>
<td>Y</td>
<td>N</td>
<td>Y</td>
<td>N</td>
<td>Y</td>
</tr>
<tr>
<td>Bootstrap</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>Overall</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>Y</td>
<td>N</td>
<td>Y</td>
</tr>
</tbody>
</table>

findings show support when p < 0.10
Effect of Communication Capability

Hypothesis 3  Communication capability will have a positive effect on the crowd’s ideation performance.

Crowds with greater communication capability were expected to outperform crowds with less communication capability. This logic was based on the assumptions that: (1) cognitive diversity has synergistic effects when concepts were freely revealed to the crowd; and (2) the negative effects of social diversity do not impact the crowds because the observability of social factors can be more readily managed on crowdsourcing platforms. This study found some support for both of these underlying assumptions. Participants with socially diverse crowd concept sets generated more useful and better overall concepts, measured by the NNU criteria. Furthermore, participants shown occupationally diverse crowd concept sets generated better overall concepts, as measured by the NNUF rating criterion. These best concepts generated by these participants also scored more highly in the need and NNU criteria. These results collectively show support for H2.

Both the BVR and bootstrap comparisons indicated that C₁ participants generated more novel concepts than C₀ participants. Beyond this finding, there were no conclusive differences found C₀ and C₁ participants. In general, this does not show strong support for H3. However, when the C₀ concepts were compared to concepts in the social diversity and cognitive diversity subgroups, significant differences between the groups were observed. For example, SD_{high} participants generated more top-rated useful concepts than C₀ participants ($p = 0.070$), and CD_{high} participants generated more top-rated concepts for the novelty ($p = 0.001$) and NNU ($p = 0.008$) criteria.

Null Hypothesis Tests

The null hypothesis could not be rejected for any of the concept quality criteria with a reasonable level of confidence. This indicates that there are no significant differences between the distributions for either communication capability group.

<table>
<thead>
<tr>
<th>TABLE 18: NULL HYPOTHESIS TESTS FOR COMMUNICATION CAPABILITY</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rejected?</td>
</tr>
<tr>
<td>------------</td>
</tr>
<tr>
<td>p-value</td>
</tr>
</tbody>
</table>
Best-Versus-Rest Comparison

Even though the null hypotheses could not be rejected, a BVR comparison was performed to see if there were any differences in the frequency of top concepts between the two groups. In general, the results, which are summarized in Table 19, are consistent with the null hypothesis tests and indicate that there are no statistically significant differences in the number of top concepts. There is one exception to this finding: the collaborative group produced a greater number of highly novel concepts than the isolated group ($p < 0.05$).

**TABLE 19: BVR COMPARISONS FOR COMMUNICATION CAPABILITY GROUPS**

<table>
<thead>
<tr>
<th></th>
<th>Need</th>
<th></th>
<th>Novelty</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$C_0$</td>
<td>$C_1$</td>
<td>$C_0$</td>
<td>$C_1$</td>
</tr>
<tr>
<td>Comm Capability</td>
<td>(n=29)</td>
<td>(n=72)</td>
<td>(n=29)</td>
<td>(n=72)</td>
</tr>
<tr>
<td>Top Concepts</td>
<td>7</td>
<td>18</td>
<td>3</td>
<td>22</td>
</tr>
<tr>
<td></td>
<td>(7.2)</td>
<td>(17.8)</td>
<td>(7.2)</td>
<td>(17.8)</td>
</tr>
<tr>
<td>Other Concepts</td>
<td>22</td>
<td>54</td>
<td>26</td>
<td>50</td>
</tr>
<tr>
<td></td>
<td>(21.8)</td>
<td>(54.2)</td>
<td>(21.8)</td>
<td>(54.2)</td>
</tr>
<tr>
<td>$p$-value</td>
<td>0.928</td>
<td></td>
<td>0.034</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Usefulness</th>
<th>Feasibility</th>
</tr>
</thead>
<tbody>
<tr>
<td>Comm Capability</td>
<td>$C_0$</td>
<td>$C_1$</td>
</tr>
<tr>
<td>(n=29)</td>
<td>(n=72)</td>
<td>(n=29)</td>
</tr>
<tr>
<td>Top Concepts</td>
<td>6</td>
<td>20</td>
</tr>
<tr>
<td></td>
<td>(7.5)</td>
<td>(18.5)</td>
</tr>
<tr>
<td>Other Concepts</td>
<td>23</td>
<td>52</td>
</tr>
<tr>
<td></td>
<td>(21.5)</td>
<td>(53.5)</td>
</tr>
<tr>
<td>$p$-value</td>
<td>0.463</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>NNU</th>
<th></th>
<th>NNUF</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Comm Capability</td>
<td>$C_0$</td>
<td>$C_1$</td>
<td>$C_0$</td>
<td>$C_1$</td>
</tr>
<tr>
<td>(n=29)</td>
<td>(n=72)</td>
<td>(n=29)</td>
<td>(n=72)</td>
<td></td>
</tr>
<tr>
<td>Top Concepts</td>
<td>4</td>
<td>21</td>
<td>8</td>
<td>17</td>
</tr>
<tr>
<td></td>
<td>(7.2)</td>
<td>(17.8)</td>
<td>(7.2)</td>
<td>(17.8)</td>
</tr>
<tr>
<td>Other Concepts</td>
<td>25</td>
<td>51</td>
<td>21</td>
<td>55</td>
</tr>
<tr>
<td></td>
<td>(21.8)</td>
<td>(54.2)</td>
<td>(21.8)</td>
<td>(54.2)</td>
</tr>
<tr>
<td>$p$-value</td>
<td>0.107</td>
<td></td>
<td>0.677</td>
<td></td>
</tr>
</tbody>
</table>
The Co concepts were also compared to the two social diversity subgroups and three cognitive diversity subgroups to see if specific crowd configurations led to differences between the Co and C1 concepts. The p-values from these comparisons are summarized in Table 20. The subgroup comparisons reveal several additional effects, which are consistent with the earlier findings regarding social and cognitive diversity.

### TABLE 20: SUMMARY OF P-VALUES FROM BVR COMPARISON

<table>
<thead>
<tr>
<th>Hypothesis</th>
<th>Need</th>
<th>Novelty</th>
<th>Useful</th>
<th>Feasible</th>
<th>NNU</th>
<th>NNUF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Co &lt; C1</td>
<td>0.928</td>
<td>0.034</td>
<td>0.463</td>
<td>0.677</td>
<td>0.107</td>
<td>0.677</td>
</tr>
<tr>
<td>Co &lt; CDlow</td>
<td>0.949</td>
<td>0.031</td>
<td>0.878</td>
<td>0.090*</td>
<td>0.166</td>
<td>0.514</td>
</tr>
<tr>
<td>Co &lt; CDmid</td>
<td>0.571</td>
<td>0.520</td>
<td>0.595</td>
<td>0.537</td>
<td>0.615</td>
<td>0.142</td>
</tr>
<tr>
<td>Co &lt; CDhigh</td>
<td>0.288</td>
<td>0.002</td>
<td>0.180</td>
<td>0.241</td>
<td>0.008</td>
<td>0.124</td>
</tr>
<tr>
<td>Co &lt; SDlow</td>
<td>0.442</td>
<td>0.125</td>
<td>0.236</td>
<td>0.286</td>
<td>0.802</td>
<td>0.159</td>
</tr>
<tr>
<td>Co &lt; SDhigh</td>
<td>0.493</td>
<td>0.023</td>
<td>0.070</td>
<td>0.879</td>
<td>0.022</td>
<td>0.713</td>
</tr>
</tbody>
</table>

Participants in CDhigh generated a greater number of concepts that scored highly in the novel ($p = 0.002$) and NNU ($p = 0.008$) criteria. No statistically significant differences were found regarding the feasibility of concepts generated by CDhigh and Co participants. Interestingly, CDlow participants were less slightly likely to generate highly feasible concepts than Co participants ($p = 0.090$). This may indicate that CDlow participants fixated on one dimension of feasibility when they generated their concepts. For example, a mechanical engineer shown crowd concepts generated by other mechanical engineers may forget to consider the financial and business viability of his or her concepts. This may lead to low feasibility scores.

Participants in SDhigh generated a greater number of concepts that scored highly in the useful ($p = 0.070$) and NNU ($p = 0.022$) criteria. These participants also generated more novel concepts ($p = 0.023$). This finding was discarded, however, because the SDhigh and CDhigh groups were coupled, and the findings regarding novelty were much more statistically significant for the CDhigh group ($p = 0.002$).

### Bootstrap Comparisons

The bootstrap comparisons corroborate some of the findings from the BVR analysis. The best C1 concepts on average score slightly higher in novelty ($p < 0.05$) than the best Co concepts. They also score significantly higher in the NNU and NNUF criteria, but confidence in these findings is low ($p = 0.185$, $p=0.236$).
The differences between \( C_0 \) concepts and concepts from the social and cognitive diversity subgroups are more significant. When \( SD_{\text{high}} \) and \( CD_{\text{high}} \) concepts are compared to \( C_0 \) concepts, both the magnitude and statistical significance of score differences increase. For example, while the best \( C_1 \) concept scored 15\% better than the best \( C_0 \) concept for the NNUF criterion \((p = 0.236)\), the best \( CD_{\text{high}} \) concept scores nearly 30\% higher than the best \( C_0 \) concept \((p = 0.011)\). Similarly, the best \( SD_{\text{high}} \) concept scores 11\% higher than the best \( C_0 \) concept for the NNU criterion \((p = 0.097)\). These findings indicate that while \( C_0 \) and \( C_1 \) crowds do not substantially differ in terms of ratings, the more collaborative crowds outperform the nominal crowd when members of the crowd are more cognitively and socially diverse.

**TABLE 21: BOOTSTRAP COMPARISON OF COMMUNICATION CAPABILITY AND SUBGROUPS**

<table>
<thead>
<tr>
<th>Difference</th>
<th>Need</th>
<th>Novelty</th>
<th>Useful</th>
<th>Feasible</th>
<th>NNU</th>
<th>NNUF</th>
</tr>
</thead>
<tbody>
<tr>
<td>( C_1 - C_0 )</td>
<td>%</td>
<td>0.7%</td>
<td>5.1%</td>
<td>-0.4%</td>
<td>-0.2%</td>
<td>8.4%</td>
</tr>
<tr>
<td>( C_0 )</td>
<td>(p)</td>
<td>(0.440)</td>
<td>(0.017)</td>
<td>(0.484)</td>
<td>(0.439)</td>
<td>(0.185)</td>
</tr>
<tr>
<td>( CD_{\text{low}} - C_0 )</td>
<td>%</td>
<td>-1.7%</td>
<td>6.7%</td>
<td>-1.9%</td>
<td>-15.5%</td>
<td>6.2%</td>
</tr>
<tr>
<td>( C_0 )</td>
<td>(p)</td>
<td>(0.246)</td>
<td>(0.032)</td>
<td>(0.176)</td>
<td>(0.051)</td>
<td>(0.282)</td>
</tr>
<tr>
<td>( CD_{\text{mid}} - C_0 )</td>
<td>%</td>
<td>-0.2%</td>
<td>2.4%</td>
<td>-1.2%</td>
<td>-6.6%</td>
<td>2.3%</td>
</tr>
<tr>
<td>( C_0 )</td>
<td>(p)</td>
<td>(0.348)</td>
<td>(0.118)</td>
<td>(0.455)</td>
<td>(0.133)</td>
<td>(0.441)</td>
</tr>
<tr>
<td>( CD_{\text{high}} - C_0 )</td>
<td>%</td>
<td>2.9%</td>
<td>6.4%</td>
<td>1.4%</td>
<td>9.3%</td>
<td>16.6%</td>
</tr>
<tr>
<td>( C_0 )</td>
<td>(p)</td>
<td>(0.128)</td>
<td>(0.009)</td>
<td>(0.277)</td>
<td>(0.099)</td>
<td>(0.033)</td>
</tr>
<tr>
<td>( SD_{\text{low}} - C_0 )</td>
<td>%</td>
<td>-1.6%</td>
<td>5.1%</td>
<td>-3.7%</td>
<td>-10.8%</td>
<td>0.8%</td>
</tr>
<tr>
<td>( C_0 )</td>
<td>(p)</td>
<td>(0.189)</td>
<td>(0.054)</td>
<td>(0.065)</td>
<td>(0.040)</td>
<td>(0.498)</td>
</tr>
<tr>
<td>( SD_{\text{high}} - C_0 )</td>
<td>%</td>
<td>1.6%</td>
<td>4.8%</td>
<td>0.9%</td>
<td>3.9%</td>
<td>11.4%</td>
</tr>
<tr>
<td>( C_0 )</td>
<td>(p)</td>
<td>(0.233)</td>
<td>(0.011)</td>
<td>(0.289)</td>
<td>(0.332)</td>
<td>(0.097)</td>
</tr>
</tbody>
</table>

The differences between \( C_0 \) concepts and concepts from the social and cognitive diversity subgroups are more significant. When \( SD_{\text{high}} \) and \( CD_{\text{high}} \) concepts are compared to \( C_0 \) concepts, both the magnitude and statistical significance of score differences increase. For example, while the best \( C_1 \) concept scored 15\% better than the best \( C_0 \) concept for the NNUF criterion \((p = 0.236)\), the best \( CD_{\text{high}} \) concept scores nearly 30\% higher than the best \( C_0 \) concept \((p = 0.011)\). Similarly, the best \( SD_{\text{high}} \) concept scores 11\% higher than the best \( C_0 \) concept for the NNU criterion \((p = 0.097)\). These findings indicate that while \( C_0 \) and \( C_1 \) crowds do not substantially differ in terms of ratings, the more collaborative crowds...
outperform the nominal crowd when members of the crowd are more cognitively and socially diverse.

The results also indicate that the C1 group may perform worse than the C0 group when diversity is low. For example, the best SD_low concept scores nearly 9% worse in the NNUF rating \((p = 0.043)\) and 11% worse in the feasibility rating \((p = 0.040)\) than the best C0 concept. Furthermore, while social diversity was found to have a positive effect on the usefulness of concepts, SD_low concepts scored nearly 4% worse in the useful rating than C0 concepts \((p = 0.065)\). This suggests that a lack of social diversity may have deleterious effects on collaborative crowds’ creative capacity. Similar effects were observed for cognitive diversity. While the best CD_high concept scores 9% better in the feasibility rating than the best C0 concept, the best CD_low concept scores 15% worse than the best C0 concept \((p = 0.051)\).

**Discussion**

On its face, the findings show little support for H3. The only statistically significant difference observed between the C0 and C1 concepts was a slight increase in the novelty ratings for the best C1 concepts. No significant differences were observed for either overall quality criteria. These findings show little evidence supporting the notion that communication capability improves the quality or quantity of the best concepts.

A closer analysis of the data, however, reveals that the best concepts generated by CD_high and SD_high participants outscore the best concepts generated by C0 participants by 10% or more for both the NNU and NNUF quality criteria. Furthermore, the best concepts generated by SD_low and CD_low participants scored 5 - 8% worse than the best C0 concepts in the NNUF criteria. These findings suggest that the social and cognitive diversity mediate the differences between nominal and collaborative crowd performance. When crowds are socially and cognitively homogenous, the nominal crowd may have a slight edge over the collaborative crowd. However, when the crowds are socially and cognitively diverse, the collaborative crowd generates higher quality concepts than the nominal crowd.

**Summary of Findings**

Concepts from different groups were compared using two methods. A best-versus-rest comparison was used to determine whether one group was more likely to generate concepts in top 75th percentile than other groups. Additionally, a bootstrap comparison was used to assess whether the top-rated concepts from one group outscored the top-rated
concepts from other groups. When both these tests reveal significant differences for a group, it indicates that the group strongly outperforms other groups, and when only one test is significant, the group is said to weakly outperform other groups.

The Effect of Social Diversity

Hypothesis 1, which asserted that social diversity would have no affect on ideation performance, was rejected. The study findings indicate that participants shown more socially diverse crowd concept sets generated more useful concepts than other participants. There was also some indication that greater social diversity led to concepts with greater overall quality, measured using the NNU and NNUF criteria.

These findings depart significantly from the mainstream creativity literature, which suggests that social diversity in a team setting degrades team performance. There are several reasons the effect of social diversity differs from the predictions in the creativity literature. First, observable social differences are predicted to degrade team performance because they lead to social categorization and a lack of cohesiveness (Hennessey and Amabile 2010; Dahlin 2005). In this study, participants could not directly interact with each other, so the notion of cohesiveness is not even relevant. While this should minimize any negative effects of social diversity on ideation performance, it does not explain why social diversity improves ideation performance.

Social diversity may improve ideation performance because it may help participants empathize with public restroom users that differ in age and gender. These differences are relatively acute in this study because the nature and frequency of public restroom usage varies across age and gender groups. For example, a young student has likely never changed a child on a baby-changing table, but a parent with a newborn is probably very familiar with the experience. Revealing concepts generated by other types of public restroom users may help the participant identify more acute needs and develop more useful solutions. The study findings that show a strong correlation between social diversity and usefulness provide some support for this notion.

The Effect of Cognitive Diversity

Cognitive diversity was predicted to have a positive effect on ideation performance. The results show strong support for this hypothesis. Both the BVR and bootstrap comparisons indicated that cognitively diversity increased the novelty and overall quality, measured using the NNUF metric, of participants’ concepts. These findings are consistent with the general position in the creativity literature, which asserts that cognitive diversity should improve ideation performance (Hennessey and Amabile 2010; Kavadias and Sommer 2009).
The Effect of Communication Capability

No significant differences were observed in the overall quality criteria for concepts from the \( C_0 \) and \( C_1 \) groups. \( C_1 \) concepts did perform slightly better, however, in the novelty ratings. When comparisons were broken down by diversity subgroups, some additional differences arose. Concepts generated by participants with more socially and/or cognitively diverse crowd concepts scored better than the \( C_0 \) concepts in both the NNU and NNUF criteria. These findings suggest that communication, and by analogy collaboration, are beneficial when the crowd is diverse. However, when the crowd was not diverse in social or cognitive characteristics, \( C_1 \) crowd performed worse than the \( C_0 \) crowd. While further research is needed to confirm these findings, they suggest that the composition of the crowd affects the relative ideation performance of nominal and collaborative crowds.
Chapter 6
Conclusions and Future Work

Crowdsourcing is increasingly being used to find creative solutions to complex social and technical problems. Examples of crowdsourcing abound in a number of design disciplines, including graphic design, product design, and engineering. Some preliminary studies indicate that, under the right conditions, the crowd may generate more creative concepts than a firm’s own employees. However, there has been little empirical research examining what factors affect the crowd’s creative capacity.

This thesis examines whether findings from the group creativity literature are applicable to the crowd. These findings suggest that collaborative teams are generally worse at generating creative concepts than nominal groups, where members of the team work independently to generate concepts. Social diversity and cognitive diversity are observed to have countervailing effects on team performance. While cognitive diversity is expected to improve the search for solutions and identify higher quality concepts (Dahlin 2005; Kavadias and Sommer 2009), social diversity can lead to social categorization which reduces team cohesiveness and degrades team performance (Hennessey and Amabile 2010). The findings also suggest that other factors, such as production blocking, reduce the performance of collaborative teams (Kavadias and Sommer 2009).

While the characteristics of crowds and teams are similar in some respects, there are also important differences, which indicate that findings from the creativity literature may not be relevant to the crowd. For example, crowds generally have more limited capability to communicate than team, so the crowd may be less susceptible to the negative effects of social diversity and production blocking.

To study the effects of social diversity, cognitive diversity, and communication capability on the crowd’s ideation performance, a design challenge to “reimagine the public restroom experience” was developed. Participants were placed into crowds with varied levels of communication capability and diversity. The findings from the study suggest that more socially and cognitively diverse crowds generate more creative concepts than less diverse crowds. Furthermore, while the creativity literature predicted that nominal crowds should outperform more collaborative crowds, the study’s findings indicate that the relative performance of nominal and collaborative crowds depends on crowd’s composition. When crowds are diverse, collaborative crowds outperform the nominal crowd. However, when crowds are highly homogenous, collaboration can reduce ideation performance. The
implications of these findings for future creativity research and the development of CCW platforms are discussed below.

Implications for Creativity Research

Studying Group Creativity with CCW Platforms

The group creativity literature limits its focus to differences between co-located ideation teams and isolated or nominal groups. When examined from a communication capability perspective, these two groups represent extreme cases on the communication capability continuum. Members of a team have full use of their communication capability while members of a nominal group have no communication capability. As teams increasingly become decentralized and distributed, it is valuable to understand how the creativity findings apply at less extreme points on the communication capability continuum.

CCW platforms represent a useful tool to study the effects of varying communication capability on group creativity because the communication features implemented on the platform control the frequency, depth, and mode of communication between individuals. For example, researchers could use CCW platforms to see whether limiting communication capability to parallel communication modes, such as text-based communication, reduces the incidence of production blocking in teams. This research could help identify levels of communication capability that more effectively manage the tradeoffs between the benefits of cognitive diversity and the deleterious effects of social diversity.

Total Diversity: A More Intuitive Measure of Diversity

In Chapter 4, the notion of total diversity – a quantitative measure of diversity that captures differences within and across attributes – was introduced and a method to measure total diversity was developed. Through several examples, it was shown that total diversity measures more closely align with the common notions of diversity than other diversity measures, such as Shannon entropy. When unobservable or composite types of diversity need to be measures, total diversity may be a more useful measure than observable diversity measures. Further research is needed to compare the utility of total diversity to the simpler observable diversity measures.
Recommendations for the Design of CCW Models

The findings from this study also have implications for the design of CCW models. While further work is needed to better understand the causal links through which social and cognitive diversity affect the crowd’s creative processes, the study shows that social and cognitive diversity increase the crowd’s creative capacity. This suggests that the designers and administrators of CCW platforms need to consider how the composition of the crowd affects ideation performance.

Identify the Types of Diversity that Create Value

In the “reimagining the public restroom” design challenge, the crowd benefitted from diversity in age, gender, and occupation. Age and gender diversity helped participants identify different users and use cases in public restrooms, while occupational diversity expanded the solution space to consider different solutions types, such as: products, services, mobile apps, and environment design. While age, gender, and occupational diversity were valuable for this design challenge, they may not be generally valuable for other types of problems.

*Designers should identify the types of diversity that are expected to benefit the crowd’s solution search and develop recruitment tactics to attract the appropriate participants.*

Manage Collaboration Failures by Reducing Participants’ Visibility

In some cases, the types of diversity that are identified as beneficial may also be expected to degrade team performance. For example, diversity in nationality could be beneficial for some types of problems, but it has also been found to degrade team performance (Dahlin 2005).

*The designers of crowdsourcing platforms can manage the negative effects of diversity by limiting the observability of participants.*

Since the benefits of diversity are closely tied to the problem framings and strategies embedded in participants’ concepts, but the negative effects of diversity are more closely linked to the participants themselves, limiting the observability of participants should not significantly degrade the benefits of diversity.
Maximize Diversity Returns by Increasing Communication Capability

The findings from the study also reveal that social diversity and cognitive diversity affect the ideation performance tradeoffs between nominal crowds and more collaborative crowds. When crowds are more diverse, crowds with small amounts of communication capability generate more high quality concepts than nominal crowds with no communication capability. These findings suggest that communication capability is a key factor that controls the ability for crowd participants to reflect upon others’ contributions and incorporate insights from those contributions into their own problem framings.

*The designers of CCW platforms should consider how varying degrees and types of communication capability affect crowd participants’ ability to utilize the benefits of diversity.*

Further research is needed to understand what level of communication capability strikes a balance between the desire to maximize diversity returns and the interest in mitigating the incidence of crowd failures.

Limitations & Future Work

As with any empirical study, this study’s findings can only be generalized to a point. In this section, the study’s limitations and opportunities for future work are explored.

Using AMT Workers as Reviewers

Amazon Mechanical Turk workers were used in this study to review concepts. While this made it possible to evaluate concepts quickly and cheaply, an analysis of the reviews revealed that inter-rater agreement was low among workers. Future work should seek to understand why there was so little agreement and whether AMT workers are fit to evaluate creative concepts. Furthermore, an expert review panel could be used to validate the research findings presented here.

Problem Type

This study examined how various factors affect the creative performance of the crowd for a single design problem. However, the examples of CCW models discussed earlier reveal that the type of design problems addressed on CCW platforms are unique and varied. Future work is needed to understand how the findings from this study apply to other types
of problems. This is difficult task because it is not obvious how problems should be characterized. Models of the solution search process by Kavadias and Sommer (2009) indicate that the complexity and multidisciplinary nature of problems are important factors that mediate the tradeoff between nominal and collaborative crowds. Other factors that may be important to consider include the crowd's familiarity with the problem and the amount of specialized knowledge required to understand the problem. Since there are many factors that would need to be accounted for, it may be impractical to conduct additional design challenges. Case studies may be a more suitable approach.


Ross, Joel, L Irani, and M Silberman. 2010. "Who Are the Crowdworkers?: Shifting 
http://dl.acm.org/citation.cfm?id=1753873.


doi:http://dx.doi.org/10.1016/S0142-694X(02)00034-0. 


e&db=cat00916a&AN=mit.001541803&site=eds-live.

Threadless. 2013. "Threadless." *SkinnyCorp, LLC.* 


Classification (SOC) System." http://www.bls.gov/soc/.


A print version of the design challenge brief is included on the following pages.
How might we improve the public restroom experience?

Public restrooms provide a basic public service, keep the streets clean, and help prevent the spread of disease. This public service has been around for millennia: there were 144 public latrines in Rome during the latter part of the Roman Empire. These Roman latrines provided both relief and economic opportunity. Urine was collected in buckets, dolia curta and sold to the fullers (cloth launderers) to clean wool and other garments.

Today, public restrooms can be found everywhere: at coffee shops, beaches and parks, train stations, and even on airplanes. While restrooms are more advanced and numerous than they were in Roman times, using a public restroom is still seldom an enjoyable experience. Public restrooms can be difficult to find, too small to accommodate a large flux of people, and are often well below our cleanliness and hygiene standards.

The Many Varieties of Public Restrooms
Meet Your Clients

Public restrooms are used by people of all ages and walks of life. Take a moment to consider who "uses" a public restroom (or perhaps who doesn't use a public restroom). Put yourself in their shoes and understand their state of mind and needs. Does the public restroom meet their needs? What opportunities might exist to make their public restroom experience better for this person?

Your Challenge: re-imagining the public restroom experience

You don't need to solve all of the problems with public restrooms. Select a problem (or two) that resonates with you and submit one to three concepts that you think will improve the public restroom experience.
A few tips to get you started:

1. Brainstorm a list of problems that specific users have in a public restroom
2. Select one or two problems that you think are the most acute or ripe for innovation
3. Research what others have done to solve these problems
4. Generate some "napkin sketch" concepts of your own that you think will address these problems

What's in it for You

Submitters of five top-rated concepts will receive a $50 Amazon gift card. After submitting your concepts, you'll be able to check back and see all of the concepts that have been submitted by others.

Judging Criteria

Concepts will be judged on three criteria:

Desirability does the concept address a real user need? If so, would the target users want to use the concept
Novelty is the concept truly unique or do similar concepts already exist?
Feasibility is the concept feasible from a business, technical, and logistic perspective?

Some Resources

Wicked Good Guide to Boston Restrooms a user-curated website documenting the location and quality of public restrooms all across Boston
If It Works, What a Relief Boston Globe article discussing "city toilets" installed across Boston
World's Greatest Public Bathrooms Travel+Leisure article detailing the world's nicest public restrooms
Appendix B
CrowdConcept Website Implementation

The following pages are screenshots of CrowdConcept, the website designed for participants to submit concepts. When an individual signed up for the study, he or she was guided through study in the manner described in Figure B-1.

FIGURE B-1: PARTICIPANT WORKFLOW
How might we improve the public restroom experience?

Think you can come up with creative ways to make the restroom experience better? Want to win a $50 Amazon Gift Card?

Take part in our Re-Imagining the Public Restroom challenge. This design challenge is part of a research project at the MIT CadLAB studying the creativity of the crowds and the applicability of crowdsourcing to early-stage design processes.

The challenge typically takes 30 minutes and can be completed online. If you choose to participate, we would ask you to:

1. Complete a short survey about your demographic background, education, and occupation.
2. Generate one or more creative concepts to improve the public restroom experience.

Individuals who submit one of the top 5 concepts will receive a $50 Amazon Gift Card.

Interested? Submissions close Wednesday, August 7th at 11:59pm.

Enter your email address to get started now

email address

Problems? Send us an email at crowdconcept@mit.edu

FIGURE B-2: SIGN UP
Provides an overview of the design challenge and a link to sign up to participate.
Consent to Participate in Non-Biomedical Research
Crowdsourcing Creative Design Concepts

You are asked to participate in a research study conducted by Jeff Mekler and David Wallace from the Mechanical Engineering Department at the Massachusetts Institute of Technology (M.I.T.). Results from this study will be used in journal publications and for the completion of a master’s thesis. You were selected as a possible participant in this study because of your educational background and professional occupation. You should read the information below, and ask questions about anything you do not understand, before deciding whether or not to participate.

PARTICIPATION AND WITHDRAWAL

Your participation in this study is completely voluntary and you are free to choose whether to be in it or not. If you choose to be in this study, you may subsequently withdraw from it at any time without penalty or consequences of any kind. The investigator may withdraw you from this research if circumstances arise which warrant doing so.

PURPOSE OF THE STUDY

Crowdsourcing is a method where a task is outsourced to a large number of unaffiliated individuals (the crowd). This study seeks to understand whether online crowdsourcing is an effective method to generate novel concepts for new

By clicking yes below, I acknowledge that I have read and understand the information contained in the Consent to Participate form (above). My questions have been answered to my satisfaction, and I agree to participate in this study. I am aware that I can download a copy of the consent form here.

Do you agree to the Consent to Participate terms?

No Yes

Problems? Send us an email at crowdconcept@mit.edu

FIGURE B-3: CONSENT FORM
Standard COUHES research consent form.
FIGURE B-4: BACKGROUND SURVEY
Responses from the survey were used to generate crowd concepts for C₁ participants.
You have submitted 1 concept. You may submit 2 more concepts.

Sample Concept
Submit a Concept

Here are several ideas that have already been submitted

Express Lanes in... TP Dispenser Straight Into the U... Sound Isolated stall

NOW SERVING
Wait Time Estimator 10 to 1 ratio of wo... The Private Stall uv to sanitize the air

Charcoal air fresh... Video Game toilet Don't re-dirty my h... A quick steam, a g...

Problems? Send us an email at crowdconcept@mit.edu

FIGURE B-5: DESIGN TASK – HOME PAGE
Page where participants can read the design brief, view concepts they have already submitted, and if they are a C1 participant, view the concepts selected for them. Participants can view the design brief at any time by clicking the "Read the Brief" button in the top-right corner of the page.
Your challenge: re-imagining the public restroom experience

If you haven't done so already, Read the Brief.

You don't need to solve all of the problems with public restrooms. Select a problem (or two) that resonates with you and submit one to three concepts that you think will improve the public restroom experience.

A few tips to get you started:
1. Brainstorm a list of problems that specific users have in a public restroom
2. Select one or two problems that you think are the most acute or ripe for innovation
3. Research what others have done to solve these problems
4. Generate some "napkin sketch" concepts of your own that you think will address these ideas

Concept Title

upload images and/or sketches to illustrate your concept and the problem or need it addresses

what problem or need does your concept address?

Who is your target user? What are their needs? Why are they not currently being met?

what is your concept and how does it work?

Tell us more about your concept so that the judges can understand what it is, how it works, and why you think it will help.

what risks or barriers might prevent your concept from being realized?

Every concept has some implementation risks. What risks might prevent this concept from being implemented? Are there

Submit Concept

Problems? Send us an email at crowdconcept@mit.edu

FIGURE B-6: DESIGN TASK – CONCEPT SUBMISSION PAGE
Web form for participants to submit concepts. Participants can upload images of their concept by clicking on the “Add Image” button.
You're almost there!
Do you have any feedback about this study?
For example, was there something in particular that was confusing? Would you have preferred to submit your concepts in a different format (e.g., video)?

Enter your feedback here.

Finish

Problems? Send us an email at crowdconcept@mit.edu

FIGURE B-7: FEEDBACK FORM
Allows participants to provide the researchers with feedback about the design study.
After participants complete the study, they can browse all concepts submitted by other participants.

FIGURE B-8: BROWSE CONCEPTS
FIGURE B-9: REVIEW FORM
Form used by Amazon Mechanical Turk workers to review concepts.
Appendix C
Sample Concepts

The following pages contain a sample of the 101 design concepts submitted by study participants. The concepts presented here were selected to emphasize the diversity of solutions generated by participants and have no bearing on the ratings concepts received. All concepts can be found online at http://crowdconcept.herokuapp.com/public/ideas.
Family Amusement/Waterpark Restroom

If the kids haven’t already peed in the pool, then mom (or dad) and little Joey have to make a mad dash to the waterpark bathroom, where they are both dripping wet and possibly barefoot – a scenario ripe for disgust in a public restroom.

what problem or need does this concept address?
A family scenario, where there are small children involved in wet clothing - possibly or likely at an age where they need assistance from mom or dad. The problems here include: soggy/wet clothing (resulting in just plain wet toilet seats for the next patron - or a dry-ish toilet seat with lots of disintegrating toilet paper), wet & dirty floors that many people will be walking on barefoot, and relatively minimal existing ventilation or means of drying and keeping these areas clean.

how does this concept address this need?
I think providing a large supply of readily biodegradable, absorbent towels for parents to clean off their kids to an extent prior to actually using the toilet will be helpful in the wetness-on-toilet-seat arena. With a significant portion of stalls that are large enough to contain both mama (or dada), child, toilet, and changing station (so not necessarily as large as a typical handicapped stall, but larger than standard tiny ones), the problem of wet seats or wasted toilet paper can be diverted to wet, biodegradable towels. As for a dirty floor, I wonder if there is space for turf-like material with significant drainage that can be cleaned with mild detergents that will not harm small, bare feet.

what risks or barriers might prevent this concept from being realized?
High upfront costs that might be strange enough to be little-accepted. With the correct staging and theme-related decoration, the floor could be well-received and used. If the environment is set up in such a way that the clientele are mostly families with very young children, then this restroom will make sense for the parent/child scenario, and users will likely acclimate to the provisions that are designed here.

Problems? Send us an email at crowdconcept@mit.edu

FIGURE C-1: SAMPLE CONCEPT 1
CRS - "Complete Restroom Sanitizer"
A complete, continuous motion, restroom cleaning system which vacuums AND scrubs floors; and is also a room deodorizer.

what problem or need does this concept address?
Target market is any business that offers a single or multi stall restroom.

how does this concept address this need?
The CRS (Complete Restroom Sanitizer) is a continuous motion, self- contained cleaning device, which when set, vacuums dirt, debris, toilet tissue and paper towels on one side; and when set on "Wash" on the opposite side, will scrub and wash floors with

what risks or barriers might prevent this concept from being realized?
Human error in setting the machine for use; Large obstacles in the path of equipment. low battery. Something stuck in brush mechanism. etc.

Problems? Send us an email at crowdconcept@mit.edu

FIGURE C-2: SAMPLE CONCEPT 2
**Toilet Paper Sensor**

Know if there's toilet paper before you enter the stall.

- **Toilet Paper Status:**
  - Good
  - Low
  - Empty

A sign on the door tells users if that particular stall has toilet paper stocked, is running low, or is completely out of toilet paper.

**what problem or need does this concept address?**

This product is designed for every stall that uses toilet paper. No one likes finding out that the stall they chose to use has no toilet paper after the fact. By allowing users to clearly see if the stall has toilet paper, they can avoid awkwardly asking their neighbor for some toilet paper. Besides from letting users know if the stall is stocked, janitors will save time by being able to quickly see which stalls need to be restocked, saving them time and their employers money. Currently, the best way to determine whether or not a stall has toilet paper is with your eyes.

**how does this concept address this need?**

The product is a sign that is installed on the outside of a stall door which lets potential users know if that stall has toilet paper. If the stall has a healthy stock of toilet paper, the status symbol will be lit up at 'good'; if running low, 'low'; if completely out or down to the last few sheets, 'empty'. The status will be determined by a sensor placed in the toilet paper holder that would send a signal. The sensor could get its readings from the weight of the toilet paper roll, the diameter of the toilet paper, a light sensor that detects when the cardboard center of the roll is visible, or any combination of the three.

**what risks or barriers might prevent this concept from being realized?**

The cost of electronics for the sensor and sign may prove costly if produced in low quantities; a power source would be needed for both the sign and sensors.

Problems? Send us an email at crowdconcept@mit.edu

**FIGURE C-3: SAMPLE CONCEPT 3**
TheBusiness
Toilets On Demand

what problem or need does this concept address?
Where is a toilet when you really need one? Often it can be hard to find toilets at critical times, especially when out in public.

how does this concept address this need?
TheBusiness is a phone app that allows customers to easily find public toilets and even provides on demand toilet services for those just-can’t-hold-it moments. TheBusiness can be downloaded onto any smartphone, and works with google maps to show a customer the nearest bathroom locations within a small range depending on what transportation is available to the user (foot, bike, public transportation, or car). The customer can also opt to hide or show pay toilets in the area. If there are no bathrooms in the area that the user finds acceptable or easy to get to, they can request an on-demand toilet to be driven to their location. These may only be available at peak bathroom-use times and locations (for example, during a football game or a 3-day concert festival). With TheBusiness, users can upload photos of each bathroom location, and can rate and comment on bathrooms they have visited, allowing other users to make an informed decision about which bathroom to choose.

what risks or barriers might prevent this concept from being realized?
The app creators would have to create maps of bathroom locations across cities, but this task could be put to the users. A feature could be added to allow users to submit new bathroom locations to the app. If revenue is an issue for the app creators, one way to benefit from the app would be to allow businesses to advertise within the app, or to feature bathrooms in locations where nearby businesses pay a small fee. This would cause people to be more likely to use those restrooms, and to use visit the businesses nearby.

Problems? Send us an email at crowdconcept@mit.edu

FIGURE C-4: SAMPLE CONCEPT 4