Distributed Interaction Design

Abdullah Ali

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington
2020

Reading Committee:
Jacob O. Wobbrock, Chair
Alexis Hiniker
Meredith Ringel Morris

Program Authorized to Offer Degree:
Information School

University of Washington

Abstract

Distributed Interaction Design

Abdullah Ali
Chair of the Supervisory Committee:
Jacob O. Wobbrock, Chair
The Information School

Creating user interfaces that are natural, guessable, learnable, memorable and accessible is a persistent challenge. Involving end users in the design process is a well-established approach to address these challenges, but traditional participatory design has limitations, especially when it comes to scaling beyond the lab and reaching diverse participants. I build on the success of a popular participatory design method called end-user elicitation. Elicitation studies work by presenting the effect of an interaction (e.g., what happens after a user makes a gesture) and asking end-user participants to perform the action that would have caused that effect (e.g., the gesture itself). Despite their success, elicitation studies have important limitations. They typically are confined to a lab setting, limiting the number of participants, their diversity and the representativeness of the results. Also, analyzing the studies' results is a laborious process. Furthermore, elicitation studies lack a formal approach to evaluate the quality of their results. My work addresses these limitations by scaling beyond the lab and conducting distributed elicitation studies with online crowds. In this dissertation, I have created an open platform and formulated the Distributed Interaction Design (DXD) process. This work overall contributes methodological extensions to elicitation studies, an open platform for the research community, and empirical studies of user-generated interactions such as gestures, voice commands, and icons. The thesis I demonstrate in this work is:

Using a custom-built platform to conduct Distributed Interaction Design (DXD) enables: creating user-elicited interactions; evaluating the guessability, learnability, and memorability of interaction designs; and the recruitment of participants through third-party services in a timely manner.

2020

DISTRIBUTED INTERACTION DESIGN

ABDULLAH X. ALI

Distributed Interaction Design

By
Abdullah X. Ali
B.S. Information Systems, University of Maryland Baltimore Country
M.S. Human-Centered Computing, University of Maryland Baltimore Country
M.S. Information Science, University of Washington

Doctor of Philosophy

Dissertation Committee
Jacob O. Wobbrock (Chair)
Alexis Hiniker
Meredith Ringel Morris
Katharina Reinecke

“Here’s to the crazy ones, the misfits, the rebels, the troublemakers, the round pegs in the square holes… the ones who see things differently — they’re not fond of rules… You can quote them, disagree with them, glorify or vilify them, but the only thing you can’t do is ignore them because they change things… they push the human race forward, and while some may see them as the crazy ones, we see genius, because the ones who are crazy enough to think that they can change the world, are the ones who do.”

-Steve Jobs

Abstract

Creating user interfaces that are natural, guessable, learnable, memorable and accessible is a persistent challenge. Involving end users in the design process is a well-established approach to address these challenges, but traditional participatory design has limitations, especially when it comes to scaling beyond the lab and reaching diverse participants. I build on the success of a popular participatory design method called end-user elicitation. Elicitation studies work by presenting the effect of an interaction (*e.g.,* what happens after a user makes a gesture) and asking end-user participants to perform the action that would have caused that effect (*e.g.,* the gesture itself). Despite their success, elicitation studies have important limitations. They typically are confined to a lab setting, limiting the number of participants, their diversity and the representativeness of the results. Also, analyzing the studies' results is a laborious process. Furthermore, elicitation studies lack a formal approach to evaluate the quality of their results. My work addresses these limitations by scaling beyond the lab and conducting distributed elicitation studies with online crowds. In this dissertation, I have created an open platform and formulated the Distributed Interaction Design (DXD) process. This work overall contributes methodological extensions to elicitation studies, an open platform for the research community, and empirical studies of user-generated interactions such as gestures, voice commands, and icons. The thesis I demonstrate in this work is:

Using a custom-built platform to conduct Distributed Interaction Design (DXD) enables: creating user-elicited interactions; evaluating the guessability, learnability, and memorability of interaction designs; and the recruitment of participants through third-party services in a timely manner.

Table of Contents

Figures List

Tables List

To Paige.

My partner, the love of my life.

Acknowledgments

I would like to thank three special women who have shaped my life, and without whom I would never be writing this document. First, my role model and inspiration, my mother, Arwa Mustafa, whose strength and perseverance lifted our family from the depths of war into safety. Second, my sister, Jood Ali, whose unconditional love taught me the compassion and empathy that I employ in my work and this dissertation specifically. Finally, my partner and love of my life, Paige B. Collins; her support and love is what kept me going through one of the most challenging chapters of my life yet.

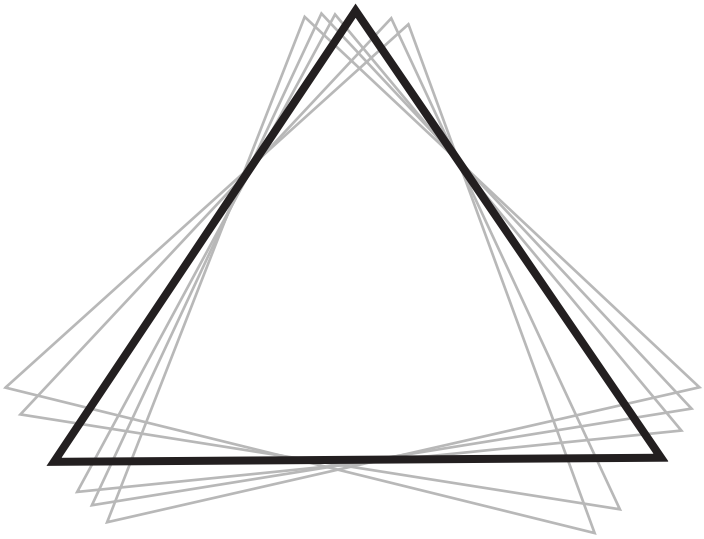
In addition, I would like to thank my advisor, Jacob O. Wobbrock, for taking a chance on me and for supporting me, collaborating and innovating with me throughout this journey. Special thanks go to the incredible scientists on my dissertation committee: Meredith Ringel Morris, Alexis Hiniker, and Katharina Reinecke. I am continuously impressed by your acheivements and humbled by your generosity.

To all those who helped me on the projects detailed in this dissertation, thank you: Erin McAweeney, Huiru Luo, Nathan Lipiarski, Aditya Harish Nayak, and my two furry housemates Nellie and Ruby Collins-Ali. Thanks to my academic family members of the SWAMP (Patrick Carrington, William Easley III, Erin Buehler, and Adegboyega Akinsiku). Fellow ACE lab (formerly MAD lab) members past and present: Martez Mott, Abi Evans, Katie O'Leary, Kristen Shinohara, Alex Mariakakis, Anne Ross, Rachel Franz, Mingrui Zhang, Ather Sharif, and Lisa Elkin; my i16 cohort; the large iSchool family; the even larger DUB family; and the biggest family of the whole HCI research community who embraced me and accepted me with no discrimination. I am grateful to have incredible mentors and colleagues who taught me, listened to me, believed in me, and challenged me intellectually including pioneers like Amy Hurst, Shaun Kane, Stacy Branham, Luz Rello, Jeffery Bigham, Aquesha Martin Hammond, Ravi Kuber, Adam Fourney, Miguel Ballesteros, and Michael Rennaker.

Thanks go out to Microsoft Research, the National Science Foundation, the Mani Charitable Foundation, and the University of Washington for funding my work. Thanks to all my participants.

Finally, I would be remiss if I don't mention all the giants before me whose shoulders I stand on. Without their contributions to science, nothing in this document would have been possible.

Foundation



One | Introduction

In this chapter, I lay out the background and motivation behind my work, outline my research questions, detail how this dissertation is structured and list the contributions I make in this dissertation.

1 Background and Motivation

Eliciting input from end users to design systems is a common practice in the field of human-computer interaction (HCI). User-centered design has involved users in many aspects of interactive system design through participatory design studies [13]. My area of interest is interaction design, which I define in more detail in the **Interaction Design** subsection of the next chapter. Perhaps the earliest example of HCI research involving end users in proposing interaction designs is Good *et al.'s* [37] work generating user-driven key terms for a command-line interface. Later, a similar approach to Good *et al.'s* was formalized by Wobbrock *et al.* [127,131] around gestural interactions. Wobbrock *et al.* set the methodology to create a conflict-free gesture set for surface computing. Their work contributed an equation to calculate agreement among participants. This method, dubbed the *end-user elicitation study*, has proven to be quite popular since 2009, with more than 300 published studies utilizing the method in many areas, such as interacting with drones, augmented and virtual reality environments, robots, and vehicles, to name a few. I provide many examples that show the versatility of the method and the ample work in the literature citing Wobbrock *et al.* [131] in the **Related Work** chapter of this dissertation.

The premise of end-user elicitation studies is that eliciting input actions from end users leads to the creation of intuitive technologies that are guessable, learnable and memorable [85]. Research on elicitation studies has shown that interactions proposed by larger groups of people tend to be preferable to those proposed by smaller groups, even those created by professional designers [80]. In short, elicitation studies work as follows: researchers invite potential users to a laboratory, present them with the effect of an interaction on a computing system, and ask them to propose the action meant to invoke that effect.

Despite elicitation studies' popularity, the method has a few important limitations. First, the status quo of running elicitation studies in a lab setting limits the number and diversity of the participants, hence limiting the representativeness and usefulness of the study results. Participants who live near the research lab and are physically able and willing to go there to partake in a research study are the only ones who propose designs shaping future technologies. Second, part of the analysis of elicitation studies consists of a laborious task that requires substantial time and effort, namely the grouping of proposals based on their similarity. Including many participants in an elicitation study is desirable; however, this practice increases the number of comparisons among proposals when analyzing the results. Finally, elicitation studies lack a formalized method to evaluate the user-elicited proposals.

2 Research Questions

My work builds on the success of the elicitation study method and extends it. I address the drawbacks outlined above by formulating methods and building tools. In addition, I have used my tools and methods to address some open research questions in the elicitation literature, investigate existing interfaces and design emerging ones.

To address the first limitation and include larger more diverse groups of participants, I created an online tool, called *Crowdlicit* [9], that translates the elicitation method to run online while incorporating the best practices published in the elicitation literature. To create Crowdlicit, I set out to answer the following research questions:

- RQ.1 How can I expand the reach of elicitation studies beyond the lab and reach a larger more diverse pool of participants?
- RQ.2 What are the requirements for a system that translates the elicitation study to run in a distributed fashion?
- RQ.3 What are the benefits and drawbacks to running elicitation studies in a distributed fashion?

The Crowdlicit tool and the answers to the above questions are discussed in detail in the chapter **Crowdlicit**.

To address the second limitation of elicitation studies, the difficulty of analyzing them efficiently, I created an online tool called *Crowdsensus* [7]. I set out to answer the following three research questions to create and evaluate Crowdsensus:

- RQ.4 How can a crowd of online workers facilitate the similarity judgments for agreement analysis in end-user elicitation studies?
- RQ.5 By using the crowd, what are the benefits, if any, in terms of cost and time compared to the status quo use of experts' judgments?
- RQ.6 How does the quality of the results produced by the crowd compare to those produced by expert researchers?

The Crowdsensus tool is discussed in detail in the **Crowdsensus** chapter.

To address the third limitation of elicitation studies and formulate a method to evaluate the user-elicited design proposals, I establish the **End-User Identification** method to answer RQ.7 in the **Crowdlicit** chapter.

- RQ.7 How can I evaluate the guessability of the interaction proposals resulting from elicitation studies on a large scale in an efficient manner?

I also capitalized on the efficiencies gained from my tools to add other measures of interaction design evaluation—learnability and memorability—by answering the following two research questions in the **"I Am Iron Man"** chapter:

- RQ.8 How can I evaluate the learnability of the interaction proposals resulting from elicitation studies on a large scale in an efficient manner?
- RQ.9 How can I evaluate the memorability of the interaction proposals resulting from elicitation studies on a large scale in an efficient manner?

My platform enabled me to investigate some open research questions in the elicitation literature such as the influence of the legacy bias reduction principles proposed by Morris *et al.* [79]. I empirically evaluated

whether these principles have an effect on the proposals collected in elicitation studies by answering the following research questions:

- RQ.10 How does production influence user-elicited interaction designs?
- RQ.11 How does priming influence user-elicited interaction designs?

I answered RQ.10 in the chapter **Anachronism by Design**. In this chapter, I evaluated existing interfaces (*i.e.,* desktop operating systems) looking at anachronistic icons—ones that represent objects which are no longer used—and their appropriateness, especially with young technology users. In addition to RQ.10, I answer the following research questions:

- RQ.12 What icons would young adults propose to trigger computer functions currently associated with anachronistic icons?
- RQ.13 How familiar are young adults with the objects represented in anachronistic icons?
- RQ.14 How identifiable is a set of icons elicited from young adults?

In the chapter **"I Am Iron Man"** I answer RQ.11 *(How does priming influence user-elicited interaction designs?),* by conducting a series of distributed studies designing gestural interactions for a video player in an MR environment. In addition to investigating the effects of priming on user-elicited gestures, I add other usability metrics to my method of distributed interaction design evaluation besides identifiability, which are learnability and memorability. I do so by answering the following research question:

- RQ.15 How can I create gestures for mixed-reality environments that are guessable, learnable, and memorable?

3 Dissertation Structure

This dissertation is made up of four sections: Foundation, Tools, DXD Applications, and DXD Impact. In the first section, Foundation, I briefly introduce the motivation for my work, what I have done in this dissertation, and the contributions of my work. Even though my work culminated in the creation of the

Distributed Interaction Design (DXD) process, I describe this process in the second chapter of the Foundation section and then work backwards. At the end of the Foundation section, I give an overview of related work, focusing especially on elicitation studies—the core method behind this dissertation—and online HCI research.

In the second section, Tools, I explain the research platform I have created called **The CROWDDESIGN engine** and the two tools that make up that platform, **Crowdlicit** and **Crowdsensus**. These chapters detail the design of the systems as well as how I evaluated their effectiveness.

In the third section, DXD Applications, I provide examples of work I have done utilizing my tools and the DXD approach. This section has three chapters showcasing how I used the DXD process. In the chapter **Anachronism by Design**, I used the DXD process to redesign computer iconography with young adult users. The chapter **"I Am Iron Man"** details how the CROWDDESIGN engine and the DXD process enabled me to run three studies designing and evaluating interactions for a Mixed Reality (MR) environment efficiently. The final chapter of the DXD Applications section, **Beyond This Dissertation**, provides two examples of how I used and continue to use the DXD process and the CROWDDESIGN engine to create usable and accessible interactions in academia and industry.

In the final section, DXD Impact, I provide a discussion, some reflections and conclusion to my work.

4 Contributions

This dissertation makes the following contributions to the field of HCI: (1) An online platform with resources teaching the end-user elicitation and identification methods and providing access to two tools: (2) the **Crowdlicit** tool to run distributed elicitation and identification studies, and (3) the **Crowdsensus** tool for researchers to utilize online crowds to find agreement in complex datasets. (4) The **End-User Identification** method, a new method that evaluates the guessability of interactions. (5) The **Distributed Interaction Design (DXD)** process. (6) The empirical results of 10 studies: three elicitation studies (eliciting voice commands, gestures, and icons), four identification studies accompanying the three elicitation studies, and three experiments (a comparison of the Crowdsensus approach to analyze elicitation studies against expert researchers, a learnability study, and a memorability study).

In this dissertation, I demonstrate the following thesis:

> *Using a custom-built platform to conduct Distributed Interaction Design (DXD) enables: creating user-elicited interactions; evaluating the guessability, learnability, and memorability of interaction designs; and the recruitment of participants through third-party services in a timely manner.*

Two | Distributed Interaction Design (DXD)

At the time of writing this dissertation, the world is grappling with the COVID-19 pandemic. Some countries are still enforcing stay-at-home orders and the world population is practicing self-isolation to slow the spread of the virus. In times like these, innovation is essential, and human-centered innovation is paramount. Despite decades of improved design and usability practices, creating systems with interactions that are highly guessable, learnable, memorable, enjoyable, and accessible is still a persistent challenge. This challenge is exacerbated even more by the number of emerging intelligent technology platforms and environments like wearable devices, drones and robots, interactive surfaces and fabrics, voice-controlled intelligent assistants, and virtual and augmented reality environments. In the field of human-computer interaction, the practice of human-centered design tackles some of these challenges, but it has not been widely adapted to remote use by physically distant practitioners and users. I have formulated a process to enable the creation and evaluation of interaction designs remotely by end users for end users. In this chapter, I provide an overview description the work I have done to take user-centered design approaches out of the lab and online, and how my work led me to formulate the Distributed Interaction Design (DXD) process.

1 Interaction Design

My definition of interaction design—inspired by early work in HCI [46]—has three components: (1) a human input, (2) a system computing function, and (3) the system's feedback or output. To best determine what inputs should trigger what functions and outputs in a system, several methodologies exist that incorporate end users into the design process, such as participatory design [104]. From 2005 – 2009, Wobbrock *et al.* [127,132] developed a related method to make interactive systems more guessable, learnable, and usable called **End-User Elicitation**. By incorporating end users of varying abilities, needs, backgrounds, and values directly in the design process, interactive systems could be made more usable and inclusive.

The end-user elicitation study works by prompting users with the output of a computing function and asking them to propose the action that would trigger that function to bring about that output. In essence, it asks users to work *backwards* from the system's response to the user's action, thereby eliciting the actions that users feel most likely would result in the responses they are shown. Over many participants in an elicitation study, patterns of similar proposals start to emerge that can be implemented in an interactive system. End-user elicitation has become popular, with over 300 published studies by researchers utilizing this method to design a wide range of interactions: gestures for interactive tabletops [131], gestures for blind users of touch screens [49], virtual and augmented reality interactions [96], smart TV controls [75], in-vehicle interactions, and human-robot interactions [92], to name a few.

2 Distributed Interaction Design

In this chapter, I extend and update the elicitation methodology to fit our current state of the world. I created an online research and design platform called **The CROWDDESIGN engine**[1]. The platform enables scaling the end-user elicitation methodology to be conducted completely online, reaching a large pool of participants, remedying the lack of access and user representation that is typical of lab-based studies, and allowing researchers to involve users in the design process of future technologies at a global

[1] http://crowddesignengine.com

scale [9]. The engine also includes a tool to analyze study results efficiently by utilizing the wisdom of the crowds and machine learning [7], drastically reducing the time it takes to conduct end-user elicitation studies and evaluate their results. In addition, I formulated a method to validate user-generated interactions in a distributed fashion, called the **End-User Identification** method. From this work, I extrapolated a six-step process for designing user-centered technologies that I call the Distributed Interaction Design (DXD[2]) process.

2.1 Stepping Out of the Lab

Over time, I identified several areas where opportunities for advancements could be pursued. To begin with, elicitation studies are traditionally run in laboratory settings with ~20 participants on average. Given the social-distancing rules we face, lab studies are impossible. Even if in-lab studies were possible, the limited number of participants leads to results that do not necessarily represent a wide range of users. Also, research has shown that interaction designs generated by large numbers of participants are preferable to those generated by one or a few professional designers [81]. To address the limited number of participants, I built a tool called **Crowdlicit** that reconceptualizes the elicitation process and its best practices to run completely in a distributed manner. Researchers conducting distributed elicitation studies can run their studies either synchronously or asynchronously with any participant in the world who has access to a web browser. Crowdlicit automates collecting, organizing, and storing user proposals in an easy-to-analyze manner.

2.2 Harnessing the Powers of the Crowd

Beyond the usual challenges of participant recruitment, study execution, and data capture, elicitation study data analysis requires a determination of whether two elicited proposals are sufficiently similar to be grouped together as if they were "the same interaction." In most elicitation studies, all elicited proposals need to be compared to each other using subjective human judgment, which requires great amounts of time and effort. In my platform, I created a tool called **Crowdsensus** to analyze elicitation studies—either by importing the data directly from Crowdlicit or manually uploading it—by harnessing the

[2] The X in DXD comes from the abbreviation IX for interaction.

power of online crowd workers and machine learning algorithms to analyze the results of elicitation studies four times faster than manual human analyses [2].

2.3 **Distributed Design Evaluation**

The literature employing elicitation studies published over the last decade shows that most studies conclude by reporting a set of user-generated interactions. However, the elicitation process lacked a formalized method to evaluate or validate these user-generated interactions. I therefore established a method called the **End-User Identification** to evaluate input actions before investing time and resources implementing these actions into interactive systems. These input actions could be new or existing actions designed by interaction designers or sets of interactions resulting from end-user elicitation studies or other participatory interaction design methodologies [3]. Identification studies reverse the elicitation process by presenting users with a human input action and asking them to propose the system function or system output they expect the input action would trigger. I built **The CROWDDESIGN engine** in a robust way to run and analyze both elicitation and identification studies in a distributed fashion.

3 The Six Steps of DXD

Putting all my work together, I created a six-step iterative process to designing interactive systems with a global pool of participants. I illustrate the six steps using an example of how the methodology works.

Figure 1 The six-step Distributed Interaction Design (DXD) process.

Step 1: Set Up the Four Pillars of a DXD Study

From my experience building **The CROWDDESIGN engine** and running distributed user-centered design studies, I found that there are four foundational elements that need to be established and communicated to remote participants properly to ensure the success of a DXD study:

- *Rules of Engagement:* Study instructions preceding the start of an elicitation study should establish the rules of engagement. These rules explain to the participant (a) the environment in which they are to imagine the system being designed; (b) the form of the system; and (c) the system's sensing capabilities. An example of this would be, "imagine you are interacting with a TV set in your living room that is able to recognize voice commands."
- *A List of Functions:* Every interactive system has a list of functions triggered by user input—*i.e.,* actions. In elicitation studies, these functions are used as prompts. For example, a media player has functions like "play" and "pause."
- *Prompt Modality:* A prompt can be presented to participants in various ways: as a text description; as still images (*e.g.,* before and after pictures of the system state); as audible feedback (*e.g.,* tones and beeps, or natural language output); or as video showing the effects on a system. All these different presentation modalities are available in the Crowdlicit tool on the CROWDDESIGN platform.

- *Proposal Modality:* The proposals collected from remote participants can take one of many forms: text descriptions of actions, like pressing a button or turning a knob on a physical interactive system; natural language commands; or actual text-based commands for command line interfaces. They could also be still images or sketches. For dynamic proposals, they could be audio or video clips. Proposals can even take the form of annotations on a wireframe or existing user interfaces.

Example: Suppose a system creator is designing a new robotic arm that performs many functions triggered by mid-air gestures. The system creator might formulate the following study instructions: "Imagine you are interacting with a robotic arm sitting on your desk. It can sense your body movements and accept them as commands." One of the functions the arm can perform is gripping an object. The creator then formulates the following prompt to present to participants in an elicitation study: "Perform a mid-air gesture that would make this robotic arm grip an object." The system creator can represent this prompt in several ways other than text, such as two images, one showing the robotic arm with open fingers—the "before" state—and one where the fingers are closed together—the "after" state. Another way would be a video showing the robotic arm closing its fingers accompanied with the instruction, "Perform a mid-air gesture that would trigger this movement." Having viewed and understood the prompt, the participant then would provide a proposal for the gesture, which the participant could describe in text, sketch as a sequence of images, or best of all, perform and record as a video.

Step 2: Collect Proposals

Human input actions can be captured in many forms, as stated above. In the example of the system creator attempting to design a mid-air gesture for triggering the grasping function in their new robotic arm, the creator then recruits tens, or maybe hundreds, of participants for an elicitation study. The participants might propose mid-air gestures such as the three shown in Figure 2.

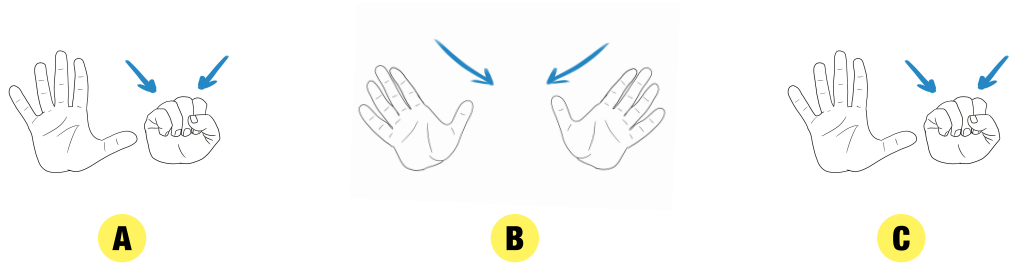


Figure 2 Three gestures by different participants in an elicitation study in response to the prompt, "Perform a mid-air gesture that would make this robotic arm grip an item." Participants "A" and "C" proposed the same gripping gesture, whereas participant "B" proposed a different gesture.

Step 3: Create Interaction Sets

Once proposals for a prompt are collected, it is time to find the proposal with the highest consensus among participants. All proposals must be compared for their pairwise similarity and put into similarity groups. In our example, the system creator would group the gestures based on their similarity and implement the gesture with the highest consensus in the new system. So, the resulting gesture here would resemble proposals "A" and "C". At the end of this step, the creator will have a list of input action designs—informed by actual end users—that map to the functions of the system.

Step 4: Test Interaction Quality

Many metrics of the ISO 9241 standard for usability lend themselves to distributed evaluations such as task, learning, and individualization suitability, and conformity with user expectations. As mentioned above, most end-user elicitation studies conclude by reporting a set of user-generated proposals, but without a decisive way to claim whether those proposals are "good" or not. I have established a method to test the quality of the proposal-prompt relationship called **End-User Identification**. An identification study is the reverse of an elicitation study: participants see a prompt of an input action and guess what the system would do or the feedback it would provide. In this example, when running the identification study, the system creator would recruit new participants and present them with the resulting gestures from the elicitation study, like the hand-closing gesture for grabbing an object (Figure 3). The creator asks participants to propose the function that the robotic arm would perform in response to this gesture.

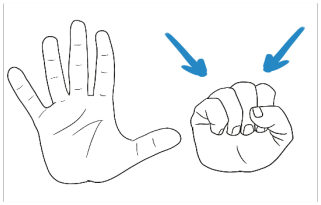


Figure 3 The gesture with the highest consensus from participants in the elicitation study of Figure 2.

The creator would then group the proposed functions based on their similarity to find the one with the highest consensus in a similar manner to Step 3 above. If the resulting proposed function matches the original function used to elicit the user-generated gesture, this indicates that the input action-system response relationship is an identifiable one, and that the proposals are a good fit for this prompt. Because of the flexibility of **The CROWDDESIGN engine**, I can utilize it to run and analyze distributed identification studies with the **Crowdlicit** and **Crowdsensus** tools. Other usability studies such as learnability and memorability studies can take advantage of Crowdlicit and use it as a tool to gather, organize, and store data. Other metrics from the ISO 9241 standard like error tolerance and controllability might be more suited for traditional usability testing as they work within the context of use with the actual system being evaluated.

∋ Step 5: Decide Whether to Repeat Steps 2 – 4

In cases where the input action-system response relationship is not easily identifiable, a new round of proposal collection and design quality testing can be conducted. The DXD process is iterative for this reason. Researchers can repeat steps 2 – 4 as necessary until they arrive at a set of interactions generated and tested by end users to implement in their system.

∋ Step 6: Recommend Interaction Designs

Finally, researchers and system creators can build systems informed by real users on a global scale, making future technologies inclusive of different users' perceptions, values, and physical abilities.

The DXD process is meant to inform designers, researchers, and system creators of actual end users' needs, abilities, and preferences. DXD sits between a wireframing tool and a code editor. After system creators have established the system's form and functions, they can get insight into how the system's users would best interact with it before investing the resources to build the system.

Having now introduced the overall DXD design process, I turn to related work, after which, I delve into the creation of my tools that enable this DXD process to efficiently execute online.

Three | Related Work

In this chapter I survey the HCI literature and present related work to this dissertation in three sections:

- Work including end users in the design process of interaction designs, how these designs compare to ones created by experts, and which design approach yields better designs. In this section I define what I believe to be a good interaction design.
- The end-user elicitation study—the methodology at the core of my work, its origins, extensions, and application areas.
- Examples of work from the two facets of online HCI, as described by Lazar et al. [61]: *Online research,* using online tools to conduct research, and *human computation*, utilizing large numbers of online users completing tasks collectively. In addition, examples of work including users in the design process in a distributed manner.

1 Designer vs. User

In this section, I discuss literature on user-defined designs and expert-defined designs in three parts: In part 1, I present work that has included users in the design process—drawing on both the literature of elicitation studies and participatory design. In part 2, I discuss specific examples in which authors evaluated user-elicited designs, and examples in which authors specifically compared user-elicited designs against ones created by experts. Based on the many definitions of "better" from the aforementioned examples, I synthesize a definition of **good design**. Finally, in part 3, I explain why I believe user-driven design is important, and where design experts fit in the process of designing better technologies.

1.1 User-Driven Interaction Design

Most of the literature shows that involving end users leads to better results in terms of ease of use and user satisfaction, however they do stress that the process is not straightforward. Much work has been done and continues and will need to continue to best figure out how to involve users in design in a successful manner. Abras *et al.* [1] outline some logistical disadvantages to involving users in the design process such as cost, time, and expertise required to conduct participatory design sessions, and the results from small groups of participants can be too specific to generalize. They do conclude that, *"Involving users in design one way or another has been shown to lead to developing more usable satisfying designs."*

In *Participatory Design: The Third Space in HCI*, Muller and Druin [84] do an excellent job of distilling the concerns with merely asking users to design a system, or as they put it "Just Add Users and Stir." They point out that the integrity of including multiple voices in design has been questioned, citing Reyman *et al.* (2005) who summarized, *"the problem from the perspective of professional designers, whose newly-won strength in systems design is challenged by the claims of users' knowledge as a crucial component of design. They note that 'designers have their own expertise,' and 'it is not yet clear which kind of user involvement is most appropriate.'"* Muller and Druin go on to mention drawbacks to the practice such as issues of disagreement among users. Finally, they outline a number of ways users can be involved in designing a system stating, *"there are four roles children [users] can play in the design process: user, tester, informant, and design partner (Druin, 2002). With each role, there is a spectrum of user involvement, at differing points in the design of new technology."*

Good *et al.* [37] talked about the dualism of approaching the disconnect between the user and a system, saying that system designers can either adapt the user to the system or adapt the system to the user with the latter as a preference. They mention an anecdote from the early days of personal computing from Larry Tesler talking about the development process of the Apple Lisa interface where the developers involved users in testing the interface. He said, *"We had a couple of real beauties where the users couldn't use any of the versions that were given to them and they would immediately say, 'Why don't you just do it this way?' and that was obviously the way to do it. So sometimes we got the ideas from our user tests, and as soon as we heard the idea we all thought, 'Why didn't we think of that?' Then we did it that way."* Good also presented the opposite view, citing Black and Morgan who stated that, *"A computer system cannot be designed by simply asking computer-naive people what they think it should be like; they are not good designers."*

In addition, Morris *et al.* [82] pointed out the limitation to relying on experts solely, saying, *"HCI researchers may not always create optimal gesture designs despite their expertise."*

It is clear that users' input is imperative to the system design process. However, this does not mean that experts are no longer needed. From the elicitation literature specifically, Pyryeskin *et al.* [99] found that the most frequent user-defined interactions—*i.e.,* the design recommendations as a result of an elicitation study—performed worse in terms of speed and accuracy than less frequent ones which were predicted to perform better by designers. Moreover, researchers need to validate the outcome of user-driven designs. An argument against the status quo

of elicitation studies comes from Nebeling *et al.* [90]. They argue that the process of elicitation appearing throughout the literature is somewhat incomplete stating, *"The majority of [elicitation] studies end with reporting a suitable interaction set."* Wobbrock *et al.* [131] concluded the paper in which they established the elicitation method itself by saying that results need to be validated and proposed running the reverse of elicitation studies to do so, by presenting end users with an input action and asking them to propose the effect that will have on the system. Although Wobbrock *et al.* [131] raised this possibility, they did not investigate it; in this dissertation, I establish and formalize this possibility as **End-User Identification** studies.

In summary, two main points appeared consistently throughout the literature: 1. the collective intelligence of end users can unearth better designs than ones created by a single designer or a small team of designers; 2. Involving end users in the design process requires special care.

1.2 A Good Interaction Design

In this section, I provide examples where the authors evaluated end-user designs, and examples where the authors pitted end-user designs against experts' ones. I conclude by defining what a **good interaction** is based on the definitions and evaluation criteria used in the aforementioned examples.

1.2.1 Evaluating User Designs

Choi *et al.* [23] conducted an elicitation study to derive a gesture set to control a smart home system. In a subsequent study, they presented participants with the entire gesture set and asked them to pick the most appropriate gesture for each command. They found that 65% of the gesture-function couplings from the elicitation study were changed when the same participants looked at all the gestures available. Choi *et al.* asked for preferences over a finite set of interactions; they did not evaluate the designs for their learnability, memorability, or reliability. Their evaluative study took a ranking approach.

Cowan and Li [26] followed Wobbrock *et al.'s* [131] advice and followed their elicitation study with another study reversing the elicitation process. Their study's aim was to determine whether participants were able to identify the effect of an action on a computing system—*i.e.,* the *discoverability* of an action. They did not provide their participants with a set of possible effects and instead took an infinite-possibility approach, in an effort to reflect a

real-world situation. My work formalized this approach in my CHI 2019 paper [9] in which I establish the **End-User Identification** method.

May *et al.* [70] followed up their elicitation study with an online survey showing videos of mid-air gestures and their effects on an in-vehicle system—an example of a distributed interaction evaluation approach. Their survey attempted to assess the workload of each gesture. They asked the respondents first to mimic the gesture themselves, imagine using the gesture while driving, then answer some questions about the physical and mental demand of the gesture, its match to the interface response, difficulty to remember, and difficulty to use while driving. In essence, May *et al.* were concerned with the ease of the interaction both physically and mentally, the action-effect fit and situational fit.

1.2.2 Comparing User-Defined and Expert-Defined Designs

Morris *et al.* [82] compared user-created gestures from an elicitation study to those created by designers by showing videos of the interaction and its effect, asking them to imitate the gesture, then asking the participants to rate the action-effect fit and ease to perform. Then for each function that included synonym gestures they asked participants to rank the gestures based on their preference. They found that gestures proposed by a large number of people were preferred to ones created by a small number of designers.

Nacenta *et al.* [86] stated that "*one of the main factors that could determine the success of gesture sets in modern interfaces is whether the gestures can be effectively learned and remembered.*" They evaluated two gesture sets—a user-defined one and a pre-designed set—by having participants learn a pre-designed gesture, create one to cause an effect on a system and perform the gesture, whether learned or self-proposed. They then invited the users to come back the following day and perform the gestures again to test their memorability. Finally, the participants rated the gestures on the following scales: concentration to learn, easy to remember, easy to articulate, fun, and ranked the gestures on their difficulty to learn, difficulty to remember, learning time, and fun. They found that user-defined gestures were easier, more fun and less effortful. I tackle the learnability of memorability of user-elicited gestures in the **"I Am Iron Man"** chapter of this dissertation.

Pyryeskin *et al.* [99] showed that designers' judgment is needed to enhance the design of interactions. Following an elicitation study they conducted to design over-surface gestures, they compared the performance of the gestures produced by end users to those created by designers. Their study measured how fast a gesture can be

performed, how accurately it can be performed, and how difficult it would be for the system to recognize the gesture. Participants performed the gesture and then answered three Likert-type scale questions about the ease, performance, and overall experience. They found that frequently proposed gestures underperformed those predicted by designers to perform better.

1.2.3 What is "Better?"

A definition of a good interaction comes from Morris *et al.* [79] in which they stated that a "good" gesture is one that meets design criteria such as discoverability, ease of performance, memorability, or reliability. From the examples in this section, I extend the ease of performance to include both mental and physical ease and add action-function fit, and situational fit, to the definition.

My own definition of a **Good Interaction Design** is one that is discoverable, learnable, memorable, mentally and physically easy to perform, matches its intended effect and is situationally appropriate.

1.3 **Toward Better-Designed Technologies**

Given the support for user-driven design in the literature, the question here becomes *how* end users can contribute to the design of systems rather than whether end users *should* contribute to the design process. I believe that users' design input is imperative to the design process of interactive systems. Participatory design methods such as elicitation studies can unlock system designers' creativity by, as Morris *et al. [79]* put it, enabling *"interaction designers to focus on end users' desires as opposed to settling for what is technically convenient at the moment."* End-user proposals will inform expert designers to create better systems.

2 Elicitation Studies

The practice of eliciting input from users to inform interaction designs has been used broadly in human-computer interaction (HCI) research. In this section, I briefly explore the **Origins** of this method. I discuss the paper **User-Defined Gestures for Surface Computing** by Wobbrock *et al.* [131] that formalized and popularized the method in HCI, and provide an overview of the types of work utilizing this method. In addition, I provide an overview of the work done to extend the method, and the areas of application in which this method has found success.

2.1 **Origins**

The earliest example of an elicitation study, to my knowledge, is Good *et al.*'s work from 1984. In their paper, Good *et al.* [37] asked users to propose key terms for their command-line-operated system. Nielsen *et al.* [91] described a similar approach in 2003 to design ergonomic, intuitive, learnable gestures. Wobbrock *et al.* [128] used the approach in 2005 to create easily guessable symbolic gestural inputs for text in EdgeWrite. Arguably, the most influential paper utilizing this approach is Wobbrock *et al.'s* paper from 2009 [131], which I discuss in detail in the next section below.

2.2 **User-Defined Gestures for Surface Computing**

In their 2009 paper, Wobbrock *et al.* [133] outlined the end-user elicitation methodology in a lab setting. For context, I will briefly describe how the method works: system designers invite participants from the target end-user population and prompt them with the effect of an action on a computing system and ask the participants to propose the action to cause the demonstrated effect. Following a proposal, a participant rates how easy it is to perform and how well it matches the function it is intended to trigger. Once the system designers collect a number of action proposals, they proceed by grouping these proposals based on their similarity, calculate the agreement among participants as to which action should trigger each function, and resolve any conflicting proposals assigned to multiple functions. Finally, the system designers match their set of functions with the corresponding action proposals. The process of end-user elicitation studies produces systems with interactions that are learnable, memorable, natural, and easily discoverable.

At the time of writing this dissertation, more than 300 papers (Figure 4) have been published that use and extend the elicitation method for myriad applications; I explore these **Method Extensions** and **Application Areas** in the two sections below.

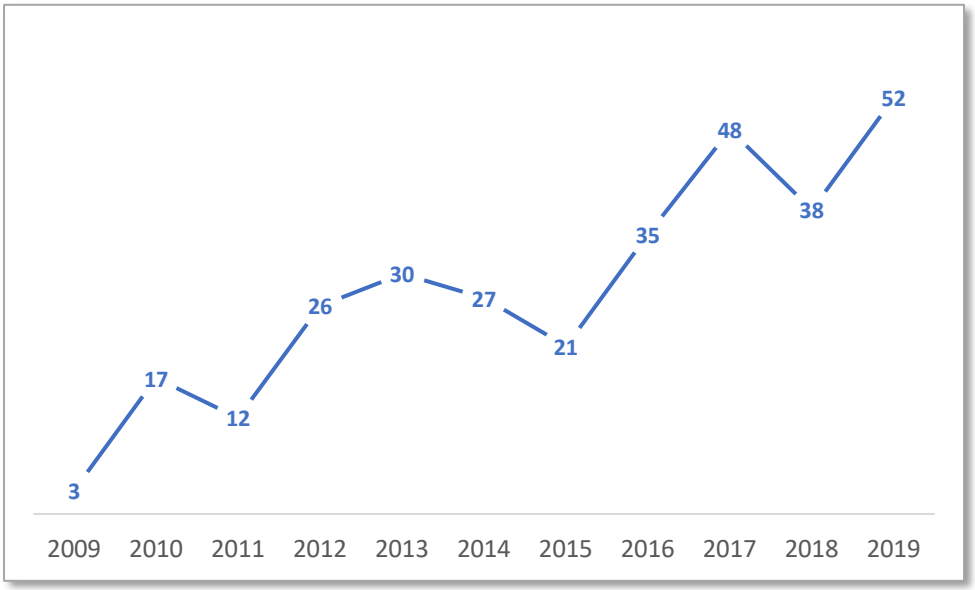


Figure 4 The 309 elicitation studies based on Wobbrock *et al.*'s [132] methodology.

2.3 **Method Extensions**

Numerous scholars published work extending the end-user elicitation method. In this section, I focus on the two main areas of extension I found in the literature: new agreement equations and methods to enhance participants' creativity.

2.3.1 Agreement Equations

Vatavu and Wobbrock [119,121] added new agreement calculations, disagreement scores, and qualitative judgments of agreement rates. Their main contribution was to distinguish agreement calculations for between-subjects elicitation studies from within-subjects elicitation studies, as there are different statistical implications for each. Findlater *et al.* [29] also introduced a version of the agreement equation, their main contribution being having an equation that ranges from 0.00 to 1.00, independent of the number of proposals generated, which Wobbrock et al.'s original equation failed to do. Tsandilas proposed their own approach to calculating proposal agreement [114], directly contesting the work of Vatavu and Wobbrock. Morris [75] proposed consensus metrics to supplement agreement scores especially for situations where the number of proposals per prompt is unequal across prompts.

2.3.2 Unlocking Creativity

Getting participants in elicitation studies to think creatively is important to get past a hurdle in these studies known as "legacy bias"[79]. Legacy bias is discussed in the end-user elicitation literature as a negative influence on the results of elicitation studies, but legacy bias also has its defenders. Drawing on the literature, I present arguments for and against legacy bias and my own take on it.

Morris *et al.* [79] defined legacy bias in elicitation studies as a potential pitfall in which *"users' proposals are often biased by their experience with prior interfaces and technologies."* Moreover, they stated that *"legacy bias limits the potential of user-elicitation methodologies for producing interactions that take full advantage of emerging application domains, form factors, and sensing capabilities."* They continued on to propose three principles—the 3P's—to reduce legacy bias in elicitation studies: Production, Priming, and Partners. Production: *"requiring users to propose multiple symbols for each referent to move them beyond legacy interactions."* Priming: *"engaging users' creativity to think of the capabilities of the new technologies and reduce the impact of legacy bias."* Finally, Partners*: "using groups of participants to engage in an elicitation task to leverage each other's ideas."*

Köpsel and Bubalo [55] presented a counter to Morris *et al.,* arguing that legacy bias helps create good interactions. Citing small participant-group sizes, they argue that the non-legacy interactions will not generalize to a wide user base. They claim that legacy gestures are simple and therefore usable. They say that the complete suspension of legacy bias would erase the advances made in HCI research and lead to unnecessary redundancies in the development of interfaces in novel systems and ask not to condemn what was invented earlier.

I believe that was not the point Morris *et al.* were making in their article. In fact, Morris *et al.* do state that legacy interactions can be discoverable, easy, and memorable, some of the traits of a **good interaction design**. The concern with legacy interactions is their potential to limit users from taking full advantage of emerging applications, form factors and sensing capabilities. Köpsel and Bubalo say that instead of rejecting legacy bias, designers should take what is good from it and invent a better solution for what is problematic and disadvantageous to make users feel more comfortable and familiar. I believe that is the same point Morris *et al.* made in their article.

Hoff *et al.* [41] experimentally tested two of the 3P principles and found that the production principle had no effect on the results and the priming principle had a small effect. They do state that their study had only 30 participants and mentioned that more work with a larger pool of participants is needed to validate these findings.

Beşevli *et al.* [15] tested legacy and non-legacy gestures for their memorability, action-function fit, situational fit and physical ease using self-reported Likert-scale ratings collected from 36 participants. Their results showed legacy gestures to have higher scores; on the other hand, users favored non-legacy gestures due to such gestures' practicality.

Connell *et al.'s* [25] work with children showed legacy bias effects on proposals, as children familiar with some touchscreen interfaces proposed whole-body navigational gestures influenced by participants' experience with touchscreens, while children with no experience with such interfaces applied the greatest variety of gestures.

Other publications in the elicitation literature have mentioned adopting legacy bias reduction techniques following Morris *et al.'s* [79] advice, but do not offer insight on how these techniques impacted their results. May *et al.* [70] and Nebeling *et al.* [89] used the production principle, Morris [75] used production and partners, Pohl and Rohs [98] used priming, and Nebeling [88] used all 3Ps.

The literature has not offered strong opinions to completely discard legacy bias nor recommendations to implement legacy interactions in all future technologies. I believe that in some situations a legacy interaction could be the best one, in the sense that is memorable, discoverable, fits its purpose and situation the best and is easy to use. That said, these measures have a subjective nature to them; what might be the easiest interaction for one person might not be for another due to ability variation. I believe that the 3P principles to reduce legacy bias should be incorporated into the process of elicitation studies. Perhaps instead of using them to reduce legacy interactions, system creators could use them as techniques to discover the best-fitting interaction given a particular use case or user population, whether this interaction is adapted from an older technology or completely original.

In my work, I have experimentally tested two of Morris's three principles—production and priming. In the chapter **Anachronism by Design**, I showed that the production principle reduced the number of legacy icons proposed by young adult technology users. In the chapter **"I Am Iron Man"** I demonstrated how priming with science fiction videos produces gestures for mixed reality environments that are easy and fast to learn and remember.

2.4 **Application Areas**

The end-user elicitation study has proven to be versatile with numerous application and interaction modalities.

2.4.1 Gestural Interaction

The majority of elicitation studies has been focused on eliciting gestural interactions. A survey by Villarreal-Narvaez *et al.* [124] shows that at least 216 gesture elicitation studies have been published as of 2020. Gestures tend to be less well defined as opposed to say voice commands or buttons labeled with text or iconography—as the latter tend to spell out their intended function. This ambiguity is the reason why gestures benefit greatly from elicitation studies to unearth interactions that are easily discoverable, learnable, and memorable.

Researchers have used elicitation studies to explore user-defined gestures to interact with a broad variety of technologies [25,33,35,69,74,92,97,113] (Figure 5). For example, Obaid *et al.* [92] used the method to elicit full-body gestures for controlling humanoid robots. Piumsomboon *et al.* [97] used it to capture user-defined interactions for augmented reality. Tan *et al.* [113] used the method to elicit micro hand gestures as inputs for cycling.

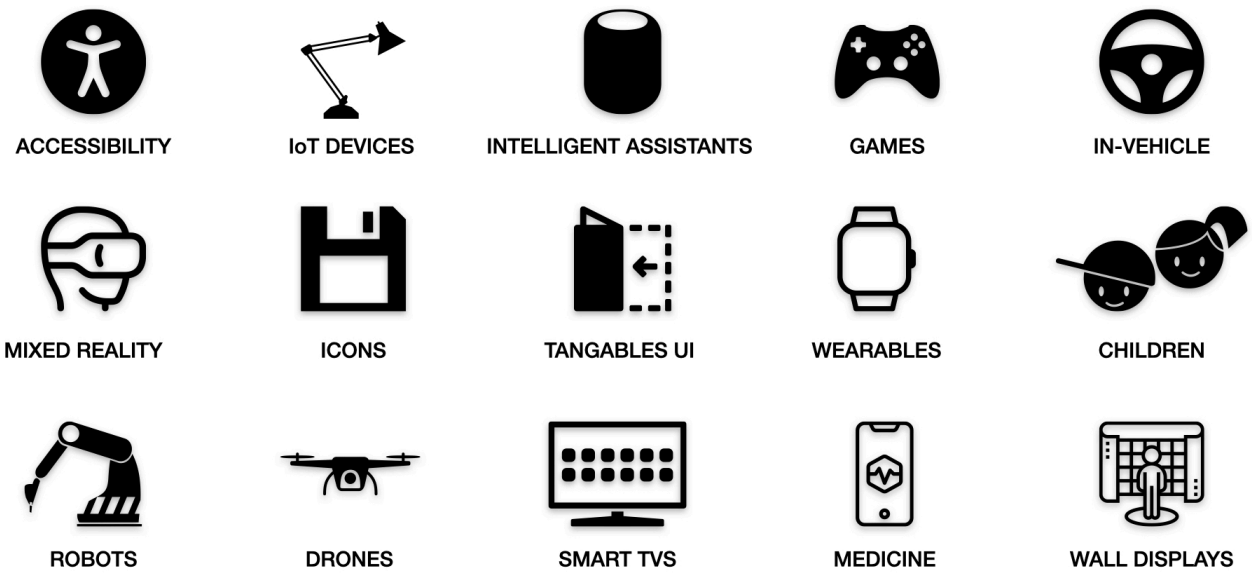


Figure 5 Examples of elicitation study application areas.

2.4.2 Beyond Gestural Interactions

Despite the fact that non-gestural interactions can be less ambiguous, they have benefited from the elicitation study approach. For example, Morris [75] used the approach to elicit speech and gesture interactions for TV-based web browsing. Nebeling *et al.* [89] used the method to elicit voice commands; I do the same in both the **Crowdlicit** and **Crowdsensus** chapters. McAweeney *et al.* [73] used the method to elicit graphical representations of gestures to create user-driven design principles for gesture representation.

To the best of my knowledge, I am the first to use the end-user elicitation as defined in this dissertation to design computer iconography in the chapter **Anachronism by Design**. The use of icons in our highly prevalent graphical user interfaces is of high importance as icons can communicate function meanings faster and more effectively than text [24,126]. Icons first appeared in David Canfield Smith's *Pygmalion* system for programmers [108]. Smith's icons combined visual depictions with behavior, in this case, the execution of computer programs. Later, Smith would join Xerox PARC and the team working on the Xerox Star. The Star employed Smith's icons, now reworked to represent office concepts rather than programming concepts. In this, the desktop metaphor was born, along with the original set of icons designed to communicate to knowledge workers about the Xerox Star's features and functions, making learning and operating the Star more intuitive [109]. According to Smith, the Star's icons were designed to be "visible, concrete embodiments of the corresponding physical objects" [109]. That is, the direct association between the real-world physical object and its computer-icon counterpart was considered a deliberate, even vital, one.

But this association is no longer maintained for many plausibly anachronistic icons today, especially for young adult users who have never encountered many icons' real-world physical objects. Young adult users today who have never relied on an analog calculator or seen a 3.5″ floppy diskette in person do not have the same familiarity with objects represented by certain icons found in today's operating systems and applications. Young adult users of today, unfamiliar with these anachronistic objects, might not draw the same intended associations as previous users once did. This disconnect between icons' functions and the objects they represent was the motivation behind the work I present in the **Anachronism by Design** chapter.

3 Online HCI

Lazar *et al.* [61] classify online HCI into two categories: online studies and human computation. In this section, I present examples of HCI studies that have been translated online from the lab, and human-computation approaches that capitalize on the wisdom of the crowd to accomplish tasks that neither a computer nor a person alone can do. Finally, I dedicate a subsection to online design specifically.

3.1 **Online Studies**

Conducting user research online has enjoyed vast success, especially in the field of HCI. Researchers have reached out from their labs, creating tools and platforms to capitalize on the benefit of using online crowds. Many researchers compared the results of online experiments and found them to be as good as the lab [22,34,38,44,54,100]. Online crowds have also acted as researchers and contributed to writing research papers [118]. Researchers have shown that online crowds are eager to engage with online research for reasons other than monetary compensation like learning about themselves or learning about research, as shown in the numerous studies conducted on the LabintheWild platform [100]. In fact, Ye *et al.* [136] have demonstrated that personalized feedback produced higher quality results in online work than monetary compensation.

Using online crowds in elicitation studies is not an entirely novel idea; some systems attempted to use crowds in parts of the elicitation-study process. For example, Speicher and Nebeling created the GestureWiz [110] tool which utilizes online crowds as gesture recognizers in gesture-based elicitation studies. Magrofuoco and Vanderdonckt created a cloud platform to facilitate gesture elicitation. In this dissertation, I demonstrate how my approach to reconceptualizing elicitation studies to run online goes far beyond eliciting gestures as a response to function prompts, but rather I have built a platform to elicit multi-modal interaction designs, analyze them efficiently, and test multiple metrics of interaction usability like identifiability, learnability, and memorability.

3.2 **Human Computation**

Human computation—also referred to as crowdsourcing—is the practice of using online crowds to perform tasks a computer is unable to complete. For example, online crowds have successfully completed computationally challenging tasks such as labeling images [4], made creative contributions to writing novels [50,51], given design critiques [65,135], and collaboratively generated ideas [106]. Researchers have created systems powered by

crowd workers like word processors [14], animating sketching for prototyping interactive interfaces [62], improving the accessibility of web-browsing [95] and answering questions about the real world for blind individuals [17]. In the **Crowdsensus** chapter, I utilize the wisdom of the crowd to analyze elicitation studies efficiently.

3.3 Distributed User-Centered Design

Ahmed *et al.* [3] studies how people collaborate online on creative tasks. Andolina *et al.* [10] created an online whiteboard to facilitate real-time collaborative ideation in the early stages of design. Bhattacharya *et al.* [16] recruited online groups of teenagers to design for stress management in an asynchronous manner. Bragget *et al.* [18] capitalized on the advantages of a distributed approach to reach a large pool of participants—including ones who are blind or low of vision—to study investigating human listening rates by running their study on the LabintheWild platform mentioned above. Briggs and Makice [19] present an approach they call "deep co-creation," a flavor of participatory design. Deep co-creation allows for a mixed online/offline participation in co-design within retail-based organizations. The work I present in this dissertation adds to the body of distributed user-centered design approaches and systems.

3.4 End-User Elicitation is a Suitable Process for Crowdsourcing

Using Law *et al.'s* [60] framework to decide whether crowdsourcing is a feasible, desired, and useful tool for research, I see that the end-user elicitation process—the core method at the center of DXD—checks the four requirements in the process uncertainties section of their framework shown in Table 1.

Process Uncertainties	*Suits crowds √*	*Less suitable ×*
Research goals	Established	Changing
Workflow	Easily decomposed	Not decomposable
Process and analysis	Separable	Inseparable
Source of serendipity	Diverse perspective	Deep knowledge

Table 1 Types of research attitudes and challenges that make crowd-sourcing feasible, desirable, useful (√) versus not (×) [60]

The process of eliciting interaction proposals from end users is suited for crowdsourcing because:

Research goals: The goal of conducting an elicitation study is established—to create interaction designs informed by users' preferences.

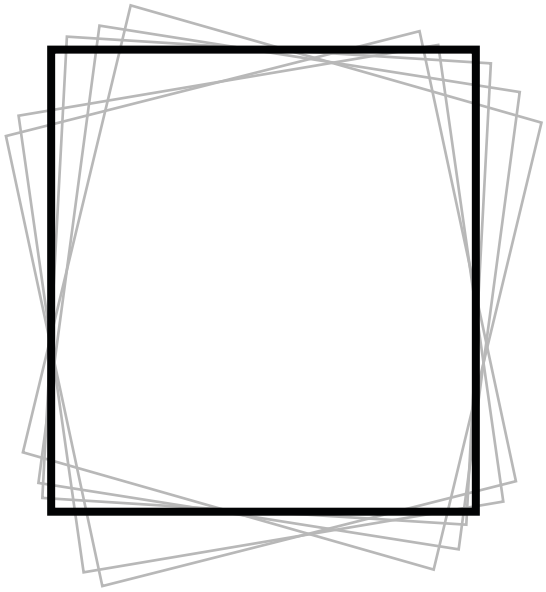
Workflow: The elicitation—and identification—study workflow can be decomposed into tasks (view prompt, submit a proposal, rate proposal).

Process and analysis: The DXD process has 6 iterative steps (see **The Six Steps of DXD**).

Source of Serendipity: Diversity of perspective is a desirable outcome of elicitation studies in order to create technologies with interactions that are appropriate for a wide range of users' preferences.

Given the advantages mentioned above and how suitable the elicitation and identification processes are to crowdsourcing according to Law *et al.'s* [60] framework, it is my opinion that the elicitation-study process can benefit from crowdsourcing, and upon this conclusion I have formulated the distributed interaction design process, enabling the creation and evaluation of guessable, learnable, and memorable user-elicited interactions from a large pool of online participants in a timely manner.

Tools



Four | **Crowdlicit**

A System for Conducting Distributed End-User Elicitation and Identification Studies

End-user elicitation studies are usually confined to a lab, limiting the number and diversity of participants, and therefore the representativeness of their results. Furthermore, the quality of the results from such studies generally lacks any formal means of evaluation. In this chapter, I address some of the limitations of elicitation studies through the creation of the *Crowdlicit* tool along with the introduction of end-user identification studies, which are the reverse of elicitation studies. Crowdlicit is a new web-based system that enables researchers to conduct online and in-lab elicitation and identification studies. I used Crowdlicit to run a crowd-powered elicitation study based on Morris's "Web on the Wall" study [75] with 78 participants, arriving at a set of interaction proposals (voice commands, and mid-air gestures) that included six new proposals different from Morris's. I evaluated the effectiveness of 49 proposals (43 from Morris and six from Crowdlicit) by conducting a crowd-powered identification study. I show that the Crowdlicit elicitation study resulted in a set of proposals that was significantly more identifiable than Morris's.

1 INTRODUCTION

As a reminder, the end-user elicitation method works as follows: researchers invite potential users to a laboratory, present those participants with the effect of an interaction on a computing system (known as a prompt), and ask the participants to elicit the input action (known as a proposal) meant to invoke that effect. Some example "actions" are mid-air or stroke gestures, button text labels or icons, command-line terms, or voice commands. The researchers then cluster the proposals into groups based on their similarity. The group with the highest consensus is chosen as the proposed interaction design to invoke its associated function.

The premise of end-user elicitation studies is that by eliciting input actions from end users, intuitive technologies that are learnable, memorable [29], and easily discoverable can be created. Research on elicitation studies has shown that interactions proposed by larger groups of people tend to be preferable to those proposed by smaller groups [28]. However, the status quo of running elicitation studies in a lab setting limits the number and diversity of the participants, limiting the representativeness and usefulness of the study results. Participants who are geographically close to the researchers and are physically able to go into a lab and partake in a research study are the only ones who propose interactions for future technologies. Also, despite the popularity of elicitation studies and the presence of some published work [28,29,40] assessing user-elicited proposals, the method has another limitation: the absence of a formal approach to evaluate such studies' results.

In this chapter, I address the limitations above. I adapted the elicitation study method to run entirely online to address the limitation of confining the studies to the lab. Web-based experiments have shown support for reaching a wide range of participants who are less WEIRD (Western, Educated, Industrialized, Rich, Democratic) [35]. Participants can partake in studies anywhere without having to take time to travel to a facility to participate in a research study. In addition to increasing participant reach, running studies online cuts down on effort and resources needed to recruit participants. Making online research with end users more accessible opens the door not only to running more studies, but also to extending and or replicating existing studies [35]. To evaluate elicitation studies and address the second limitation, I present the end-user identification method, which reverses aspects of the elicitation study methodology. Participants in identification studies are shown an input action and asked to suggest the system function invoked by it. Researchers are then able to assess the identifiability of their interaction designs.

To conduct elicitation and identification studies online efficiently, I created a system called *Crowdlicit*, making it available[3] to researchers, developers, and designers interested in creating user-centered interactive systems. Crowdlicit provides a centralized way to design, run, and manage elicitation and identification studies online or in the lab. The system allows technology creators to store, organize, and view their study results. Crowdlicit enables system creators to reach participants all over the globe with diverse experiences, backgrounds, and abilities. I built Crowdlicit to flexibly support studies that present prompts in different formats (*e.g.,* text, images, videos, and audio) and collect input actions—or their representations—in varying modalities (*e.g.,* gestures, voice commands, icon sketches, image annotations). Also, Crowdlicit provides a centralized way to organize study results and export them for analysis, either as a csv file for local analysis or ported directly into the **Crowdsensus** system. The Crowdlicit system aims to increase the scalability, accessibility, and efficiency of elicitation and identification studies; to facilitate new studies; and to easily replicate or extend existing ones.

To put Crowdlicit through its paces, I conducted a distributed elicitation study based on Morris's lab-based "Web on the Wall" elicitation study [26]. The Crowdlicit study had 78 participants recruited from Amazon's Mechanical Turk (mTurk). I asked participants to propose free-form gestures or voice commands to interact with a TV-based web browser. I arrived at 15 proposed actions for the 15 functions Morris identified for controlling a web browser on a TV. Morris's proposal set had 43 proposals because it included synonym proposals for each function (*i.e.,* different actions to invoke the same effect). Crowdlicit's 15 proposals had six proposals different than Morris's. I evaluated the identifiability of all 49 proposals (43 from Morris, six new ones from Crowdlicit) by running an end-user identification study using the Crowdlicit system with 24 new participants. I found that Crowdlicit's proposal set was significantly more identifiable than Morris's. I also report on participants' feedback on their experience using Crowdlicit.

This chapter contributes the following: (1) the Crowdlicit system; (2) the new end-user identification method, which evaluates the identifiability of elicitation-study results; and (3) the empirical results of two studies—(i) a distributed elicitation study of gesture and voice commands for a web browser, based on prior work [26], and (ii) an identification study comparing the identifiability of that original study [26] and the Crowdlicit-based distributed elicitation study.

[3] Crowdlicit is part of **The CROWDDESIGN engine**

2 METHODOLOGY

This section details the methods on which the Crowdlicit system is built, the end-user elicitation and identification studies.

2.1 **End-User Elicitation**

An elicitation study is a user-centered interaction design methodology in which end users are presented with the effect of an action on a computing system, known as a prompt, and are asked to propose the action, known as a proposal, meant to invoke the effect. Researchers collect proposals, and other data such as subjective ratings, demographic information, and study notes from participants representing the target end-user population. Researchers then cluster similar proposals into groups to find the proposals with the maximum consensus to trigger each prompt for the computing system they are designing.

2.2 **End-User Identification**

An end-user identification study is a new evaluation method for the input actions that could or do appear in a user interface, including those generated by elicitation studies. Conceptually, identification studies are the reverse of elicitation studies. In identification studies, researchers prompt end users with input actions (*e.g.,* mid-air or stroke gestures, command-line or voice commands, button icons or labels, etc.). Researchers then ask users to propose the system response, usually without giving knowledge of the commands available in the target system. Researchers aggregate the user-generated system responses in groups based upon similarity and proceed by either confirming the input action-system response appropriateness or assigning new responses to actions that had low identifiability.

3 THE CROWDLICIT SYSTEM

This section details the requirements Crowdlicit had to meet to successfully adapt the elicitation study methodology to be online. I outlined these requirements by gaining first-hand experience running an in-lab (traditional) elicitation study exploring young adult technology users' perceptions of computer iconography. I detail the results of the iconography study, which I ran in parallel to the development of Crowdlicit, in the **Anachronism by Design** chapter.

3.1 System Requirements

I defined six requirements Crowdlicit had to satisfy to be able to author studies, collect data from end users, and view and organize results. These requirements allow for a system flexible enough to conduct both elicitation and identification studies.

R1. Study definition. Each study has a unique, dedicated URL distributed to participants. A single study contains prompts and holds all elicited proposals.

R2. Prompt presentation. As prior work shows that elicitation studies have used various prompt formats (e.g., text [7], videos [72]), it is important to maximize prompt-presentation flexibility for researchers by allowing them to choose from different formats.

R3. Legacy bias reduction. Capture natural interactions by allowing researchers to employ legacy bias reduction techniques as put forth by Morris *et al.* [78].

R4. Proposal modality. Prior work has demonstrated that elicitation studies can be applied to various fields (*e.g.,* AR environments [96], in-vehicle interactions [71]). It is imperative to maximize proposal-type flexibility for researchers.

R5. Contextual richness. Prior studies have gathered proposal ratings, think-alouds, and other study notes (e.g., [87,133]). It is important to gather information besides action proposals to provide researchers with rich study results.

R6. Data analysis. Analyzing the results of elicitation studies is a complex and time-consuming process [7]; hence, it is important that my system facilitates this aspect of the elicitation methodology by allowing researchers to clean and organize results for analysis.

3.2 Creating a Study

System creators can create a study in Crowdlicit, as shown in Figure 6 below, with a title, description, an optional password field and post-study survey link, and a dedicated unique URL. A study serves as a container to hold prompts and elicited proposals. Participants who receive the study URL see the title and description. The description of the study can serve as an introduction and instruction manual on how to participate. The option to

include a password is aimed at researchers working on projects protected with intellectual property rights and who might want to restrict access to the study. The option to include a post-study survey allows researchers to enter a URL to an external survey—*e.g.,* SurveyMonkey—they would like participants to complete upon finishing the elicitation study. These options satisfy the first requirement, R1-Study definition.

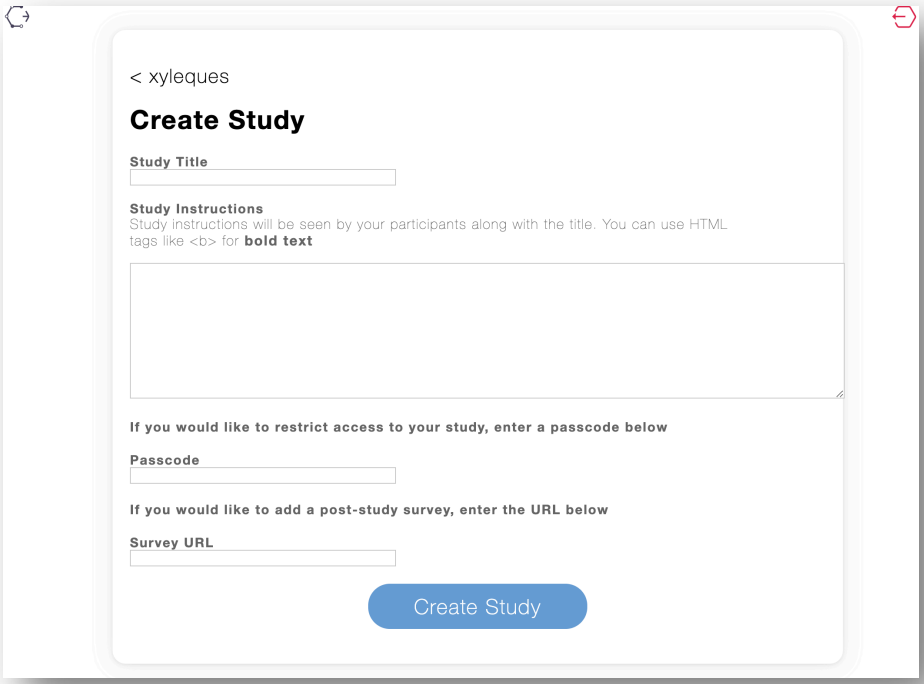


Figure 6 The Crowdlicit interface to create a study. A study requires a title and a description and has the option to be protected by a password and include a post-study link to a survey.

3.2.1 Prompt Presentation

Crowdlicit is designed to provide maximum flexibility when creating a prompt. Each prompt has a title, instructions, and the prompt itself. Crowdlicit offers four ways to present a prompt: (1) A text string describing the effect of an action on a computing system. (2) An audio clip of the prompt itself—used when designing interactions for systems with voice user interfaces (*e.g.,* voice assistant responses). An audio clip can be used to describe a prompt as well. This modality can be beneficial when conducting experiments with individuals who are blind or who have low vision. (3) An image showing the effect of a prompt (*e.g.,* two screenshots side-by-side

showing the before and after states of a system). (4) A video showing the prompt (*e.g.,* screen recordings in the case of an elicitation study in which the prompt represents a computing function). Having multiple ways to present a prompt satisfies the R2-prompt presentation requirement. The interface to add a prompt to a Crowdlicit study is shown below in Figure 7.

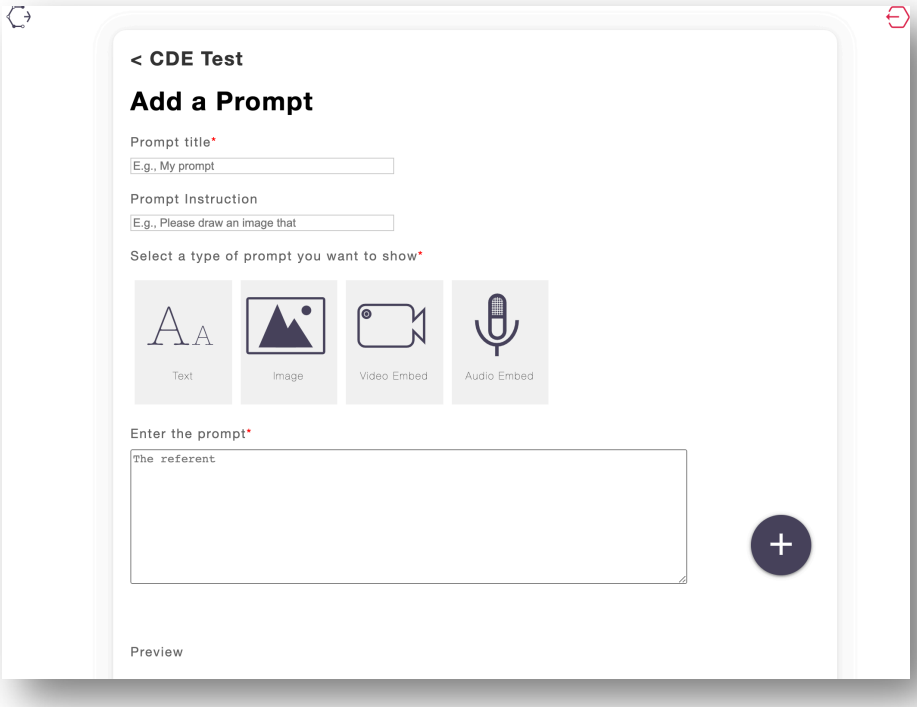


Figure 7 A screenshot of the Crowdlicit interface.

Adding a prompt requires adding a title, instructions, and the prompt itself as text, an image, a video, or an audio clip.

3.2.2 Proposal Preferences

In elicitation studies, researchers collect proposals to inform the design of interactive systems. The majority of work on elicitation studies has centered around eliciting gestural interactions, with a few exceptions in which researchers obtained speech commands [76,87]. But the elicitation method applies beyond just gestural interactions to other input modalities, such as sketches of icons—as I demonstrate in the **Anachronism by Design** chapter. Crowdlicit allows for the flexibility to collect different input modalities: (1) text strings; (2) images; (3) video recordings; (4) drawings, using a canvas element as a drawing pad for sketches; (5) stroke gestures; (6)

user-dictated, *i.e.,* an option to elicit the most appropriate modality as decided by the end user. The flexibility in choosing the type of proposals to collect from participants satisfies R4-Symbol modality from the list of requirements I outlined for the Crowdlicit system.

3.2.3 Reducing Legacy Bias

Morris *et al.* [78] published an extension to the elicitation method explaining that participants tend to propose commands they are familiar with before proposing ones that may be more intuitive. To combat this legacy bias, Morris *et al.* suggest using one or more of the "3P" principles: Production, Priming, and Partners. Crowdlicit implements the first two principles. The capabilities to select a production option and add priming satisfy R3-Legacy bias reduction.

3.2.4 Post-Task Questions and Post-Study Surveys

Crowdlicit includes the option to add proposal-rating Likert-type scale questions assessing the ease and fit of proposals derived from Wobbrock *et al.'s* original method [133], with an option to add an additional custom question. Having participants rate their proposals provides more in-depth insight into the appropriateness of their proposals, making the study results richer. The inclusion of post-task questions, and the researchers' ability to collect demographic as well as other information by including a post-study survey link, satisfy the R5-Contextual richness requirement.

3.3 **Running a Study**

It is possible to run elicitation studies in either collocated or distributed situations using Crowdlicit. When running an in-lab elicitation study, which is the status quo, Crowdlicit allows researchers to collect data from their participants and store it in one convenient location for analysis. Crowdlicit's web-based infrastructure also enables researchers to extend beyond their labs to reach remote participants. Reaching remote participants widens the pool of participants providing interaction-design proposals. In a distributed setting, data collection can be supervised or unsupervised. In an unsupervised setting, researchers have the option to provide their participants with a post-study survey to collect more data, adding context to their results, satisfying R5-Contextual richness. Researchers can also supervise the elicitation session by being in contact with their participants while they are partaking in the study, recoding the session for think-alouds and collect study notes—a practice I employ

in the **"I Am Iron Man"** chapter to run a distributed supervised learnability and memorability study. Each Crowdlicit study has a Welcome page, a Task Manager page, an Elicitation Interface page, and a Thank You page as shown in Figure 8 below.

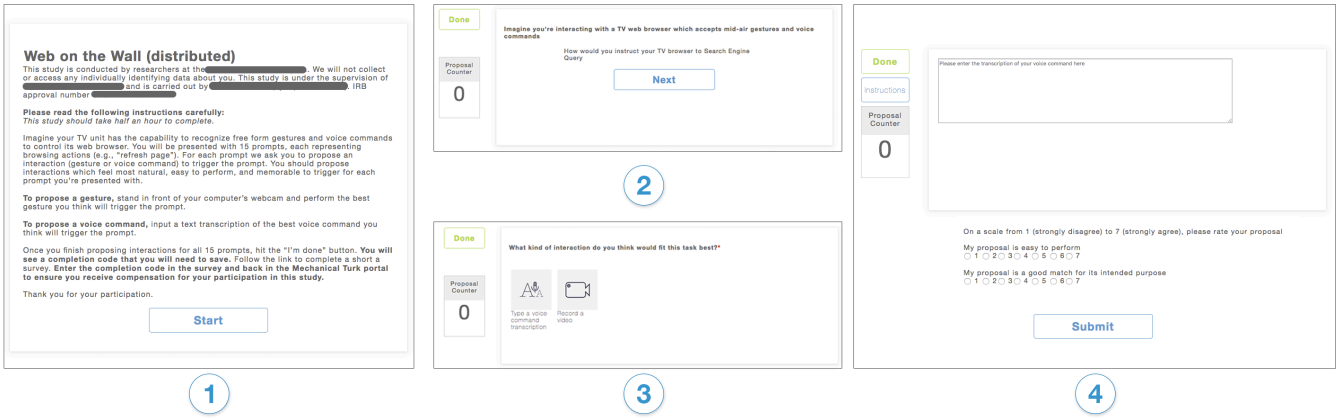


Figure 8 Screenshots of a study created with Crowdlicit. (1) The Welcome page shows the instructions for participating in a study entitled, "Web on the Wall (distributed)." (2) A text prompt. (3) An interface allowing participants to choose between proposing a voice-command or a gesture. (4) A text-based symbol elicitation interface. On the left there are two buttons: Done, which navigates back to the Task Manager; and Instructions, which brings up the prompt and its instructions. Below the buttons there is a proposal counter. The interface shows two Likert-type rating scales and a Submit button.

3.3.1 Welcome Page

This page shows the title, study description, and instructions. A "Start" button is at the bottom that leads to the Task Manager page (Figure 8.1).

3.3.2 Task Manager Page.

This page shows a list of prompts (i.e., set of functions in the case of an elicitation study, or set of actions in the case of an identification study) that the participants have to complete.

3.3.3 Elicitation Interface

This page displays the prompt and priming content, if included. It collects proposals from participants, asks them to rate their proposals, and shows them their progress. On this page, participants see instructions and the actual prompt in whatever modality the researchers choose, *i.e.,* text, video, audio, or an image (Figure 8.2). The participants click "Next" to bring up the priming content, if any has been provided by the researchers. Clicking the "Next" button again dismisses the priming content and displays the proposal-input interface. The proposal-

input interface changes based on the type of proposal the researchers want their participants to propose. For text input, the participants see a text area input element (Figure 8.4). For images and videos, the participants capture an image or video using their device's camera. To draw a proposal or perform a stroke gesture, participants see a canvas element. If the researchers want participants to dictate the modality of the proposal in addition to the proposal itself, they can choose the "User-dictated" option when setting proposal-modality preferences. For such tasks, the participants see a screen asking, "What kind of interaction do you think would fit this task best?" (see Figure 8.3). Participants click on the modality that they deem to be most appropriate for the task, and then the appropriate type of proposal-input screen appears.

3.3.4 Post-Task Questions.

Below the proposal input interface are the Likert-type scale questions, if desired by the researchers, and a "Submit" button that records the participant's proposal and ratings.

3.3.5 Thank You Page

Upon completing a study, participants see the Thank You page. This page contains a link to the post-study survey, if included, and displays the participant's unique identifying code.

3.4 **Analyzing a Study**

Each study in Crowdlicit has a Results page. The page displays all the proposals collected for a specific study, organized by the prompt for which they were proposed. Proposals are listed along with their elicitor's unique identification code and ratings. The researchers have the option to delete proposals. They can also export the study results as a *.csv file to either conduct the agreement analysis themselves [129,133], or utilize an online crowd to handle the analysis for them by using my **Crowdsensus** tool [7] for crowdsourcing similarity judgments for agreement analysis. The ability to store, organize, and export results in Crowdlicit satisfies R6-Data analysis requirement.

4 EVALUATING CROWDLICIT

To test the feasibility of running elicitation studies with Crowdlicit, I ran an elicitation study based on Morris's "Web on the Wall" study [76], which has previously been the focus of replication studies in this genre (*e.g.,* [8,87]). I also ran an identification study, the reverse of an elicitation study, to evaluate the results of my elicitation study and that of Morris's original study. Crowdlicit's flexibility in presenting prompts and collecting proposals of different formats allowed me to run my identification study online. I asked all of my participants from both studies to complete a post-study survey to gather some basic demographic information, data about their technology use and participation in research studies, and their feedback on the Crowdlicit interface.

4.1 **Web on the Wall: Distributed**

I ran the Crowdlicit study with 78 participants from mTurk. Fifty participants completed the entire study—double that of Morris [76]—and 28 gave partial answers, which I include in my analysis. Thirty-three of the 50 participants who completed the entire study filled out the post-study survey. The study required ~30 minutes to complete and paid $6 USD, based on Washington state's $11/hour minimum wage at the time of this study. After participants submitted a minimum of one proposal per prompt for all 15 functions, the system allowed them to click the "I'm Done" button to go to the Thank You page. Participants received their completion code and a link to complete the post-study survey. Participants entered the completion code in both the survey and in the mTurk portal. I used the completion codes to link the demographic information to study answers and identify which participants completed the entire study. Table 2 below shows the demographic information of the participants.

Demographic		Study 1 n=33	Study 2 N=22
Gender	Male	61%	68%
	Female	39%	32%
Age	18–25	18%	23%
	26–40	67%	68%
	41–55	15%	9%
	56 or older	0	0
Highest level of education	< High school	0	0
	High school degree	6%	23%
	Technical degree	9%	9%
	Associate degree	21%	27%
	Bachelor's degree	52%	41%
	Master's degree	9%	0
	Doctoral degree	3%	0
Nationality	USA	85%	95%
	India	12%	5%
	Canada	3%	0
Native language	English	88%	95%
	Other	12%	5%
No. previous research studies	0	12%	23%
	1–3	9%	5%
	4–6	3%	0
	More than 6	76%	73%
No. previous online research studies	0	0	5%
	1–3	6%	0
	4–6	9%	5%
	More than 6	85%	90%

Table 2 Demographic information for 33 of 78 participants from my elicitation study (study 1) and 22 of 24 participants from my identification study (study 2).[4]

4.1.1 End-User Identification Study

I recruited 24 new participants from mTurk for an identification study. They provided open-ended function proposals for all 49 input action prompts (43 from Morris's study plus six new ones unearthed by my elicitation study). Of the 24 participants, 22 completed the post-study survey; Table 2 shows their demographic information. Participants who accepted the mTurk HIT (Human Intelligence Task) went to a Crowdlicit page, which was structured like an elicitation study except that in each task, participants viewed a text prompt describing a gesture or voice command instead of a computing function. Study instructions asked participants to imagine they were interacting with a TV-based web browser. For every prompt (Table 3), participants were asked to freely propose one function in text form. The HIT required about an hour to complete and paid $11 USD, Washington state's minimum wage.

[4] Some participants did not complete the demographic information.

Function	Morris Proposal	#	A	Crowdlicit Proposal	#	A
Open Browser	1. hand-as-mouse to select browser icon	8	0.30	2.“open browser”	76	0.77
	2.“open browser”	5	0.77			
	3.“internet”	3	0.38			
	4.“<browser name>” (e.g., “Internet Explorer,” “Firefox,” “Chrome”)	3	0.84			
Search Engine Query	5.“<query>”	6	0.30	6.“search <query>”	40	0.25
	6.“search <query>”	5	0.25			
Click Link	7.hand-as-mouse to select link	13	0.39	44.“click <link name>” *	37	0.77
	8.“<link #>” (assumes all links have a number assigned to them)	3	0.40			
Go Back	9.“back”	7	0.92	9.“back”	52	0.92
	10.flick hand from right to left	7	0.23			
	11.hand-as-mouse to select back button	5	0.66			
	12.f1lick hand from left to right	4	0.18			
Go Forward	13.“forward”	6	0.58	13.“forward”	34	0.58
	14.flick hand from right to left	5	0.33			
	15.flick hand from left to right	5	0.25			
	16.hand-as-mouse to select forward button	3	0.77			
Open Link in Separate Tab	17.hand-as-mouse hovers on link until context menu appears, then hand-as-mouse to select menu option	3	0.43	45. “open <link> in a new tab” *	35	0.92
Switch Tab	18.hand-as-mouse selects tab	7	0.56	46.“switch tab” *	35	0.92
	19.“next tab”	4	0.84			
	20.“tab <#>” (assumes all tabs have a number assigned to them)	3	0.84			
	21.flick hand	3	0.18			
Find in Page	22.“find <query>”	4	0.77	22.“find <query>”	36	0.77
	23.hand-as-mouse to select a find button, then type on virtual keyboard	3	0.30			
Select Region	24.hand-as-mouse sweeps out diagonal of bounding box	6	0.12	47. “select <region>” *	44	0.64
	25.hand-as-mouse acts as highlighter, sweeping over each item to be included in region	3	0.77			
Open New Tab	26.hand-as-mouse to select new tab button	6	0.44	28.“open new tab”	40	0.92
	27.“new tab”	5	0.92			
	28.“open new tab”	5	0.92			
Enter URL	29.“<url>” (e.g., “its2012conf.org”)	7	1.00	48. “enter <URL>” *	23	0.92
	30.type on virtual keyboard	5	0.36			
	31.“go to <url>”	3	0.92			
Reload Page	32.“refresh”	9	0.92	49.“reload” *	38	0.84
	33.“refresh page”	9	0.92			
	34.move finger in spiral motion	3	0.28			
Bookmark Page	35.hand-as-mouse selects bookmark button	7	0.43	36.“bookmark page”	33	0.71
	36.“bookmark page”	5	0.71			
Close Tab	37.“close tab”	5	0.92	37.“close tab”	42	0.92
	38.hand-as-mouse to select close button on tab	4	0.92			
	39.“close tab <#>” (assumes all tabs have a number assigned to them)	3	0.77			
Close Browser	40.hand-as-mouse to select close button on browser	6	0.77	41.“close browser”	36	0.84
	41.“close browser”	3	0.84			
	42.“exit”	3	0.70			
	43.“exit all”	3	0.25			

Table 3 Morris's 43 proposals [76]. Crowdlicit' 15 proposals. “*” are new proposals from the Crowdlicit study. The proposals describe gestures; proposals in quotes (“”) are voice commands. The # column shows the number of participants who proposed the proposal. The “A” column shows the function agreement score for each proposal from the identification study.

4.1.2 Post-Study Survey

Each of my participants completed a survey asking demographic questions and about previous research participation (see Table 2). Participants also reported on their technology use and experience with the Crowdlicit interface. I based some of the survey questions on Finstad's [31] usability metric for user experience. Other questions asked participants about their willingness to participate in research studies either by going to a physical facility or by participating online. The last question was open-ended to collect any comments about the interface.

5 RESULTS

I evaluated my proposals and Morris's [76] by conducting an identification study. Participants from both my elicitation and identification studies completed a post-study survey.

5.1 **Crowdlicit Proposals**

I grouped the proposals for each prompt based on their similarity and generated 15 voice commands to trigger my15 prompts (Table 3). For each prompt, participants collectively proposed an average of five gestures that had little agreement; this led me to discard gesture proposals and focus solely on the elicited voice commands. Of my 15 proposals, nine were the same voice-command proposals Morris arrived at in her study [76], and six were new. The number of participants who proposed the selected proposals ranged from 23 – 76. Since I used the production principle to reduce legacy bias in my study [78], my participants were free to elicit multiple proposals per prompt. Having several proposals from a single participant leads to an unequal number of proposals among prompts, making Wobbrock *et al.'s* [129,133] agreement equation unsuitable. Instead, I used Morris's [76] max-consensus to compare agreement between prompts. Max-consensus is the percentage of participants suggesting the most popular proposal for a prompt [76]. My participants gave proposals with high consensus (M=59%, SD=9%). Figure 9 shows the max-consensus for the 15 proposals.

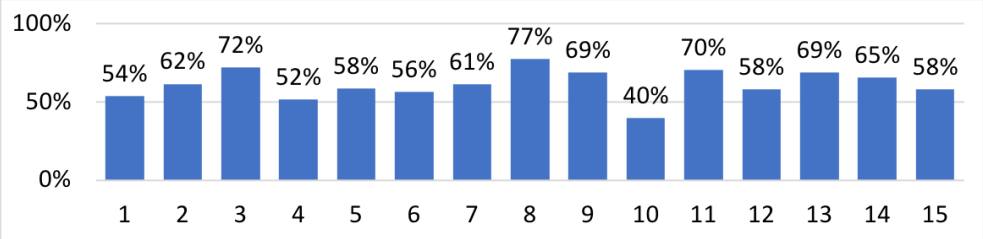


Figure 9. Max-consensus for function proposals 1–15.

5.2 **Action Identification**

In the same manner as an elicitation study, I grouped the proposed function for each one of the 49 action prompts based on their similarity. For each prompt, I selected the function proposal with highest consensus as the triggered function by that action. I arrived at a set of 19 distinct function proposals, which identified Morris's original 15 and added four new ones: open menu, next tab, scroll, and open keyboard.

5.2.1 Function Agreement

For each prompt (Table 3), I calculated the proposed-function agreement using Wobbrock *et al.'s* [129,133] original agreement equation:

$$A_s = \sum_{P_i \subseteq P_s} \left(\frac{|P_i|}{|P_s|} \right)^2$$

Equation 1

In Equation 1, A_s is the agreement of functions proposed for prompt s, P_s is the set of all functions proposed for prompt s, and P_i is a subset of similar functions in P_s. Table 3 lists all 49 function agreement scores.

5.2.2 Accuracy

I compared the functions with highest consensus from the identification study to the original functions. The original function is the one used as a prompt in the elicitation studies (Morris's and Crowdlicit). Identification study participants were able to correctly identify the function for each one of the 15 actions used as prompts in the Crowdlicit interaction set. In this case, "identify" means that the function with the highest consensus from the list of functions proposed for an action prompt matched the original function for that action.

Action	Original Function	Accuracy	New Function	Max-Consensus
15. flick hand from left to right	Go forward	21%	Go back	38%
17. hand-as-mouse hovers on link until context menu appears, then hand-as-mouse to select menu option	Open link in a separate tab	4%	Open menu	63%
19. "next tab"	Switch tab	0%	Next tab	92%
21. flick hand	Switch tab	8%	Scroll	29%
30. type on virtual keyboard	Enter URL	4%	Open keyboard	54%

Table 4 Five action prompts; their original functions from Morris's study, the accuracy % of the original function, the new function, and the max-consensus % of the new functions. Action prompts in quotes are voice commands.

For Morris's set of 43 action prompts, participants were able to correctly identify the functions for 38 prompts and assigned new functions for five prompts. Table 4 lists the five action prompts with new functions from Morris's study, the original function, and the percentage of proposals of the original function–I refer to it here as "accuracy." The table also lists the newly assigned functions, and their percentage of the total number of proposed functions.

5.2.3 Comparability

I compared the agreement scores of Morris's action proposals to Crowdlicit's action proposals by conducting a two-tailed Welch two-sample unpaired t-test. (Due to unequal sample sizes, using a Student's t-test would be inappropriate.) I found that the prompts' function-agreement scores resulting from the Crowdlicit elicitation study were significantly higher than Morris's prompts ($t(37) = 2.99, p < .005$). In addition to higher agreement scores, the fact that all of Crowdlicit's action prompts were correctly identified as a result of the identification study leads me to believe that the action proposal set resulting from the Crowdlicit study was more identifiable than Morris's.

5.3 **Interaction Habits + Interface Usability**

More than half the participants, 56.5%, had never used mid-air gestures to interact with technologies (*e.g.,* a Microsoft Kinect). On the other hand, only 9.1% of participants had never used voice commands. Figure 10 shows the frequency of voice and gesture use for the 55 participants who completed the post-study survey (33 from the elicitation study, 22 from the identification study). The popularity of voice use in participants' daily lives is a possible reason why the new set of action proposals resulting from the Crowdlicit elicitation study is made up entirely of voice commands.

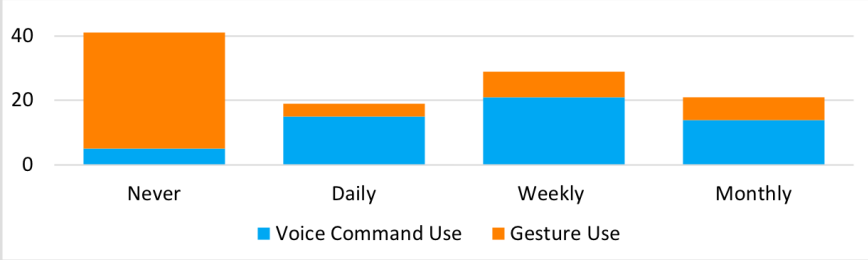


Figure 10 The frequency of using voice vs. mid-air gestures to interact with technology from 55 participants: 33 from the elicitation study and 22 from the identification study.

Participants generally had a positive experience interacting with Crowdlicit itself (Table 5). A Wilcoxon signed-rank test found that participants' willingness to participate in online studies was significantly greater than their willingness to physically go and partake in one ($p < .05$). The majority of the optional feedback was positive. A comment to improve the interface came from P18, who wanted the option to review her proposals after submitting. One positive comment from P16 from the elicitation study stands out: "The interface for this study is EXTREMELY well made, and it only took me about 30 seconds to fully understand how to use it. This is a memorable one, and I definitely will be doing more studies for you."

Question	Mean (n=55)
Using the study interface was a frustrating experience.	2.2 (SD=1.3)
The study interface was easy to use.	6.1 (SD=1.3)
I spent too much time correcting things with the study interface.	2.1 (SD=1.7)
I would participate in online studies (like this one) in the future.	6.9 (SD=0.5)
I would go into a physical facility (lab, university) to participate in research studies.	4.8 (SD=1.9)

Table 5. Fifty-five participants' ratings of the Crowdlicit interface and willingness to participate in research studies. Scores range from 1-strongly disagree to 7-strongly agree.

6 DISCUSSION

To understand my findings in context, I consider the benefits and drawbacks of Crowdlicit, identification studies, the limitations of my work, and directions for the future.

6.1 **Crowdlicit Benefits**

I demonstrated that running an elicitation study using Crowdlicit is not only possible but yields more identifiable proposals than lab-based elicitation studies. Using Crowdlicit allows researchers to scale up their studies and access a large number of participants easily and quickly. For my first study, an elicitation study based on Morris's "Web on the Wall" study [76], it took six hours to recruit and collect data from 78 participants. Fifty of my 78 participants completed the entire study, twice the number of participants as Morris's study [76], which had 25. Twenty-eight of my participants gave partial answers. Morris's study took about 12 hours to run (12 groups of participants × 1 hour per session as reported in [76]). Crowdlicit allowed me to double Morris's number of participants and cut the time in half. Previous work in elicitation studies shows that having more participants in elicitation studies generally yields better results [80].

My work also showed that participants are more willing to participate in online studies than come to a lab, although this could be a self-selection effect, since I only asked people who were already participating in my online study. However, participants' willingness to partake in online studies over lab-based ones complements findings by Zyskowski et al. [138] that some participant groups, such as people with disabilities, prefer crowd-work platforms over in-person studies due to the ability to avoid travel.

6.2 **Crowdlicit Drawbacks**

I collected a number of spam answers in my elicitation study. The first task in the study (open browser) had 58 unusable proposals out of 224 total elicited proposals. Examples of spam answers were text strings saying "nice" or "good" repeatedly. An explanation for the higher number of spam responses in the first task than later tasks is that spammers dropped out after the first task. I attempted to limit spam answers by assigning a minimum time threshold of 10 seconds to answer, but that did not identify all spam answers and risked eliminating some legitimate answers that were provided in less than 10 seconds (*e.g.,* proposing the voice command "Open Browser" took nine seconds). Collecting spam answers is a drawback of the Crowdlicit approach. I added further

measures to limit spam later by randomizing the order of tasks in the Task Manager page. By limiting spam, researchers can collect data efficiently from a large number of remote participants.

My participants reported that they used voice-commands more frequently than gestures in their daily lives, which could explain why the set of proposals I gathered in my Crowdlicit elicitation study mostly comprised voice-commands. Another contributing factor for the popularity of voice-commands in my study may be that proposing voice-commands was faster than proposing gestures, and crowd-workers may have been trying to maximize their efficiency. Future work might explore how to structure Crowdlicit tasks to mitigate attempts by the crowd to game the system to maximize earnings.

6.3 **Identification Studies**

I established and formalized the end-user identification study method to evaluate the action proposals from two elicitation studies. I calculated function proposal agreement using a form of Wobbrock *et al.'s* [129,133] original symbol agreement equation (Equation 1). Agreement can also be measured using one of the other agreement measures proposed by the extensive prior work published on elicitation studies [29,115,119,121].

My comparability analysis can be used to compare two sets of actions meant to invoke the same set of functions. Comparability can be used to assess the outcome of two elicitation studies with different conditions (*e.g.,* lab vs. online), or the outcome of studies conducted with two different populations. Comparability can also be used to evaluate multimodal interactions. Another way the comparability analysis can be used is to evaluate the results of different levels of a dependent variable in an elicitation study. I utilize this approach to evaluate the effects of Priming on user-elicited gestures in the **"I Am Iron Man"** chapter.

I showed in this work that voice commands were more identifiable than gestures for this particular use case. In this case, I attribute the voice preference over mid-air gestures to two factors: (1) Voice commands often spell out their intended purpose and provide a more concrete action than gestures, which tend to be more abstract; (2) Voice-enabled technologies have enjoyed a recent rise in popularity. The majority of my participants had more experience interacting with voice-enabled technologies (see Figure 10) than devices that accept mid-air gestures like the Microsoft Kinect, which was used in Morris's study [76]. This preference may indicate the influence of legacy bias on users' preferences.

Crowdlicit's set of action proposals was more identifiable than Morris's as participants were able to correctly identify all 15 functions for the Crowdlicit action set. For Morris's action set, the identification study participants were able to correctly identify 38 functions for the 43 action proposals and assigned new functions to five actions. In addition to the correct identification of all functions for the Crowdlicit symbol set, the agreement rates for the Crowdlicit set were significantly higher than for Morris's set. I attribute Crowdlicit's symbol set's high identifiability to the larger pool of participants proposing the actions—a testament to Crowdlicit's effectiveness. Another factor worth noting is that six years have passed since Morris's study, which might be reflected in its results (*i.e.,* norms around the interpretation of voice or gesture commands may have shifted due to changes in users' exposure to new commercial technologies).

6.4 **Limitations and Future Work**

For this study, I decided to run my elicitation study unsupervised; the lack of supervision gave me the advantage of collecting a large amount of data in a short amount of time. On the other hand, I did not have any study notes or added insight into the participants' answers besides the proposals they offered, some basic demographic information, and the feedback they gave me on Crowdlicit.

7 Summary

This chapter reports on Crowdlicit, a system for conducting distributed elicitation and identification studies. I also introduced end-user identification studies, which are the reverse of elicitation studies, as studies that evaluate the identifiability of interface actions and how well they map to intended functions. My work demonstrated that it is possible to run elicitation studies online and get quality results. Using Crowdlicit cuts down on resources required to conduct elicitation studies, especially time, opening the door to expanding, replicating, or extending such studies, as well as increasing the quality of user-driven designs by conducting identification studies using the flexible Crowdlicit system. It is my hope that researchers, designers, and developers will use Crowdlicit to efficiently run crowd-powered end-user elicitation and identification studies, gaining quality data in little time.

Five | **Crowdsensus**

Crowdsourcing Similarity Judgments for Agreement Analysis in End-User Elicitation Studies

Analyzing design proposals collected in an elicitation study requires substantial resources. In this chapter, I present *Crowdsensus*, a crowd-powered tool that enables researchers to efficiently analyze the results of elicitation studies using subjective human judgment and automatic clustering algorithms. In addition to my own analysis, I asked six expert researchers with experience running and analyzing elicitation studies to analyze an end-user elicitation dataset of 10 functions for operating a web-browser, each with 43 voice commands elicited from end users for a total of 430 voice commands. I used Crowdsensus to gather similarity judgments of these same 430 commands from 410 online crowd workers. The crowd outperformed the experts by arriving at the same results for seven of eight functions and resolving a function where the experts failed to agree. Also, using Crowdsensus was about *four times faster* than using experts.

1 INTRODUCTION

End-user elicitation studies are laborious to run and analyze, especially grouping elicited proposals based on similarity. Although including a large and diverse group of end users in elicitation studies is desirable, this practice vastly increases the number of comparisons among proposals when analyzing the results of the study, making the workload immense. In this chapter, I demonstrate using a combination of online crowds and automatic clustering algorithms to determine the similarity of elicited proposals from elicitation studies, thereby eliminating the burden of manually comparing proposals.

To support efficient elicitation study analysis, I created *Crowdsensus*. Crowdsensus generates web interfaces that present crowd workers with interaction design proposals collected in an elicitation study and asks them to vote on their similarity. After collecting the votes, Crowdsensus employs automatic clustering algorithms to find agreement between proposals using the votes, thereby resolving the underlying set of proposals. I used the Crowdsensus system to explore research questions 4–6 that I outlined in the **Introduction** chapter of this dissertation:

- RQ4. How can the crowd facilitate similarity judgments for agreement analysis in end-user elicitation studies?
- RQ5. By using the crowd, what are the benefits, if any, in terms of cost and time compared to the status quo use of experts' judgments?
- RQ6. How does the quality of the results produced by the crowd compare to those produced by expert researchers?

In pursuing answers to these questions, I address challenges including the design of the interface the crowd workers should interact with, the number of action proposals that should be presented, the best phrasing of the instructions, how to detect spam answers, and which clustering algorithms are suited to this use case.

I deployed a study on the mTurk platform to validate my approach. In the study, 410 crowd workers were asked to find the similarity among 43 proposed text representations of voice commands for 10 functions of a voice-activated web browser, for a total of 430 voice command proposals. I found that a crowd of non-expert workers, in combination with automatic clustering algorithms, was able to select the same commands for seven out of the

eight prompts for which a group of experts' judgments converged. As for the remaining two prompts, the crowd successfully converged upon one command for which the experts' opinions diverged. Also, the crowd was about four times faster than the expert researchers.

This work contributes a method to crowdsource the analysis of end-user elicitation studies, cutting effort and time. I also contribute the Crowdsensus system itself (as part of **The CROWDDESIGN engine**), a tool for researchers to utilize online crowds to find agreement in complex datasets (or analyze data from elicitation studies themselves). Crowdsensus could be used equally well to find agreement among gestures, icon sketches, text commands, or voice commands—any situation where action inputs can be used to invoke commands.

2 Crowdsensus: A Similarity Judgement Tool

I developed Crowdsensus to capture and analyze proposal-similarity votes from online crowds. Crowdsensus generates custom web interfaces that facilitate the collection of similarity votes from online crowd workers. The interfaces are platform-agnostic and run on any device with a web browser.

2.1 **Importing Proposals**

After conducting an elicitation study, a Crowdsensus user uploads a set of proposals collected for a prompt. This set can be imported directly from **Crowdlicit**, or takes the form of a comma separated value file (CSV). The proposals are stored in the Crowdsensus database to be used in generating similarity-judgment webpages for the uploaded data.

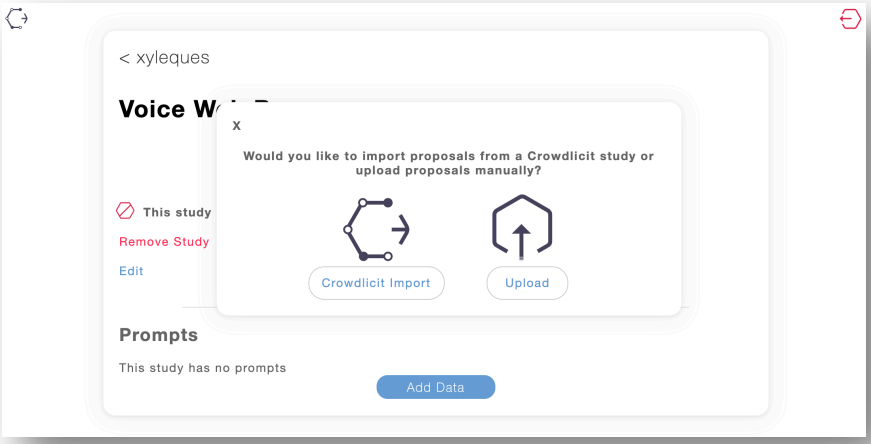


Figure 11 A screenshot of the Crowdsensus interface asking a user if they would like to import proposals from a Crowdlicit study or upload them manually.

2.2 **Interface Designs for Similarity Judgments**

The Crowdsensus interface (Figure 12.B) first shows an instruction page that includes a training video on how to complete the task of similarity voting. Once crowd workers are familiar with the task, they click "Start." The main screen shows a prompt asking them to select similar symbols. There is a "Help" button that brings up the instructions screen with the training video. At the bottom, a "Done" button takes the workers to the next task, and a progress bar indicates how far through the process they are. I designed three different approaches to present the proposals to the workers for comparison, described below.

2.2.1 The Direct Comparison Interface "(1:1)"

The first design (Figure 12.A) presented direct comparisons between two proposals, *i.e.,* a 1:1 comparison. With this approach, crowd workers see two proposals for a given prompt displayed side-by-side, with two buttons under them labeled "Yes" and "No."

2.2.2 The List Comparison Interface "(1:N)"

The second design was the comparison of a symbol to a list (Figure 12.B), where one proposal for a given prompt is being compared to a subset of the other proposals for that same prompt, *i.e.,* a 1:N comparison. This design

shows a prompt in a box in the middle of the screen and a checklist of other proposed prompts from which to select.

2.2.3 The Grouping Interface "(N:N)"

The final similarity judgment interface, the Grouping interface, presents all of the proposals for a given prompt inside draggable elements. Crowd workers compare all prompts to each other in a many-to-many approach, *i.e.,* an N:N comparison. On the right-hand side of the screen, there is a create a new group drop zone. The workers have to drag-and-drop elements onto the drop zone to create a new group (Figure 12.C). The workers repeat the operation, dragging-and-dropping proposals either onto established groups or to create new groups. On the right side of every group there is a count showing the number of proposals in that group. Clicking on a group will bring a pop-up to the front of the screen that displays all of the proposals belonging to that group.

2.2.4 Preference for the List Comparison Interface "(1:N)"

After thoroughly exploring the benefits and downsides of these three interfaces, I eliminated the Direct Comparison interface (1:1) due to the large volume of tasks required to get adequate votes for analysis. For each prompt there had to be N^2 1:1 comparisons, where N is the number of proposals for that prompt. For instance, in the study I analyze in the next section, I gathered 43 proposals for 10 prompts, which would need 18,490 comparisons to get a single similarity vote for every pair of proposals. Also, it was hard to create vote validation mechanisms for this interface to eliminate spam voting. On the flip side, I eliminated the Grouping interface (N:N) because pilot testing showed it was too complex and frustrating for the workers when working with a large dataset. I therefore decided to make the List Comparison interface (1:N) the default interface for Crowdsensus. By providing a list of options, the 1:N interface provides more context for the user than the 1:1 interface.

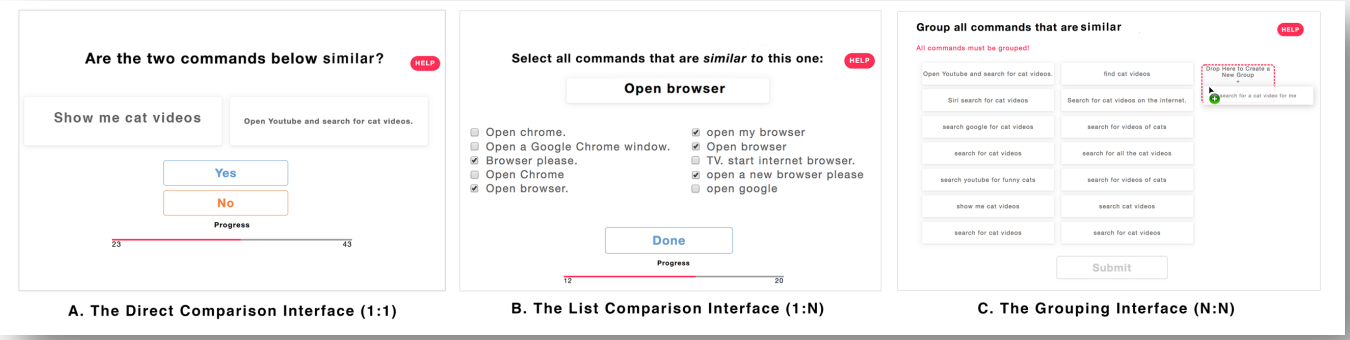


Figure 12 The three comparison interfaces in Crowdsensus. (A) The Direct Comparison interface. A prompt asks crowd workers to vote on the similarity of two commands. (B) The List Comparison interface. The worker selects from a list of proposals those she thinks are similar to the one highlighted in the middle of the screen. (C) A list of draggable proposals, with one dragged over the "create new group" drop zone.

2.3 **Instructions for Similarity Judgments**

At the top of the List Comparison interface there is a crucial prompt instructing the crowd workers to vote on the similarity of the presented proposal to those appearing in the list. I tested four different phrasings of this instruction. The instruction was phrased with the following variations: "Select all commands that are ('essentially the same as,' 'substantially similar to,' 'similar to,' 'kinda similar to') this one." From pilot testing, I found that the phrase "essentially the same" and "substantially similar" yielded very strict, inflexible similarity voting from the crowd. Conversely, the phrase "kinda similar" gave very loose similarity votes, resulting in an agreement score of 1.00, meaning that the crowd voted all proposals to be "kinda the same." I found that the best prompt to use was simply, "Select all commands that are similar to this one." Such neutral phrasing struck a nice balance between promoting strict and permissive comparisons.

2.4 **Number of Proposals Presented**

In the List Comparison interface, the number of proposals listed could conceivably vary from as few as two proposals up to N, the entire set of proposals for a given prompt. I tested out three different configurations of the number of proposals shown: 10 proposals, 20 proposals, and all 43 proposals that I used in the study presented in this chapter, below. In my pilot testing, I found that the number of proposals presented did not affect the crowd's performance or the results, making it possible to present large sets of data in smaller subsets that fit on a single screen.

2.5 **Vote Validation**

To eliminate spam votes, I devised two procedures for vote validation: time thresholds and the identical-proposal test.

2.5.1 Time Threshold

From pilot testing, I saw that the average time spent voting on a single List Comparison task with 43 proposals in the list was a little over 30 seconds. I decided to put a threshold of 15 seconds per task to accept valid data. Data coming from voters who did not spend at least 15 seconds looking at the proposals were disqualified and not recorded.

2.5.2 Identical Symbol Test

As stated, in the List Comparison interface, there is a list of proposals being compared to the primary proposal. That list includes the primary proposal itself. Any answers submitted where the primary proposal was not selected were discarded, since the failure to self-match indicates the worker was not paying adequate attention to the task.

3 Evaluating Crowdsensus

To answer my research questions around using online crowds to generate similarity judgments for agreement analysis in end-user elicitation studies, I created a set of prompts and proposals. I then analyzed this data with two methods: manual analysis by elicitation study experts, and crowd-based analysis using Crowdsensus.

3.1 **Data Collection**

I used **Crowdlicit** to run a distributed elicitation study. I based the prompts on Morris's "Web on the Wall" study [77]. I asked 43 crowd workers to type in text strings representing voice commands that they felt would be the most intuitive way to perform each one of my 10 prompts to interact with a voice-controlled web browser at a distance.

3.2 **Data Analysis**

3.2.1 Grouping of the Proposals by Elicitation Experts

I requested groupings of these proposals by a set of elicitation experts (the status quo method by which such data are analyzed) to compare the experts' output to that of the crowd via Crowdsensus. In addition, I analyzed the data as an expert would. For every one of the 10 prompts, the experts grouped the elicited proposals based on their similarity, and calculated the agreement score using Equation 1. For each prompt, the experts elected the group with the largest number of similar proposals to trigger the given function prompt.

My elicitation experts were considered experts because they had previously published research using elicitation studies. All experts had Ph.D. degrees and were external to my own university. Experts did all groupings before any output from Crowdsensus was available. Table 6 gives the experts' demographic information.

I sent each of the experts an Excel spreadsheet that included a separate tab for each of the 10 function prompts. Within each tab, 43 separate rows contained the text of the proposals for that prompt. I asked them to group the proposals for each prompt by entering a group number in a column adjacent to each proposal, and to record how long it took them to complete the task using provided timing software. As compensation for their time, each expert received a $50 Amazon gift card. I chose this compensation level based on the amount of time it took me to analyze the same data (approximately 30 minutes), estimating that faculty are compensated at approximately $100 / hour based on typical faculty salaries in the U.S.

3.2.2 Grouping of the Proposals by the Crowd

I recruited 410 workers from Amazon Mechanical Turk. The majority of them were between the age of 26 and 40. Sixty-three percent of the workers had bachelor's degrees, and 85% of them considered English to be their primary language. Table 6 provides the workers' demographic information. Workers on Mechanical Turk who accepted the human intelligence task (HIT) went to a webpage on the Crowdsensus server by following the link in the HIT itself. The page provided instructions, human subjects study approval details, and researchers' contact information. Before starting the task, the workers had to fill out a brief, pre-task survey that collected demographic information, as well as details about their familiarity with voice-operated technologies.

Demographic		Experts	Crowd
Gender	Male	6 (86%)	224 (55%)
	Female	1 (14%)	186 (45%)
	Non-binary	0	0
Age	18-25	0	155 (38%)
	26-40	6 (86%)	214 (52%)
	41-55	1 (14%)	28 (7%)
	56 and over	0	13 (3%)
Education level	Less than high school	0	3 (1%)
	Graduated high school	0	65 (16%)
	Technical school	0	14 (3%)
	Associate degree	0	38 (9%)
	Bachelor's degree	0	257 (63%)
	Advanced degree	7 (100%)	33 (8%)
Country	USA	6 (86%)	136 (33%)
	India	0	250 (61%)
	Other	1 (14%)	24 (6%)
English is primary language	Yes	4 (57%)	350 (85%)
	No	3 (53%)	60 (15%)
Frequency of voice command use	Never	1 (15%)	184 (45%)
	Daily	5 (70%)	88 (21%)
	Weekly	0	105 (26%)
	Monthly	1 (15%)	32 (8%)

Table 6 Demographic information for the academic experts, including myself, and the crowd workers who analyzed the web-browser-voice-command elicitation data set.

After completing the background survey, workers moved on to the page generated by the Crowdsensus tool. Pilot testing showed workers would be able to finish 20 runs of the List Comparison interface (i.e., 1:N for N = 43 symbols) within about 15 minutes. Therefore, I compensated the workers $2.75 per HIT. I based my pricing on the recommendation of Silberman et al.[107], equaling a rate of $11/hour (Washington state's minimum wage). After finishing the 20 voting tasks, the participants moved on to a page that thanked them for participating and provided them with a unique code to enter back into the HIT page on Mechanical Turk—similar to the **Thank You Page** in **Crowdlicit**. I used those unique codes to track which participants finished the HIT.

3.2.3 Grouping Algorithms

A crucial step after the crowd has provided votes indicating perceived similarity among proposals is to group these proposals. For a given prompt, all pairs of proposals have a level of similarity expressed by the number of "yes" votes given by the crowd in response to the instruction, "Select all commands that are similar to this one." The more "yes" votes, the more similar the crowd felt two proposals were. In grouping like proposals, the challenge is to determine how many distinct proposals emerged for a given prompt.

This challenge amounts to a graph clustering problem where, for a single prompt, the proposals are nodes in a fully-connected graph with weighted edges between all nodes being the ratio of "yes" similarity votes to total votes (Figure 13). The clustering problem is to form sets of nodes such that nodes within the same set have maximal similarity while nodes across different sets have minimal similarity. This problem is a version of the correlation clustering problem for general weighted graphs, which is of class APX-HARD [27].

I wrote[5] various optimization algorithms for this problem: hill climbing, shotgun hill climbing, simulated annealing, a genetic algorithm, and correlation clustering. I briefly describe the formulation of each algorithm and then present results of testing the algorithms on our crowd-supplied data.

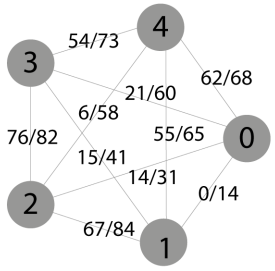


Figure 13 A small example of a fully-connected vote-weighted graph with five nodes. Weights are visible at the midpoints of the edges between nodes. The challenge of clustering is evident with nodes 2, 3, and 4, as nodes 2 & 3 and 3 & 4 have strong affinity, but 2 & 4 do not. So should {2,3,4} be grouped? The same problem exists for nodes 0, 1 and 4, with 0 & 4 and 1 & 4 having strong affinity, but 0 & 1 do not. Should {0,1,4} be grouped? The actual data in this study had 43 nodes per referent, not just 5.

Hill Climbing

I implemented steepest-ascent hill climbing as a baseline optimization algorithm. A solution state was represented as a set-of-sets, with each subset containing the nodes (proposals) deemed similar. Thus, if all proposals were deemed similar, the set-of-sets would contain one set of all nodes. If no proposals were deemed similar, the set-of-sets would contain a separate set each with only one node.

A fitness function for the hill climber's state was defined as follows. In the set-of-sets, for each pair of proposals *within* a set, if the "yes" votes (out of all votes) passed a one-sided binomial test,[6] I added the percentage of "yes" votes to the fitness score. If, on the other hand, the binomial test failed, I subtracted one minus the percentage of "yes" votes from the fitness score. Conversely, for each pair of proposals *across* different sets, I did the opposite:

[5] I implemented versions of these clustering algorithms from my advisor's software library.
[6] The one-sided binomial test examined whether there was a statistically significant proportion of "yes" votes out of all votes indicating similarity between two symbols at the $\alpha = .05$ level.

passing the binomial test subtracted the percentage of "yes" votes from the fitness score, while failing the binomial test added one minus the percentage of "yes" votes to the fitness score. Defined in this way, the fitness score drove the climber towards states having more similar proposals grouped together and less similar proposals grouped separately.

An essential component to any hill climber is formulating the possible moves to neighboring states. For a climber in a given state represented by a set-of-sets, neighboring states were created by taking each proposal and placing it in every other set, including the empty set. Thus, for a state with N sets containing a total of M items, $N \times M$ possible neighboring states were considered by the climber at each move. As a steepest-ascent hill climber, the move yielding the greatest fitness gain was always chosen. A concern with this approach, of course, is getting stuck on local maxima within the search space.

Shotgun Hill Climbing

I also implemented a variant of hill climbing called shotgun hill climbing, which is equivalent to random-restart hill climbing. In shotgun hill climbing, multiple climbers are "shot" randomly into the search space and climb from wherever they land, thereby increasing the chances of at least one climber reaching the global maximum. I initially utilized 1,000 climbers but surprisingly found that even five climbers performed just as well (and much faster), suggesting a search space with few local maxima. The shotgun hill climber therefore used five climbers.

Simulated Annealing

A more sophisticated iterative improvement algorithm than hill climbing is simulated annealing. In my implementation, a random move was chosen from among neighboring states. If that move was to a better state, it was always taken. If it was not, then unlike hill climbing, it still might have been taken depending on a probability that starts high and decreases over time, like the temperature of a cooling metal. Following Kirkpatrick *et al.* [52], I automatically set the initial and minimum temperatures via stochastic sampling of the search space. I set the cooling rate to 3% and enabled 100 possible moves at each temperature.

Genetic Algorithm

Genetic algorithms have been used to form clusters within graphs [137]. I implemented a genetic algorithm that began with 1,000 randomly generated "organisms," each encoding a set-of-sets solution state, and evolved them

for up to 10,000 generations, mutating offspring by moving a random proposal into a random set, including a new set. Offspring mutated on a decreasing schedule, with more mutations in early generations than in later generations. At each generation, 20% of the fittest organisms survived to populate the next generation. An exception rate of 5% was used to retain less-fit organisms for subsequent generations to avoid local maxima.

Correlation Clustering

Unlike general iterative improvement algorithms, correlation clustering is specific to the problem of clustering nodes in graphs that have "affinity," while avoiding clustering nodes that have "aversion." It was first formalized by Bansal *et al.* [11] for graphs with binary weighted edges represented as $\langle +, - \rangle$. Ailon *et al.* [5] introduced a correlation clustering approximation algorithm for general weighted graphs in which each edge had a positive weight w+ and a negative weight w–. However, my problem is different, as edges are not negatively weighted; all are positively weighted, some more than others. I implemented the KwikCluster approximation algorithm of Ailon *et al.* [6] (p. 23:13) but defined "affinity" as passing the one-sided binomial test described above, *i.e.,* node pairs that did not pass this test were deemed to have "aversion." As this algorithm is highly dependent on the selection of random nodes, I ran it with 1,000 random restarts, taking the best outcome.

3.2.4 Performance Results

To assess the performance of each algorithm in grouping my proposals, I performed a simple experiment. I first used trial-and-error to find the settings for each algorithm that seemed to perform best within reasonable time limits. I then ran the grouping algorithms on all 43 proposals for each of our 10 prompts. I did this 10 times and averaged the results. Thus, I was left with 5 algorithms × 10 prompts = 50 fitness scores and execution times (Table 7). Ultimately, I selected shotgun hill climbing for Crowdsensus because of its high fitness scores and fast execution times for the analysis in this chapter. However, Crowdsensus allows for the selection of the grouping algorithm and its parameters in the clustering web-based interface shown here in Figure 14.

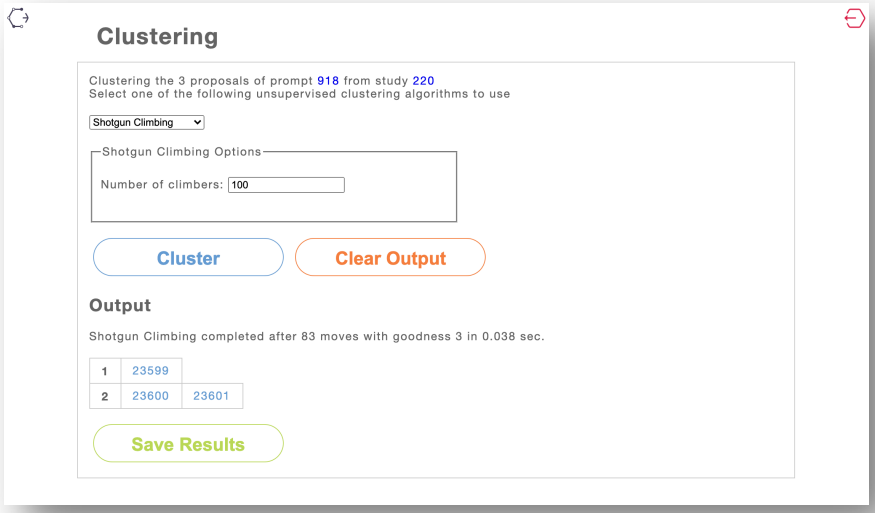


Figure 14 A screenshot of the Crowdsensus clustering interface. The page shows that the shotgun climbing algorithm clustered 3 proposals for a prompt.

3.2.5 Solution Fitness Scores

Statistical analysis shows that fitness scores conformed to the assumptions of analysis of variance. A repeated measures ANOVA using the Greenhouse-Geisser correction [39] indicated no significant effect of the algorithms on fitness scores ($F(1.2, 10.7) = 2.32$, n.s.).

3.2.6 Algorithm Execution Times

Algorithm execution times violated the normality assumption of ANOVA and were therefore analyzed with nonparametric tests. A Friedman test indicated a significant effect of algorithm on execution time ($\chi^2(4,N=50) = 37.60$, $p < .0001$). *Post hoc* Wilcoxon signed-rank tests corrected with Holm's sequential Bonferroni procedure for multiple comparisons [42] indicated that all execution times were significantly different under this analysis ($p < .02$), except the genetic algorithm and hill climbing.

On balance, given the parity in fitness scores and the desire for fast execution times, it seems everything but simulated annealing is a viable option. As noted, I chose Shotgun Hill Climbing for its peak fitness score and reasonable execution time. It was noteworthy that hill climbing performed so well, indicating that the search space

must be relatively smooth. Also, the fast speed of correlation clustering, even with 1,000 random restarts, was impressive.

Algorithm	Solution Fitness	Execution Time (sec)
Hill climbing	561.3 (141.8)	14.5 (6.1)
Shotgun hill climbing	562.6 (141.1)	100.3 (47.7)
Simulated annealing	562.6 (141.1)	282.0 (241.8)
Genetic algorithm	560.5 (141.1)	17.1 (1.2)
Correlation clustering	559.9 (144.3)	2.8 (1.8)

Table 7 Means (and standard deviations) of the quality of the solutions produced by the proposal-grouping algorithms, and how long it took to produce them, in seconds. Higher fitness scores indicate better performance. Lower execution times are preferred.

4 Results: Validating the Crowdsensus Approach

I used data from 410 crowd workers who provided votes that passed the validation tests out of 461 workers who began the study. The 51 workers whose data I did not use triggered the spam detectors. I used the shotgun hill climbing algorithm, described above, to cluster symbols based on the crowd's votes. I scrutinize Crowdsensus by:

- Calculating the agreement score for each prompt (see Equation 1). I wanted to understand how much the crowd and the experts each thought the users who elicited the proposals were in agreement.
- Quantitatively measuring the similarity of the proposal groupings produced by the crowd and by the experts.
- Finding the definitive proposal for each prompt. I wanted to see what proposal the crowd chose for each prompt, which ones the experts' judgments converged on, and what overlap there was between the experts and the crowd.
- Calculating the cost of using Crowdsensus to analyze the proposals, comparing this to the cost of using experts.
- Calculating the time it took the crowd to analyze the proposals, comparing this to the time for experts.

4.1 **Agreement Scores**

I calculated the agreement scores from the crowd and from the experts. Vatavu and Wobbrock [122] recommend using qualitative judgments for agreement scores and provided a guide for these judgements. According to their guide, low agreement scores are less than 0.1, medium scores are between 0.1 – 0.3, high scores are between

0.3 – 0.5, and very high agreement scores are above 0.5. Figure 15 shows how each of the experts, including myself, and the crowd (via Crowdsensus) grouped the proposals for each prompt. High or very high agreement scores indicated that whoever grouped the proposals thought the end users who provided these proposals exhibited some consensus. The experts tended to generate fewer groupings with larger numbers of proposals leading to high agreement scores, meaning the experts thought there was high similarity among the proposals for most prompts. The crowd (plus clustering algorithm, *i.e.,* Crowdsensus) had low agreement scores for five of the 10 prompts, meaning the crowd was stricter than experts in its assessment of which proposals were similar to each other. This strictness led to the crowd making small groups of very similar proposals.

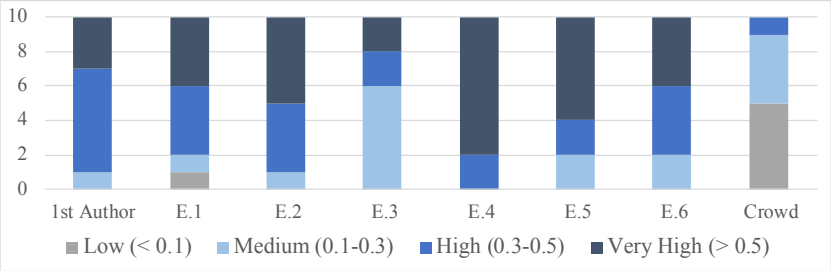


Figure 15 The amount of agreement among proposals elicited for each of the 10 prompts as grouped by myself, the experts (E.1-6), and the crowd via Crowdsensus.

4.2 **Similarity of Groupings**

To determine how similar the various groupings were, I devised a distance metric by which to compare groupings. For each prompt, I compared the groupings generated with Crowdsensus to the groupings generated from the experts' judgments by converting each grouping into a 2-D matrix of 43 rows and 43 columns, each row or column representing one proposal. Each cell was populated with one of three values: (-1, 0, +1). A -1 was for unused cells like the intersection of a proposal with itself. A 0 meant the proposals intersecting at this cell belonged to different groups. A +1 meant the two proposals intersecting at this cell belonged to the same group.

An example will help illustrate. Consider the grouping {(0,2), (1)}, where proposals 0 and 2 belong to the same group and proposal1 is in a group by itself. I would represent this grouping as shown in Figure 16.A. Another grouping of these same proposals could be {(0), (1,2)}, where proposal 0 is in a group by itself, and proposal 1

and 2 are in a group. This second grouping is represented in Figure 16.B. To measure the distance between the two groupings, I can use Equation 2 [45], below:

$$d_2(A,B) = \sqrt{\sum_{i=1}^{n-1}\sum_{j=0}^{i-1}(a_{ij} - b_{ij})^2}$$

Equation 2

In Equation 2, the 2-D distance d2 between 0-based indexed matrices A and B is given by taking the root of the sum of squared differences between all cells a_{ij} and b_{ij} "above the diagonal." A cell a_{ij} is a cell in n × n matrix A and indicates whether items i and j are in a set together (+1) or not (0). The outer summation causes index i to start in column 1, and the inner summation causes index j to start at row 0. Thus, rows are iterated within columns for cells for which j < i.

	A				B		
j\i	0	1	2	j\i	0	1	2
0	-1	0	+1	0	-1	0	0
1	-1	-1	0	1	-1	-1	+1
2	-1	-1	-1	2	-1	-1	-1

Figure 16 (A) Matrix A represents the grouping {(0,2), (1)}. (B) Matrix B represents the grouping {(0), (1,2)}. This type of matrix representation allows me to compute the distance between two sets-of-sets. Note that the matrices have 0-based indices and assume that their items are indexed likewise from zero.

Using Equation 2, the distance d2 between matrices (A, B) in Figure 16 is 0.67, which makes intuitive sense because 1/3 of the elements in A must be moved to create B (*i.e.,* move the 2). In general, then, the distance d2 between two matrices is a value ranging from 0, if they are identical, to 1, if the two matrices are entirely different.

Table 8 shows the mean pairwise distances (and standard deviations) among Crowdsensus, myself, and each one of the experts averaged over 10 prompts. The crowd's groupings via Crowdsensus are similar to the experts', with E.3 being the closest on average to the crowd's groupings at 0.27 and E.4 the furthest at 0.59. The average distance for all the experts among themselves was 0.29.

4.3 **Definitive Proposal Selection**

The third measure I used to judge the crowd's performance relative to the experts was to find the definitive proposal generated by the crowd workers, *i.e.,* the largest group of similar symbols, and compare that to the

definitive symbol that emerged from the experts for each prompt. I used an "experts' group" to achieve my goal. The experts' group for each prompt was made up of proposals that appeared across all of the largest proposal groups per prompt generated by the first author and the experts.

Figure 17 shows the number of proposals in the experts' group for every prompt, the number of proposals the crowd elected, and the number of overlapping proposals between the two. For prompts 2, 3, and 10, all of the proposals selected by the crowd appeared in the experts' group, with the crowd's proposals making up 73%, 42%, and 56% of the experts' group, respectively. For prompts 1, 5, 8, and 9, the crowd's proposals made up 81%, 100%, 30%, and 59% of the experts' group, respectively.

Prompt 4 was a case where there was no overlap in proposals between the crowd and the experts, *i.e.,* Crowdsensus converged on a different set of proposals than the experts' merged judgements to invoke "switch between pages in a backward direction." For prompts 6 and 7, I was unable to find a single proposal that all experts had in their individual definitive proposal groups, which resulted in empty experts' groups for these two prompts. The crowd's votes generated a definitive proposal for prompt 6, but the crowd's votes did not cluster any two proposals into a group to be used as the definitive proposal for prompt 7. Table 9 presents the experts' and crowd's sets of definitive proposals (*i.e.,* elicited voice commands).

	Crowd	1st Author	Expert 1	Expert 2	Expert 3	Expert 4	Expert 5	Expert 6
Crowd	0	.39 (.14)	.40 (.23)	.41 (.17)	.27 (.18)	.59 (.14)	.41 (.21)	.39 (.19)
1st Author		0	.39 (.14)	.36 (.10)	.20 (.15)	.40 (.12)	.39 (.10)	.37 (.10)
Expert 1			0	.20 (.10)	.36 (.14)	.27 (.17)	.18 (.11)	.14 (.11)
Expert 2				0	.37 (.90)	.29 (.16)	.19 (.07)	.15 (.08)
Expert 3					0	.48 (.13)	.36 (.15)	.36 (.12)
Expert 4						0	.24 (.15)	.26 (.18)
Expert 5							0	.18 (.10)
Expert 6								0

Table 8 The mean distances over 10 referents between Crowdsensus, the first author, and the experts. The distance values are in [0.0, 1.0], where 0.0 means the two groupings are identical, and 1.0 means they are entirely different. Standard deviations are in parentheses.

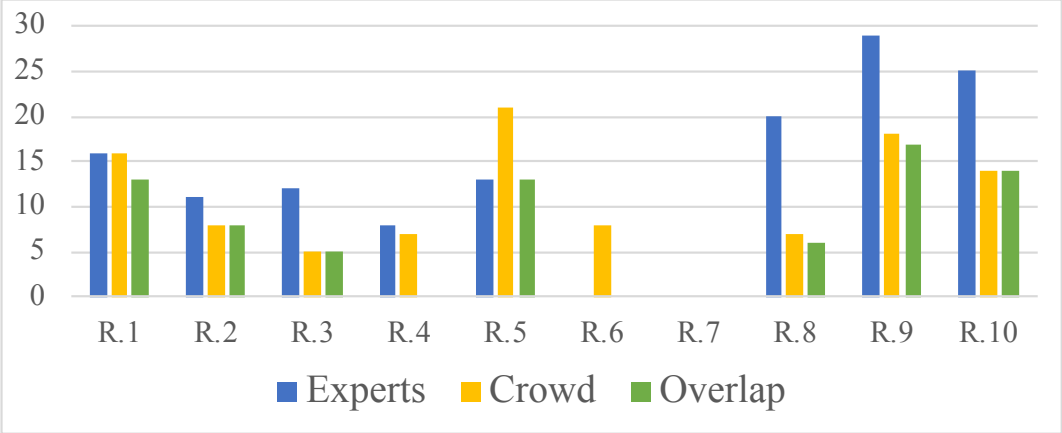


Figure 17. Comparison of proposal counts in the definitive group for each prompt (R1-R10). Blue bars are the number of proposals in the experts' group. Yellow bars are for the crowd. Green bars are the number of overlapping proposals in the experts' and crowd's groups.

Prompt **How would you phrase a voice command that would…**	**Experts' proposal**	**Crowd's proposal**
1. open the browser?	"Open browser"	"Open browser"
2. perform a search for cat videos?	"Search for cat videos"	"Search for cat videos"
3. click the first link on the page?	"Click the first link"	"Click the first link"
4. switch between pages in a backward direction?	"Go back 'one / to the last / to the previous' page"	"Go to the previous page"
5. switch between pages in a forward direction?	"Go to the next page"	"Go to the next page"
6. switch between multiple open tabs?	N/A	"Switch tabs"
7. select a region on the page?	N/A	N/A
8. request the same page you are browsing to be loaded again?	"Reload the page"	"Reload the page"
9. bookmark a page?	"Bookmark this page"	"Bookmark this page"
10. close the browser?	"Close browser"	"Close browser"

Table 9. A list of all function prompts used in this study, and the proposals that would invoke them, as chosen by experts and non-expert crowd workers. An "N/A" indicates a lack of convergence.

4.4 **Cost**

I paid each researcher $50 to analyze my dataset of 10 prompts. Therefore, the cost of one prompt to be analyzed by one expert was $5.00. For this study, with its 43 proposals per prompt, I needed 2.15 Mechanical Turk HITs to complete one prompt because one HIT contained 20 List Comparison voting tasks, and I needed 43 such tasks to compare all of the proposals to each other. My Turk tasks were priced at $2.75. Thus, the cost for a crowd worker to analyze a prompt was $5.90. In summary, then, the cost of Crowdsensus was similar to the cost of the experts.

4.5 **Time**

On average, it took the experts, including myself, 3.17 (*SD*=1.2) minutes to group the proposals for a single prompt. For the crowd workers, one HIT consisted of 20 List Comparison tasks. HITs were batched into groups of 31. Therefore, a batch resulted in 31 × 20 = 620 List Comparison tasks. To analyze a single prompt, I needed 43 List Comparison tasks. In 11.22 minutes, the crowd analyzed 14.4 prompts, meaning, on average, this took about 0.78 minutes per prompt. This result was about four times faster than the experts (*i.e.,* 3.17 / 0.78 = 4.06).

5 Discussion

To understand my findings in context, I revisit the research questions I posed in the introduction.

5.1 **RQ4. How can the crowd facilitate similarity judgments for agreement analysis in end-user elicitation studies?**

I was able to successfully utilize Crowdsensus to produce agreement analysis for my study. Crowdsensus takes a dataset of end-user-generated proposals and creates custom web interfaces to deploy online. The web interfaces Crowdsensus generates facilitate the gathering of similarity judgment votes. Clustering algorithms group proposals during the agreement analysis phase of elicitation studies.

I pilot tested potential interfaces for Crowdsensus, examining the possible effects of simple user interface choices—displaying 1:1, 1:N, or N:N simultaneous comparisons; the specific phrasing of instructions; and the number of proposals displayed within a single HIT. Iterative comparative testing and refinement of the interface helped me identify a set of interface parameters that produce high-quality crowd judgements, particularly the use of a 1:N selection mechanism and the crucial phrase "similar to" in the instructions. I discovered that it is safe to present large proposal sets as smaller, manageable subsets.

From my study, I discovered that the average time a crowd worker spent voting on the similarity of one proposal compared to 43 others was 38.23 seconds. I recommend using 15 seconds as a threshold to accept valid data. I also used an approach to ensure the crowd workers were paying attention by testing whether they voted-as-similar the exact proposal to which they were comparing all other proposals.

5.2 **RQ5. By using the crowd, what are the benefits, if any, in terms of cost and time, compared to the status quo use of experts' judgments?**

In this study, the crowd cost $5.90 per prompt, while the experts cost $5.00. Therefore, the crowd cost was similar to the experts. Specific monetary costs may vary in practice depending on the exact wages of experts and crowd workers and on the desired level of redundancy in crowd judgments.

The average time it took the crowd to analyze a single prompt was 0.78 minutes. The average time it took the expert researchers we recruited to examine our data was 3.17 minutes per prompt. Hence, the crowd can provide similarity judgments to analyze end-user elicitation studies using our List Comparison interface about four times faster than expert researchers.

Having the crowd analyze the results of elicitation studies, a very laborious task, cuts time significantly. The researcher is also free to conduct further analysis and tweak the crowd's groupings, capitalizing on her own expertise and whatever supplementary material (*e.g.,* study notes) she has available. By cutting down on resources needed to conduct elicitation studies, I open up numerous possibilities to advance this methodology, like scaling up the number of proposals collected to create more inclusively-designed technologies—especially in distributed elicitation studies using **Crowdlicit**, and lowering barriers to conduct replication studies or reanalyze published studies.

5.3 **RQ6. How does the quality of the results produced by the crowd compare to those produced by expert researchers?**

Similarity judgments for proposals generated in elicitation studies are subjective. I expected to find differences in the way people grouped proposals. I therefore discuss the quality of the crowd's results compared to those of the experts along three lines: agreement scores, grouping similarity, and the definitive proposal for each prompt.

5.3.1 Agreement Scores.

Overall, the crowd's grouping-agreement scores were lower than those of the experts. The experts seemed to believe that the proposals were more similar to each other than did the crowd. The crowd was stricter than the

experts in considering which proposals belonged with each other in the same group, thereby creating smaller sets of more-similar proposals.

5.3.2 Grouping Similarity

The average distances between the crowd's grouping and each one of the experts' was similar to those of the experts among themselves for every prompt. This finding means that the crowd produced groupings comparable to those produced by the experts.

5.3.3 Definitive Proposal Selection

Due to the lower agreement scores mentioned above, the crowd's groupings were of smaller similar sets of proposals. Thus, a definitive proposal chosen by the crowd contained less variance than those from the experts. For instance, for prompt 10, both the crowd and experts chose "close browser;" however, the experts' grouping included variants such as "please close this browser for me" as being an equivalent, while the crowds' grouping required a stricter match. This situation is where the experience of the experts and their familiarity with elicitation studies gives them an advantage over the crowd. Experts conducting elicitation studies might choose to create synonyms for a given prompt, such as in the case where there are two or more popular groups of proposals proposed for a given prompt. In our analysis, we made the simplifying assumption that a designer would choose a single proposal for each prompt. However, the output of Crowdsensus could easily be analyzed by designers to form proposal synonyms.

5.4 **Limitations**

The proposals I elicited were text strings of voice commands, a decision I made to simplify the process of capturing proposals and being able to share them easily with external experts. I suspect that richer proposals like audio clips or gesture videos will require more time to analyze. Also, the experts I recruited worked independently. Typically, in an elicitation study, if more than one researcher is analyzing the results, they work together to reach consensus on a definitive proposal. Also, my dataset of prompts and proposals was relatively small, at 10 prompts and 430 proposals. Some studies are larger, such as Wobbrock *et al.'s* [134], which had 27 prompts and 1,080 proposals, or Kane *et al.'s* [49], which had 22 prompts and 880 proposals. The scope of my dataset was limited to the standard functionality of a web browser, which made the proposals elicited easier to analyze than novel

interactions for an emerging technology; it is unknown whether expertise beyond that of crowd workers is required in more complex domains.

6 Summary

In this work, I developed *Crowdsensus*, a system that extends the popular design method of end-user elicitation studies by allowing researchers to crowdsource the crucial similarity judgments so central to agreement analysis. My work demonstrated that it is possible to use a crowd of non-experts, in conjunction with automatic clustering algorithms, to successfully analyze the results of an elicitation study. I also showed that using the crowd comes with benefits like saving time; it also produces results similar to those obtained with experts. It is my hope that using the crowd can propel end-user elicitation further by lowering barriers to running these studies at scale and with diverse audiences.

Six | The CROWDDESIGN engine

The CROWDDESIGN engine (CDE) is a website that serves two functions: (1) a DXD educational resource. (2) A software service that provides access to the **Crowdlicit** and **Crowdsensus** tools. The engine is available at the URL: **CROWDDESIGNengine.com**. The landing page is shown below in Figure 18.

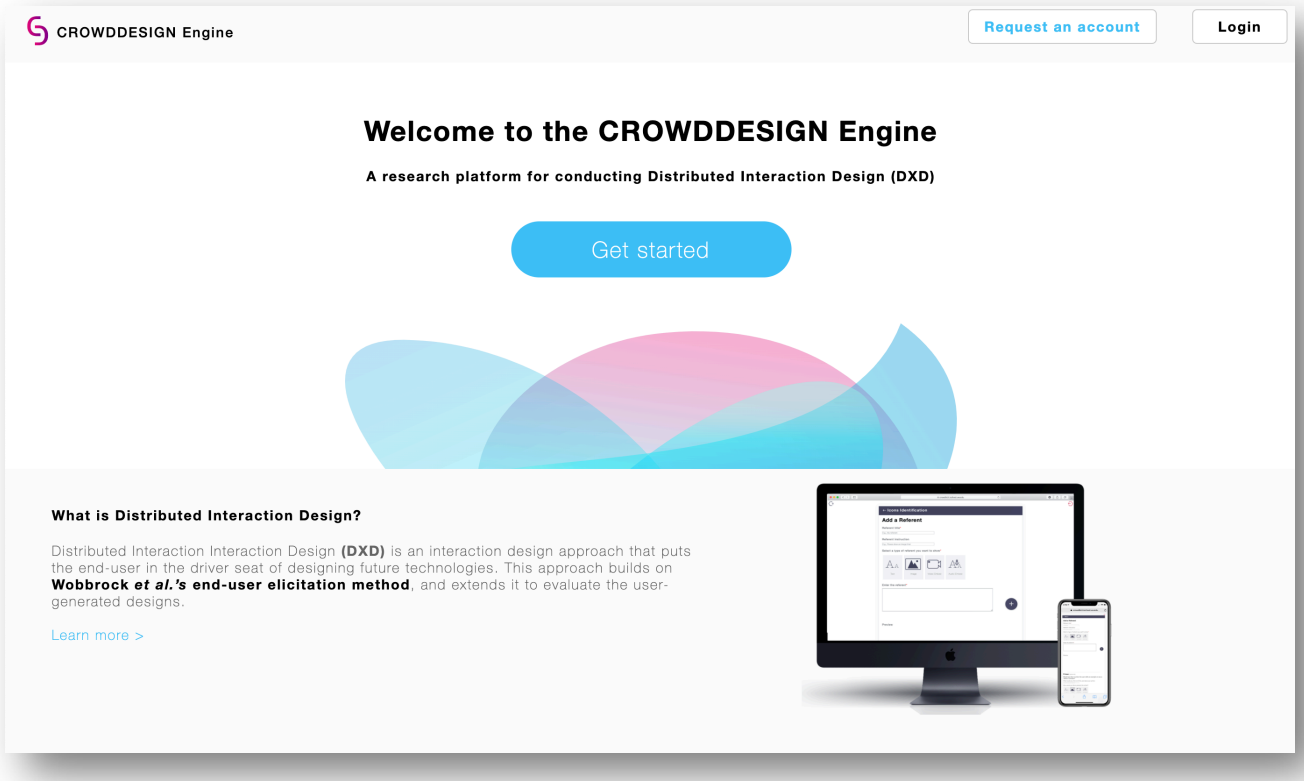


Figure 18 The CROWDDESIGN engine homepage.

1 Education

In this section, I explain the educational benefits of the CDE website.

1.1 **What is Distributed Interaction Design?**

This section of the CDE provides information about the DXD process. An online version of the **Distributed Interaction Design (DXD)** chapter from this dissertation.

1.2 **The Science**

This section gives background information about the methods behind DXD, starting with Good *et al.'s* [37] paper—the earliest example of user-driven interaction design. The section also shows the versatility of the approach displaying examples of a wide range of applications.

1.3 **What is the CROWDDESIGN engine?**

This section provides an overview of the tools (**Crowdlicit**, **Crowdsensus**). It also includes instructional videos on how to use the tools—shown in Figure 19.

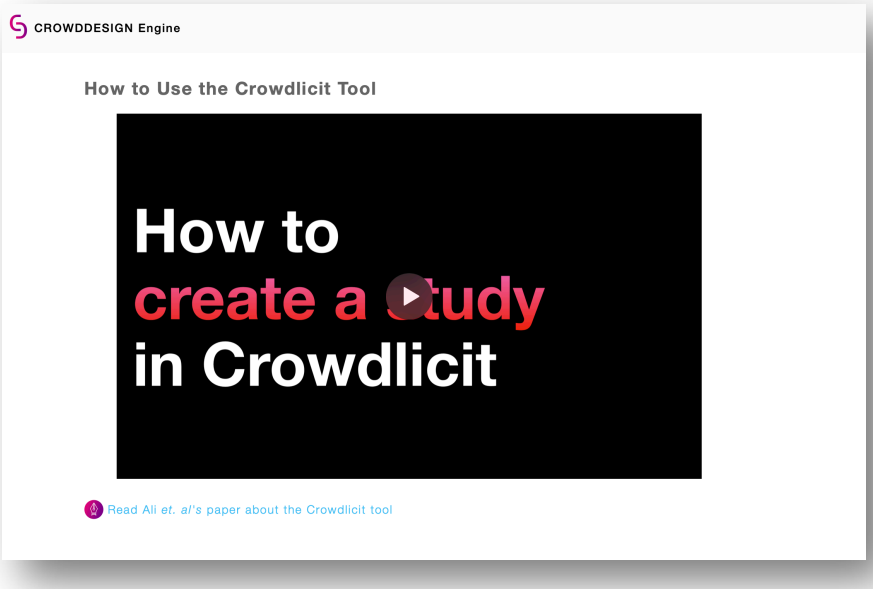


Figure 19 A screenshot from the CDE showing an instructional video of how to use the Crowdlicit tool.

1.4 **Features at a glance**

The landing page of the CDE shows a list of features (Figure 20) the CDE offers, like the flexibility of an online platform that does not require any installation, the ability to use mobile devices to participate in studies, and the use of machine learning to analyze study results efficiently.

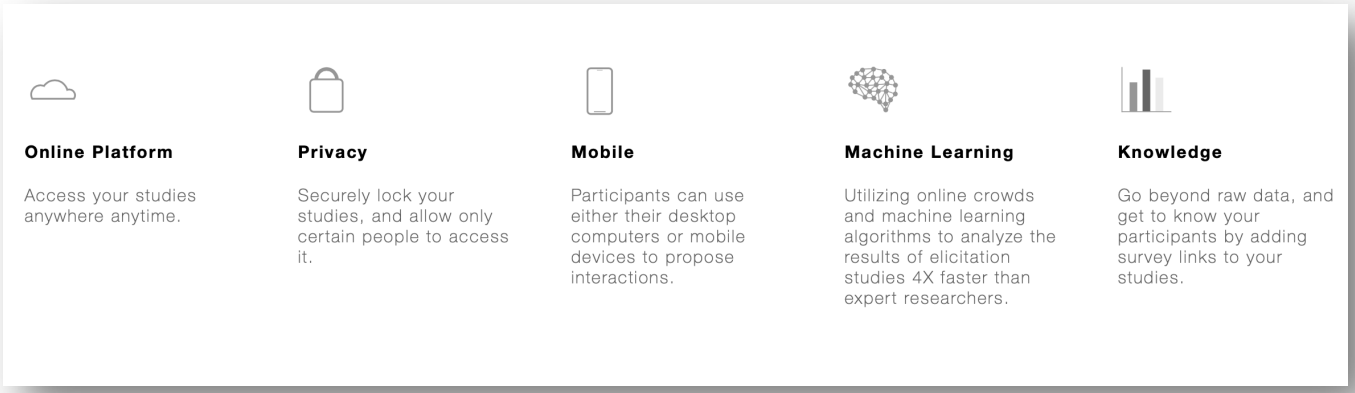


Figure 20 A list of the CDE features at a glance.

1.5 **Contact and Team**

Of course, the CDE was not a solo effort. I dedicate a page in the platform acknowledging collaborators and funding sources that made the platform possible.

1.6 **Research Studies**

This section currently lists a number of the most-influential papers in the DXD research area. This page will be continuously updated with noteworthy publications.

2 Tool Access

In this section, I explain the current state of public access to the CDE and my future plans for it.

2.1 **Current Access**

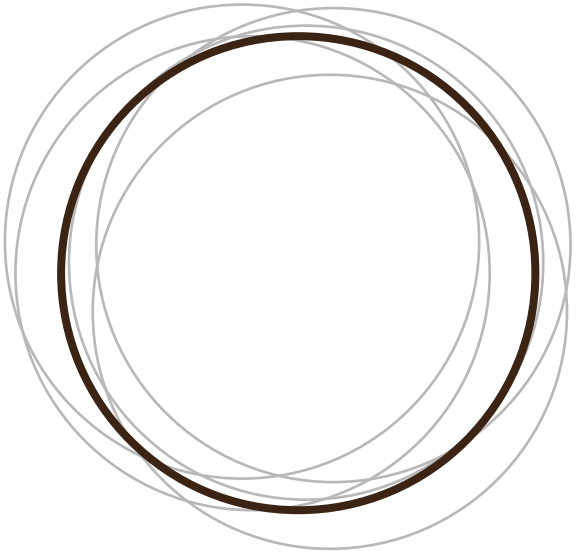
At the time of writing this dissertation, the CDE is hosted on a small shared server belonging to the Information School at the University of Washington. This current setup has limited computing resources. Due to these limitations, the engine is behind a registration request wall. Any potential user of the engine needs to fill out a registration form making the case for their need to use the system. I personally review the requests and grant access to the tools. I have observed the CDE engine's use in person during a guest lecture I gave in the Research Method's class at The Information School in the University of Washington. During my lecture, I taught the DXD process and introduced the CDE engine. Within 40 minutes of learning the process, 90 undergraduate students in that class, working in groups, signed into the platform, created 30 different studies designing interactive systems ranging from a rice cooker, to a voice assistant, to a smart toilet. After a peer-participation task, the 30 research groups generated more than 160 unique interaction proposals.

2.2 **Future Plans**

Starting in 2021, I will be hosting the CDE on my own server. The structure of service will change. It will no longer be a complete service storing its users' data. The new setup will ask users to log into the system using credentials for a cloud storage service (*e.g.,* Google's Cloud drive, Microsoft's OneDrive, etc.). This setup will allow users to store any data they upload and collect on their own private cloud storage. In doing so, (1) the CDE will not bear the costs of storing large multimedia files, and (2) private sensitive data collected from participants will not be stored on the CDE server.

Moving forward, the CDE will be an educational tool and a structured software that generates studies to facilitate all steps of the DXD process.

DXD Applications



Seven | **Anachronism by Design**

Understanding Young Adults' Perceptions of Computer Iconography

Computer iconography in desktop operating systems and applications has evolved in style but, in many cases, not in substance for decades. Many of today's young adult computer users grew up without direct physical experience of many objects—like floppy diskettes—represented by legacy icons. In this chapter, I describe a multi-part study conducted to gain an understanding of young adults' perceptions of computer iconography, and to update that iconography based on young adults' current mental models. To pursue this work, I gathered a set of icons (for 39 functions) found on common desktop operating systems and applications. I recruited 30 young adults for my study. In the first part, lab-based elicitation study, which did not use Crowdlicit, I asked the participants to propose sketches of icons they deemed most appropriate to trigger the 39 functions. I elicited a total of 3,590 individual icon sketches and grouped these into a set of 39 participant-generated icons. In the second part, a lab-based identification study, which also did not use Crowdlicit, I showed participants the current icons and asked them to: name the computing functions triggered when those icons were clicked, identify the real-world objects those icons represented, and answer questions regarding their personal experiences with those objects. Finally, I conducted another identification study, this time using my **Crowdlicit** tool within **The CROWDDESIGN engine**, with 60 new participants from mTurk on the set of 39 participant-generated icons I obtained from the first part of my study to see how recognizable the young adults' icons were. Generally, my study results highlight 20 anachronistic icons currently found on desktop operating systems in need of a redesign. My results also show that with increased icon production during the elicitation process, the chances for anachronism significantly decrease, supporting the "production principle" [79] in elicitation studies. Furthermore, my results include an updated set of icons derived from my young adult participants. This work contributes an

approach to using end-user elicitation studies to understand users, user interface design, and specifically, icon design. It is also, to the best of my knowledge, the first-time end-user elicitation has been used to generate and evaluate iconography.

Figure 21 A tweet with an image of a physical floppy diskette. The tweet reads, "In the 'I'm getting old' department, a kid saw this and said, 'oh, you 3D-printed the 'Save' Icon.'"

1 INTRODUCTION

The graphical representation of computing functions as icons has been prevalent since their inception in David Canfield Smith's *Pygmalion* [108] and then their commercial adoption in the personal computer and its graphical user interface in the Xerox Star [47]. The Star used everyday objects familiar to office desktops at the time as icons to trigger machine functions. Despite the technological strides made since the advent of these early icons, many of the same icons persist even in today's desktop computer systems. Some of these icons are even graphical representations of objects that are no longer used in most people's daily lives. Examples of such icons are the 3.5″ floppy diskette—representing the "Save" function—and a compact disk—representing the Windows "Program Manager." Today's young adult technology users have never interacted with many such objects, which could complicate guessability and learnability, and raises interesting questions about such users' mental models. An example of the disconnect between the anachronistic objects represented by some of today's interface icons and young adults' perceptions of the objects themselves can be seen in Figure 21, which is a tweet from a person who, holding a floppy diskette, was told by a youth that he had "3D-printed the 'Save' icon." Rather than the physical diskette informing the meaning of the computer icon, the computer icon had informed the meaning of the physical object: it was merely a plastic model of the icon. The computer icon was now the prevalent object in the world, and the physical diskette was now subject to *its* meaning.

Along with text, icons are of central importance for guessing and learning functions in any graphical user interface. Yet surprisingly, such a fundamental aspect of user interfaces has rarely been reconsidered. Stylistic updates are common, but a certain "stickiness" pervades the conceptual underpinnings behind icon design. Presumably this stickiness is so that existing users, upon receiving a software update or installing a new application, can leverage their pre-existing knowledge of icons' meanings. At the same time, newer generations of computer users are encountering more and more "objects" as icons that they have never encountered in the physical world, like the anachronistic floppy diskette.

To understand young adults' perceptions of the objects represented by the icons in current desktop operating systems and applications, I assembled a set of 39 icons found on Windows 10 and Mac OS X that feature plausibly anachronistic objects. Based on the set of 39 icons and the functions they are

associated with, I sought to answer the following research questions that I outlined in the **Introduction** of this dissertation:

- RQ.10 How does production influence user-elicited interaction designs?
- RQ.12 What icons would young adults propose to trigger computer functions currently associated with anachronistic icons?
- RQ.13 How familiar are young adults with the objects represented in anachronistic icons?
- RQ.14 How identifiable is a set of icons elicited from young adults?

To answer these questions, I conducted a multi-part study. First, I recruited 30 young adult technology users ages 18–22 (people born in 1994 or after) to participate in a two-part study. The first part was an icon elicitation study. In this study, I presented my participants with 39 descriptions of computing functions that currently have an icon representing a *plausibly* anachronistic object. I asked my participants to sketch (and describe) icons that would trigger these functions. I elicited a total of 3,590 icons from the 30 participants, or an average of about 3 icon proposals per function from each participant. I then clustered similar icons to arrive at a set of 39 participant-generated icons.

In the second part of the multi-part study, I conducted an identification study with the same 30 participants based on my end-user identification method [9]. In this part, I presented the participants with cards (Figure 22), each showing one plausibly anachronistic icon, and asked them to identify the computing function that the icon would trigger when clicked. I also asked the participants how familiar they were with the represented real-world objects themselves. Finally, to assess the set of participant-generated icons that I derived from the elicitation study, I conducted a second identification study with 60 participants recruited online from mTurk using **Crowdlicit**.

Figure 22 A deck of 38 custom-made cards for the icon identification study. Each card had a number and a plausibly anachronistic icon.

I conducted open-coding [112] analysis on all 3,590 icons that I collected in the elicitation study and formulated a taxonomy of computing iconography. I found that almost half of the 3,590 participant-generated icons were of new concepts, while the other half were drawings of existing icons. When assembling the participant-generated set of icons from the elicitation study, I found that 17 out the 39 icons remained anachronistic, and 22 icons were of new concepts. I also found that 73% of all elicited icons were representations of physical objects. The remaining 27% were made up mostly of text and abstract shapes. I also found that the propensity for elicited icons to be anachronistic decreased significantly from the first elicited icon to subsequent icons, confirming that the "production principle" does, indeed, increase the novelty of elicited proposals, at least in this context.

In the laboratory-based identification study, I found that there were only 16 plausibly anachronistic physical objects that all 30 participants had used. None of the remaining 22 physical objects were used by all young adult participants. Furthermore, the identification study with young adult participants resulted in the correct identification of the functions triggered by 31 of the 39 plausibly anachronistic icons. Of the eight icons that had their functions identified incorrectly by consensus from the young adult participants, five of them had *new concept* icons from the elicitation study. These five new concept icons had improved identifiability as they were identified correctly by the second identification study—an online study with 60 participants from mTurk. In this second identification study, the 60 online participants were able to correctly identify 34 of the 39 user-generated icons from the elicitation study. Three of the five incorrectly identified icons were *new concept icons*, and two were anachronistic ones.

Finally, as a result of all three parts of this multi-part study (in-lab elicitation and identification, and an online identification study), I derived a set of 39 participant-generated icons that included three types: *Anachronistic by Elicitation*, *Anachronistic by Identification*, and *New Concept Icons*. This set of icons contained 20 *New Concept Icons* and 19 anachronistic ones–15 *Anachronistic by Elicitation* and four *Anachronistic by Identification.* In addition, while I was analyzing the results of the elicitation study, I found that the Production principle to reduce legacy bias as put forth by Morris *et al.* [79] indeed has an effect on the lowering legacy inspired interactions–in this case, anachronistic icons. This production principle's effect on user-generated interactions answers my tenth research question outlined in the chapter **Introduction** of this dissertation:

- RQ.10 How does production influence user-elicited interaction designs?

This chapter contributes empirical results of an elicitation study, two identification studies, a taxonomy of computer iconography, a set of participant-generated icons based on the results of the three studies, and an investigation into the effects of Morris *et al.'s* [79] production principle to reduce legacy bias in elicitation studies. Generally, this work can be of use to researchers understanding young adult users, user interface designers, and specifically, iconographers looking to design the next generation of icons for tomorrow's graphical user interfaces.

2 ANACHRONISTIC ICONS

I scrutinized current desktop operating systems–both Windows 10 and Mac OS X–and assembled a set of 39 plausibly anachronistic icons that represent 38 real-world physical objects[7] no longer as widely used as they once were. (I call these icons "*plausibly* anachronistic," because for some young adult users, the physical objects might still be conceivably familiar.)

Figure 23 displays 39 icons and their associated functions. A prevalent example of a plausibly anachronistic icon is the 3.5″ floppy diskette (icon #20), the physical version of which is rarely still used except on antiquated computers, but whose semblance still pervades user interfaces as the *'save'* icon. Other possibly outmoded real-world objects for young adults might include the magnifying glass (#2, #3),

[7] Functions #2. search and #3. zoom shared the magnifying glass object.

"snail mail" (#6, #26), print photographs (#8), printed books (#13, #14), paper calendars (#17, #18), analog clocks and watches (#17, #24, #38), compact discs (#31), filament light bulbs (#34), and analog magnetic compasses (#39), among others. Table 10 displays the current and original origins of each icon.

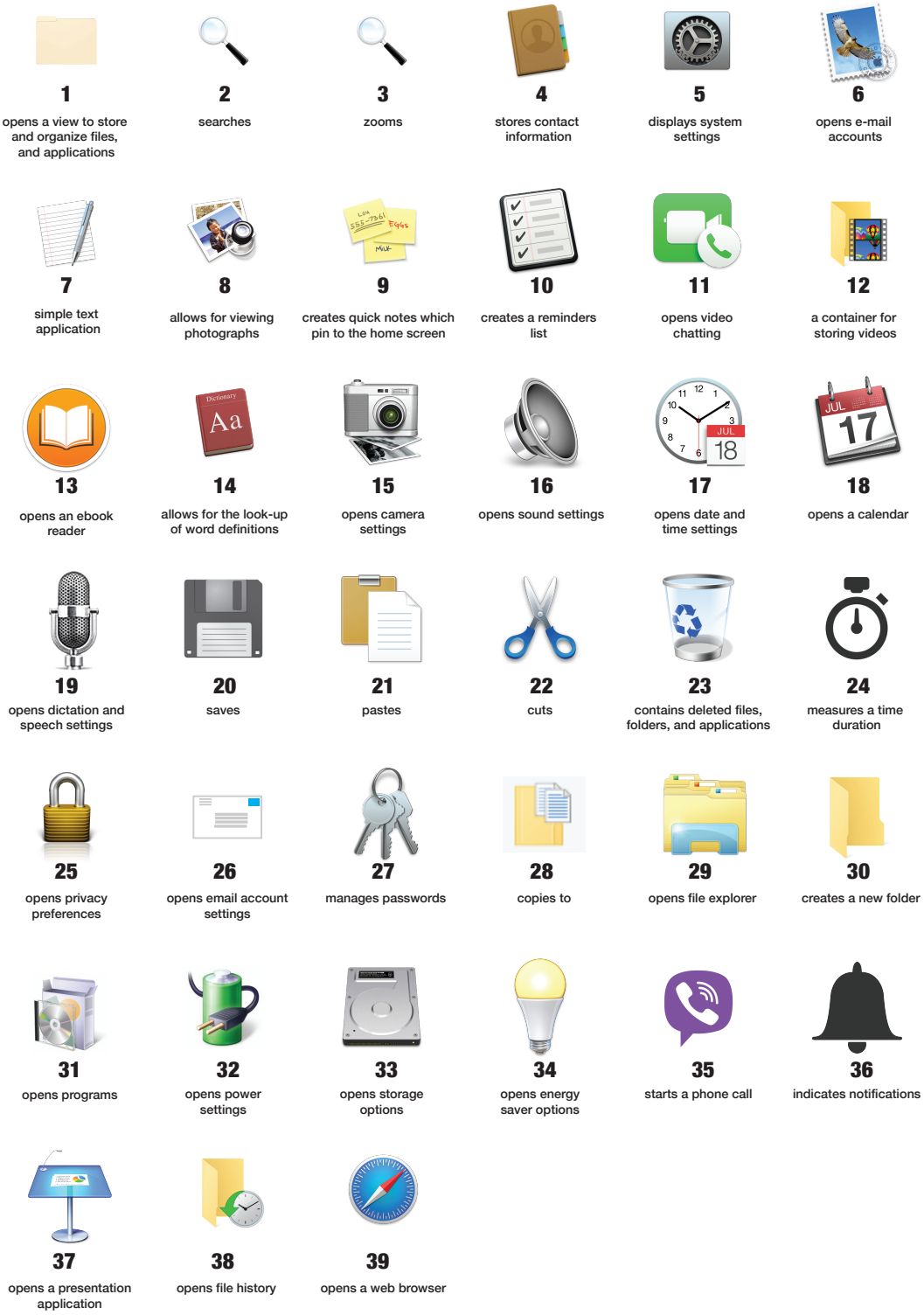


Figure 23 The 39 plausibly anachronistic icons and the functions they trigger.

#	Function ("An icon that …")	Source	Origin, Year first appeared in GUIs
1	...opens a view to store and organize documents, files, and applications	Windows 10	Xerox Star, 1981
2	...searches	Windows 7	Windows 95, 1995
3	…zooms	Windows 7	Windows 95, 1995
4	...stores contact information	Mac OS X 10.11.6	Windows 1, 1985
5	...displays system settings	Windows 10	Windows 95, 1995
6	...opens e-mail accounts	Mac OS X 10.11.6	Mac OS X, 2001
7	...is a simple text application	Mac OS X 10.11.6	Xerox Star, 1981
8	...allows for the viewing of photographs	Mac OS X 10.11.6	NeXTSTEP, 1989
9	...creates quick notes which pin to the home screen	Mac OS X 10.11.6	Mac OS 7, 2013
10	...creates a reminders list	Mac OS X 10.11.6	OS X 10.8 "Mountain Lion", 2012
11	...opens video chatting	Mac OS X	Mac OS X 10.6, 2010
12	...is a container for storing videos	Windows 10	Windows 98, 1998
13	...opens an eBook reader	Mac OS X 10.11.6	Mac iOS 4, 2010
14	...allows for the look-up of word definitions	Mac OS X 10.11.6	OpenStep, 1994
15	...opens camera settings	Mac OS X 10.11.6	Mc OS X 10, 2001
16	...opens sound settings / volume	Mac OS X 10.11.6/Windows 10	Xerox Star, 1981
17	...opens date & time settings	Mac OS X 10.11.6	Apple Lisa, 1983
18	...opens a calendar	Mac OS X 10.11.6	Windows 1 1985
19	...opens dictation and speech settings	Windows 10	Xerox Star, 1981
20	...saves	Mac OS X 10.11.6	Xerox Star, 1981
21	...pastes	MacOS X 10.11.6	Apple Lisa, 1983
22	...cuts	MacOS X 10.11.6	Xerox Star, 1981
23	...contains deleted files, folders, and applications	Windows 7	Xerox Star, 1981
24	...measures a time duration	Windows 8	Apple Lisa, 1983
25	...opens privacy preferences	Windows 7	Macintosh System 7, 1991
26	...opens email account settings	Windows 10	Windows 3.1, 1992
27	...manages passwords	Windows 10/ Mac OS X 10.11.6	Windows 3.1, 1992
28	...copies to	Windows 10	Xerox Star, 1981
29	...opens file explorer	Windows 7	Xerox Star, 1981
30	...creates a new folder	Windows 10	Xerox Star, 1981
31	...opens programs	Windows 10	Windows 3, 1995
32	...opens power settings	Windows 10	Windows NT 3.1, 1992
33	...opens storage options	Windows 10	Xerox Star, 1981
34	...opens energy saver options	Mac OS X 10.11.6	Mac OS X, 2001
35	...starts a phone call	Google Gmail	Gmail, 2009
36	...indicates notifications	Google	Google, 2003
37	...opens a presentation application	Mac OS X 10.11.6	NeXTSTEP, 1989
38	...opens file history	Windows 10	Xerox Star, 1981
39	...opens a web browser	Mac OS X 10.11.6	Mac OS X Panther, 2003

Table 10 The 39 computer functions of our plausibly anachronistic icons. Also shown are the systems from which our icons were taken and the systems in which they first appeared.

3 UNDERSTANDING YOUNG ADULTS' ICON PERCEPTIONS

I conducted a two-part study with 30 young-adult technology users to understand what icons they would create to trigger functions that are currently triggered with the *plausibly* anachronistic icons. I also sought to understand young adults' perceptions of anachronistic icons and the experiences they have of the real-world objects those icons portray. My study also resulted in a set of user-generated icons that I validated by conducting a distributed identification study using **The CROWDDESIGN engine** with a set of 60 adult participants to see how recognizable the new user-generated icons might be to general computer users.

3.1 **Young-Adult Participants**

I recruited 30 young adults to participate in our two-part study using flyers on and around the University of Washington (UW) campus. At the time of my study, the 18–22-year-old participants were born between the years 1994–1998. All participants were students at UW majoring in a wide range of areas including computer science, bioengineering, informatics, political science, and humanities. About half (57%) of the participants were from the United States, 23% were from China, and others were from myriad countries such as Switzerland and Nepal. Fifty-seven percent of the participants identified as female; the rest identified as male. Most of the participants (87%) did not have a professional design background. About half of the participants self-reported spending 6–10 hours a day using a computer, 3–5 hours using a mobile device, and about 70% of them spent less than 2 hours a day using physical objects (*i.e.,* books, wrenches). Table 11 reports additional demographic information for the participants.

Demographic		N = 30	Demographic		N = 30
Gender	Male	13 (43%)	First device owned	Desktop	13 (43%)
	Female	17 (57%)		Laptop	3 (10%)
	Non-binary	0		Mobile phone	6 (20%)
Country of origin	USA	17 (57%)		Other	8 (27%)
	China	7 (23%)	Total hours per day using a computer (self-reported)	< 1 hour	0
	Switzerland	1 (3%)		1–2 hours	7 (23%)
	South Korea	1 (3%)		3–5 hours	7 (23%)
	Nepal	1 (3%)		6–10 hours	14 (47%)
	Indonesia	1 (3%)		> 10 hours	2 (7%)
	India	1 (3%)	Total hours per day using a mobile device (self-reported)	< 1 hour	1 (3%)
	Bangladesh	1 (3%)		1–2 hours	9 (30%)
Age	18	8 (27%)		3–5 hours	17 (57%)
	19	8 (27%)		6–10 hours	3 (10%)
	20	4 (13%)		> 10 hours	0
	21	7 (23%)	Total hours per day using a physical object (books, paint brushes, wrenches, etc.) (self-reported)	< 1 hour	9 (30%)
	22	3 (10%)		1–2 hours	12 (40%)
Any professional design background	Yes	4 (13%)		3–5 hours	7 (23%)
	No	26 (87%)		6–10 hours	2 (7%)
Age started using computers	3–5	8 (27%)		> 10 hours	0
	6–8	16 (53%)	Preferred desktop operating system	Mac OS	16 (53%)
	9–11	5 (17%)		Windows	14 (47%)
	12–16	1 (3%)	Preferred mobile operating system	Apple iOS	20 (67%)
	17+	0		Google Android	9 (30%)
First operating system	Mac OS	3 (10%)		Blackberry OS	1 (3%)
	Windows	26 (87%)			
	MS-DOS	1 (3%)			

Table 11 Demographic information for our 30 young-adult participants.

3.2 **New Icons Elicitation Study**

The vast majority of elicitation studies have been centered around gestural interactions [124]. However, some studies elicited other forms of input actions like voice commands (Morris [75], Nebeling *et al.* [89], and my work on **Crowdlicit** and **Crowdsensus**), and sketches of gesture representation [72].

In an attempt to reduce "legacy bias" [79], where participants propose only familiar interactions, I conducted the elicitation study first before the identification study, thus limiting exposing my participants to plausibly anachronistic icons in the identification study.

The elicitation study session took about two hours to complete. In a session, participants were given a card that displayed text descriptions of a computer function (*e.g.*, "open a calendar") and were asked to

sketch as many icons as they could devise to trigger that function. Having participants propose multiple icons, rather than just one, for each referent in an elicitation study is a legacy bias reduction technique known as the "production principle" set forth by Morris *et al.* [79].

I asked the participants to rate each sketch they proposed on how well they felt it matched its intended function, prompted by the following: *"The icon I drew is a good match for its intended purpose."* Likert-type ratings were on a scale from 1–7, with 1 being "strongly disagree" and 7 being "strongly agree." I also asked the participants to rate the familiarity of the icon they drew on a 1–3 scale as such: 1 was "I am not familiar. I have never seen it before;" 2 was "I am somewhat familiar. I am not sure where I have seen it before;" and 3 was "I am very familiar. I know where I have seen it before."

3.3 **Anachronistic Icons Identification Study**

I asked the 30 young-adult participants to come back at a later date—after completing the elicitation part of the study—and complete an identification study in order to receive their compensation of $50 for their participation in both parts of the study. In the second part of the study, I handed the participants a deck of cards (Figure 22). The cards were numbered, and each card had a single plausibly anachronistic icon on it (Figure 23). Upon viewing a card, participants answered questions regarding the icon on the card. In addition to asking what function the icon on the card triggered when clicked, I asked participants to identify what real-world object was depicted by the icon, and whether participants had ever used the object itself, and when. Thus, I gathered data making it possible to examine whether there is any relationship between young adults' ability to identify what an icon *does*, and what the icon's real-world object *is*, for plausibly anachronistic icons.

3.4 **Young-Adult-Generated Icons Identification Study**

The first part of my study, the end-user elicitation study, resulted in a set of user-generated icons. To assess this set of new icons, I conducted a second end-user identification study online using **Crowdlicit**. I recruited 60 participants mTurk. Participants who accepted the human intelligence task (HIT) on mTurk were directed to a unique URL created by the Crowdlicit system in which they participated in a 15-minute study. In the study, participants had a total of 39 tasks, where each task showed one user-generated icon in isolation with no further context. Upon viewing an icon (the prompt), participants were

asked to propose the computing function that would be triggered by clicking that icon. Upon completing all 39 tasks, participants received a code that they entered back in the mTurk portal to indicate their completion of the study and ensure their payment of $3.75. (I based our payment on Washington state's minimum wage of $15/hour.) Participants also filled out a demographics survey upon completing the 39 tasks.

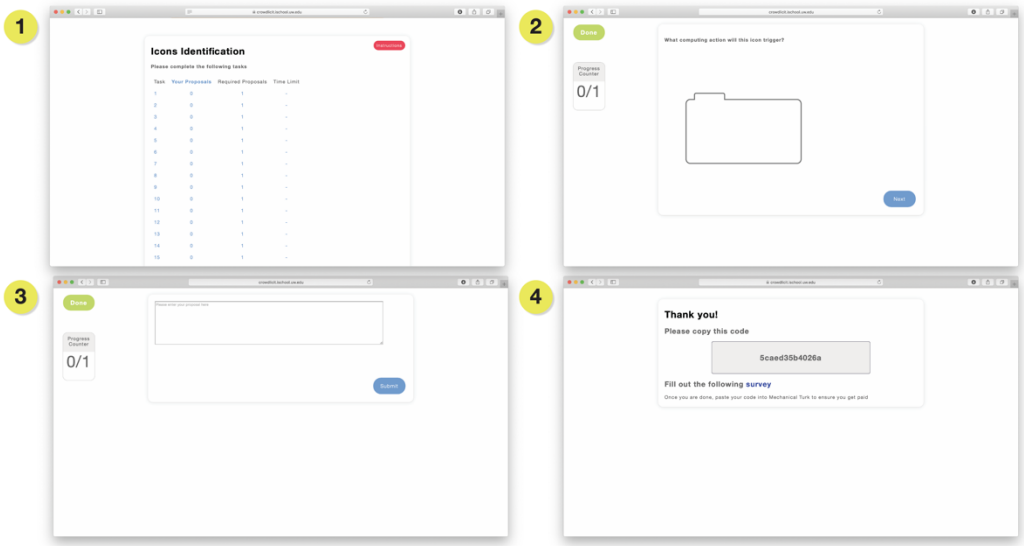


Figure 24 The Crowdlicit interface. (1) The task list of 39 user-generated icons to identify. (2) The prompt "what computing action will this icon trigger?" and a basic image of the icon. (3) The interface to identify the function triggered by the icon. (4) A thank you page with unique completion code and a link to the demographics survey.

Of the 60 participants who completed the HIT, 42 filled out the demographics survey. More than half of the participants (64%) identified as male. Table 12 shows the demographics of the online participants. The majority of the participants (69%) were from the United States; 21% were from India, and the rest were from other countries including Italy and Canada. The average age of the participants was 31.1 (*SD*=7.97) years. Most of our participants (72%) indicated that they spend more than 6 hours a day using a computer. Seventy-nine percent of the participants spend between 1–5 hours per day using a mobile device. Seventy-two percent of the participants spend 2 hours or less using physical objects.

Demographic		N = 42	Demographic		N = 42
Gender	Male	27 (64%)	Total hours a day using a computer (self-reported)	< 1 hour	1 (2%)
	Female	15 (36%)		1–2 hours	2 (5%)
	Non-binary	0		3–5 hours	9 (21%)
Country of origin	USA	29 (69%)		6–10 hours	18 (43%)
	India	9 (21%)		> 10 hours	12 (29%)
	Italy	1 (2%)	Total hours a day using a mobile device (self-reported)	< 1 hour	4 (10%)
	Canada	1 (2%)		1–2 hours	21 (50%)
	Pakistan	1 (2%)		3–5 hours	12 (29%)
	Ireland	1 (2%)		6–10 hours	2 (5%)
Average age	31.1 (*SD*=7.97) years			> 10 hours	3 (7%)
Professional design background	Yes	11 (26%)	Total hours a day using a physical object (books, paint brushes, wrenches, etc.) (self-reported)	< 1 hour	10 (24%)
	No	31 (74%)		1–2 hours	21 (50%)
Age started using computers	3–5	10 (24%)		3–5 hours	7 (17%)
	6–8	13 (31%)		6–10 hours	2 (5%)
	9–11	11 (26%)		> 10 hours	2 (5%)
	12–16	2 (5%)	Preferred desktop operating system	Mac OS	8 (19%)
	17+	6 (14%)		Windows	33 (79%)
First operating system	Mac OS	4 (10%)		Other	1 (2%)
	Windows	35 (83%)	Preferred mobile operating system	Apple iOS	17 (40%)
	Other	3 (7%)			
First device owned	Desktop	35 (83%)		Android	25 (60%)
	Laptop	6 (14%)			
	Other	1 (2%)			

Table 12 Demographic information for our 42 participants recruited from Amazon's Mechanical Turk.

4 RESULTS: ICONS AND PERCEPTIONS

I present the results of my studies in the following subsections, including the icons proposed by the 30 young-adult participants in the elicitation study, their perceptions of plausibly anachronistic icons, and their experiences, if any, of the real-world objects those icons represent. I also present the results of an additional crowdsourced identification study I conducted with 60 adult Mechanical Turk participants on the set of user-generated icons that resulted from our in-lab elicitation study with the young-adult participants. I take each of these in turn.

4.1 **Eliciting User-Generated Icons**

The 30 young-adult participants offered a total 3,590 icon sketches for 39 function prompts. Participants were encouraged to propose as many icon sketches as they wanted for each prompt. On average, 3.07 (*SD*=0.68) icons were sketched per prompt by each participant (Figure 25). Given the unequal number of icons elicited per prompt, the average number of icon sketches elicited per prompt therefore was 91.92 (*SD*=3.83). I conducted open-coding analysis [112] on the entire set of icons, and then followed Wobbrock *et al.'s* [131] method of similarity-based clustering to derive a set of 39 user-generated icons.

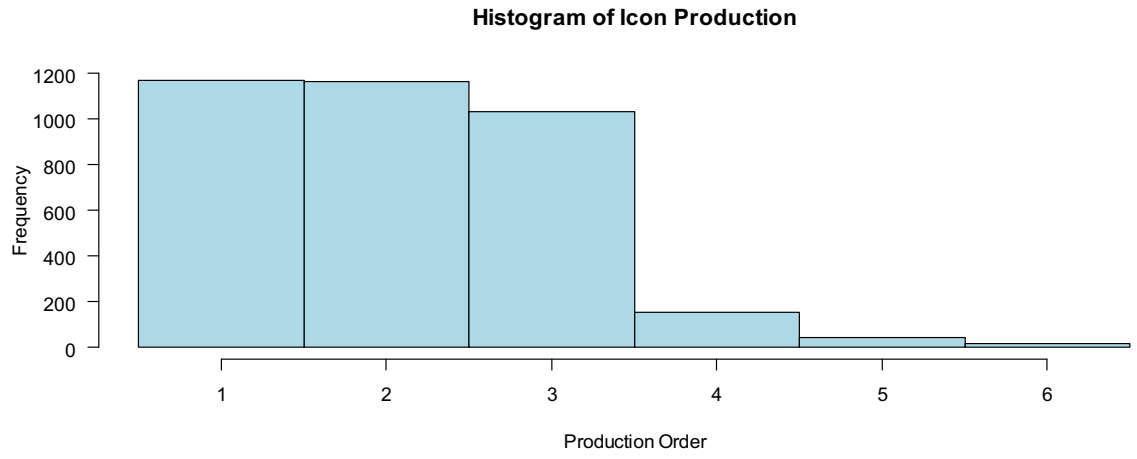


Figure 25 The total number of icons proposed by 30 participants in the study by production order (1st, 2nd, 3rd, etc.). All participants proposed at least two icons for each computing function, but very few proposed as many as five or six icons for a computing function.

4.1.1 Understanding the Set of Elicited Icons.

I took an open-coding [112] approach to analyze the entire set of 3,590 icon sketches elicited for 39 prompts from the 30 participants. I generated a set of 14 codes to describe the icon sketches I collected. The codes described whether an icon sketch represented a physical object, contained a metaphor from the desktop, or contained a metaphor from newer technologies such as mobile devices and their interfaces. The codes also included anachronisms of both objects and practices, representations of computer hardware and actions, representations of natural elements, and physical activities or body parts. Other codes described shapes present in sketches like abstract squares or lines. The codes included design conventions—*i.e.,* hamburger menu. Table 13 shows the code book I formulated and used to analyze the icons, as well as the number of icons and percentage classified for each code, and an illustrative example of the code.

I found that the majority of icons (73%) proposed by the young-adult participants were representations of physical objects. In addition, 60% of all icons extended the desktop metaphor, and 55% of all proposed sketches were of existing plausibly anachronistic icons. The sketches were also influenced by new technologies, as 20% of icons had metaphors based on mobile platforms and others such as "the cloud." Thirty percent of all icons were of abstract shapes such as arrows and rectangles. Finally, 20% of all icons had alphanumeric text in them.

Code	Description	N = 3,590	Example
Object	Representation of a physical object for its literal meaning	2,612 (73%)	Book
Metaphor	Representation of a real-world concept or practice for an associated meaning		
Desktop	Furthering the desktop metaphor	2,157 (60%)	Use of file folders
New Technology	Use of new technology metaphors	574 (16%)	The cloud
Mobile	Furthers metaphors from mobile	150 (4%)	Home button
Anachronistic Practice	Something we used to do	56 (2%)	Store contacts in a physical address book
Computer Hardware	A non-anachronistic object used for computer function	162 (5%)	USB Drive
Computer Action	Something that happens on the computer screen	17 (0.4%)	Typing, clicking
Rare	An object that is rarely used	200 (6%)	Microscope; not a common household object like a knife or notebook
Body Part	A part of the human body	309 (9%)	An eye
Physical Activity	Represents physical activity	39 (1%)	Hand motion
Text	Text including alphanumerical characters	729 (20%)	The word "save"
Shape			
Abstract	General shape	1,088 (30%)	Circle, rectangle, arrow
Design Convention	Familiar design convention	365 (10%)	Hamburger menu, 3 dots for typing
Nature	Natural elements	171 (5%)	Animals, plants, and other natural elements
Brand			
Brand Name	Text use of a brand name	26 (0.7%)	"Skype", "Gmail"
Logo	Use of image representation of a brand	137 (4%)	"Apple logo"
Brand Influence	A visual representation of an element influenced by a brand practice	73 (2%)	Apple's use of groups of apps
Meta GUI	Use of existing UI elements in creating new icons	332 (9%)	Dropdown menu item; mouse cursor
Anachronistic object	The use of an anachronistic object		
Existing Icon	Use of an existing anachronistic icon	1,965 (55%)	A floppy disk
New Anachronism	Use of a new anachronistic icon	237 (7%)	Binoculars

Table 13 The codebook and the breakdown of our coding of the 3,590 elicited icons.

4.1.2 Creating a Set of User-Driven Icons

For each of our 39 prompts, I grouped participants' icon sketches based on their similarity. I took the sketches with the highest consensus for each prompt and illustrated them as clean representative line drawings as shown in Figure 26. The figure shows three example sketches provided by three separate participants (P007, P020, and P027.) The three sketches were in response to prompt 21 *"draw an icon that pastes."* These examples were of a glue bottle, which was the icon concept of the highest consensus

among our participants which led me to select it as the new icon to represent the paste function. Based on the sketches I collected, I illustrated the icon as shown on the right of Figure 26.

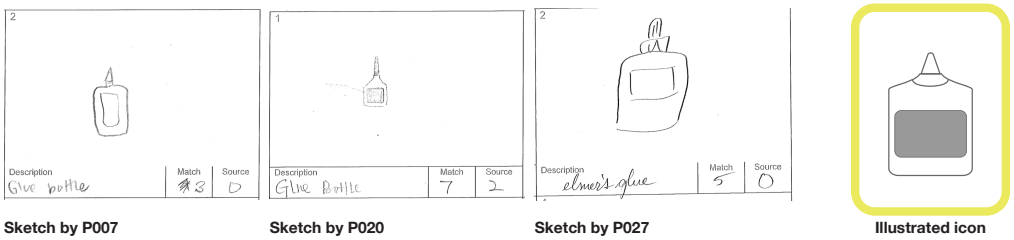


Figure 26 An example of the illustrated icon I crated based off participants' sketches.

Following the same approach, I produced a set of 39 user-generated icons (Figure 27). I list the prompts associated with each icon in Table 14, distinguish whether the user-generated icon is of a new concept or of the same plausibly anachronistic icon, and report the median "match" and "novelty" scores. (The match score is a user-reported score on a 7-point Likert scale assessing the proposed sketch on how well it matches the prompt. The novelty score is a 3-point scale assessing the novelty of the concept that the sketch depicts.)

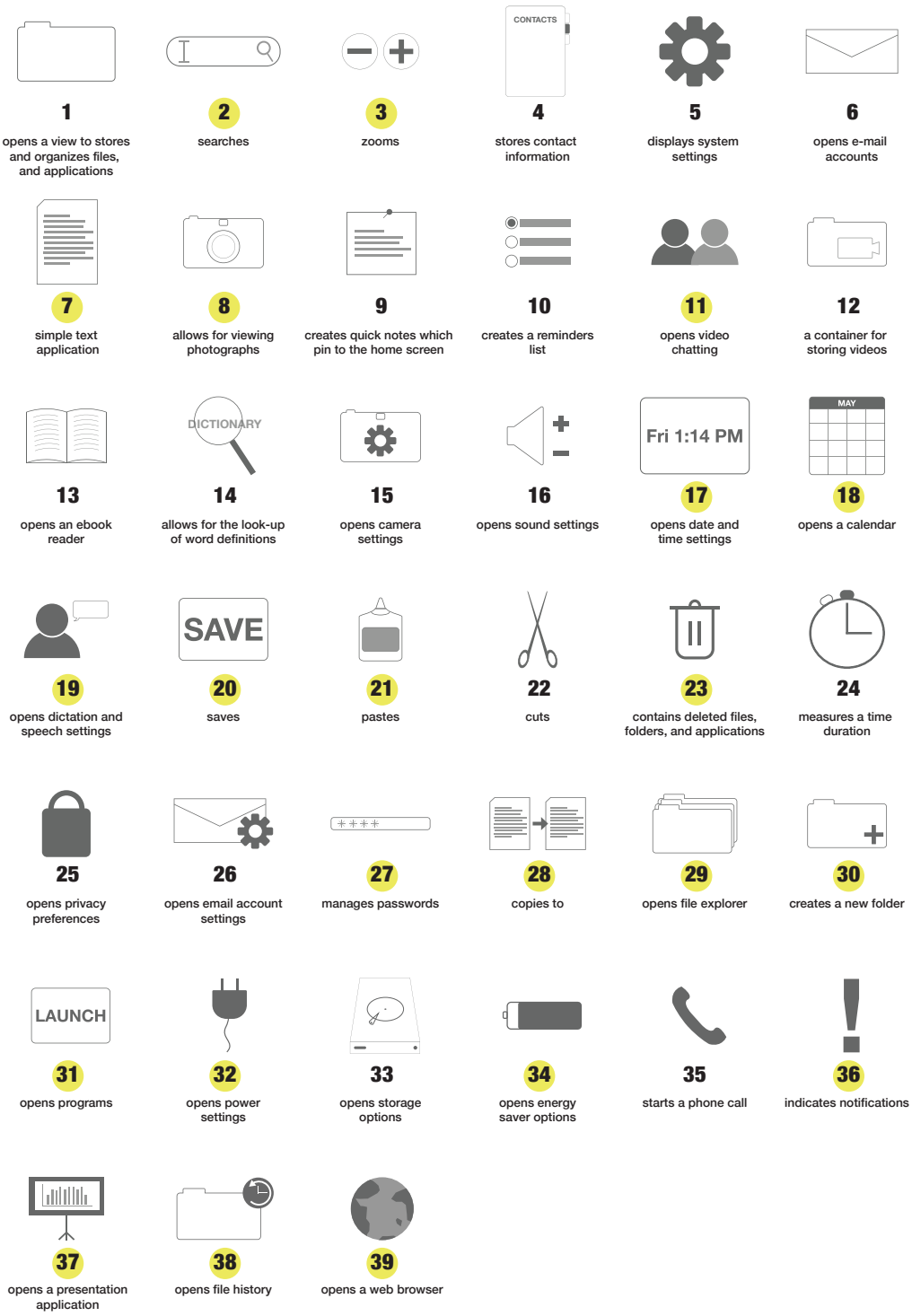


Figure 27 The set of user-generated icon concepts emerging from our end-user elicitation study. Icons marked in yellow are new concepts different from the plausibly anachronistic icons we assembled (22 of 39 here are new).

Draw an icon that...	New Concept	Median Match Score (1st, 3rd)	Median Novelty Score (1st, 3rd)
1. ...opens a view to store and organize files, and applications	No	7 (5,7)	3 (2,3)
2. ...searches	Yes	5.5 (5,6)	2 (2,3)
3. ...zooms	Yes	7 (6, 7)	3 (3,3)
4. ...stores contact information	No	6 (5,7)	2 (2,3)
5. ...displays system settings	No	6 (5.5,7)	3 (3,3)
6. ...opens e-mail accounts	No	6 (5,7)	1.5 (1,2)
7. ...is a simple text application	Yes	7 (6.5,7)	3 (3,3)
8. ...allows for the viewing of photographs	Yes	5 (5,6)	3 (2,3)
9. ...creates quick notes which pin to the home screen	No	5.5 (4.25, 6)	1 (1, 2.75)
10. ...creates a reminders list	No	6 (5, 6.75)	3 (2,3)
11. ...opens video chatting	Yes	4 (4,5)	1 (1,1)
12. ...is a container for storing videos	No	6 (5,6.75)	2 (1.25,2)
13. ...opens an eBook reader	Yes	7 (6,7)	2 (2,3)
14. ...allows for the look-up of word definitions	No	5.5 (4.25, 6)	2 (1.25, 2.75)
15. ...opens camera settings	No	6 (5, 7)	2 (1,3)
16. ...opens sound settings / volume	No	7 (6,7)	3 (3,3)
17. ...opens date & time settings	Yes	6 (5,6)	1 (1,2.5)
18. ...opens a calendar	Yes	7 (6,7)	3 (2,3)
19. ...opens dictation and speech settings	Yes	5 (4, 6)	1.5 (1, 2)
20. ...saves	Yes	5 (4, 7)	2 (1, 2.75)
21. ...pastes	Yes	5 (3.5, 6)	1 (1, 2)
22. ...cuts	No	7 (6, 7)	3 (2, 3)
23. ...contains deleted files, folders, and applications	Yes	7 (5, 7)	3 (2, 3)
24. ...measures a time duration	No	5 (5, 6)	3 (3, 3)
25. ...opens privacy preferences	No	7 (6, 7)	3 (2, 3)
26. ...opens email account settings	No	7 (5, 7)	1.5 (1, 2)
27. ...manages passwords	Yes	6 (4, 6)	1 (1, 2)
28. ...copies to	Yes	6 (4, 6)	1 (1, 2)
29. ...opens file explorer	Yes	5.5 (4.25, 6)	2 (1, 2)
30. ...creates a new folder	Yes	6 (5, 7)	2 (2, 3)
31. ...opens programs	Yes	5.5 (4.75, 6.25)	1 (1, 1.5)
32. ...opens power settings	Yes	6 (6, 7)	2 (1, 2.75)
33. ...opens storage options	Yes	6 (6, 7)	2 (1, 3)
34. ...opens energy saver options	Yes	5 (5, 6)	2 (2, 3)
35. ...starts a phone call	No	6 (5, 7)	2 (2, 3)
36. ...indicates notifications	Yes	6 (6, 7)	3 (2, 3)
37. ...opens a presentation application	Yes	6 (5, 7)	2 (1, 3)
38. ...opens file history	Yes	5 (3, 5)	1 (1, 1.5)
39. ...opens a web browser	Yes	6 (5,7)	3 (2,3)

Table 14 The function each user-generated icon would trigger in a computing system. Columns describe whether the icon is of a new concept or of an anachronistic object, the median match score and the 1st and 3rd quartile on 1–7 Likert-type scale (higher is a better perceived proposal-function match), and the median novelty score and the 1st and 3rd quartile on 1–3 Likert-type scale (higher is greater perceived novelty). See text for details.

When assembling the set of user-generated icons, I omitted the first elicited icon for each prompt by each participant because participants tended to propose legacy interactions first [79]—in other words, their first impulses were to simply provide an icon with which they were already familiar. Because I did not show participants the current—plausibly anachronistic—icon being used for each function prompt and did not tell them they were not allowed to draw this icon, 70% of the first-elicited icons were plausibly anachronistic, compared to the second elicited icons, which contained 56% anachronistic icons. In general, with more icons proposed for each referent, the chances that the icon was anachronistic,

compared to the first icon, went down significantly (Figure 28). A mixed logistic regression model [111,140] for the chances of anachronism by icon production order shows a significant main effect ($\chi^2(5, N=3{,}590) = 65.87$, $p<.0001$). *Post hoc* pairwise comparisons corrected with Holm's sequential Bonferroni procedure [43] show that the first icon was significantly more likely to be anachronistic than icons proposed second, third, fourth, or fifth, but icons in these positions were not significantly different. (Icons proposed in the sixth position were not significantly less likely to be anachronistic than the first icons proposed, but this result is unreliable due to having only $N=19$ icons out of 3,590 proposed sixth.)

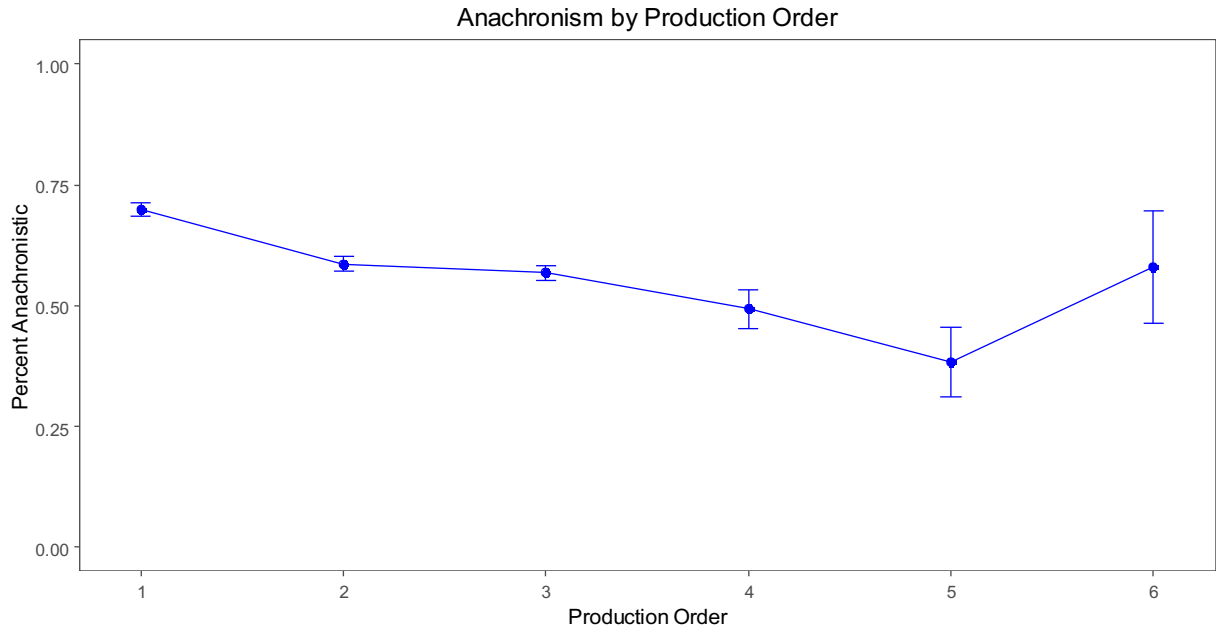


Figure 28 The percent of icons that were anachronistic by production order. Higher percentage is more likely to be anachronistic, while lower is less likely. Error bars represent ±1 standard error.

When calculating icon agreement rates for each of the 39 prompts, because I had an unequal number of icons elicited for each prompt, I did not use Equation 1; instead I used Morris's Max-Consensus score [75]. The Max-Consensus score refers to the percentage of participants who proposed the most frequent icon. The scores are between 0–1, with 1 meaning that all the participants proposed the same icon and were in complete agreement.

Figure 29 shows the Max-Consensus rates for all 39 user-generated icons. On average, a quarter of the participants agreed on what the icon to trigger a function should be. The mean Max-Consensus score was

0.25 (*SD*=0.09). The icon for prompt #22—the function "Cut"—had the highest Max-Consensus score at 0.52, *i.e.,* half the participants proposed the same icon—a pair of scissors, which I included as plausibly anachronistic icon because I considered it extends the desktop metaphor from an era where papers were more prevalent on physical desktops than technologies like laptops. The icon for prompt #31—open program—had the lowest score of 0.07. The icon for prompt #31 is merely the word "Launch."

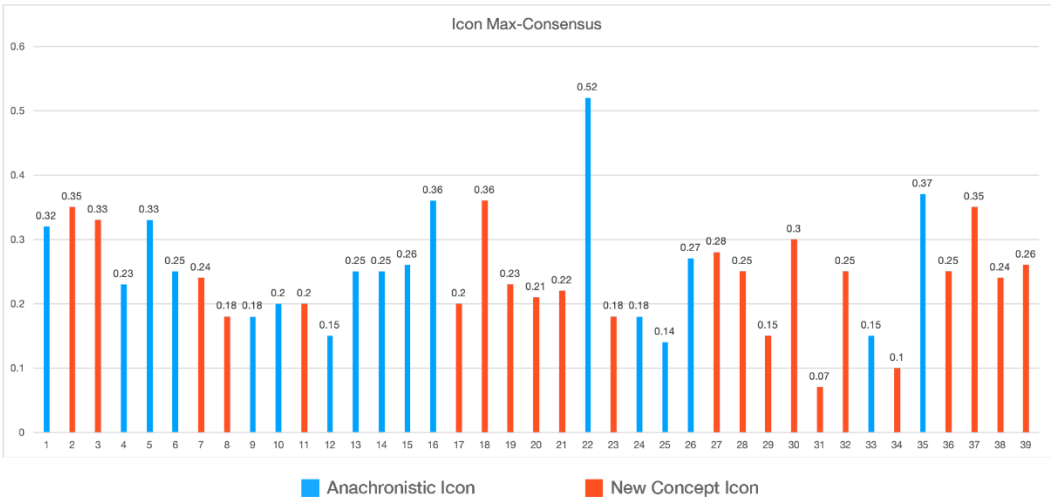


Figure 29 Max-consensus scores for the 39 icons elicited from the 30 young-adult participants. The scores are between 0–1, with 1 being total agreement, i.e., all participants proposed the same icon. The blue bars represent anachronistic icons, and the orange bars represent new concept icons (22 of 39 are new).

4.2 **Anachronistic Icon Identification**

I asked my participants to identify the objects depicted in each one of the plausibly anachronistic icons, whether they have used the depicted real-world object and when, and what computer feature or function would be triggered if the icon were clicked. Following my end-user identification method, I grouped the proposed functions based on their similarity, calculated function agreement rates and function accuracy.

4.2.1 Function Agreement

I used the function proposal agreement formula that I established [9] based on Wobbrock *et al.'s* [128,131] agreement formula. The function agreement scores are reported in Figure 30.

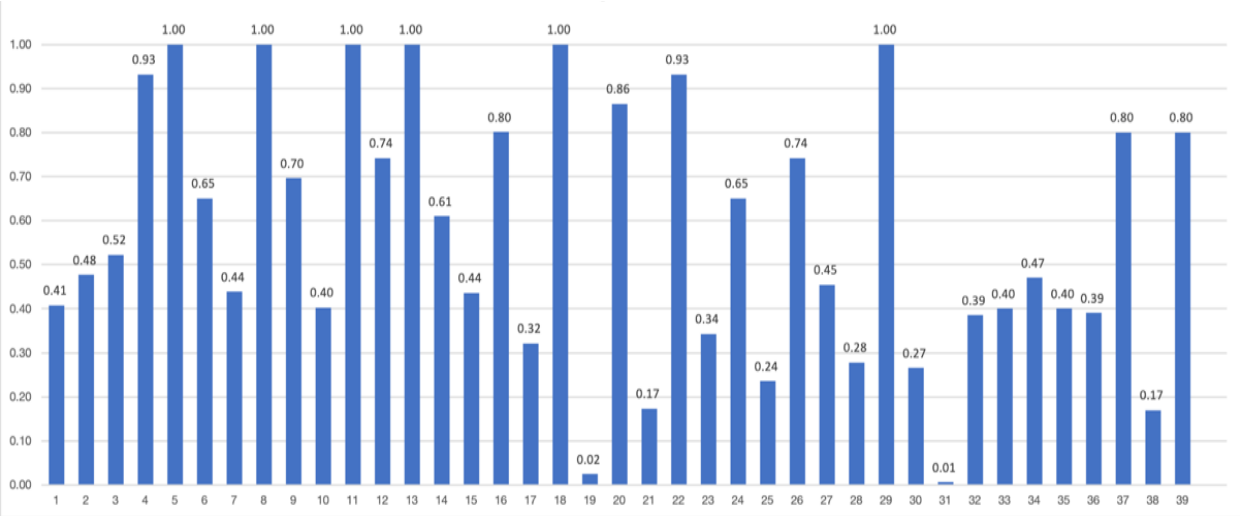


Figure 30. The function agreement scores for each of the 39 plausibly anachronistic icons in the in-lab identification study with the 30 young-adult participants. Eight of them had agreement at 0.90 or above.

There were six icons with perfect agreement–*i.e.,* all 30 young-adult participants guessed the same function for the icon. Despite getting a perfect identification score of 1.0, two of these icons had new icon design concepts as a result of the elicitation study. One was prompt #8, viewing photographs, and the other was prompt #11, making a video call. The eight icons with above 0.9 function agreement, or identification scores, are listed in Table 15 along with the function identified by the participants and the user-generated icon for each function that resulted from the elicitation study.

No.	Function	Anachronistic Icon	User-Driven Icon
4	Stored contact information		
5	Display system settings		
8	Viewing photographs		
11	Video call		
13	eBook reader		
17	Calendar		
22	Cuts		
29	File Explorer		

Table 15. Eight icons with high agreement scores of 0.9. The functions they trigger, and the participant-generated icon created for these functions.

There were seven icons with function agreement scores lower than 0.30. They were: #19. Speech and dictation setting; #21. Paste; #25. Privacy settings; #28. Copies to; #30. Create a new folder; #31. Open program; and #38. File history. Six of these seven icons had new icons in our user-generated icon set. Table 16 lists these icons, the functions they trigger, their plausibly anachronistic icons, and the user-generated icon elicited from the 30 young-adult participants.

No.	Function	Anachronistic Icon	User-Generated Icon
19	Speech and dictation settings		
21	Paste		
25	privacy settings		
28	copies to		
30	create a new folder		
31	opens program		
38	file history		

Table 16. Seven icons with agreement scores lower than 0.3, the functions they trigger, and the user-generated icon created for these referents.

4.2.2 Function Accuracy

With 30 young-adult participants and 39 plausibly anachronistic icons, I obtained 1,170 attempted icon identifications for the computer functions those icons trigger. After grouping function proposals by similarity, nine functions of 39 were not identified correctly from their anachronistic icons, *i.e.,* the function proposed with highest consensus by the participants did not match the function currently triggered by the icons in the actual operating systems from which those icons were extracted. The functions that were not identified correctly were: #7. Simple text application; #15. Open camera settings; #19. Open dictation and speech settings; #21. Paste; #26. Open email account settings; #30. Create a new folder; #31. Open program; #34. Open energy saver options; and #36. Indicate notifications. In Table 17, I list the incorrectly identified icons, the function with the highest consensus proposed by the participants, the number of participants who proposed that function, the actual function associated with the icon, and the number of participants who proposed the actual function.

No.	Icon	Identified Referent	No. Participants	Actual Referent	No. Participants
7		Take notes	17	is a simple text application	7
15		open webcam	17	opens camera settings	6
19		record voice	14	opens dictation and speech settings	1
21		copy	9	pastes	5
26		send email	24	opens email account settings	2
30		open folder	11	creates a new folder	8
31		install software	9	opens programs	1
34		control brightness	19	opens energy saver options	1
36		set an alarm	14	indicates available notifications	10

Table 17. Nine of 39 plausibly anachronistic icons were identified incorrectly by our 30 young-adult participants. The table lists the identified function, the number of participants who proposed it, the actual function, and the number of participants who proposed that.

4.2.3 Experience with Real-World Objects

I also asked the 30 young adult participants additional questions beyond having them identify computing functions from icon representations. I asked participants to identify the real-world objects represented in the plausibly anachronistic icons, what those objects are used for, and when they last interacted with them. Twenty-five objects out of 38 (icons #2 search and #3 zoom shared the same object, a magnifying glass) were identified correctly by all 30 participants, and the remaining 13 objects were identified correctly by an average of 86% (*SD*=11%) of participants. I also asked participants if they had used the real-world object depicted by each icon. Twelve real-world objects in 16 icons—four icons shared the same object, a folder—were, in fact, used by all of our young-adult participants. (And by extension, 23 objects had not ever been used by all participants.) The real-world objects that had been used were: folders, pens and paper, print photographs, paper sticky notes, paper books, scissors, a recycling bin, a padlock, a paper envelope, jagged metal keys, a compact disk, a wall plug, and a filament light bulb. It is fair to say that these

real-world objects had, therefore, not become anachronistic to our young-adult participants yet. (One can speculate as to how many years from now these real-world objects *will* become unfamiliar to young adults.)

By contrast, the real-world objects depicted by 23 of the 39 plausibly anachronistic icons had *never* been used by at least some of our participants. Figure 31 shows the percentage of participants who had never used each of the real-world objects. On average, 25% (*SD*=20%) of all participants never used these 23 objects. These objects were: a magnifying glass, an address book, a gear, a stamp, a reminder list, a video camera, a film strip, a paper dictionary, a dedicated camera, an audio speaker, a wall clock, a wall calendar, a microphone, a 3.5″ floppy diskette, a physical clipboard, an analog stopwatch, a manila folder organizer, a computer hard drive, a land-line telephone, a bell, a podium, and a magnetic compass.

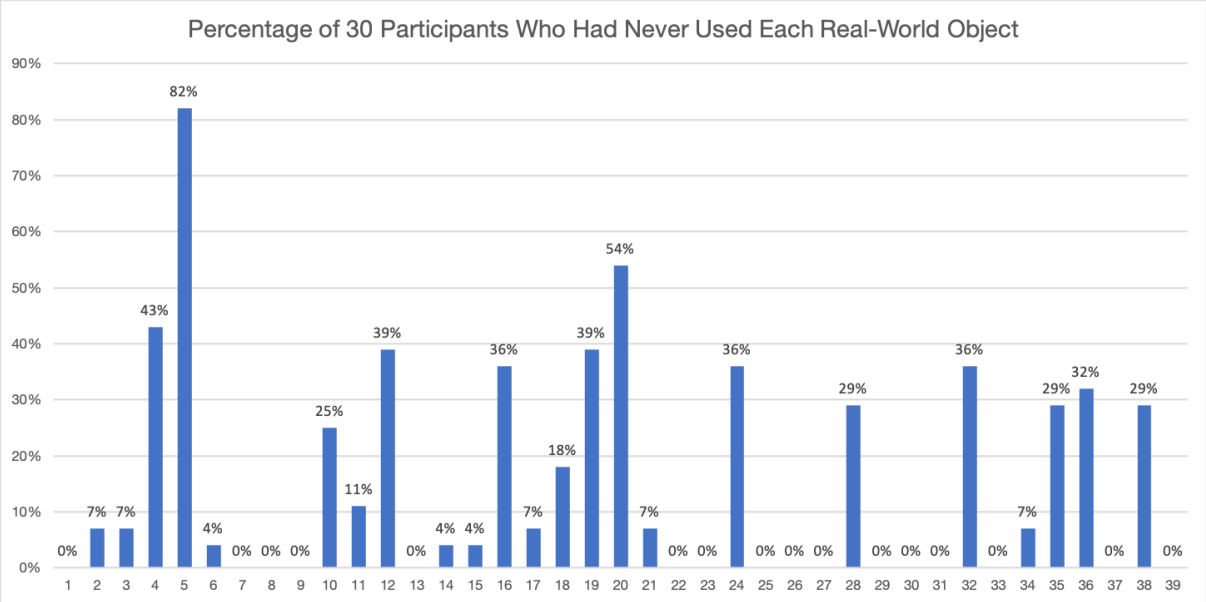


Figure 31. The percentage of 30 participants who had never used the real-world object shown in each one of the 39 plausibly anachronistic icons. For example, 82% of participants had never used icon #5, which is a mechanical gear cog. See Figure 2 for the set of plausibly anachronistic icons.

4.3 **Identification of Icons Generated by Young Adults**

To test the identifiability of the set of user-generated icons from the young adults, I conducted an identification study using **Crowdlicit** with 60 new online participants recruited from Amazon's Mechanical Turk (MTurk). I report on the function agreement rates reached by the 60 participants and their accuracy in identifying the functions associated with each user-generated icon.

4.3.1 Function Agreement

I calculated the function agreement for the proposals collected in the online identification study using Equation 1. There were 22 new-concept icons and 17 icons that matched my plausibly anachronistic icons. The mean function agreement score for the new concept icons was 0.53 (*SD*=0.24); the mean function agreement score for the plausibly anachronistic icons was almost identical at 0.54 (*SD*=0.24). The function agreement rates can be seen in Figure 32 below.

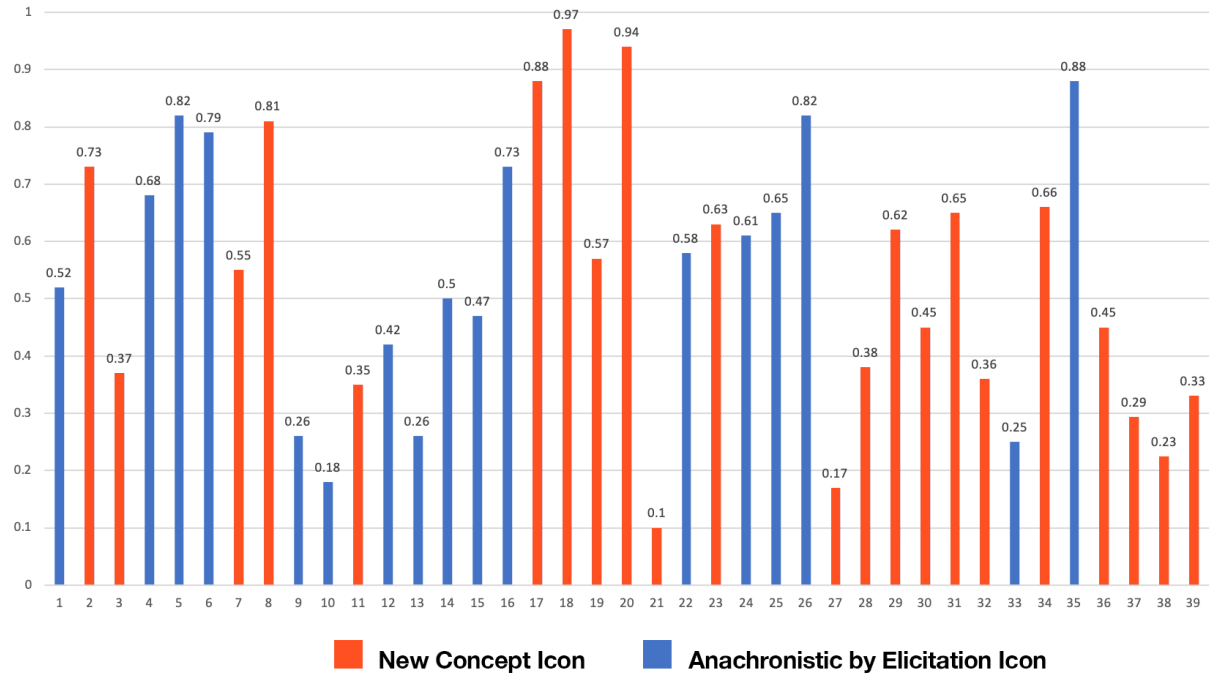


Figure 32. The function agreement scores for each of the 39 icons in the online identification study with our 60 participants. Blue bars represent the 17 icons that remained plausibly anachronistic; orange bars represent the 22 new concept icons.

4.3.2 Function Accuracy

I collected a total of 2,340 function proposals for the 39 user-generated icons from 60 online participants. Of the collected function proposals, 61.58% (1,441) were correct–*i.e.,* the participants were able to identify the actual function the icon should trigger. For each user-generated icon I found the prompt with the highest consensus among the 60 participants. The participants from mTurk were able to correctly identify 34 of the 39 user-generated icons. Five user-generated icons had an incorrect function with the highest consensus among the 60. The five icons identified incorrectly were: #11. Open video chat; #19. Open dictation and speech settings; #24. Measure time duration; #25. Open privacy preferences; and #39. Open web browser. Table 18 lists the icons, the function with the highest consensus resulting from our identification study, the correct function associated with the icon, the number of participants who proposed the function with the highest consensus, and the number of participants who proposed the correct function.

No.	Icon	Identified Function	No. Participants	Actual Function	No. Participants
11*		Open contacts	28	Open video chatting	21
19*		Open messages	45	Open dictation and speech settings	0
24		Show time	47	Measure a time duration	7
25		Lock computer	48	Open privacy preferences	0
39*		Open map	26	Open a web browser	22

Table 18. The five icons were identified incorrectly in the online identification study. A star (*) indicates a New Concept icon. The table lists the identified function, the number of participants out of 60 who proposed the identified function, the actual function, and the number of participants who proposed that.

5 A FINAL SET OF ICONS

To review, I conducted a multi-part study: (1) a lab-based elicitation study with 30 young-adult participants eliciting icon sketches for 39 computing functions, (2) a lab-based identification study of 39 plausibly anachronistic icons from current operating systems and applications, and (3) a distributed identification study of 39 user-generated icons with 60 online adult participants.

Drawing on the results of these studies, I have crafted what I consider to be the best-suited set of icons for 39 computing functions (Figure 33). By "best-suited" I mean icons that have high identifiability of their

functions regardless of whether they represent a plausibly anachronistic object or not. The 39 computing functions are currently portrayed by at least some icons that depict anachronistic objects. Twenty-seven of the real-world objects depicted by these icons were *never* used by a quarter of our young adult participants, on average. This final set of 39 user-generated icons has three categories: *Anachronistic by Elicitation* icons are those generated in the elicitation study that represent anachronistic objects; *New Concept* icons are of new concepts that resulted from the end-user elicitation studies; and *Anachronistic by Identification* icons are those that had new icon concepts as a result of the elicitation study, but they were not identified correctly in the online identification study—hence, I chose the anachronistic icons rather than the new concepts.

The *Anachronistic by Elicitation* subset has 15 icons for the following functions: #1. Opens a view to store and organize files and applications; #4. Store contact information; #5. Display system settings; #6. Open email account settings; #9. Create quick notes that pin to the home screen; #10. Create a reminder list; #12. Open a container for storing videos; #13. Open an eBook; #14. Allow for the look-up of word definitions; #15. Open camera settings; #16. Open sound settings; #22. Cut; #26. Open email account settings; #33. Open storage options; and #35. Start a phone call.

The *Anachronistic by Identification* list has four icons for the following functions: #11. Open video chat; #24. Measure a time duration; #25. Open privacy preferences; and #39. Open a web browser.

Finally, the *New Concept* list contains icons for the remaining 20 computing functions, all of which were identified correctly by our 60 online participants in the identification study. One notable exception is the icon for function #19—open dictation and speech settings—that had a *New Concept* icon as a result of the elicitation study. However, the original anachronistic icon for function #19 was not identified correctly in the lab-based identification study, and the *New Concept* icon was not identified correctly in the online identification study, either. The reason I include the *New Conc*ept icon in the final icon set, rather than its anachronistic counterpart, is that the new icon had a higher function agreement rate of 0.57 compared to the 0.02 function agreement rate the anachronistic icon received. Despite this new concept icon being an unsuitable one for the function, it is a more identifiable icon on its own. Finding a more highly identifiable icon for function #19 remains a possibility for a future end-user elicitation study.

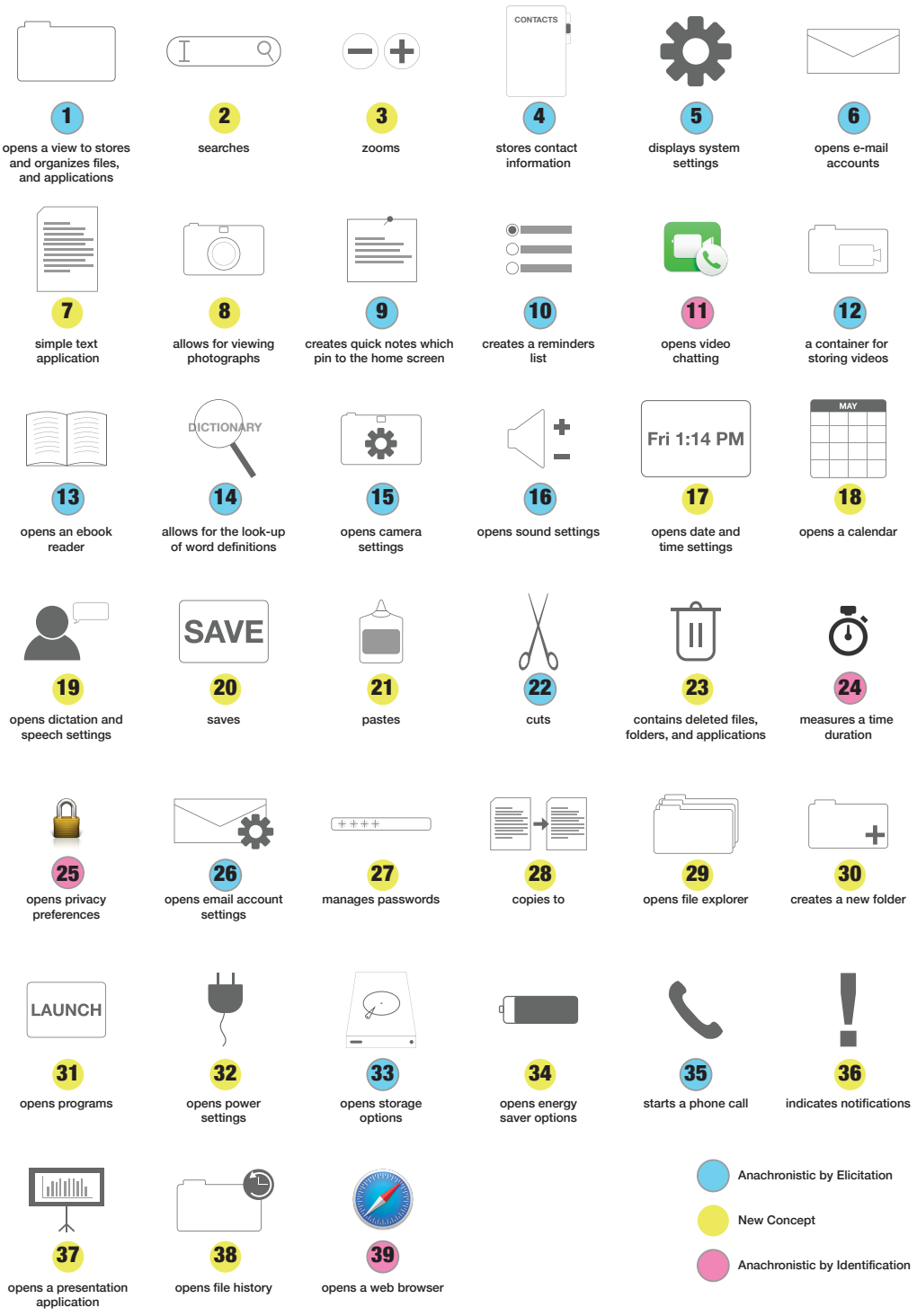


Figure 33. The final set of icons based on the elicitation and identification studies.

I established a set of codes when I conducted the open-coding analysis on the entire set of 3,590 icon sketches that I elicited from the 30 young-adult participants in the lab-based elicitation study. I used these codes on the final set of icons (Table 19). Most icons in the final set (77%) represent a physical object. Forty-one percent of the icons utilize the desktop metaphor. About a quarter (26%) of the icons incorporating GUI elements had a "Meta GUI" code. Finally, about half (49%) of the icons were of plausibly anachronistic objects from the original set (Figure 23).

Code	Icon Referent No.	N = 39
Object	1, 2, 4, 5, 6, 7, 8, 9, 11, 12, 13, 14, 15, 16, 21, 22, 23, 24, 25, 26, 28, 29, 30, 32, 33, 34, 35, 37, 38, 39	30 (77%)
Metaphor		
Desktop	1, 4, 6, 7, 9, 13, 14, 21, 22, 24, 26, 28, 29, 30, 35, 38	16 (41%)
New Technology		0 (0%)
Mobile		0(0%)
Anachronistic Practice		0 (0%)
Computer Hardware	33	1 (3%)
Computer Action	10, 27, 28, 34	4 (10%)
Rare		0 (0%)
Body Part	19	1 (3%)
Physical Activity		39 (1%)
Text	3, 4, 14, 16, 17, 18, 20, 30, 31, 36	10 (26%)
Shape		
Abstract		0 (0%)
Design Convention	7, 9, 10, 13, 28	5 (13%)
Nature		0 (0%)
Brand		
Brand Name		0 (0%)
Logo		0 (0%)
Brand Influence		0 (0%)
Meta GUI	2, 10, 18, 19, 27, 37	6 (15%)
Anachronistic object		
Existing icon		19 (49%)
New Anachronism		0 (0%)

Table 19 Analysis of the final set of icons.

6 DISCUSSION

To discuss my findings in context, I revisit the three research questions I outlined in the introduction section of this chapter. I looked over current desktop operating systems and applications and assembled a list of 39 icons that represented plausibly anachronistic objects. Based on those icons, I set out to answer the following research questions:

- RQ.10 How does production influence user-elicited interaction designs?
- RQ.12 What icons would young adults propose to trigger computer functions currently associated with anachronistic icons?
- RQ.13 How familiar are young adults with the objects represented in anachronistic icons?
- RQ.14 How identifiable is a set of icons elicited from young adults?

I address each of these questions in turn. In addition, I highlight 20 anachronistic icons due for a redesign, I discuss the limitations of my work and identify possible directions for future work.

6.1 **User-Generated Icons**

I conducted open-coding analysis on the entire set of 3,590 icon sketches I gathered in the lab-based elicitation study with 30 young-adult participants. The icon sketches proposed by participants were heavily influenced by physical objects. I found that 73% of all sketches represented a physical object. A similar number was found in the final set of 39 icons, as 77% of these icons represented physical objects, too.

The use of alphanumeric characters, or "Text" as I categorized it in the coding scheme (Table 13), was prevalent at 26% of the final set of 39 icons. In four cases, the icons were entirely made up of text. These four icons were for the following functions: #17. Open date and time settings; #20. Save; #31. Open programs; and #36. Indicate notifications. All four icons were identified correctly in the online identification study and had high function agreement rates—above 0.60—except for prompt #36, which still had a decent agreement rate of 0.45. This is not surprising, as text icons spell out the function they trigger. A blatant example of using text was for function #20. Save, which was only the outlined word "Save," and had a function agreement rate—or identifiability score—of 0.94. Such a choice stands in stark

contrast to the 3.5″ floppy diskette icon so prevalent in today's desktop interfaces and begs the question of whether it is time for an overhaul of modern computer iconography.

For icons that were *anachronistic by elicitation*, none of them had a max-consensus score of 0.10 or less. In fact, five *anachronistic by elicitation* icons had max-consensus above 0.30, which were #1 opens a view to store and organize files, and applications (0.32), #5 displays system settings (0.33), #16 opens sound settings (0.36), #22 cuts (0.52), and #35 starts a phone call (0.37). Initially I had considered the objects represented in these icons–folder, engine cog, speaker, scissors, landline handset–to be plausibly anachronistic. I consider the anachronism of the objects depicted in the set of 39 plausibly anachronistic icons in the next subsection.

New concept icons also had five icons with max-consensus scores above 0.30; they were #2 searches (0.35), #3 zooms (0.33), #18 opens a calendar (0.36), #30 creates a new folder (0.30), #37 opens a presentation application (0.35). However, there were two *new concept icons* with 0.10 or less max-consensus scores, #31 opens programs (0.07) and #34 opens energy saver options (0.10).

The *new concept icons* seem to be a viable update for the anachronistic icons in some cases like the anachronistic icon #2 searches that was represented by the anachronistic magnifying glass–7% of the participants never used a magnifying glass. The icon had an identifiability score of 0.48 from our young adults, yet they elicited a *new concept icon* for it that had a 0.73 identifiability score in the online identification study. For icon #18 opens a calendar, even though the anachronistic icon representing a desk calendar was identified correctly by all 30 participants and had a perfect function agreement score of 1.00, it had a new concept icon that had similarly high function agreement score of 0.97 in the online identification study. Given the fact that 18% of our young adult participants had never used a desk calendar, the new concept of a digital calendar seems like an appropriate update to this icon. Icon #30 create a new folder, had the icon that represents a folder, an object that all our participants had used. The young adult participants updated this icon by incorporating the folder and adding a (+) sign to it. This update seems to have enhanced the icon given that its referent agreement score in the online identification study went up to 0.45.

New concept icons improved the identifiability scores even when the consensus on what the new icon should be was low. This was the case for icons #31 opens programs and #34 opens energy saver options; both had low agreement scores in the elicitation study of 0.70 and 0.10 respectively. The *new concept icons* had higher identifiability than their anachronistic counterparts, as #31 had an identifiability score of 0.01—even though the objects, a box and a compact disk, was used by all the participants—that identifiability increased to 0.65 for the #31 *new concept icon*. Icon #34 opens energy saver options that was represented by a light bulb had 0.47 identifiability that increased to 0.66 for the *new concept icon* of a battery.

On the other hand, there were cases where the young adult participants elicited a *new concept icon* that was identified correctly in our online identification study, which warranted its inclusion in our final set of icons, but its anachronistic counterpart was more identifiable. An example is icon #37 opens a presentation, which had an elicitation max-consensus score of 0.35 depicting a screen showing a bar graph. In the online identification study, the new concept icon had a 0.29 identifiability score. The anachronistic icon for #37, a podium, an object that was used by all the participants, had high identifiability of 0.80. Another example is icon #3 zooms, represented by a magnifying glass. It had a new concept icon of two circles with the signs (-) and (+). This new concept icon had an identifiability score of 0.37. Its anachronistic counterpart, the magnifying glass, had an identifiability score of 0.52 among our young adult participants.

6.2 **Anachronism**

The results of my investigations culminated in a final set of icons informed by three user studies (Figure 33). About half of the final set of icons—19 of 39—fit what I initially thought of as a plausibly anachronistic icon. The number of anachronistic icons that ended up in the final set is consistent with the total number of anachronistic icons elicited in the lab-based elicitation study with the 30 young adult participants, 1,965 of the 3,590 total icons (55%.) Despite the fact that the 30 young-adult participants identified all of the 38 real-world objects represented in the plausibly anachronistic icons—25 objects were identified correctly by all participants, and the remaining 13 were identified by more than half of the participants—all of the young adult participants had only personally used 16 of these objects. On average, a quarter of the young adults had never used the objects depicted in 23 of the 39 anachronistic icons.

I further explore anachronism as it relates to the set of 39 icons and divide it into three categories: (1) *truly anachronistic icons* are those that represent objects that some of my young adult participants have not used and proposed new concept icons to trigger their functions; (2) *usable anachronistic icons* are icons that represent objects that some of the young adults have not used but proposed icons that still represent the same anachronistic icon; (3) *not anachronistic icons* are ones which I have initially classified as anachronistic, but I found that at least one of our young adult participants had used the object depicted in the icon.

Truly anachronistic icons are those that represent objects that some of the young adults have not used and proposed new concept icons to trigger their functions. There were 13 truly anachronistic icons: #2. searches, #3. zooms, #11. opens video chat, #17. opens date and time settings, #18. opens a calendar, #19. opens dictation and speech settings, #20. saves, #21. pastes, #28. copies to, #32. opens power settings, #34. opens energy saver options, #36. indicates notifications, and #38. opens file history. These *New Concept Icons* on average had a max-consensus rate of 0.25 in the elicitation study. The anachronistic icons had an average of 0.47 function agreement in the in-lab identification study. Their *New Concept Icons* had an increased average function agreement score of 0.54.

Usable anachronistic icons are icons that represent objects that some of the young-adult participants had not used, but for which they proposed icons that still represented the same anachronistic icon. There were 10 *usable anachronistic icons* that depicted anachronistic objects: #4. stores contact information, #5. displays system settings, #6. opens e-mail accounts, #10. creates a reminders list, #12. is a container for storing videos, #14. allows for the look-up of word definitions, #15. opens camera settings, #16. opens sound settings / volume, #24. measures a time duration, and #35. starts a phone call. These *usable anachronistic icons* had an average max-consensus rate of 0.26. The in-lab and online identification function agreement rates for these *usable anachronistic icons* were close at 0.66 and 0.61, respectively. I called these icons *usable anachronistic icons* because even though young-adult participants had never used these objects, the anachronistic icons that depict these objects are still highly identifiable. For example, even though 82% of my participants had never used a cog in real life for icon #5. Displays system settings, all 30 participants were able to identify the correct function associated with the icon. Also, because icon #5 had a 0.82 function agreement in the distributed identification study with the 60

online participants, it seems that having a cog to "display system settings" is still a good icon choice. The same goes for icon #16. open sound settings / volume, which is represented by a speaker. Even though 36% of the young adults never used such a speaker in real life, the function agreement in the identification study with the young adults was 0.80; in the distributed identification study it was 0.73. For icon #35, which used a landline handset to represent the "starts a phone call" function, 29% of the participants had never used a landline in real life. The function agreement rate in the identification study with the young adults was 0.40, less than half of that for the distributed identification study, which was 0.88. For these three icons, I believe their high identifiability and the reason they remained *anachronistic by elicitation* could be attributed to the fact that these same icons represent the same functions in smartphone operating systems. So even though, on average, more than a quarter of the young-adult participants had never used these three objects in real life, they still perceived them to be the best iconic representation of their functions.

Not anachronistic icons are icons that I initially classified as anachronistic but found that at least one of the young-adult participants had used the object depicted in the icon. These 16 icons were: #1. opens a view to store and organize files, and applications, #7. is a simple text application, #8. allows for the viewing of photographs, #9. creates quick notes which pin to the home screen, #13. opens an eBook reader, #22. cuts, #23, contains deleted files, folders, and applications, #25. opens privacy preferences, #26. opens email account settings, #27. manages passwords, #29. opens file explorer, #30. creates a new folder, #31. opens programs, #33. opens storage options, #37. opens a presentation application, and #39. opens a web browser. For the *not anachronistic icons*, the average max-consensus score in the elicitation study was 0.24. The in-lab identification study had a function agreement rate of 0.60, and the online function agreement rate was 0.49.

I discovered that folders and scissors are items still being used by today's young adults: all of the 30 participants reported that they have used these physical objects. These two objects also work well as icons associated with the computing functions they represent: folders for #1. opens a view to store and organize documents, files, and applications, and scissors for #22. cuts. From the lab identification study, icon #1 had 0.41 function agreement and icon #22 had 0.93 function agreement. Their online identifiability

scores were 0.52 and 0.58, respectively. These two icons are examples of icons that do not need to change due to a lack of identifiability.

Even though they were *not anachronistic*, 8 of 16 *not anachronistic icons* had *new concept icons*, which indicates, to me, that our computer iconography needs updating even if some of these icons represent physical objects we still use. These icons were: #7. is a simple text application, #8. allows for the viewing photographs, #23. contains deleted files, folders, and applications, #27. manages passwords, #29. opens file explorer, #30. creates a new folder, #31. opens programs, and #37. opens a presentation application.

6.3 Identifiability of the User-Generated Set of Icons

The set of user-generated icons that resulted from the elicitation study was highly identifiable, at an average function agreement rate—or identifiability score—of 0.54. In fact, despite that in the distributed identification study conducted using the Crowdlicit system I showed the icons alone without any context or labels, only five icons' computing functions were *not* identified correctly. I reverted these icons to their original plausibly anachronistic icons in the final set of icons (Figure 33) because their anachronistic counterparts were identified correctly in the lab-based identification study. A single exception was the icon for function #19. open speech and dictation settings. It was not identified correctly in either identification study. Perhaps this is because the function of opening speech and dictation settings is not as widely used as other functions.

The icons with the highest identifiability scores were #18. opens a calendar and #20. save, with over 0.90 function agreement rates. Icons #18 and #20 were *new concept icons* that included alphanumeric text, which explains the high identifiability scores. Other user-generated icons with relatively high function agreement scores of 0.80 and above included #17. opens date and time settings, which was a *new concept icon* from the *truly anachronistic* group that had a 0.88 identifiability score. Icon #35. starts a phone call, a *usable anachronistic icon*, had a 0.88 identifiability score as well. These icons with high identifiability scores support the motivation for this work and demonstrate that user-generated icons are viable updates to anachronistic icons.

The icon with the lowest identifiability score (0.10) was #21. pastes, represented by a glue bottle, which was a *new concept icon* with 0.22 proposal agreement rate among the young-adult participants. The glue

bottle replaced the clipboard icon, which had a 0.17 identifiability score. Clearly, further design work is needed to create a highly identifiable icon for the paste function.

Other icons with relatively low identifiability scores were #27. manages passwords, at 0.17, and #10. creates a reminders list, at 0.18. Icon #27 was a *new concept icon*—a password field that could be found on websites—elicited with 0.28 max-consensus. It replaced a set of metallic keys, familiar anachronistic objects, which had a 0.47 identifiability score. This is a case where the original icon (the keys)—which I had initially classified as anachronistic—was more identifiable than the user-generated *new concept icon*, and probably should remain unchanged. As for icon #10, it had a *usable anachronistic icon*, that of a bulleted checklist. The young adults were able to identify it with a 0.40 function agreement rate. The young-adult participants had a max-consensus of 0.22 when eliciting this anachronistic icon, and 25% of them had never used a physical "to do" checklist. In this case, I might attribute the lowered identifiability score to my own simplistic illustration of the icon.

6.4 Icons Due for a Redesign

In this section, I highlight 20 icons currently found in desktop operating systems and applications that could benefit from an update as a result of this work. First, *truly anachronistic icons* need an update immediately as they represent anachronistic objects that young adults have never used and proposed *new concept icons* for. These *truly anachronistic icons* are: #2. searches, #3. zooms, #11. opens video chat, #17. opens date and time settings, #18. opens a calendar, #19. opens dictation and speech settings, #20. saves, #21. pastes, #28. copies to, #32. opens power settings, #34. opens energy saver options, #36. indicates notifications, and #38. opens file history. These icons were effectively redesigned in my elicitation study, the results appearing in Figure 33.

The second set of icons that could benefit from an update in the near future is the *usable anachronistic icons*. While they are currently identifiable icons for the functions they trigger, they still represent objects that young adults no longer use. As a result, they might be headed towards "truly anachronistic" status with the passage of more time. These icons are: #5. displays system settings, #4. stores contact information, #10. creates a reminders list, #12. is a container for storing videos, #16. opens sound settings / volume, #24. measures a time duration, and #35. starts a phone call.

The final set of icons (Figure 33) could serve as inspiration for redesigning these icons. I generated this final set of icons to exhibit the results of my multi-part study, but not as a definitive redesign of the 39 plausibly anachronistic icons that served as the basis for this work. Future elicitation studies with larger more diverse populations would be beneficial in informing the design of the 20 icons I highlighted in this section. Perhaps a distributed elicitation study—using **Crowdlicit**—involving end users of different backgrounds could inform the design of these icons to be identifiable to a wide population of potential users. In addition, platform-specific considerations and guidelines would be important to consider in any redesign effort.

6.5 **Limitations and Future Work**

As with any study, and especially in the case of three studies, there are certain limitations. Despite the diversity of the 30 young-adult participants' countries of origin, the participants were all people attending a major university in the United States. Most of the participants used technology quite heavily: 6–10 hours per day using a computer, 3–5 hours per day using a mobile device, and less than 2 hours per day using certain physical objects. It would be interesting to replicate this study in the future with a larger, more diverse set of young users.

I utilized one of three possible techniques from Morris *et al.* [79] to reduce legacy bias in the elicitation study, which was the "production principle," which asked the participants to sketch as many icons as they could imagine, rather than only one. I also discovered some statistical support that employing this principle did, indeed, reduce the chances for anachronism. A future study could utilize Morris *et al.*'s other legacy bias reduction principles such as "priming"—which I investigate in the **"I Am Iron Man"** chapter—or "partners." For example, the plausibly anachronistic icons could be shown as *prohibited* examples that participants would not be allowed to propose. Or, participants could work together to brainstorm additional icon possibilities. My results showed that 55% of the 3,590 icons we collected and 19 of the 39 icons I had in the final icon set were still plausibly anachronistic. By using additional techniques to prioritize novelty, I might reduce potential anachronism even further. The viability of an icon set resulting from such a study would be a question best answered by an identification study, which could compare that set of icons to the final set of icons I derived in this chapter (Figure 33).

As noted above, none of the computing functions or icons were contemplated in context by participants. Rather, they were always described or depicted in isolation, apart from any graphical user interface in which they might be used. Further validation of any icon set requires contextualizing that set in real user interfaces, whether visual mockups, functional prototypes, or even commercial products. Contextualized guessability studies can then ask participants to predict the behavior of computing systems before icons are clicked. Similarly, if participants are ever surprised at the computing function that results from using an icon, they can suggest a different, perhaps better, alternative.

7 Summary

In this chapter, I sought to understand young adult technology users' perceptions of yesterday's and today's computer iconography. To this end, I conducted a multi-part study that resulted in a set of 39 user-generated icons for common computing functions. Half of these icons were made up of new concept icons and the other half remained anachronistic, indicative of objects or concepts from an earlier era but which nonetheless still have meaning for today's young-adult computer users. My work provides insight into young adults' high reliance on physical objects to depict computer iconography. I provide a taxonomy of computer iconography and highlight 20 anachronistic icons in need of redesign. It is my hope that researchers can utilize the findings from this work to better understand the mental models and perceptions of today's young-adult computer users. I also hope that user interface designers can employ my findings to direct their efforts at updating computer icons for the next generation of computer users. It is no longer the floppy disk that means "Save." Rather, it is "Save" that is represented by a small abstract square.

Eight | "I Am Iron Man"

Priming with Sci-Fi Videos Improves Learnability and Memorability of User-Elicited Gestures

Priming is a technique that has been proposed [79] as a way of increasing the creativity and diversity of proposals from end-user elicitation studies, but to date, priming has not been investigated thoroughly in this context. I conduct an online distributed end-user gesture elicitation study with 167 participants, which had three priming groups: a no-priming control group, a sci-fi priming group, and a creative mindset priming group. Then, I evaluated the gestures proposed by these groups in a distributed learnability and memorability study with 18 participants to see if priming made a difference. I found that the user-elicited gestures from the sci-fi group were significantly faster to learn, requiring an average of 1.22 viewings to learn compared to 1.60 viewings required to learn the control gestures, and 1.56 viewings to learn the gestures elicited from the creative mindset group. In addition, 80% of the sci-fi-primed gestures were recalled correctly after one week without practice, compared to 43% of the control group gestures and 73% of the creative mindset gestures.

1 INTRODUCTION

This chapter focuses on extending the work of Morris *et al.* [79] in which they point out a pitfall of elicitation studies they call *legacy bias*. This pitfall, as they define it, is when "participants in elicitation studies propose familiar interactions from current technologies that might not be best suited for new technologies' form factors or sensing capabilities." Morris *et al.* [79] proposed principles to reduce legacy bias in elicitation studies. One of their principles is *priming,* a practice from the field of psychology used to enhance creative thinking. The effects of priming in elicitation studies have been explored a little in prior work. In one example, Hoff *et al.* [41] found that priming results in fewer legacy gestures and quicker generation of ideas; however, their results were not statistically significant and they stated that given the typically small number of participants in (traditionally lab-based) elicitation studies, they do not recommend the use of priming.

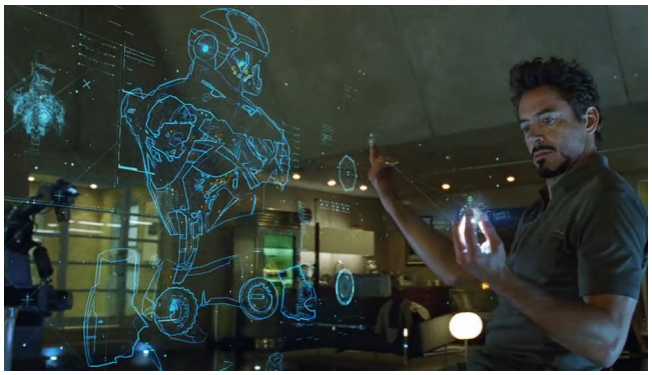


Figure 34. A still from the movie Iron Man 2 [48], showing the main character Tony Stark interacting with an augmented reality hologram interface with hand gestures.

I utilized my methods and tools to run distributed elicitation studies efficiently to explore the effects of priming with a large number of participants. For this exploration, I used mixed reality (MR) environments as a use case due to their novelty as a technology for end users. These environments are on the cusp of becoming a mainstream but have yet to be widely adopted by the average technology user. I used priming to push beyond legacy interactions informed by desktop or mobile computing and to elicit interactions for an MR environment that are easily learnable and memorable. I employed priming in two ways: (1) viewing of sci-fi clips, (2) having a creative mindset. For the sci-fi group, I primed participants by having them view a montage of sci-fi films depicting characters interacting with technologies using gestures. For

this primer, I drew inspiration from science fiction movies such as Iron Man—the namesake of this chapter—which depict fascinating human-computer interactions. It is no wonder there exists a bidirectional relationship between sci-fi and HCI [105], where each endeavor, at its best, showcases brilliant new possibilities for people's use of technology. Schmitz *et al.* [103], in their survey of HCI in sci-fi movies, reported that there exists a collaboration between filmmakers and scientists regarding the use of HCI in film. They mentioned that director Steven Spielberg consulted with HCI researchers to develop the system and interactions shown in his movie *Minority Report*. Larson [57] stated that sci-fi depictions of technologies mirror trends in real life computing. Aaron Marcus, in his article "The History of the Future: Sci-Fi Movies and HCI" [68], stated that sci-fi movies can be a useful material to inform designers of possible future technological, social, or cultural contexts. Mubin *et al.* [83] cited many examples of devices and products that have roots that can be traced back to sci-fi movies, like mobile phones inspired by the communicators from *Star Trek*, and video conferencing similar to that depicted in *2001: A Space Odyssey*. The recurring "Future Tense" section of *Communications of the ACM* often features sci-fi writers like David Brin who, in the words of the magazine, "present stories and essays from the intersection of computational science and technological speculation" (*e.g.,* [20]). Not incidentally, David Brin also gave the ACM CHI 2002 keynote address, drawing on themes explored in his sci-fi writings to inspire an audience of HCI researchers. The work presented in this paper contributes to this body of literature by investigating the effects of using sci-fi as a primer to influence the design of future interactions.

For my second primer, I looked to the field of psychology where the concept of priming originates back to 1951 [59]. I followed Sassenberg and Moskowitz's [102] practice of using a "creative mindset" to suppress stereotyping and encourage participants to "think different." In two experiments Sassenberg and Moskowitz observed that automatic activation of stereotyping was not observed in a creative mindset.

I conducted a between-subjects elicitation study with priming as our independent variable. I recruited 167 participants—more than eight times the number of participants in a typical lab-based study (~20) [125]—from Amazon's Mechanical Turk (mTurk) platform. I randomly assigned participants to one of three groups: control (no priming), sci-fi priming, and creative-mindset priming. Participants were prompted with 10 functions of a media player in an MR environment and were asked to propose a mid-air

gesture to trigger each of the 10 functions. As a result, I formulated three sets of user-elicited gestures. I found overlap in four gestures across all three groups. The sci-fi gesture set had only five gestures in common with each of the other two gesture sets. On the other hand, seven of the ten gestures in the creative-mindset set were identical to the control. The number of unique gestures in the sci-fi set indicates that sci-fi priming enhances creativity and leads to the creation of more novel gestures than no priming or the creative-mindset technique [102].

I followed my elicitation study with a distributed end-user identification study with 50 new participants from mTurk. The identification study showed that priming had no statistically significant impact on the number of correctly identified gesture-function relationships of user-elicited gestures.

I capitalized on **Crowdlicit** to run supervised distributed interaction-evaluation studies, expanding the fourth step of the DXD process, "test interactions' quality," by moving beyond interaction identifiability to include learnability and memorability. I recruited 18 new online participants for a two-part study to evaluate the three gesture sets. In the first part of this study, participants were randomly assigned to view one of the three user-proposed gesture sets (control, sci-fi, creative mindset). After viewing a video clip of each gesture in the set once, participants were prompted with a function and asked to perform the corresponding gesture they had just learned. After going through all 10 functions, participants were allowed to go back and view the video clips of the gestures that they got wrong, if any. I repeated this process until the participants learned and were able to correctly perform all 10 gestures in their set. Furthermore, after one week, I contacted the same participants and asked them to go through the testing protocol only once to assess the memorability of the gestures—without allowing them to review the video clips of the gestures beforehand. I found that sci-fi-primed gestures were faster to learn, as they required an average of 1.22 viewings to learn. Non-primed gestures required an average of 1.60 views and the creative-mindset primed gestures required 1.56 viewings to learn. After a single viewing, 80% of the sci-fi primed gestures were learned compared to 65% of the gestures from the control and 58% from the creative-mindset sets—sci-fi gestures were learned most quickly. Additionally, sci-fi gestures had a higher memorability accuracy compared to the other two gesture sets, with 80% of the sci-fi gestures recalled correctly compared to 73% and 43% for the creative mindset and non-primed sets, respectively.

The work in this chapter contributes: (1) an empirical study of the effects of *priming* in elicitation studies, (2) a user-elicited gesture set for a media player in a mixed-reality environment, and (3) methods to evaluate the learnability and memorability of interaction designs in a distributed manner.

2 The Effects of Priming on User-Elicited Gestures

To evaluate the effects of priming on the learnability and memorability of user-elicited gestures, I first conducted a distributed elicitation study accompanied by an identification study. The elicitation study resulted in three gesture sets—one for each level of priming (control, sci-fi, and creative mindset). The identification study subsequently evaluated how guessable the gestures were in each priming group.

2.1 **Creating a User-Elicited Gesture Set**

I conducted a between-subjects distributed elicitation study with 167 online participants to design gestures for a media player in an MR environment.

2.1.1 Participants

I recruited 167 participants in total from Amazon's Mechanical Turk (mTurk), to provide video recordings of their proposed gestures in response to 10 prompts showing functions (see Table 21) of a futuristic media player using **Crowdlicit**. I followed the elicitation study with a distributed identification study. To do so, I recruited 50 new participants from mTurk. In both studies, each participant filled out a demographic survey upon completing the study. Table 20 shows the demographic information for both studies. Participants needed to have a device with a camera (*i.e.,* a webcam or a smart phone) to participate in the elicitation study. Participants in the elicitation study were compensated $7.50 for participation in the half-hour study. Participants in the 15-minute identification study were compensated $3.75. I based the compensation on Seattle, Washington's $15/hour minimum wage rate.

Demographic		*Elicitation (N=84)*	*Identification (N=33)*
Gender	Man	59	20
	Woman	24	13
	Non-binary	1	0
Age	Mean (*SD*)	30.69 (5.61)	38.03 (10.39)
Nationality	USA	71	24
	India	7	4
	Brazil	3	3
	Canada	1	1
	Germany	1	0
	Pakistan	1	0
	Italy	0	1
Do you own an MR device?	Yes	15	9
	No	69	24
How often do you use an MR device?	Never	29	14
	Daily	2	4
	Monthly	20	5
	Once or twice ever	33	10
Do you use mid-air gestures?	Yes	7	32
	No	77	1
Favorite movie genre	Comedy	19	7
	Action & adventure	15	5
	Drama	5	4
	Horror	6	1
	Sci-fi	18	9
	Documentary	7	2
	Thriller	10	2
	Epic/ historical	2	1
	Musicals	2	1
	Western	0	1
Have you seen this movie before?	Minority Report	39	19
	Iron Man 2	62	28
	Black Mirror	43	16
	Gamer	11	4
	Enders Game	31	10

Table 20 Participants' demographic information from the elicitation study with 167 participants (84 filled out demographics survey) and identification study with 50 participants (33 filled out the survey).

2.1.2 Apparatus

I used **Crowdlicit** to run the distributed elicitation and identification studies. I created video clips of 10 functions to interact with a media player in an MR environment. The 167 online participants viewed these video clips as prompts to propose gestures that would trigger the functions shown in the videos. Participants recorded their gesture proposals using the web interface and their personal computer's webcam or via the camera on their mobile device, then uploaded the footage to Crowdlicit. This process

resulted in 1,168 totally gestures. For each one of the ten prompts, I found the gesture with the highest consensus. This process resulted in 15 unique gestures across the three priming groups.

I then used **Crowdlicit** again in the identification study with 50 new online participants. The identification study showed the video clips of the 15 elicited gestures as prompts. For each gesture, participants entered text descriptions of the functions they anticipated the gesture would trigger in an MR media player. I also captured self-reported Likert-type ratings after every gesture or function proposal. In addition, **Crowdlicit** also captured the time participants needed to submit a proposal (in seconds).

To prime participants in the elicitation study who were assigned to the sci-fi priming condition, I created a short montage film from movies and TV shows like *Iron Man*, *Minority Report*, and *Black Mirror*. I created this montage film using the open catalog of gestures in sci-fi movies by Figueiredo *et al.* [28]. Their catalog[8] has tags of what task is being performed in the clip (*e.g.,* play, previous, etc.).

2.1.3 Procedure

Following the first four steps of **The Six Steps of DXD**, I conducted a distributed elicitation study and followed it up with a distributed identification study to produce three user-generated gesture sets from the three levels of *Priming*.

Distributed End-User Elicitation

The 167 participants who accepted the human intelligence task (HIT) on mTurk were directed to a custom webpage that randomly assigned them to one of three priming groups (control, sci-fi, and creative mindset). Based on the assignment, participants were automatically directed to a Crowdlicit study URL. This setup allowed me to organize the elicited gestures into three groups (control, sci-fi, creative mindset). Upon navigating to the unique Crowdlicit study link, participants were presented with instructions for the study explaining that they were about to watch 10 video clips of functions for a media player in a mixed reality (MR) environment (Table 21) and that such a system responds to mid-air gestures. The participants were required to propose a gesture of their choosing to trigger each of the 10 functions depicted in the video clips. The instructions also showed a diagram instructing the participants

[8] http://goo.gl/XSX5fn

on how to position themselves in front of the camera in such a way that would exclude their face from showing in the recording to protect their privacy.

The participants in the creative mindset group were asked to provide three examples of a time they were creative before starting the elicitation session. I borrowed this technique from Sassenberg and Moskowitz [102].

All three variations (control, sci-fi, and creative mindset) of the elicitation study on Crowdlicit were identical except for the instructions section for the sci-fi group and a pre-session task for the creative mindset group. The instructions for the sci-fi group included a section stating: "The environment should respond to gestures like, but not limited to, ones shown in this video." Below that message was a montage film of sci-fi clips[9]. I chose one clip from a tag that represents each function in our list of functions (Table 21) and compiled them into the montage film.

No.	*Function*
1.	Play a video
2.	Pause a video
3.	Fast forward
4.	Rewind
5.	Next video
6.	Previous video
7.	Close a video
8.	Pin view to a surface
9.	Bring view into field of vision
10.	Activate headtracking

Table 21 A list of 10 functions to control a media player in a mixed-reality environment. "View" refers to the video element. The last three functions are specific to an augmented reality environment.

Distributed End-User Identification

Whereas the elicitation study enabled me to show 10 MR functions and elicit 15 unique gestures to trigger them, an identification study enabled me to test how guessable those 15 gestures were. The 50 new participants who accepted the HIT on mTurk for the identification study were redirected to a custom study webpage created by the Crowdlicit system. The identification study had 15 unique gesture prompts. These gestures were the results of the elicitation study, which resulted in 15 unique gestures for the 10

[9] The video material for this study can be found here: http://crowdlicit.ischool.uw.edu/Crowdesign/ironmanproject.php

functions with overlap across the three priming levels. Figure 35 shows these gestures and which ones overlap. Each prompt in the identification study showed a video of a person—myself—performing one of the 15 gestures that resulted from the elicitation study. After viewing a video of a gesture, the participants were asked to propose the function they thought a system would trigger in the context of interacting with an MR video player.

2.1.4 Design and Analysis

The elicitation study was a single-factor between-subjects design, whose factor was *Priming*, which had three levels: no priming (control), sci-fi priming, and creative-mindset priming. In this study, I collected 1,167 gesture proposals from a total of 167 online participants. Due to a server error, 12 gestures were not recorded from the control group, leaving the control group with a total of 381 gestures; sci-fi and creative-mindset groups each had 393 gestures.

The identification study was a single-factor within-subjects design, with the same priming factor and levels as the elicitation study. In this study, I collected 750 function proposals from our 50 online participants. As a within-subjects study, the identification study showed each participant gestures from all three priming groups.

I investigated the effect of *Priming* on the results of the distributed elicitation and identification studies on three dependent variables: (1) Agreement scores for proposed gestures and functions calculated using Equation 1. (2) Self-reported Likert-type satisfaction ratings (ease, match, enjoyment). (3) Proposal time—*i.e.,* the time it took participants to come up with a gesture or a function. In addition, I explored subjective differences across the three user-elicited gesture sets that we created from the elicitation study. I also compared identification accuracy across the three gesture sets in our identification study.

To calculate a gesture agreement score, I used Equation 1, which was updated in these publications [30,120,123]; however, due to the fact that I collected a single gesture proposal per prompt in the same manner as Wobbrock's *et al.'s* original paper [131], I opted to use the original equation in my analysis. Furthermore, I used that equation to calculate function agreement in identification studies. To provide a sense of uniformity for this chapter's analysis, I decided that Equation 1 is the best fit. Agreement scores

have an upper limit of 1.0. The upper limit represents total agreement in which all the proposals collected in response to a prompt match each other.

Because agreement scores are bounded, I used a non-parametric Kruskal-Wallis test [56] to assess the differences in agreement scores among the three levels of *Priming*. I followed a significant omnibus test with *post hoc* comparisons using a pairwise Mann-Whitney *U* test [67], corrected with a Tukey HSD test [12]. To investigate differences in the ordinal Likert-type self-reported ease, match, and enjoyment ratings, I used mixed ordinal logistic regression [2,40]. Then, I conducted *post hoc* analyses on any significant omnibus tests using a Tukey HSD test [12]. To investigate the effect of *Priming* on the time it took participants to propose a gesture or a function, I used a linear mixed model analysis of variance [32,64]. I log-transformed the time response prior to analysis, as is common [63], to comply with the assumption of conditional normality. I followed up any significant omnibus test with a *post hoc* Tukey HSD test [12].

2.1.5 Results

I identified the gesture with highest agreement for each one of the 10 functions in each of the priming groups (control, sci-fi, and creative mindset). If all gestures were unique, this would result in $10 \times 3 = 30$ gestures, but due to substantial overlap among the groups, there were 15 unique gestures in all. These gestures are shown in Figure 35, and served as prompts in the subsequent identification study.

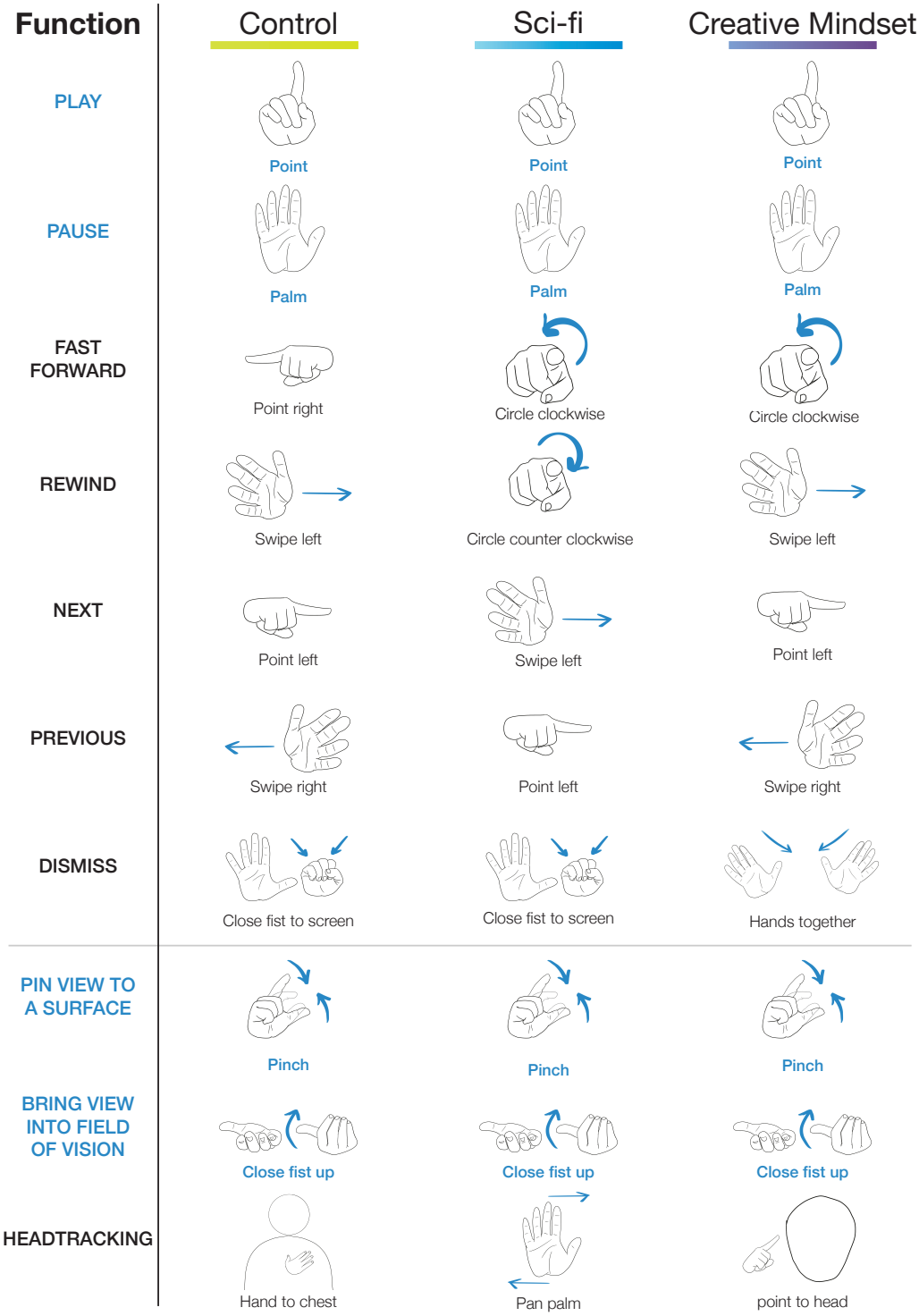


Figure 35 Three gesture sets for 10 functions. Functions in blue show that the same gesture among the three sets was proposed by the majority of participants in that group. The last three functions are specific to an augmented reality environment

2.1.6 User-Elicited Gesture Sets

I compared gestures for each function across the three sets for their similarity and found that four functions (play, pause, pin view to a surface, and bring view into field of vision) had the exact same gestures across the three gesture sets (point, palm, pinch, close fist up, respectively). The sci-fi gesture set had five gestures in common with the control set that triggered the same function. The creative-mindset set had seven in common with the control set, leaving only three functions triggered by a different gesture (fast forward, dismiss and headtracking). The sci-fi and creative-mindset gesture sets had five gestures in common (the four gestures found across all three gesture sets and the gesture "circle clockwise" to trigger the fast forward function).

I compared the gestures in the sci-fi group against the gestures appearing in the sci-fi montage video that I used as a primer to see if any of the gestures in the video were replicated by participants. The "circle finger clockwise" and "counterclockwise" gestures were present in the montage clip (*Black Mirror* clip under the rewind tag) and so was the "pan palm to screen" gesture (*Enders Game* clip under the play tag). Other gestures from the montage like "close fist up", and "swipe" were present throughout all three gesture sets and not limited to the sci-fi group.

2.1.7 Agreement Scores

For the control level of *Priming,* the mean agreement score—the degree to which the participants agreed on a gesture to trigger a function—was $A = .182$ ($SD = .077$); for sci-fi, it was $A = .178$ ($SD = .088$); and for creative mindset it was $A = .186$ ($SD = .084$). A Kruskal-Wallis test found no statistical significance among these scores ($\chi^2(2, N=30) = 0.267$, *n.s.*).

For the identification study, the function agreement scores were also similar across *Priming* levels. For control the mean agreement was $A = .170$ ($SD = .094$); for sci-fi, it was $A = .191$ ($SD = .086$); for creative mindset it was $A = .173$ ($SD = .067$). A Kruskal-Wallis test found no statistical significance in the differences among these scores ($\chi^2(2, N=30) = 0.937$, *n.s.*).

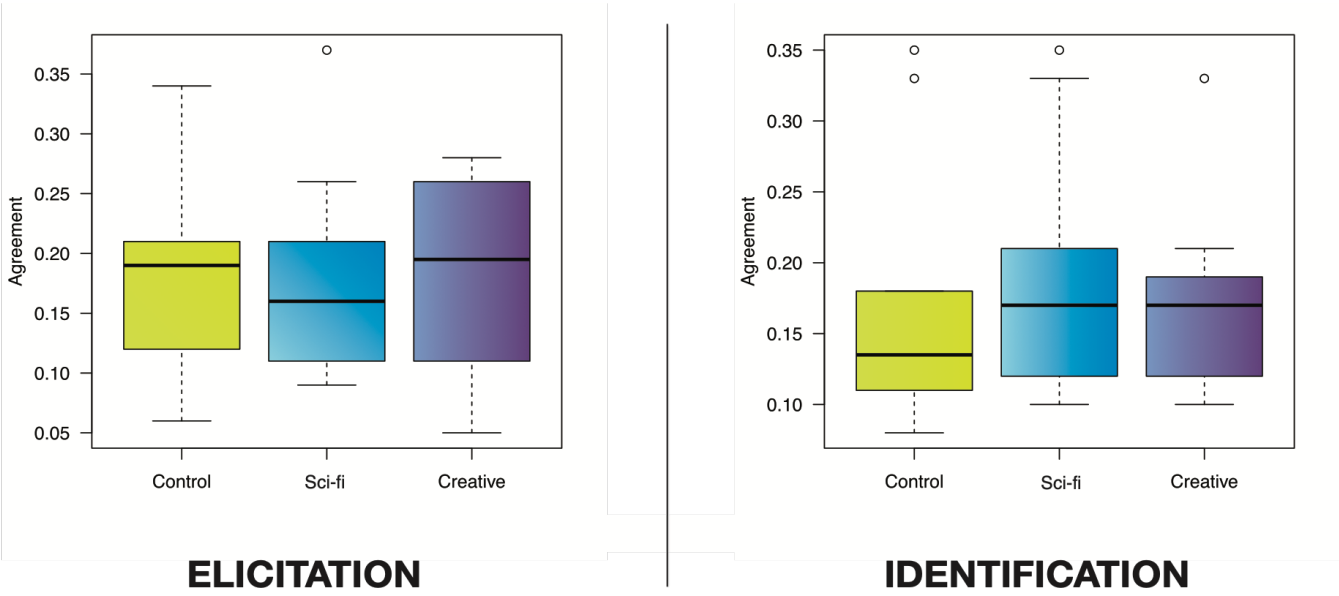


Figure 36. Agreement scores for three gesture sets created under three *Priming* levels (control, sci-fi, and creative mindset).

2.1.8 Priming Effect on Self-Reported Ratings

I collected Likert-type ratings on a scale of 1–7 (1. Strongly disagree, 7. Strongly agree). The scales assessed the following statements: 1. **Ease**, *my proposal is easy to perform*. 2. **Match**, *my proposal is a good match for its intended purpose.* 3. **Pleasure**, *my proposal is enjoyable to perform.* Admittedly, Likert-type ratings are ordinal in nature and their numeric markings (1–7) cannot be taken as scalar responses. That said, for illustrative purposes and for the reader's appreciation, I report the means and standard deviations of the Likert-type ratings in Table 22, in addition to the median scores and interquartile ranges.

	Priming	Rating	Median (IQR)	Mean (SD)	Priming	Rating	Median (IQR)	Mean (SD)	
ELICITATION	Control	Ease	6 (3)	5.12 (2.47)	Control	Ease	7 (1)	6.07 (1.25)	IDENTIFICATION
	Sci-fi		6 (1)	6.06 (1.19)	Sci-fi		6 (2)	6.00 (1.27)	
	Creative mindset		7 (1)	5.69 (2.21)	Creative mindset		6 (2)	6.02 (1.27)	
	Control	Match	6 (3)	4.97 (2.43)	Control	Match	6 (2)	5.30 (2.15)	
	Sci-fi		6 (2)	5.71 (1.29)	Sci-fi		6 (2)	5.30 (2.13)	
	Creative mindset		6 (2)	5.48 (2.20	**Creative mindset ***		**6 (2)**	**5.87 (1.27)**	
	Control	Pleasure	5 (3)	4.62 (2.34)	Control	Pleasure	5 (3)	4.84 (2.11)	
	Sci-fi		6 (3)	5.48 (1.31).	Sci-fi		5 (3)	4.86 (2.09)	
	Creative mindset		6 (3)	5.08 (2.16)	**Creative mindset ***		**6 (3)**	**5.36 (1.40)**	

Table 22. The self-reported ease, match, and pleasure scores from the 167 participants in the elicitation study and 50 participants in the identification study. Higher numbers mean "easier," "better matches," and "more enjoyable," respectively. *Bold font indicates statistically significant differences using mixed ordinal logistic regression [2,40] ($p < .05$).

The numeric **ease** ratings were higher, on average, for sci-fi (6.06) compared to control (5.12) and creative mindset (5.69) as reported by the 167 participants in the elicitation study. On the other hand, the ease ratings were nearly identical across all three levels as reported by the 50 participants in the identification study. An analysis of variance based on mixed ordinal logistic regression indicated no statistically significant effect on ease ratings of *Priming* in either the elicitation study ($\chi^2(2, N=1167) = 4.41$, *n.s.*) or identification study ($\chi^2(2, N=750) = 3.22$, *n.s.*).

The numeric **match** ratings for sci-fi (5.71) were also higher, on average, than control (4.97) and creative mindset (5.48), according to the participants who proposed the gestures in the elicitation study. However, sci-fi (5.30) and control (5.30) match scores were similar and lower than creative mindset (5.87) as rated by the participants who attempted to identify the function associated with the gesture in our identification study. These differences were not detectably different in the elicitation study ($\chi^2(2, N=1167) = 1.98$, *n.s.*). However, the creative mindset match ratings in the identification study were significantly higher than the other priming groups ($\chi^2(2, N=750) = 8.60, p < .05$). *Post hoc* pairwise comparisons using Tukey's HSD correction indicated that creative mindset *vs.* control ($Z = 2.44, p < .05$) and creative mindset *vs.* sci-fi ($Z = -2.60, p < .05$) were statistically significant. The sci-fi *vs.* control match ratings were not detectably different ($Z = -0.14$, *n.s.*).

Finally, the numeric **pleasure** ratings had a similar outcome to the match ratings, with sci-fi gestures (5.48) rated higher than control (4.62) and creative mindset (5.08) gestures in the elicitation study, but showing no detectable differences ($\chi^2(2, N=1167) = 3.42$, *n.s.*). In the identification study, the creative mindset gestures (5.36) had significantly higher pleasure ratings ($\chi^2(2, N=750) = 8.67, p < .05$) than

the almost identical ratings of sci-fi (4.86) ($Z = -2.457$, $p < .05$) and control gestures (4.84) ($Z = 2.60$, $p < .05$).

2.1.9 Priming Effect on Elicitation Time

The mean time needed by participants to provide a gesture in the control group was the highest at 58.3 seconds (*SD* = 59.2). The creative-mindset group was second at an average 49.3 seconds (*SD* = 41.8), and the sci-fi group had the fastest elicitation time with an average 46.5 seconds (*SD* = 46.1). Despite the effect of priming on lowering the mean elicitation time, a linear mixed-effects model analysis of variance indicated no statistical significance in time differences ($F(2, 123.81) = 1.617$, *n.s.*).

As for the time needed to identify the function associated with a gesture in the identification study, the results were very close across all three priming groups—control: 38.0 seconds (*SD* = 34.6); sci-fi: 38.7 (*SD* = 41.7); creative mindset: 39.4 (*SD* = 39.0). There were no statistically significant differences among these results ($F(2, 14.973) = 0.70$, *n.s.*).

2.1.10 Priming Effect on Identifiability

The percentage of correctly identified gestures did not differ much across the three levels of *Priming*. The control gesture set had 22.2% (*SD* = 41.6%) of its gestures identified correctly. The sci-fi gesture set had 24.2% (*SD* = 42.8%) identified correctly, and the creative-mindset gesture set had 22.2% (*SD* = 41.47%) of its gestures identified correctly by the 50 participants in the distributed identification study. Priming had no statistically significant effect on the chances of a gesture being correctly identified across the three levels, according to an analysis of variance based on mixed logistic regression [36] ($\chi^2(2, N=750) = 3.417$, *n.s.*).

3 Distributed Learnability and Memorability

Of crucial importance to system designers is how *learnable* a set of gestures is—and once learned, how *memorable* those gestures are. In this phase of my investigation, I sought to find out whether priming affects gesture learnability and memorability. Accordingly, I conducted a two-part supervised distributed study again using **Crowdlicit** to evaluate the learnability and memorability of the three gesture sets I created as a result of the distributed elicitation study.

3.1 **Participants**

I recruited 18 new participants using convivence and snowball sampling by advertising the study on the university's Slack channels, and on social media platforms. Two of my participants failed to complete the demographics survey. Of the 16 participants who did complete the demographics survey, nine were women, six were men, and one non-binary. The mean age was 27.3 years (SD = 4.84). The participants' nationalities were mostly from the United States (10/16); other nationalities included India, China, and Kazakhstan. Seven participants had never used an MR device, and five had only used one once or twice. Two participants used an MR device on a monthly basis and two others used one on a daily basis. As for participants' use of mid-air gestures to interact with technologies, only two participants reported having used mid-air gestures to interact with a desktop music app and an Xbox Kinect.

3.2 **Apparatus**

I used **Crowdlicit**, once again utilizing its web-based video recording capabilities to collect, store, and organize the data in our study. I used Google Meet to video-call my participants and guide them through the procedures of the studies. For the first part of the study, which was devoted to learnability, I created a custom learning website for participants, which came in three versions corresponding to the three gesture sets I created as a result of our elicitation study (control, sci-fi, and creative mindset). In each version, the website displayed 10 videos, each depicting a gesture from the corresponding set.

3.3 **Procedure**

Once I connected with a participant over Google Meet, I asked them to share their screen and directed them to the custom learning site. The page displayed all 10 videos of one of the three gesture sets shown

in Figure 35 in a random order. Participants were asked to view each video once and then navigate via a Crowdlicit-generated link to perform all 10 gestures. I used **Crowdlicit** to prompt participants with a function and asked them to record a video of themselves performing the gesture corresponding to that function, which they had just learned on the page with 10 videos. I monitored the video recordings as they were being uploaded to the Crowdlicit system and assessed their correctness. I then refreshed the custom learning page, removing the videos of the correctly performed gestures, leaving only videos of the gestures that the participant missed. Participants repeated the learning and performing process until they learned all 10 gestures successfully. I recorded how many times the videos had to be viewed before all 10 gestures were learned successfully to measure gesture learnability.

To assess memorability, all 18 participants from the learnability study were informed they would be contacted via Google Meet one week after the learnability session to perform the gestures again using **Crowdlicit**. Participants did not have access to the learning page to practice the videos in the interim period.

3.4 Design and Analysis

Both the learnability and memorability studies used single-factor between-subjects designs, with a factor for *Priming* having three levels: no priming (control), sci-fi priming, and creative-mindset priming. I collected a total of 263 gesture trials from our 18 online participants. The learnability study had the following dependent variables: (1) initial learnability—the number of gestures learned after a single viewing of all 10 gesture videos; (2) overall learnability—the total video views needed to learn a gesture; and (3) learned-gesture performance time.

I used mixed logistic regression [36] to analyze the dichotomous results of the first trial, i.e., whether a gesture was performed correctly or incorrectly after viewing each of the 10 gesture videos once. I carried out post hoc testing using Tukey's HSD test [116,117] for multiple comparisons. For the total number of views required to learn a gesture, overall learnability, I used a nonparametric Aligned Rank Transform procedure [130], corrected with Tukey's HSD correction [116,117]. Again, post hoc pairwise comparisons were conducted using Tukey's HSD correction. For analyzing gesture performance time, I used the same analysis approach as in the elicitation study.

In the memorability study, I collected 18 × 10 = 180 gesture trials from our 18 online participants. I used the same analysis approach from the learnability study to evaluate two dependent variables: (1) the number of correctly recalled gestures, analyzed in the same manner as initial learnability; (2) the time to recall and perform a gesture, analyzed in the same manner as the learned-gesture performance time.

3.5 **Results**

I evaluated the effects *Priming* had on the learnability of user-elicited gestures in terms of initial learnability, overall learnability, and learned-gesture performance time. For the effects of priming on memorability, I evaluated the gestures based on the number of correctly remembered gestures and gesture-recall time.

3.6 **Initial Learnability**

Initial learnability is the number of gestures performed correctly (out of 10) after one viewing each of the 10 gesture videos in each priming set. The number of learned gestures after a single viewing for the control set was 39 gestures (out of 60 total gestures, or 65%); for the sci-fi set it was 48 gestures (80%); and for the creative mindset set it was 35 (58%). These differences were marginally significant for *Priming*'s overall effect on initial learnability ($\chi^2(1, N=180) = 5.75, p = .056$). *Post hoc* pairwise comparisons using Tukey's HSD correction indicated that creative mindset gestures were significantly less learnable initially than sci-fi gestures ($Z = -2.36, p < .05$).

3.7 **Overall Learnability**

Overall learnability is the total number of gesture video views required to learn all 10 gestures from a given priming group. For the control gesture set, the mean count of viewings required to learn a gesture was 1.60 ($SD = 1.06$); for the sci-fi gesture set it was 1.22 ($SD = 0.45$); and for the creative-mindset gesture set it was 1.56 ($SD = 0.87$). *Priming* had a statistically significant effect on overall learnability ($F(2, 38.5) = 3.35, p < .05$). *Post hoc* pairwise comparisons using Tukey's HSD correction showed that the difference between sci-fi and creative mindset was marginally significance ($t(267) = -2.40, p = .062$). Specifically, sci-fi gestures seemed easier to learn than creative-mindset gestures.

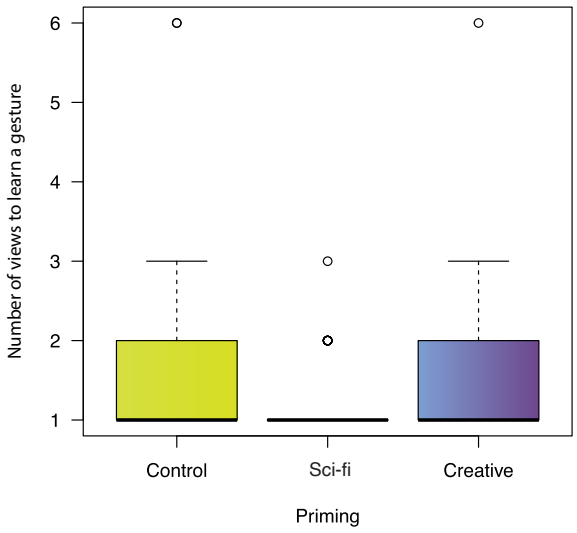


Figure 37 A boxplot of the number of viewings required to learn a gesture by Priming.

3.8 **Memorability**

For each level of *Priming* (control, sci-fi, and creative mindset), I assessed the memorability of the gesture set—*i.e.,* whether or not a gesture was recalled and performed correctly one week after the learning session, without continued exposure between sessions. Each priming condition had 6 participants (3 conditions × 6 participants) who were asked to recall 10 gestures one week after learning them. Out of the 60 gestures for the control group (6 participants × 10 gestures), only 26 gestures (43%) were recalled correctly. For the primed groups, the percentage of correctly recalled gestures increased significantly, with 48 of the 60 sci-fi gestures (80%) recalled correctly, and 44 of the 60 (73%) creative mindset gestures recalled correctly. Indeed, *Priming* had a statistically significant effect on memorability ($\chi^2(1, N=180) = 11.54, p < .01$). *Post hoc* pairwise comparisons using Tukey's HSD correction indicated that the control group gestures were recalled significantly less than either the sci-fi gestures ($Z = 3.17, p < .01$) or the creative-mindset gestures ($Z = 2.57, p < .05$).

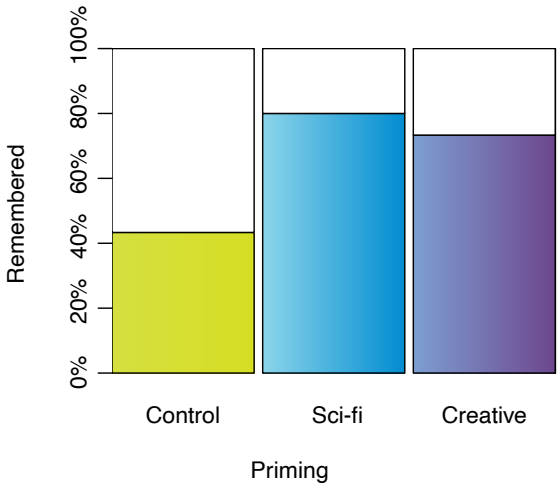


Figure 38 A bar chart of the percentage of correctly remembered gestures by Priming.

3.9 **Gesture Performance and Recall Time**

In the learnability study, the mean time to perform a gesture successfully after learning it by watching a video clip was 17.3 seconds (*SD* = 10.3) for the control gesture set; for the sci-fi gesture set it was 16.6 seconds (*SD* = 11.9); and for the creative-mindset gesture set it was 16.8 seconds (*SD* = 12.5). These differences were not statistically significant ($F(2, 14.97) = 0.25$, *n.s.*).

In the memorability study, the mean time to recall and perform a gesture from the control gesture set was 18.8 seconds (*SD* = 10.6); for the sci-fi gesture set, it was 14.0 seconds (*SD* = 7.3); and for the creative mindset gesture set, it was 13.7 seconds (*SD* = 6.7). These differences were not detectably different ($F(2, 14.99) = 2.19$, *n.s.*).

4 Discussion

In this section, I discuss my findings about the effects of priming across our various studies (elicitation, identification, learnability, and memorability). In particular, I highlight how priming through sci-fi gestures seems to have had the largest effects in some of our studies.

4.1 **Effects of Priming on User-Proposed Gestures**

Sci-fi priming had half of its gestures (5 out of 10) unique only to it. By contrast, creative mindset priming and no priming (control) had 7 out of 10 gestures in common. The sci-fi gesture set showed more cohesion, despite the gestures for each function being elicited independently. This cohesion was visible in functions that have inverse associations like fast-forward and rewind. The sci-fi gesture set contained the gestures "circle counterclockwise" and "circle clockwise." These gestures were present in our sci-fi movie montage clip (*Black Mirror*) that I showed our participants as a primer, and they mapped to the same functions. The control gesture set had "point right" and "swipe right" for fast forward and rewind, respectively. I noticed this lack of cohesion during the learnability and memorability studies, as these gestures were the hardest to learn—*i.e.,* they required the largest number of viewings.

On average, priming seemed to increase the match and pleasure ratings of gestures, both in the elicitation study and in the identification study, although only creative mindset gestures were shown to have statistically significantly higher match and pleasure scores in the identification study. Although the results from the Likert-type ratings were largely non-significant, taken together, there is a trend suggesting that priming produces gestures that are perceived to be a better fit to their functions and more pleasurable to perform. Further research is needed, perhaps with a different set of participants or different types of priming, to confirm this trend.

4.2 **Gestures' Learnability and Memorability**

Sci-fi priming significantly increased the initial learnability of gestures, as 80% of the sci-fi gestures were performed correctly after a single viewing, compared to only 58% of creative mindset gestures and 65% of the control gestures. Sci-fi priming also significantly increased the overall learnability of gestures, with an average of 1.22 views required to learn each sci-fi gesture, compared to 1.60 views for control gestures

and 1.56 views for creative mindset gestures. Furthermore, both sci-fi and creative mindset priming resulted in gestures that were more memorable than control group gestures, with 80% of sci-fi gestures and 73% of creative mindset gestures recalled correctly, but only 43% of control group gestures recalled correctly. Thus, it seems again that priming, and sci-fi priming especially, has advantages in gesture learnability and memorability. Perhaps Sci-fi gestures provided a sense of familiarity that lead to their higher learnability and memorability.

4.3 **Distributed Design Studies**

In this chapter, I demonstrated multiple methods of conducting distributed user-centered design studies typically carried out in a lab. I added two more usability metrics to the distributed interaction-evaluation approach I introduced in the **Crowdlicit** chapter: learnability and memorability. Due to the requirement of providing feedback in learnability studies to participants—so participants would know which gestures they learned and which ones they needed to review and attempt to perform again—I had to conduct the distributed learnability study in a supervised manner. A limitation of supervised distributed user studies is they are slower than unsupervised distributed studies like the elicitation and identification studies. It took me a few hours to recruit and collect data for the unsupervised distributed studies compared to the supervised studies that required multiple days to conduct—plus the one week that separated the learnability and memorability studies. Supervised studies, like my learnability and memorability studies, are hard to run in parallel like the unsupervised elicitation and identification studies. This limitation is the reason why my learnability and memorability studies had only 18 participants like an in-lab study—a number that was sufficient for the investigation in this chapter. Having multiple researchers conducting a supervised study could increase the number of participants. Distributed learnability and memorability studies enjoy other benefits of distributed studies like increased diversity of participants—in terms of both geographical distribution and physical abilities—and discarding physical requirements such as testing labs. In the **Crowdlicit** chapter, I reported that participants are more willing to participate in online studies than lab-based ones. In this chapter, I was able to capitalize on **Crowdlicit** to facilitate all of the studies, collecting, organizing and storing study data, validating this platform's versatility. The platform also provided the participants with an easy-to-use interface to participate in the studies.

4.4 **Limitations and Future Work**

These studies aim to inform the design of future devices with users' preferences, cutting down on the resources required to build, deploy, test, and adjust interaction designs. A limitation of my approach is that I was relying on my participants' imaginations to interact with a system. Interacting with an actual system would be a different experience that takes into account gesture-recognition errors, among other experiences. My study only tested two types of priming: the viewing of a sci-fi montage and recalling times of creativity to be in a creative mindset. Priming with other sci-fi clips might produce different results and have a different impact on use cases other than a gesture-controlled mixed reality video player. Another limitation to this work with sci-fi priming is that participants could have mimicked gestures popularized by the movies which led to the creation of a more learnable and memorable gesture set. It is hard to tell from the results of this study whether the impact on the sci-fi gestures was because participants were primed to think creatively or mimicked the gestures from the primer.

Future work would be to take the gesture sets recommended in this chapter and test them either in a usability study that mimics an actual system, like a Wizard of Oz type of study, or invest the resources to build an interactive prototype of a system to test those interactions. Other future work might test the effects of the sci-fi clips I used in this chapter on other modes of user-elicited interactions like voice commands or graphical interface elements like icons. Other priming approaches like having participants do physical activities before an elicitation session could have different effects on the mid-air gestures proposed to trigger the functions to control an MR video player.

5 Summary

In this chapter, I conducted the largest investigation, to the best of my knowledge, into the effects of priming on user-elicited gestures in a distributed end-user elicitation study. I evaluated the effects of priming by viewing science fiction clips and having a creative mindset on user-elicited gestures, with a novel approach of running supervised distributed learnability and memorability studies. I showed that priming with science fiction videos produces user-elicited gestures that are significantly faster and easier to learn and remember. Besides the empirical investigation into the effects of priming, the distributed

learnability and memorability methodological contributions, this chapter contributes a user-elicited gesture set for a media player in a mixed reality environment. I recommend using priming in elicitation studies to unlock participants' creativity and elicit interactions that are learnable and memorable.

Nine | Beyond This Dissertation

In this chapter, I provide two examples of my work using the DXD process and CROWDDESIGN engine in projects that fall outside the scope of this dissertation.

1 My Time at Apple

During the summer of 2019, I joined Apple's Siri advanced development group in Cupertino, CA. I was the directly responsible individual for a project designing multimodal interactions for a prototype of a future device. Due to the novel nature of the prototype, I insisted on including actual users in the design process. I believed that empathy–while a great starting point for design–would not be enough to understand how users would interact with this technology due to its futuristic nature. I wanted to take a user-centered design approach. Initially, I was faced with resistance from the head of research for my organization. During a meeting, I was able to convince them to grant me a pilot with 5 participants. After the pilot, I presented the results to the research group's director. This in turn green-lit a full-scale study that changed the trajectory of my project from designing interactions and building a prototype to a user-driven research approach to designing the prototype utilizing the DXD process[10]. At the end of the summer, I was granted the opportunity to present my work and explain my research to the SVP of the organization at Apple.

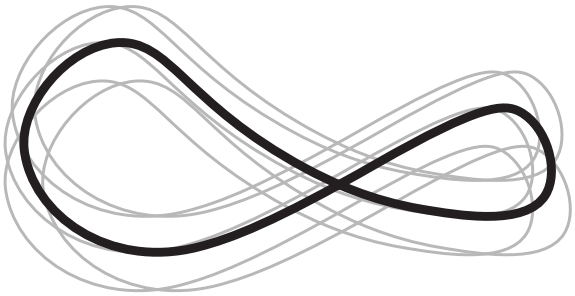
2 DXD + Accessibility

I am using the CROWDDESIGN engine and DXD as part of a collaborative research effort that falls outside the scope of this dissertation. I have been collaborating with Martez Mott (Microsoft Research) and Patrick Carrington (assistant professor at Carnegie Mellon University) to study how individuals in wheelchairs would interact with Virtual Reality (VR) interfaces. VR headsets typically require movements that are impossible for wheelchair users–*e.g.,* walking around, crouching, etc. In this research project, we are capitalizing on the distributed opportunities offered by the CDE to gain an understanding of how

[10] Due to the secretive nature of the project, I was not able to use my CDE and conducted lab-based studies. However, I did follow the six steps of DXD.

individuals in wheelchairs would utilize their wheelchairs as an input device to interact with a VR environment. This work extends Carrington *et al.'s* [21] work on Chairable Computing—a research area that uses power wheelchairs as interfaces to interact with myriad technologies.

DXD IMPACT



Ten | **Discussion**

In this chapter, I revisit the research questions I outlined in the **Introduction** chapter of this dissertation. In addition, I present: the advantages and disadvantages of DXD; ways to limit the effects of the disadvantages; my thoughts on unlocking creativity in DXD studies; my reflections on my work designing with the world; and in closing, future work directions I intend to pursue after my Ph.D.

1 Answers to Research Questions

In the **Crowdlicit** chapter, I answered RQ.1 *(How can I expand the reach of elicitation studies beyond the lab and reach a larger more diverse pool of participants?)* by building the Crowdlicit system and putting it through its paces. I used Crowdlicit to replicate a published elicitation study and surveyed the participants on the system's usability.

I answered RQ.2 *(What are the requirements for a system that translates the elicitation study to run in a distributed fashion?)* by drawing on the experience I gained conducting the in-lab elicitation study detailed in chapter **Anachronism by Design** and putting forth six requirements for Crowdlicit.

My answer to RQ.3 *(What are the benefits and drawbacks to running elicitation studies in a distributed fashion?)* was detailed in the discussion section of the **Crowdlicit** chapter. I expand on this answer to reflect on the advantages and disadvantages of the DXD process in general in the next sections of this chapter. As for RQ.7 (*How can I evaluate the guessability of the interaction proposals resulting from elicitation studies on a large scale in an efficient manner?*), I formulated the **End-User Identification** method.

In the **Crowdsensus** chapter, I answered RQ.4 *(How can a crowd of online workers facilitate the similarity judgements for agreement analysis in end-user elicitation studies?)* by building the **Crowdsensus** tool and testing it against expert researchers. This testing answers the next couple of research questions. For RQ.5 (*By using the crowd, what are the benefits, if any, in terms of cost and time compared to the status quo use of experts' judgments?*), I found that Crowdsensus was four times faster at a comparable cost. And for RQ.6 (*How does the quality of the results produced by the crowd compare to*

those produced by expert researchers?), **Crowdsensus** matched, and in some cases provided better, results that individual expert researchers.

In the **Anachronism by Design** chapter I answered RQ.10 *(How does production influence user-elicited interaction designs?)* by showing that the production principle—requiring participants to provide multiple proposals in an elicitation study—reduces the number of anachronistic or legacy icons produced. I also formulated a set of icons that included 20 (out of 38 total) new concept icons (Figure 33), in turn answering RQ.12 *(What icons would young adults propose to trigger computer functions currently associated with anachronistic icons?)*. When answering RQ.13 *(How familiar are young adults with the objects represented in anachronistic icons?)*, I showed that while young adults were mostly familiar with the objects represented in anachronistic icons, they had never used most of the objects. That begs the question, would icons depicting these objects be usable for the upcoming generation of users? And finally, I answered RQ.14 *(How identifiable is a set of icons elicited from young adults?),* demonstrating that young-adult-elicited icons were easily identifiable.

In the chapter **"I Am Iron Man"** I answered RQ.8 (*How can I evaluate the learnability of the interaction proposals resulting from elicitation studies on a large scale in an efficient manner?*) by running a supervised distributed learnability study. I answered RQ.9 (*How can I evaluate the memorability of the interaction proposals resulting from elicitation studies on a large scale in an efficient manner?*) by running a supervised distributed memorability study. From the results of the learnability and memorability studies, I showed that priming with sci-fi videos produces gestures that are easier and faster to learn answering RQ.11 (*How does priming influence user-elicited interaction designs?).* I answered RQ.15 (*How can I create gestures for mixed-reality environments that are guessable, learnable, and memorable?*) by running a DXD study eliciting gestures from users using priming techniques and evaluating the user-elicited gestures in distributed identification, learnability, and memorability studies.

2 Advantages of DXD

From the work I have presented in this dissertation and drawing on the literature of online HCI research, I present *five advantages* to conducting online design studies.

2.1 **A Large Pool of Participants**

One of the benefits of online studies is the ease of recruiting a large number of participants efficiently. Participants can take part in the study anywhere, without the need to deal with transportation or scheduling a time with the research team conducting the study. For elicitation studies specifically, recruiting a significant number of participants to propose interaction designs is desirable. Both Morris *et al.* [82] and Nacenta *et al.* [86] show that interaction designs created by a larger number of people are favorable to those created by one or a few expert designers. Also, many published elicitation studies [23,25,79,93] have concluded their findings by stating that future work is required with additional participants and that the generalization of their results is limited due to their small sample size. Of course, the potential pool of participants is dependent on the recruitment methods followed by the researcher. In the work presented in this dissertation, I capitalized on Amazon's mechanical Turk to reach my participants. While this allowed me speedy access to participants, the majority of them were from north America and India—an improvement still over the reach of lab-based studies. The focus of my work is to provide a tool that makes conducting studies online possible, the number and diversity in terms of location and physical abilities of the participants is only as good as the recruitment platform or method of choice.

2.2 **A Geographically Distributed Pool of Participants**

A drawback of lab-based studies is the demographics of the participants is often WEIRD—Western, Educated, Industrialized, Rich, Democratic. Running research studies online has the potential to collect data from participants nationally and globally. Designing interactions benefits from a culturally diverse global pool of participants. Mauney *et al.* [69] found differences across cultures, specifically for symbolic gestures, in an elicitation study they conducted in 9 countries with 340 participants by deploying 12 different research teams. Obaid *et al.* [93] cited the need for more culturally diverse participants to affirm their framework of user-generated human-robot interactions. Reinecke and Gajos [100] state that cultural diversity is needed to generalize the results of HCI research citing literature showing differences between western and eastern societies' use of technology. Malizia and Belluci [66] talked about a challenge facing the creation of natural interactions that is *"performing experimental evaluations for*

validating gestures in multicultural and multidisciplinary environments instead of classic controlled experiments in laboratory settings."

While I did not focus on investigating cultural differences in the work presented in this dissertation, using **The CROWDDESIGN engine** allowed me to reach participants who were not local to Seattle, WA as is shown in Figure 39. **The CROWDDESIGN engine** has the potential to enable future research directions investigating interaction designs elicited from international participants.

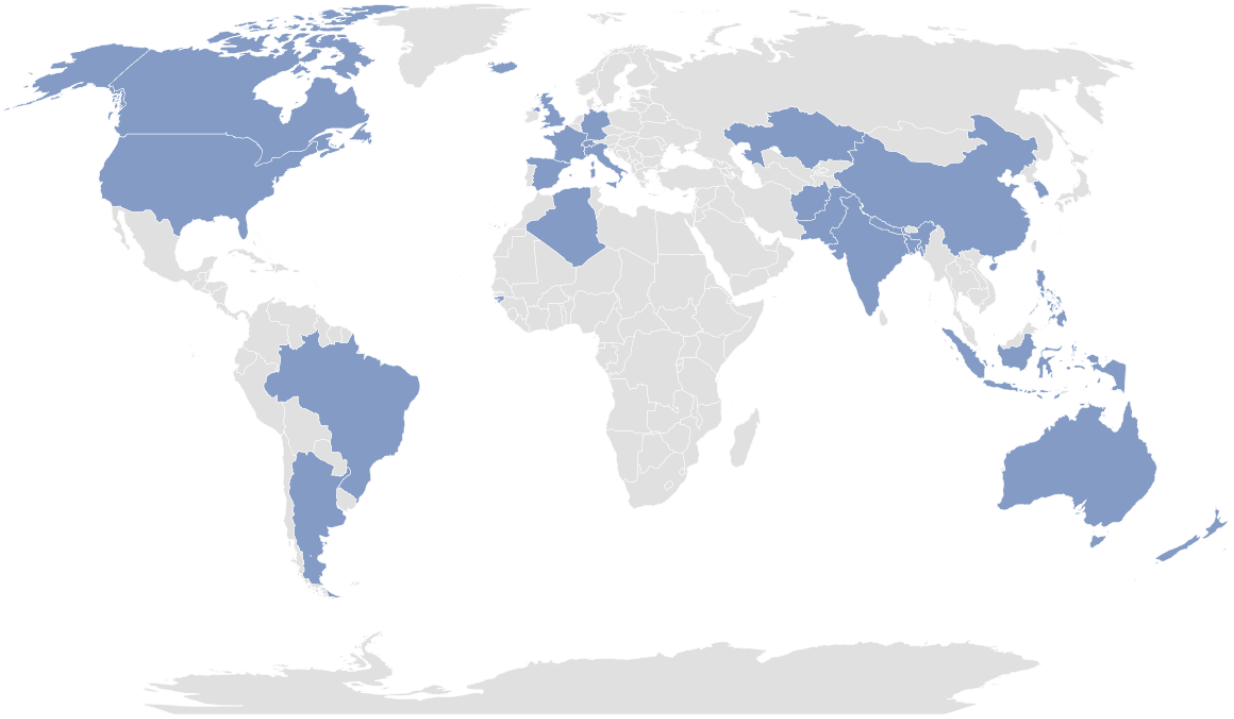


Figure 39 A world map highlighting the country of origin of the participants in the different studies in this dissertation. A country is highlighted even if a single participant reported it as their country of origin. This map does not reflect density of distribution.

2.3 **Inclusivity**

It is imperative that future technologies be inclusive of people's different abilities. The DXD process is meant to be an informative design approach, so for future work utilizing the DXD approach, including participants of diverse abilities is vital to make inclusive technologies. In addition, Zyskowski *et al.* [139] show that online crowd work is a desirable option for individuals with disabilities. While the work presented in this dissertation has not advanced this issue scientifically, I am already using **The**

CROWDDESIGN engine to reach participants with mobility limitations as I stated in the **DXD + Accessibility** section of the **Beyond This Dissertation** chapter. Of course, online HCI studies, specifically ones appearing in this dissertation are not all-inclusive. Participants needed to meet inclusion criteria such as speaking English, access to the internet, and in some cases access to computing devices equipped with cameras.

2.4 **Speed**

Because online crowd workers can work unsupervised in parallel, it is possible to collect hundreds or even thousands of proposals in an online study in a matter of days, a process that could take years in a traditional lab setting. I found this to be true from my own experience running in-lab and online elicitation and identification studies for the work described in the **Anachronism by Design** chapter. It took me and a graduate research assistant almost six months to recruit, schedule, and run 30 participants to study the perceptions of computer iconography in a lab-based elicitation study. As a result of the elicitation study, I created a set of user-elicited icons. I tested the identifiability of that icon set with 60 online participants—double the number of participants in the elicitation study—by running a distributed identification study using **Crowdlicit** that was completed in a single day.

2.5 **Collective Intelligence**

Online crowds have demonstrated time and again their ability to collectively complete tasks, even creative ones, as shown in the **Related Work** chapter of this dissertation. I capitalized on online crowds' collective intelligence to analyze elicitation and identification study results efficiently by increasing the analysis speed four fold, as described in the **Crowdsensus** chapter.

3 Disadvantages of DXD

In this subsection, I highlight *four disadvantages* to conducting distributed interaction design studies.

3.1 **Missing the Personal Touch**

In addition to recording interaction proposals, researchers conducting elicitation studies often take notes of participants' reactions and thoughts and sometimes ask participants to do a think-aloud practice.

Conducting an elicitation study online can make that practice challenging, as the researcher is not sitting in the same room as the participant. A disadvantage to crowdsourcing work is the issue of clarification, as mentioned by Horton *et al.* [44]. If the instructions are not clear, it is hard for participants to ask questions and clarify things promptly.

3.2 **Quality Control and Safety**

The quality of crowdsourced DXD studies is vulnerable to be compromised. Spam answers to collect monetary rewards could be collected as some crowd workers may attempt to game the system. Online systems are also open to attacks. Lasecki *et al.* [58] mention that crowd-powered systems can have their outcomes maliciously manipulated. In the **Crowdsensus** system, I implemented same-design match, and timing threshold measures to combat spam answers. In the **Crowdlicit** system, I implemented randomizing prompt orders to avoid collecting uneven number of proposals for the first few prompts in case of participants' drop off. I also tested timing thresholds in **Crowdlicit**. However, this practice was not as effective in detecting spam answers as it was for **Crowdsensus** because some legitimate answers were proposed within the time threshold. This led me to remove timing threshold from the **Crowdlicit** system.

3.3 **Efficiency Over Quality**

The quality of DXD studies can face an issue of breadth vs. depth. Collecting a vast amount of data may diminish the ability to gain deeper insight into the motivation as to why a participant proposed a particular interaction. In addition, the advantages of a distributed, or crowdsourcing, approach might entice researchers to rely on it more often than lab-based studies. Kitture *et al.* [53] argue that crowd work has the potential to displace skilled labor, with unskilled labor tasks decomposed into smaller pieces even in some complex and expert tasks such as writing and product design. Running DXD with large crowds might displace interaction designers because crowds are faster, cheaper, and often more creative than a single expert. However, I will reiterate here that the aim of DXD studies is to inform interaction designers of the values, abilities and preferences of their potential users so they can be reflected in their designs, not to replace designers.

3.4 **Infrastructure**

It is possible to conduct parts of the DXD process, like an elicitation or an identification study, using existing tools. Survey tools, for instance, allow participants to respond to a prompt with a text string or in some cases a video recording of themselves performing input-action proposals. However, these tools are restrictive and do not allow the same freedom allowed by in-person lab-based studies. Because of the drawbacks to utilize existing tools to conduct DXD, investments in creating a custom infrastructure had to be made to realize the potential of distributed interaction design fully.

3.5 **Majority Rule**

The way elicitation studies are analyzed is by finding consensus from participants as to which action design should trigger a function. This consensus is reached under a majority rule. Typically, the most popular interaction "wins", and the researcher conducting the elicitation study uses their best judgment to decide what interaction has the highest consensus. With this approach, the "winning" interaction might not be the most suitable, or representative interaction for all potential users of the system.

4 Mitigating the Disadvantages of DXD

To reflect, Table 23 distills the advantages and disadvantages of the distributed interaction design process.

Advantages	Disadvantages
A Large Pool of Participants	Personal Connections
A Geographically Distributed Pool of Participants	Quality Control and Safety
Inclusivity	Efficiency Over Quality
Speed	Infrastructure
Collective Intelligence	Majority Rule

Table 23 The advantages and disadvantages of DXD

In this section, I offer possible ways to mitigate the disadvantages of DXD.

4.1 **Personal Connections**

The DXD process starts with **Step 1: Set Up the Four Pillars of a DXD Study,** which are: rules of engagement, a list of functions, prompt modality, and proposal modality. Having these four foundational elements clearly defined limits—or at least lessens—participants' confusion and provides clear

instructions. In addition, it is possible to run supervised DXD studies, *i.e.,* the researcher would be in contact with the participant using video conferencing software and records the session—as I demonstrated in the **"I Am Iron Man"** chapter. This practice would reduce the speed advantage and scale down the number of participants in a DXD study but allows for a higher level of personal connection. Supervised DXD studies can still capitalize on the other advantages of online studies such as the geographic and ability diversity of online participants. Not to mention, the ability to conduct user-centered design research during global pandemics.

4.2 **Quality Control and Safety**

One way to eliminate spam answers is to conduct studies on platforms that offer personalized feedback other than monetary rewards for participating in an online study, such as the LabintheWild platform [100] as opposed to mTurk. Oliveira *et al.* [94] have shown that participants are motivated by learning about themselves and the research process. Ye *et al.* [136] showed that personalized feedback produces higher-quality results than monetary compensation in crowdsourced studies.

4.3 **Efficiency Over Quality**

Using Morris *et al.'s* [79] legacy bias reduction principles can enhance the quality of online participants' proposals. Pyryeskin *et al.* [99] showed that in addition to eliciting interactions from end users, professional designers' input could aid in choosing the best-performing interactions for emerging technologies. In their case, creating a mid-air gesture set to interact with an interactive tabletop. Pyryeskin *et al.'s* findings support my view that DXD studies do not aim to displace professional interaction designers with cheap inexpert users; rather, designers are needed to create the final system informed by the users' proposals.

4.4 **Infrastructure**

In this dissertation, I have demonstrated the need to and benefits of building custom tools to facilitate the DXD process. It is out of this need that I plan to keep **The CROWDDESIGN engine** available online for as long as possible.

4.5 **Majority Rule**

My work in the **Crowdsensus** chapter improved the drawback of reaching consensus by one or two researchers by utilizing the wisdom of the crowd to find the interaction with the highest consensus for a given function in an elicitation study. As for the suitability of the most popular interaction for all potential users of a system, I touch on that in the **Future Directions** subsection of this **Discussion** chapter.

5 Unlocking Creativity in DXD

In the **Related Work** chapter, in the section titled **Unlocking Creativity**, I stated that the literature has not offered strong opinions to completely discard legacy bias interactions or to recommend implementing them in future technologies. And I believe that in some situations, a legacy interaction could be the best one—based on my definition of a **good interaction design**, in the sense that it is memorable, discoverable, fits its purpose and situation, and is easy to use. That said, these measures have a subjective nature to them; what might be the easiest interaction for one person might not be for another due to ability variation, preference differences, cultural differences, differences in experience or levels of education, and so on.

In the **Anachronism by Design** chapter, I investigated the effects of the production principle on reducing the proposal of anachronistic icons. I found that new-concept icons were indeed discoverable. A drawback to using the production principle is the increased number of proposals collected, which require additional resources to analyze. However, this drawback can be curbed by utilizing the **Crowdsensus** tool to analyze that data efficiently.

In the **"I Am Iron Man"** chapter, I investigated the effects of priming on user-elicited gestures by using sci-fi videos, finding that gestures elicited from participants who were primed with sci-fi videos were easier and faster to learn and remember than those elicited without priming. From my work, I believe that Morris *et al.'s* [79] 3P principles to reduce legacy bias should be incorporated into the proposal elicitation step of DXD studies.

6 Reflections on Designing with The World

In this section, I offer some reflections on my work on distributed user-centered design.

6.1 **Cloud Computing**

I am a believer in cloud computing and web-based applications due to their flexibility and low barrier to use. Web-based applications work across devices without the need to write additional code, cutting down on development resources. From the user's standpoint, no app installation is needed. I found great success developing web-based applications in the past. For example, I co-invented Dytective [101] as a web-based game that uses machine learning to detect the risk of having dyslexia. The web approach proved especially beneficial in helping us reach tens of thousands of children who had access to old computers that had a web browser but who probably would not have been able to access a smartphone app. My experience with Dytective influenced my approach to building **The CROWDDESIGN engine**; I knew I needed to create a tool easily accessible by as many potential participants as possible without platform or device-specific restrictions.

6.2 **Stepping Out of the Lab**

From my work, I have arrived at the conclusion that a lot of exploratory design should not be done in labs. We have the technology to run distributed design and evaluation studies. Distributed user-centered design and evaluation leads to better technologies.

6.3 **Looking for Black Swans**

To me, the main reason to include end users in the design process of interactive systems can be summed up in a single concept: Black Swans. A black swan is a term used to describe unknown unknowns. In the context of design, these unknown unknowns cannot be accounted for by designers without the inclusion of actual future users of the technology who might unearth these black swans, whether they are a physical ability, a mental model, or a preference. I believe empathy is the starting point of inclusive design, but true inclusive design cannot be achieved without the inclusion of actual users in the design process.

6.4 **The Axe-Maker Metaphor**

As a technologist, I view my role in the world as an axe-maker. What is an axe but a technology? Technology, as I see it, is an artifact that augments our abilities to aid us in achieving our goals. Just like the axe, any technology can be utilized for good or bad. We use it to chop wood and build shelter, burn fire to provide warmth, or make paper to spread knowledge—perhaps the most noble of an axe's utilities. There do exist axe murderers in our world. Bad agents will utilize any tool to achieve their mischievous goals – an axe, a knife; tools as sophisticated as a gun or primitive as a rock. We, as computing technologists, are fortunate to be able to instill intelligence in our creations. That can restrict the harm bad agents might impose with our technology. This is why it is extremely important to understand the users of our technologies, their values, needs, abilities and preferences, before we build our technology. This is to understand what harm to them might mean, and not impose our own world views and biases onto them with our technologies.

7 Future Directions

As a research topic, distributed interaction design opens numerous interesting future directions. The third principle of reducing legacy bias [79], partners—the practices of eliciting interactions from a pair of participants rather from an individual—has yet to be explored in a distributed fashion. Future work might look at how pairing participants from different cultural backgrounds could impact the interaction designs resulting from an elicitation study.

Another research topic of interest would be the approach to analyzing the results of an elicitation study. The way elicitation studies work is by recommending interaction designs with the highest consensus. The interaction with the highest consensus is dependent on the participants in the elicitation study. The work I presented in this dissertation takes steps towards increasing the diversity of participants—*i.e.,* geographic distribution. However, the way the elicitation method is limited fundamentally in representing all its participants. Perhaps future work could look into expanding the third step of DXD "create interaction sets" to include creating participant personas and tying the recommendations of step six "recommend interaction designs" to these personas. Resulting in multiple recommended sets of

interactions for different types of future users rather than a one-size-fits-all approach as is the status quo in elicitation studies.

As for me personally, I will primarily use **The CROWDDESIGN engine** to collaborate on research projects by reaching out to my numerous contacts at universities and research labs. I will mentor students and provide them access to the CDE as a platform to help answer research questions, especially those dealing with inclusive design, accessibility, and future technologies like mixed reality, voice-based user interfaces, Internet of Things devices, and AI-powered applications.

I intend to widen the reach of the DXD approach by teaching the method at workshops at academic conferences like the ACM's CHI conference and industry-focused conferences like IXDA's Interactions. I also intend to give guest lectures on distributed design and to design a semester-long course based on the DXD process to teach at universities as an adjunct lecturer.

Finally, I will focus some of my future research efforts on investigating distributed approaches to other user-centered design methods besides interaction design.

Eleven | **Conclusion**

In this dissertation I have demonstrated the following thesis statement:

> *Using a custom-built platform to conduct Distributed Interaction Design (DXD) enables: creating user-elicited interactions; evaluating the guessability, learnability, and memorability of interaction designs; and the recruitment of participants through third-party services in a timely manner.*

As demonstrated in chapters **Crowdlicit**, **Crowdsensus**, and **The CROWDDESIGN engine,** custom-built systems enabled a distributed approach to creating user-elicited interactions. These custom-built systems also enabled the investigation of interactions' guessability, learnability, and memorability as shown in chapters **Anachronism by Design**, **"I Am Iron Man"**, and **Beyond This Dissertation**. All the chapters mentioned above demonstrated the ability to conduct elicitation and interaction evaluation studies online in a timely manner.

I hope that the work in this dissertation inspires researchers to broaden their reach beyond their physical labs and to reach an online pool of participants to create technologies that are more guessable, learnable, memorable, and ultimately usable.

References

1. Chadia Abras, Diane Maloney-Krichmar, and Jenny Preece. 2004. User-Centered Design. *Encyclopedia of Human-Computer Interaction*: 14.
2. Alan Agresti. 2010. *Analysis of Ordinal Categorical Data*. John Wiley & Sons, Inc.
3. Faez Ahmed, Nischal Reddy Chandra, Mark Fuge, and Steven Dow. 2019. Structuring Online Dyads: Explanations Improve Creativity, Chats Lead to Convergence. In *Proceedings of the 2019 on Creativity and Cognition* (C&C '19), 306–318. https://doi.org/10.1145/3325480.3325486
4. Luis von Ahn and Laura Dabbish. 2004. Labeling images with a computer game. In *Proceedings of the 2004 conference on Human factors in computing systems - CHI '04*, 319–326. https://doi.org/10.1145/985692.985733
5. Nir Ailon and Moses Charikar. 2005. Aggregating Inconsistent Information : Ranking and Clustering.
6. Nir Ailon, Moses Charikar, and Alantha Newman. 2008. Aggregating inconsistent information. *Journal of the ACM* 55, 5: 1–27. https://doi.org/10.1145/1411509.1411513
7., 8. Abdullah X. Ali, Meredith Ringel Morris, and Jacob O. Wobbrock. 2018. Crowdsourcing Similarity Judgments for Agreement Analysis in End-User Elicitation Studies. In *Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology - UIST '18* (UIST '18), 177–188. https://doi.org/10.1145/3242587.3242621
9. Abdullah X Ali, Meredith Ringel Morris, and Jacob O. Wobbrock. 2019. Crowdlicit: A System for Conducting Distributed End-User Elicitation and Identification Studies. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 255. https://doi.org/10.1145/3290605.3300485
10. Salvatore Andolina, Daniel Lee, and Steven Dow. 2013. Crowdboard: an augmented whiteboard to support large-scale co-design. In *Proceedings of the adjunct publication of the 26th annual ACM symposium on User interface software and technology - UIST '13 Adjunct*, 89–90. https://doi.org/10.1145/2508468.2514724
11. Nikhil Bansal, Avrim Blum, and Shuchi Chawla. 2004. Correlation Clustering. *Machine Learning* 56, 1–3: 89–113. https://doi.org/10.1023/B:MACH.0000033116.57574.95
12. Yoav Benjamini and Henry Braun. 2002. John W. Tukey's Contributions to Multiple Comparisons. *The Annals of Statistics* 30, 6: 1576–1594.
13. Birgitta Bergvall-K. 2008. Participatory design: one step back or two steps forward? *Proceedings of the tenth anniversary conference on participatory design*: 102–111.
14. Michael S Bernstein, Greg Little, and Robert C Miller. 2010. Soylent: a word processor with a crowd inside. In *Proceedings of the 23rd Annual ACM Symposium on User Interface Software and Technology - UIST '10*, 313–322.
15. Ceylan Beşevli, Oğuz Turan Buruk, Merve Erkaya, and Oğuzhan Özcan. 2018. Investigating the Effects of Legacy Bias: User Elicited Gestures from the End Users Perspective. In *Proceedings of the 19th International ACM SIGACCESS Conference on Computers and Accessibility - DIS '18*, 277–281. https://doi.org/10.1145/3197391.3205449
16. Arpita Bhattacharya, Calvin Liang, Emily Y. Zeng, Kanishk Shukla, Miguel E. R. Wong, Sean A. Munson, and Julie A. Kientz. 2019. Engaging Teenagers in Asynchronous Online Groups to Design for Stress Management. In *Proceedings of the Interaction Design and Children on ZZZ - IDC '19*, 26–37. https://doi.org/10.1145/3311927.3323140
17. Jeffrey P. Bigham, Tom Yeh, Chandrika Jayant, Hanjie Ji, Greg Little, Andrew Miller, Robert C. Miller, Aubrey Tatarowicz, Brandyn White, and Samuel White. 2010. VizWiz: nearly real-time answers to visual questions. In *Proceedings of the 2010 International Cross Disciplinary Conference on Web Accessibility (W4A) - W4A '10*, 1. https://doi.org/10.1145/1805986.1806020
18. Danielle Bragg, Cynthia Bennett, Katharina Reinecke, and Richard Ladner. 2018. A Large Inclusive Study of Human Listening Rates. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems - CHI '18*, 1–12. https://doi.org/10.1145/3173574.3174018
19. Christian Briggs and Kevin Makice. Bricks and clicks: participatory organizational design through microparticipation. 4.
20. David Brin. 2010. Future Tense: How the Net ensures our cosmic survival. *Communications of the ACM* 53, 6: 120–ff. https://doi.org/10.1145/1743546.1743576
21. Patrick Carrington, Amy Hurst, and Shaun K. Kane. 2014. Wearables and chairables: inclusive design of mobile input and output techniques for power wheelchair users. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '14), 3103–3112. https://doi.org/10.1145/2556288.2557237
22. Yan Chen and Joseph Konstan. 2015. Online field experiments: a selective survey of methods. *Journal of the Economic Science Association* 1, 1: 29–42. https://doi.org/10.1007/s40881-015-0005-3
23. Eunjung Choi, Sunghyuk Kwon, Donghun Lee, Hojin Lee, and Min K. Chung. 2012. Can User-Derived Gesture be Considered as the Best Gesture for a Command?: Focusing on the Commands for Smart Home System. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 56, 1: 1253–1257. https://doi.org/10.1177/1071181312561222
24. Belinda L. Collins and Neil D. Lerner. 1982. Assessment of Fire-Safety Symbols. *Human Factors* 24, 1: 75–84. https://doi.org/10.1177/001872088202400108
25. Sabrina Connell, Pei-Yi Kuo, Liu Liu, and Anne Marie Piper. 2013. A Wizard-of-Oz elicitation study examining child-defined gestures with a whole-body interface. In *Proceedings of the 12th International Conference on Interaction Design and Children - IDC '13*, 277–280. https://doi.org/10.1145/2485760.2485823
26. Lisa G. Cowan and Kevin A. Li. 2011. ShadowPuppets: supporting collocated interaction with mobile projector phones using hand shadows. In *Proceedings of the 2011 annual conference on Human factors in computing systems - CHI '11*, 2707–2716. https://doi.org/10.1145/1978942.1979340
27. Erik D. Demaine, Dotan Emanuel, Amos Fiat, and Nicole Immorlica. 2006. Correlation clustering in general weighted graphs. *Theoretical Computer Science* 361, 2–3: 172–187. https://doi.org/10.1016/j.tcs.2006.05.008

28. Lucas S. Figueiredo, Mariana G.M. Gonçalves Maciel Pinheiro, Edvar X.C. Vilar Neto, and Veronica Teichrieb. 2015. An Open Catalog of Hand Gestures from Sci-Fi Movies. In *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems* (CHI EA '15), 1319–1324. https://doi.org/10.1145/2702613.2732888
29., 30. Leah Findlater, Ben Lee, and Jacob Wobbrock. 2012. Beyond QWERTY: augmenting touch screen keyboards with multi-touch gestures for non-alphanumeric input. In *Proceedings of the 2012 ACM annual conference on Human Factors in Computing Systems - CHI '12*, 2679–2682. https://doi.org/10.1145/2207676.2208660
31. Kraig Finstad. 2013. Response to commentaries on "The usability metric for user experience." *Interacting with Computers* 25, 4: 327–330. https://doi.org/10.1093/iwc/iwt005
32. Brigitte N. Frederick. 1999. *Fixed-, Random-, and Mixed-Effects ANOVA Models: A User-Friendly Guide for Increasing the Generalizability of ANOVA Results*. Retrieved July 1, 2020 from https://eric.ed.gov/?id=ED426098
33. Dustin Freeman, Nathan LaPierre, Fanny Chevalier, and Derek Reilly. 2013. Tweetris: a study of whole-body interaction during a public art event. In *Proceedings of the 9th ACM Conference on Creativity & Cognition - C&C '13*, 224. https://doi.org/10.1145/2466627.2466650
34. Laura Germine, Ken Nakayama, Bradley C. Duchaine, Christopher F. Chabris, Garga Chatterjee, and Jeremy B. Wilmer. 2012. Is the Web as good as the lab? Comparable performance from Web and lab in cognitive/perceptual experiments. *Psychonomic Bulletin & Review* 19, 5: 847–857. https://doi.org/10.3758/s13423-012-0296-9
35. Bogdan-Florin Gheran, Jean Vanderdonckt, and Radu-Daniel Vatavu. 2018. Gestures for Smart Rings: Empirical Results, Insights, and Design Implications. In *Proceedings of the 2018 on Designing Interactive Systems Conference 2018 - DIS '18*, 623–635. https://doi.org/10.1145/3196709.3196741
36. A. R. Gilmour, R. D. Anderson, and A. L. Rae. 1985. The analysis of binomial data by a generalized linear mixed model. *Biometrika* 72, 3: 593–599. https://doi.org/10.1093/biomet/72.3.593
37. Michael D. Good, John A. Whiteside, Dennis R. Wixon, and Sandra J. Jones. 1984. Building a user-derived interface. *Communications of the ACM* 27, 10: 1032–1043. https://doi.org/10.1145/358274.358284
38. Sandy J J Gould, Anna L Cox, Duncan P Brumby, and Sarah Wiseman. 2015. Home is Where the Lab is: A Comparison of Online and Lab Data from a Time-sensitive Study of Interruption. *Human Computation* 2, 1. https://doi.org/10.15346/hc.v2i1.4
39. Samuel W. Greenhouse and Seymour Geisser. 1959. On methods in the analysis of profile data. *Psychometrika* 24, 2: 95–112. https://doi.org/10.1007/BF02289823
40. Donald Hedeker and Robert D. Gibbons. 1994. A Random-Effects Ordinal Regression Model for Multilevel Analysis. *Biometrics* 50, 4: 933–944. https://doi.org/10.2307/2533433
41. Lynn Hoff, Eva Hornecker, and Sven Bertel. 2016. Modifying Gesture Elicitation: Do Kinaesthetic Priming and Increased Production Reduce Legacy Bias? In *Proceedings of the TEI '16: Tenth International Conference on Tangible, Embedded, and Embodied Interaction - TEI '16*, 86–91. https://doi.org/10.1145/2839462.2839472
42., 43. Sture Holm. 1979. A Simple Sequentially Rejective Multiple Test Procedure. *Scand J Statist* 6: 65–70. https://doi.org/10.2307/4615733
44. John J. Horton, David G. Rand, and Richard J. Zeckhauser. 2011. The online laboratory: conducting experiments in a real labor market. *Experimental Economics* 14, 3: 399–425. https://doi.org/10.1007/s10683-011-9273-9
45. Https://math.stackexchange.com/questions/507742/distance-similarity-between-two-matrices. https://math.stackexchange.com/questions/507742/distance-similarity-between-two-matrices. Retrieved from https://math.stackexchange.com/questions/507742/distance-similarity-between-two-matrices
46. Edwin L Hutchins, James D Hollan, and Donald A Norman. Direct Manipulation Interfaces. 28.
47. J. Johnson, T.L. Roberts, W. Verplank, D.C. Smith, C.H. Irby, M. Beard, and K. Mackey. 1989. The Xerox Star: a retrospective. *Computer* 22, 9: 11–26. https://doi.org/10.1109/2.35211
48. Jon Favreau. 2010. *Iron Man II*. Paramount Pictures.
49. Sk Kane, Jo Wobbrock, and Re Ladner. 2011. Usable gestures for blind people: understanding preference and performance. *Proceedings of the 2011 annual conference on Human factors in computing systems - CHI '11*: 413–422. https://doi.org/10.1145/1978942.1979001
50. Joy Kim, Maneesh Agrawala, and Michael S. Bernstein. 2017. Mosaic: Designing Online Creative Communities for Sharing Works-in-Progress. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing - CSCW '17*, 246–258. https://doi.org/10.1145/2998181.2998195
51. Joy Kim, Justin Cheng, and Michael S. Bernstein. 2014. Ensemble: exploring complementary strengths of leaders and crowds in creative collaboration. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing - CSCW '14*, 745–755. https://doi.org/10.1145/2531602.2531638
52. S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi. 1983. Optimization by Simulated Annealing. *Science* 220, 4598: 671–680. https://doi.org/10.1126/science.220.4598.671
53. Aniket Kittur, Jeffrey V. Nickerson, Michael Bernstein, Elizabeth Gerber, Aaron Shaw, John Zimmerman, Matt Lease, and John Horton. 2013. The future of crowd work. In *Proceedings of the 2013 conference on Computer supported cooperative work - CSCW '13*, 1301. https://doi.org/10.1145/2441776.2441923
54. Steven Komarov, Katharina Reinecke, and Krzysztof Z. Gajos. 2013. Crowdsourcing performance evaluations of user interfaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems - CHI '13*, 207. https://doi.org/10.1145/2470654.2470684
55. Anne Köpsel and Nikola Bubalo. 2015. Benefiting from legacy bias. *interactions* 22, 5: 44–47. https://doi.org/10.1145/2803169
56. William H. Kruskal and W. Allen Wallis. 1952. Use of Ranks in One-Criterion Variance Analysis. *Journal of the American Statistical Association* 47, 260: 583–621. https://doi.org/10.2307/2280779
57. Jerrod Larson. 2008. Limited imagination: Depictions of computers in science fiction film. *Futures* 40, 3: 293–299. https://doi.org/10.1016/j.futures.2007.08.015
58. Walter S. Lasecki, Jaime Teevan, and Ece Kamar. 2014. Information extraction and manipulation threats in crowd-powered systems. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing - CSCW '14*, 248–256. https://doi.org/10.1145/2531602.2531733
59. K. S. Lashley. 1951. The Problem of Serial Order in Behavior. *New York: Wiley.*: PP. 112-131.

60. Edith Law, Krzysztof Z. Gajos, Andrea Wiggins, Mary L. Gray, and Alex Williams. 2017. Crowdsourcing as a Tool for Research: Implications of Uncertainty. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing - CSCW '17*, 1544–1561. https://doi.org/10.1145/2998181.2998197
61. Jonathan Lazar, Jinjuan Heidi Feng, and Harry Hochheiser. 2017. *Research Methods in Human-Computer Interaction*. Morgan Kaufmann.
62. Sang Won Lee, Yujin Zhang, Isabelle Wong, Yiwei Yang, Stephanie D. O'Keefe, and Walter S. Lasecki. 2017. SketchExpress: Remixing Animations for More Effective Crowd-Powered Prototyping of Interactive Interfaces. In *Proceedings of the 30th Annual ACM Symposium on User Interface Software and Technology - UIST '17*, 817–828. https://doi.org/10.1145/3126594.3126595
63. Eckhard Limpert, Werner A. Stahel, and Markus Abbt. 2001. Log-normal Distributions across the Sciences: Keys and CluesOn the charms of statistics, and how mechanical models resembling gambling machines offer a link to a handy way to characterize log-normal distributions, which can provide deeper insight into variability and probability—normal or log-normal: That is the question. *BioScience* 51, 5: 341–352. https://doi.org/10.1641/0006-3568(2001)051[0341:LNDATS]2.0.CO;2
64. R. C. Littell, P. R. Henry, and C. B. Ammerman. 1998. Statistical analysis of repeated measures data using SAS procedures. *Journal of Animal Science* 76, 4: 1216–1231. https://doi.org/10.2527/1998.7641216x
65. Kurt Luther, Jari-Lee Tolentino, Wei Wu, Amy Pavel, Brian P. Bailey, Maneesh Agrawala, Björn Hartmann, and Steven P. Dow. 2015. Structuring, Aggregating, and Evaluating Crowdsourced Design Critique. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing - CSCW '15*, 473–485. https://doi.org/10.1145/2675133.2675283
66. Alessio Malizia and Andrea Bellucci. 2012. The artificiality of natural user interfaces. *Communications of the ACM* 55, 3: 36. https://doi.org/10.1145/2093548.2093563
67. H. B. Mann and D. R. Whitney. 1947. On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. *The Annals of Mathematical Statistics* 18, 1: 50–60.
68. Aaron Marcus. 2013. The History of the Future: Sci-fi Movies and HCI. *interactions* 20, 4: 64–67. https://doi.org/10.1145/2486227.2486240
69. Dan Mauney, Jonathan Howarth, Andrew Wirtanen, and Miranda Capra. 2010. Cultural similarities and differences in user-defined gestures for touchscreen user interfaces. In *Proceedings of the 28th of the international conference extended abstracts on Human factors in computing systems - CHI EA '10*, 4015. https://doi.org/10.1145/1753846.1754095
70. Keenan R. May, Thomas M. Gable, and Bruce N. Walker. 2017. Designing an In-Vehicle Air Gesture Set Using Elicitation Methods. In *Proceedings of the 9th International Conference on Automotive User Interfaces and Interactive Vehicular Applications - AutomotiveUI '17*, 74–83. https://doi.org/10.1145/3122986.3123015
71. Keenan R. May, Thomas M. Gable, and Bruce N. Walker. 2017. Designing an In-Vehicle Air Gesture Set Using Elicitation Methods. *Proceedings of the 9th International Conference on Automotive User Interfaces and Interactive Vehicular Applications - AutomotiveUI '17*: 74–83. https://doi.org/10.1145/3122986.3123015
72., 73. Erin McAweeney, Haihua Zhang, and Michael Nebeling. 2018. User-Driven Design Principles for Gesture Representations. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems - CHI '18*, 1–13. https://doi.org/10.1145/3173574.3174121
74. Gourav Modanwal and Kishor Sarawadekar. 2018. A Gesture Elicitation Study with Visually Impaired Users. In *HCI International 2018 – Posters' Extended Abstracts*, Constantine Stephanidis (ed.). Springer International Publishing, Cham, 54–61. https://doi.org/10.1007/978-3-319-92279-9_7
75., 76., 77. Meredith Ringel Morris. 2012. Web on the wall: insights from a multimodal interaction elicitation study. In *Proceedings of the 2012 ACM international conference on Interactive tabletops and surfaces - ITS '12*, 95. https://doi.org/10.1145/2396636.2396651
78., 79. Meredith Ringel Morris, Andreea Danielescu, Steven Drucker, Danyel Fisher, Bongshin Lee, C. Schraefel, and Jacob O. Wobbrock. 2014. Reducing legacy bias in gesture elicitation studies. *Interactions* 21, 3: 40–45. https://doi.org/10.1145/2591689
80., 81., 82. Meredith Ringel Morris, Jacob O. Wobbrock, and Andrew D. Wilson. 2010. Understanding users' preferences for surface gestures. In *Proceedings of Graphics Interface 2010*, 261–268. https://doi.org/10.1016/j.actamat.2009.07.058
83. Omar Mubin, Mohammad Obaid, Philipp Jordan, Patricia Alves-Oliveria, Thommy Eriksson, Wolmet Barendregt, Daniel Sjolle, Morten Fjeld, Simeon Simoff, and Mark Billinghurst. 2016. Towards an Agenda for Sci-Fi Inspired HCI Research. In *Proceedings of the 13th International Conference on Advances in Computer Entertainment Technology* (ACE '16), 10:1–10:6. https://doi.org/10.1145/3001773.3001786
84. Michael J Muller and Allison Druin. Participatory Design: The Third Space in HCI. 70.
85., 86. Miguel A. Nacenta, Yemliha Kamber, Yizhou Qiang, and Per Ola Kristensson. 2013. Memorability of pre-designed and user-defined gesture sets. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems - CHI '13*: 1099. https://doi.org/10.1145/2470654.2466142
87., 89. Michael Nebeling, Alexander Huber, David Ott, and Moira C. Norrie. 2014. Web on the Wall Reloaded: Implementation, Replication and Refinement of User-Defined Interaction Sets. In *Proceedings of the Ninth ACM International Conference on Interactive Tabletops and Surfaces - ITS '14*, 15–24. https://doi.org/10.1145/2669485.2669497
88. Michael Nebeling. 2017. XDBrowser 2.0: Semi-Automatic Generation of Cross-Device Interfaces. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems - CHI '17*, 4574–4584. https://doi.org/10.1145/3025453.3025547
90. Michael Nebeling, David Ott, and Moira C. Norrie. 2015. Kinect analysis: a system for recording, analysing and sharing multimodal interaction elicitation studies. In *Proceedings of the 7th ACM SIGCHI Symposium on Engineering Interactive Computing Systems - EICS '15*, 142–151. https://doi.org/10.1145/2774225.2774846
91. Michael Nielsen, Moritz Störring, Thomas B. Moeslund, and Erik Granum. 2004. A Procedure for Developing Intuitive and Ergonomic Gesture Interfaces for HCI. In *Gesture-Based Communication in Human-Computer Interaction* (Lecture Notes in Computer Science), 409–420. https://doi.org/10.1007/978-3-540-24598-8_38
92. Mohammad Obaid, Markus Häring, Felix Kistler, René Bühling, and Elisabeth André. 2012. User-Defined Body Gestures for Navigational Control of a Humanoid Robot. In *Social Robotics*, Shuzhi Sam Ge, Oussama Khatib, John-John Cabibihan, Reid Simmons and Mary-Anne Williams (eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 367–377. https://doi.org/10.1007/978-3-642-34103-8_37

93. Mohammad Obaid, Felix Kistler, Markus Häring, René Bühling, and Elisabeth André. 2014. A Framework for User-Defined Body Gestures to Control a Humanoid Robot. *International Journal of Social Robotics* 6, 3: 383–396. https://doi.org/10.1007/s12369-014-0233-3
94. Nigini Oliveira, Eunice Jun, and Katharina Reinecke. 2017. Citizen Science Opportunities in Volunteer-Based Online Experiments. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems - CHI '17*, 6800–6812. https://doi.org/10.1145/3025453.3025473
95. Steve Oney, Alan Lundgard, Rebecca Krosnick, Michael Nebeling, and Walter S. Lasecki. 2018. Arboretum and Arbility: Improving Web Accessibility Through a Shared Browsing Architecture. In *The 31st Annual ACM Symposium on User Interface Software and Technology - UIST '18*, 937–949. https://doi.org/10.1145/3242587.3242649
96., 97. Thammathip Piumsomboon, Adrian Clark, Mark Billinghurst, and Andy Cockburn. 2013. User-defined gestures for augmented reality. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 282–299. https://doi.org/10.1007/978-3-642-40480-1_18
98. Henning Pohl and Michael Rohs. 2014. Around-device devices: my coffee mug is a volume dial. In *Proceedings of the 16th international conference on Human-computer interaction with mobile devices & services - MobileHCI '14*, 81–90. https://doi.org/10.1145/2628363.2628401
99. Dmitry Pyryeskin, Mark Hancock, and Jesse Hoey. 2012. Comparing elicited gestures to designer-created gestures for selection above a multitouch surface. In *Proceedings of the 2012 ACM international conference on Interactive tabletops and surfaces - ITS '12*, 1. https://doi.org/10.1145/2396636.2396638
100. Katharina Reinecke and Krzysztof Z. Gajos. 2015. LabintheWild: Conducting Large-Scale Online Experiments with Uncompensated Samples. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing - CSCW '15*, 1364–1378. https://doi.org/10.1145/2675133.2675246
101. Luz Rello, Abdullah Ali, and Jeffrey P. Bigham. 2015. Dytective: Toward a Game to Detect Dyslexia. In *Proceedings of the 17th International ACM SIGACCESS Conference on Computers & Accessibility - ASSETS '15*, 307–308. https://doi.org/10.1145/2700648.2811351
102. Kai Sassenberg and Gordon B. Moskowitz. 2005. Don't stereotype, think different! Overcoming automatic stereotype activation by mindset priming. *Journal of Experimental Social Psychology* 41, 5: 506–514. https://doi.org/10.1016/j.jesp.2004.10.002
103. Michael Schmitz, Christoph Endres, and Andreas Butz. 2007. A Survey of Human-computer Interaction Design in Science Fiction Movies. In *Proceedings of the 2Nd International Conference on INtelligent TEchnologies for Interactive enterTAINment* (INTETAIN '08), 7:1–7:10. Retrieved February 20, 2019 from http://dl.acm.org/citation.cfm?id=1363200.1363210
104. Douglas Schuler and Aki Namioka. 1993. *Participatory Design: Principles and Practices*. CRC Press.
105. Nathan Shedroff and Chris Noessel. 2012. Make It So: Learning from Sci-fi Interfaces. In *Proceedings of the International Working Conference on Advanced Visual Interfaces* (AVI '12), 7–8. https://doi.org/10.1145/2254556.2254561
106. Pao Siangliulue, Joel Chan, Steven P. Dow, and Krzysztof Z. Gajos. 2016. IdeaHound: Improving Large-scale Collaborative Ideation with Crowd-Powered Real-time Semantic Modeling. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology - UIST '16*, 609–624. https://doi.org/10.1145/2984511.2984578
107. M. S. Silberman, B. Tomlinson, R. LaPlante, J. Ross, L. Irani, and A. Zaldivar. 2018. Responsible research with crowds. *Communications of the ACM* 61, 3: 39–41. https://doi.org/10.1145/3180492
108. David Canfield Smith. 1977. Pygmalion: A Computer Program to Model and Stimulate Creative Thought. Retrieved November 5, 2019 from https://link.springer.com/book/10.1007/978-3-0348-5744-4
109. David Canfield Smith. Designing the Star User Interface. 21.
110. Maximilian Speicher and Michael Nebeling. 2018. GestureWiz: A Human-Powered Gesture Design Environment for User Interface Prototypes. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems - CHI '18*, 1–11. https://doi.org/10.1145/3173574.3173681
111. Robert Stiratelli, Nan Laird, and James H. Ware. 1984. Random-Effects Models for Serial Observations with Binary Response. *Biometrics* 40, 4: 961–971. https://doi.org/10.2307/2531147
112. Anselm Strauss and Juliet M. Corbin. 1997. *Grounded Theory in Practice*. SAGE.
113. Yanke Tan, Sang Ho Yoon, and Karthik Ramani. 2017. BikeGesture: User Elicitation and Performance of Micro Hand Gesture As Input for Cycling. In (CHI EA '17), 2147–2154. https://doi.org/10.1145/3027063.3053075
114., 115. Theophanis Tsandilas. 2018. Fallacies of Agreement: A Critical Review of Consensus Assessment Methods for Gesture Elicitation. *ACM Transactions on Computer-Human Interaction* 25, 3: 1–49. https://doi.org/10.1145/3182168
116. John W. Tukey. 1949. Comparing Individual Means in the Analysis of Variance. *Biometrics* 5, 2: 99–114. https://doi.org/10.2307/3001913
117. John W. Tukey. 1953. The Problem of Multiple Comparisons. *Princeton, NJ: Princeton University*.
118. Rajan Vaish, Snehalkumar (Neil) S. Gaikwad, Geza Kovacs, Andreas Veit, Ranjay Krishna, Imanol Arrieta Ibarra, Camelia Simoiu, Michael Wilber, Serge Belongie, Sharad Goel, James Davis, and Michael S. Bernstein. 2017. Crowd Research: Open and Scalable University Laboratories. In *Proceedings of the 30th Annual ACM Symposium on User Interface Software and Technology* (UIST '17), 829–843. https://doi.org/10.1145/3126594.3126648
119., 120. Radu-Daniel Vatavu and Jacob O. Wobbrock. 2015. Formalizing Agreement Analysis for Elicitation Studies: New Measures, Significance Test, and Toolkit. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems - CHI '15*, 1325–1334. https://doi.org/10.1145/2702123.2702223
121., 122., 123. Radu-Daniel Vatavu and Jacob O. Wobbrock. 2016. Between-Subjects Elicitation Studies: Formalization and Tool Support. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems - CHI '16*, 3390–3402. https://doi.org/10.1145/2858036.2858228
124. Santiago Villarreal-Narvaez, Jean Vanderdonckt, Radu-Daniel Vatavu, and Jacob O Wobbrock. A Systematic Review of Gesture Elicitation Studies: What Can We Learn from 216 Studies? 26.

125. Tijana Vuletic, Alex Duffy, Laura Hay, Chris McTeague, Gerard Campbell, and Madeleine Grealy. 2019. Systematic literature review of hand gestures used in human computer interaction interfaces. *International Journal of Human-Computer Studies* 129: 74–94. https://doi.org/10.1016/j.ijhcs.2019.03.011
126.Colin Ware. 2010. *Visual Thinking: For Design.* Morgan Kaufmann. Retrieved November 13, 2019 from https://www.researchgate.net/profile/Colin_Ware/publication/200027485_Visual_Thinking_For_Design/links/53ce927c0cf2aada06e6a4b4.pdf
127., 128., 129. Jacob O. Wobbrock, Htet Htet Aung, Brandon Rothrock, and Brad A. Myers. 2005. Maximizing the guessability of symbolic input. In *CHI '05 extended abstracts on Human factors in computing systems - CHI '05*, 1869. https://doi.org/10.1145/1056808.1057043
130. Jacob O. Wobbrock, Leah Findlater, Darren Gergle, and James J. Higgins. 2011. The aligned rank transform for nonparametric factorial analyses using only anova procedures. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '11), 143–146. https://doi.org/10.1145/1978942.1978963
131., 132., 133., 134. Jacob O. Wobbrock, Meredith Ringel Morris, and Andrew D. Wilson. 2009. User-defined gestures for surface computing. In *Proceedings of the 27th international conference on Human factors in computing systems - CHI 09*, 1083. https://doi.org/10.1145/1518701.1518866
135. Anbang Xu, Shih-Wen Huang, and Brian Bailey. 2014. Voyant: generating structured feedback on visual designs using a crowd of non-experts. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing - CSCW '14*, 1433–1444. https://doi.org/10.1145/2531602.2531604
136. Teng Ye, Katharina Reinecke, and Lionel P. Robert. 2017. Personalized Feedback Versus Money: The Effect on Reliability of Subjective Data in Online Experimental Platforms. In *Companion of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing - CSCW '17 Companion*, 343–346. https://doi.org/10.1145/3022198.3026339
137. Zhenya Zhang, Hongmei Cheng, Shuguang Zhang, Wanli Chen, and Qiansheng Fang. 2008. Clustering aggregation based on genetic algorithm for documents clustering. In *2008 IEEE Congress on Evolutionary Computation, CEC 2008*, 3156–3161. https://doi.org/10.1109/CEC.2008.4631225
138., 139. Kathryn Zyskowski, Meredith Ringel Morris, Jeffrey P. Bigham, Mary L. Gray, and Shaun K. Kane. 2015. Accessible Crowdwork?: Understanding the Value in and Challenge of Microtask Employment for People with Disabilities. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing - CSCW '15*, 1682–1693. https://doi.org/10.1145/2675133.2675158
140. analysis of binomial data by a generalized linear mixed model | Biometrika | Oxford Academic. Retrieved February 3, 2020 from https://academic.oup.com/biomet/article-abstract/72/3/593/253948?redirectedFrom=fulltext

Biographical Sketch

Abdullah was born in Baghdad, Iraq, in 1989. His mother was the CEO of their family business, FOMEDA medical supplies, the second largest medical supplies company in Iraq. His father was a diplomat who served as a council at the Iraqi embassy in Rome, Italy, before the Iraq war in 2003. Abdullah has a younger sister who is a mechanical engineer, and global climate change activist. Abdullah attended the distinguished boys high school in Baghdad after placing 8th in a country-wide admission test. Abdullah arrived in the United States in 2008 as a war refugee. He started his academic journey at the Community College of Baltimore county and transferred to the University of Maryland Baltimore County. There, he joined Dr. Amy Hurst's Prototyping and Design lab where he discovered his passion for computing research. He started pursuing a research-focused master's degree in Human-Centered Computing thanks to a two-year NSF fellowship. During his master's degree, he collaborated with numerous researchers and published many papers focused on accessibility research. Most notably, he co-invented the game Dytective while spending a summer at Carnegie Mellon University.

Abdullah moved to Seattle, WA, in 2016 to start his Ph.D. at The Information School of the University of Washington. During the following four years, Abdullah completed the work described in this dissertation, spent a summer at Microsoft Research examining how people with dyslexia search the web, and a summer working at Apple in Cupertino, CA, as a lead interaction designer. In 2017, Abdullah met his partner Paige B. Collins. Abdullah and Paige will be moving to New York City, NY, to start a new life chapter. Paige is a product manager at The New York Times, and Abdullah will be joining Amazon's Web Services to start a design studio offering design services to AWS's enterprise and startup clients.