Crowdlicit

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

Abdullah X. Ali
The Information School
DUB Group
University of Washington
Seattle, WA 98195 USA
xyleques@uw.edu

Meredith Ringel Morris
Microsoft Research
Redmond, WA 98052 USA
merrie@microsoft.com

Jacob O. Wobbrock
The Information School
DUB Group
University of Washington
Seattle, WA 98195 USA
wobbrock@uw.edu

ABSTRACT

End-user elicitation studies are a popular design method. Currently, such 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 paper, we address some of the limitations of elicitation studies through the creation of the *Crowdlicit* system 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. We used Crowdlicit to run a crowd-powered elicitation study based on Morris's "Web on the Wall" study (2012) with 78 participants, arriving at a set of symbols that included six new symbols different from Morris's. We evaluated the effectiveness of 49 symbols (43 from Morris and six from Crowdlicit) by conducting a crowd-powered identification study. We show that the Crowdlicit elicitation study resulted in a set of symbols that was significantly more identifiable than Morris's.

CCS CONCEPTS

• Human-centered computing → HCI design and evaluation methods

KEYWORDS

End-user elicitation study; end-user identification study; user-driven design; crowdsourcing; Mechanical Turk.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from Permissions@acm.org.
CHI 2019, May 4–9, 2019, Glasgow, Scotland UK.
© 2019 Association for Computing Machinery.
ACM ISBN 978-1-4503-5970-2/19/05…$15.00
https://doi.org/10.1145/3290605.3300485

ACM Reference format:
Abdullah X. Ali, Meredith Ringel Morris, Jacob O. Wobbrock. 2019. Crowdlicit: A System for Conducting Distributed End-User Elicitation and Identification Studies. In *2019 CHI Conference on Human Factors in Computing Systems Proceedings (CHI 2019), May 4–9, 2019, Glasgow, Scotland, UK.* ACM, New York, NY, USA. https://doi.org/10.1145/3290605.3300485

1 INTRODUCTION

Eliciting input from end-users to design system interactions is a common practice. Perhaps the earliest example is Good *et al.*'s [14] work generating user-driven commands for command-line interfaces. Wobbrock *et al.* [40,41] formalized the method of end-user elicitation studies in the lab. The 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 *referent*), and ask the participants to propose the action (known as a *symbol*) 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 proposed symbols into groups based on their similarity. The group with the highest consensus is chosen as the representative symbol to invoke its associated referent.

Elicitation studies have gained popularity in recent years, with more than 170 published studies employing the method. They have been used to design gesture interactions for touchscreens [16,41], virtual and augmented reality interactions [21,34], TV controls [11,37], in-vehicle interactions [23], drone navigation controls [7], interactions for Internet-of-Things devices [19], and human-robot interactions [33]. Elicitation studies have also been used to explore interaction designs with populations like people who are blind [16] and children [9].

The premise of end-user elicitation studies is that by eliciting symbolic input 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, hence 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-generated symbols, the method has another limitation: the absence of a formal approach to evaluate such studies' results.

In this work, we address the limitations above. First, we 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, we present the *end-user identification method*, which reverses aspects of the elicitation study methodology. Participants in identification studies are shown a symbol and asked to suggest the referent invoked by it. Researchers are then able to assess the identifiability of their symbols.

To conduct elicitation and identification studies online efficiently, we created a system called *Crowdlicit*, making it available [1] 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. We built Crowdlicit to flexibly support studies that present referents in different formats (*e.g.*, text, images, videos, and audio) and collect symbols of varying modalities (*e.g.*, gestures, voice commands, icon sketches). Also, Crowdlicit provides a centralized way to organize study results and export them for analysis. 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, we conducted a distributed elicitation study based on Morris's lab-based "Web on the Wall" elicitation study [26]. Our study had 78 participants recruited from Amazon's Mechanical Turk (mTurk). We asked participants to propose free-form gestures or voice commands to interact with a TV-based web browser. We arrived at 15 symbols for the 15 referents Morris identified for controlling a web browser on a TV. Morris's symbol set had 43 symbols because it included synonym symbols for each referent (*i.e.*, different actions to invoke the same effect). Our 15 symbols had six symbols different than Morris's. We evaluated the identifiability of all 49 symbols (43 from Morris, six new ones from Crowdlicit) by running an end-user identification study using the Crowdlicit system with 24 new participants. We found that Crowdlicit's set of symbols was significantly more identifiable than Morris's. We also report on participants' feedback on Crowdlicit.

This paper 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.

2 RELATED WORK

Relevant prior work includes numerous studies eliciting user input, methodological extensions to end-user elicitation studies, and work done conducting online HCI research.

2.1 End-User Elicitation Studies

Good *et al.*'s [14] user-driven command-line interface work is possibly the earliest example of an elicitation study, although the term was never used. Wobbrock *et al.* [40,41] formalized the end-user elicitation method around gesture interactions and presented conflict resolution techniques and agreement calculations. Many researchers have utilized Wobbrock *et al.'s* method to design interactions for different types of technologies for various populations. For example, Morris [26] used the method to design a Kinect-based TV web browser, eliciting both gestures and voice

[1] Crowdlicit is available at http://depts.washington.edu/madlab/proj/crowdlicit/

commands. Nebeling *et al.* [30] replicated Morris's study and built a Kinect-based system [32] to capture and classify interactions. Nebeling also used the method to design cross-device interactions [31]. May *et al.* [23] used the method to develop in-vehicle mid-air gestures. Kühnel *et al.* [19] and Desolda *et al.* [10] have used the method to design interactions with Internet-of-Things devices. Leng *et al.* [22] designed gestures for music interaction in a virtual environment and Piumsomboon *et al.* [34] used the approach to develop gestures for augmented-reality environments. Dim *et al.* [11] and Vatavu [37] used the method to design interactions with TV systems. Cauchard *et al.* [7] employed the method to explore natural human-drone interactions. Connell *et al.* [9] used this approach with children to define whole-body gestures. Other researchers capitalized on the method's inclusivity of end-users' abilities to design interactions for blind populations [5,11,16,25].

2.2 Extending the End-User Elicitation Methodology

There has been work extending Wobbrock *et al.'s* [41] original method. Some of this work proposes updated agreement measures [12,36,38,39]. There is also work on the role legacy bias plays in elicitation studies. Whereas Morris *et al.* [27] and Nebeling *et al.* [31] argue that legacy bias should be minimized, Köpsel and Bubalo [18] and Hoff *et al.* [15] maintain that there are benefits to having participants suggest interactions similar to ones employed in existing technologies. There is also some work evaluating user-defined interactions and demonstrating the benefits of the elicitation study methodology by showing that user-defined interactions were more memorable than and preferable to those created by designers; see, *e.g.*, Wobbrock *et al.* [40], Morris *et al.* [28] and Nacenta *et al.* [29].

There are also tools addressing aspects of gesture elicitation studies. However, the work we present in this paper aims to streamline the entire process of carrying out an elicitation study by providing a web-based interaction-and-platform-agnostic system that enables researchers to conduct elicitation studies efficiently at scale in the crowd. To our knowledge, our work is the first attempt to conduct online elicitation studies, let alone build a general-purpose tool to facilitate such studies. We sought to understand how to conduct elicitation studies online, formalized a method to evaluate elicitation study results, and used the method to compare the results of two elicitation studies—online and in the lab.

2.3 HCI Research and Online Crowds

Using online crowds in HCI research is a regular practice. Many impactful papers have utilized online crowds in their research [1,17]. Due to the diversity of online participants [6] and the improved accessibility of online work [42], online participants have been used to edit documents [3], ideate solutions for social paradigms [8], prototype interactive interfaces [20], and aid blind individuals in understanding their surroundings [4].

We have published work adapting other aspects of elicitation studies to be online. We have utilized online crowd-workers to provide similarity judgments for symbol agreement analysis in elicitation studies [2]. We coupled the crowd-workers' votes with machine learning algorithms to conduct agreement analyses. Our approach provided results of the same quality as experts and was four times faster.

The work presented in this paper adds to the field of online crowd-powered HCI research by adapting the symbol elicitation aspect of end-user elicitation studies from the lab to the online crowd.

3 THE CROWDLICIT SYSTEM

This section details the requirements Crowdlicit had to meet to successfully adapt the elicitation study methodology to be online.

3.1 System Requirements

We 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.

3.1.1 What is an Elicitation Study? 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 *referent*, and are asked to propose the action, known as a *symbol*, meant to invoke the effect. Researchers collect symbols, and other data such as subjective ratings, demographic information, and study notes from participants representing the target end-user population. Researchers then cluster similar symbols into groups to find the symbols with the maximum consensus to trigger each referent for the computing system they are designing.

3.1.2 What is an Identification Study? An end-user identification study is a new evaluation method for the symbols 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 present end users with symbols (actions for invoking effects on a computing system, *e.g.*, mid-air or stroke gestures, command-line or voice commands, button icons or labels, etc.). Researchers then ask users to propose the referent (the effect on the computing system, *i.e.*, what the symbol would *do*), usually without giving knowledge of the commands available in the target system. Researchers aggregate the user-generated referents in groups based upon similarity and proceed by either confirming the symbol-referent appropriateness or assigning new referents to symbols that had low referent-identifiability. Wobbrock *et al.* [41] called for such studies in the limitations and next steps section of the paper that formulated the elicitation method saying, *"An important next step is to validate our user-defined gesture set. Unlabeled video clips of the gestures can be shown to 20 new participants, along with clips of designers' gestures, to see if people can guess which gestures perform which commands. (This, in effect, reverses the current study to go from signs to referents, rather than from referents to signs.)"*

3.1.3 Crowdlicit Requirements. The requirements R1-R6 below are phrased in terms of an elicitation study for simplicity. The same requirements apply to identification studies, but with the role of referents and symbols reversed.

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

R2. Referent presentation. As prior work shows that elicitation studies have used various referent formats (*e.g.*, text [2], videos [24]), it is important to maximize referent-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.* [27].

R4. Symbol modality. Prior work has demonstrated that elicitation studies can be applied to various fields (*e.g.*, AR environments [34], in-vehicle interactions [23]). It is imperative to maximize symbol-type flexibility for researchers.

R5. Contextual richness. Prior studies have gathered symbol ratings, think-alouds, and other study notes (*e.g.*, [30,41]). It is important to gather information besides symbols to provide researchers with rich study results.

R6. Data analysis. Analyzing the results of elicitation studies is a complex and time-consuming process [2]; hence, it is important that our 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 with a title, description, an optional post-study survey link, and a dedicated unique URL. A study serves as a container holding referents and elicited symbols. 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 post-study survey allows researchers to enter a URL to an external survey they would like participants to complete upon finishing the elicitation study. These options satisfy the first requirement, *R1-Study definition,* from our list above.

3.2.1 Referent Presentation. Crowdlicit is designed to provide maximum flexibility when creating a referent. Each referent has a title, instructions, and the referent itself. Crowdlicit offers four ways to present a referent. (1) A text string describing the effect of an action on a computing system. (2) An audio clip of the referent 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 referent 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 (*e.g.*, two screenshots side-by-side showing the before and after states of a system). (4) A video showing the referent (*e.g.*, screen recordings). Having multiple ways to present a referent satisfies the *R2-Referent presentation* requirement.

3.2.2 Symbol Preferences. In elicitation studies, researchers collect symbols 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 [26,30]. But the elicitation method applies beyond just gestural interactions to other input modalities, such as sketches of icons. 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 symbols to collect from participants satisfies *R4-Symbol modality* from the list of requirements we outlined for the Crowdlicit system.

3.2.3. Reducing Legacy Bias. Morris *et al.* [27] 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 with plans to add support for partner-based elicitation studies in the future.

The capabilities to select a production option and add priming capabilities satisfy *R3-Legacy bias reduction*.

3.2.4 Post-Task Questions. Crowdlicit includes the option to add symbol-rating Likert scale questions assessing the ease and fit of symbols derived from Wobbrock *et al.'s* original method [41], with an option to add an additional "custom" question. Having participants rate their proposed symbols provides more in-depth insight into the appropriateness of their symbols, 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 current 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. Each Crowdlicit study has a Welcome page, a Task Manager page, an Elicitation Interface page, and a Thank You page.

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 1.1).

3.3.2 Task Manager Page. This page shows a list of tasks (*i.e.*, set of referents in the case of an elicitation study, or set of symbols in the case of an identification study) that the participants have to complete.

3.3.3 Elicitation Interface. This page displays the referent and priming content, if included. It collects symbol input from participants, asks them to rate their symbols, and shows them their progress. On this page, participants see instructions and the actual referent in whatever modality the researchers choose, *i.e.*, text, video, audio, or an image (Figure 1.2). 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 symbol-input interface. The symbol-input interface changes based on the type of symbol the researchers want their participants to propose. For text input, the participants see a text area input element (Figure 1.4). For images and videos, the participants capture an image or video using their device's camera.[2] To draw a symbol or perform a stroke gesture, participants see a canvas element. If the researchers want participants to propose the modality of the symbol in addition to the symbol itself, they can choose the "User-dictated" option when setting symbol-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 1.3). Participants click on the modality that they deem to be most appropriate for the task, and then the appropriate type of symbol-input screen appears.

3.3.4 Post-Task Questions. Below the symbol input interface are the Likert 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 symbols collected for a specific study, organized by the referent for which they were proposed and meant to invoke. Symbols are listed along with their

[2] Currently, the beta version of Crowdlicit uses webcams to capture images and videos. Future work will allow users to upload multimedia files or use smartphones to capture images and videos.

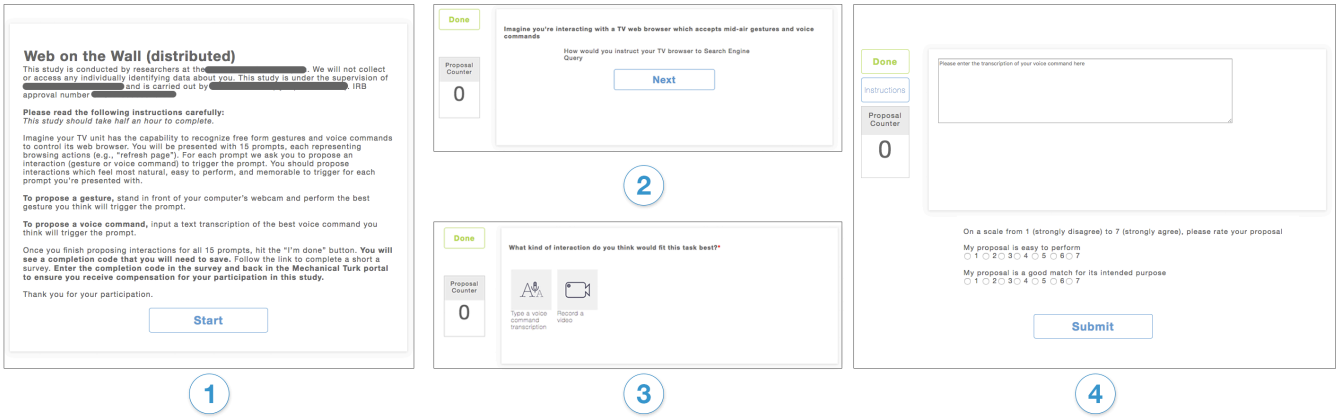

Figure 1. Screenshots of a study created with Crowdlicit. (1) The Welcome page shows the instructions for participating in a study entitled, "Web on the Wall (distributed)." (2) A text referent. (3) An interface allowing participants to choose between proposing a voice-command or a gesture. (4) A text-based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

elicitor's unique identification code and symbol ratings. The researchers have the option to delete specific symbols. They can also export the study results as a `*.csv` file to either conduct the agreement analysis themselves [40,41], or utilize an online crowd to handle the analysis for them by using our Crowdsensus tool [2] 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, we ran an elicitation study based on Morris's "Web on the Wall" study [26], which has previously been the focus of replication studies in this genre (*e.g.*, [2,30]). We also ran an identification study, the reverse of an elicitation study, to evaluate the results of our elicitation study and that of Morris's original study. Crowdlicit's flexibility in presenting referents and collecting symbols of different formats allowed us to run our identification study online. We asked all of our 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

We ran a study using Crowdlicit based on Morris's "Web on the Wall" lab-based elicitation study [26] with 78 participants from Amazon's Mechanical Turk (mTurk). Fifty participants completed the entire study—double that of Morris [26]—and 28 gave partial answers, which we include in our 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 our state's $11/hour minimum wage.

Table 1. Demographic information for 33 of 78 participants from our elicitation study (study 1) and 22 of 24 participants from our identification study (study 2).[3]

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%

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

After participants submitted a minimum of one symbol per referent for all 15 referents, 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. We used the completion codes to link the demographic information to study answers and identify which participants completed the entire study.

4.2 End-User Identification Study

We recruited 24 new participants from mTurk for an identification study. They provided open-ended referent proposals for all 49 symbols (43 from Morris's study plus six new ones unearthed by our elicitation study). Of the 24 participants, 22 completed the post-study survey; Table 1 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 symbol describing a gesture or voice command instead of a referent. Study instructions asked participants to imagine they were interacting with a TV-based web browser. For every symbol (Table 2), participants were asked to freely propose one referent in text form. The HIT required about an hour to complete and paid $11 USD, our state's minimum wage.

Table 2. Morris's 43 symbols [26]. Crowdlicit' 15 symbols. "*" are new symbols from the Crowdlicit study. The symbols describe gestures; symbols in quotes ("") are voice commands. The # column shows the number of participants who proposed the symbol. The "A" column shows the referent agreement score for each symbol from the identification study.

REFERENT	MORRIS SYMBOL	#	A	CROWDLICIT SYMBOL	#	A
1. 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			
2. Search Engine Query	5."<query>"	6	0.30	6."search <query>"	40	0.25
	6."search <query>"	5	0.25			
3. 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			
4. 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			
5. 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			
6. 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
7. 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			
8. 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			
9. 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			
10. 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			
11. 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			
12. 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			
13. Bookmark Page	35.hand-as-mouse selects bookmark button	7	0.43	36."bookmark page"	33	0.71
	36."bookmark page"	5	0.71			
14. 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			
15. 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			

4.3 Post-Study Survey

Each of our participants completed a survey asking demographic questions and about previous research participation (see Table 1). Participants also reported on their technology use and experience with the Crowdlicit interface. We based some of the survey questions on Finstad's [13] 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

We evaluated our symbols and Morris's [26] by conducting an identification study. Participants from both our elicitation and identification studies completed a post-study survey.

5.1 Crowdlicit Symbols

We grouped the symbols proposed for each referent based on their similarity and generated 15 voice commands to trigger our 15 referents (Table 2). For each referent, participants collectively proposed an average of five gestures that had little agreement; this led us to discard gesture proposals and focus solely on the elicited voice commands. Of our 15 symbols, nine were the same voice-command symbols Morris arrived at in her study [26], and six were new. The number of participants who proposed the selected symbols ranged from 23 – 76. Since we used the production principle to reduce legacy bias in our study [27], our participants were free to propose multiple symbols per referent. Having several symbols from a single participant leads to an unequal number of proposals among referents, making Wobbrock *et al.'s* [40,41] agreement equation unsuitable. Instead, we used Morris's [26] *max-consensus* to compare agreement between referents. Max-consensus is the percentage of participants suggesting the most popular symbol for a referent [26]. Our participants proposed symbols with high consensus (M=59%, SD=9%). Figure 2 shows the max-consensus for our 15 symbols.

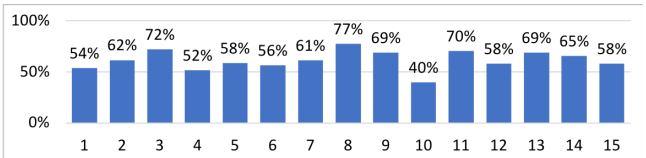

Figure 2. Max-consensus for referents 1–15.

5.2 Symbol Identification

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

5.2.1 Referent Agreement. For each symbol (Table 2), we calculated the referent agreement using Wobbrock *et al.'s* [40,41] original agreement equation:

$$A_s = \sum_{P_i \subseteq P_r} \left(\frac{|P_i|}{|P_r|} \right)^2 \quad (1)$$

In Eq. 1, A_s is the agreement of referents proposed for symbol s, P_r is the set of all referents proposed for symbol s, and P_i is a subset of similar referents in P_r. Table 2 lists all 49 referent agreement scores.

5.2.2 Accuracy. We compared the referent with highest consensus from the identification study to the original referent. The original referent is the one used in the elicitation studies (Morris's and Crowdlicit). Identification study participants were able to correctly identify the referent for each one of the 15 symbols in the Crowdlicit symbol set. In this case, "identify" means that the referent with the highest consensus from the list of referents proposed for a symbol matched the original referent for that symbol.

Table 3. Five symbols; their original referents from Morris's study, the accuracy % of the original referent, the new referent, and the max-consensus % of the new referent. Symbols in quotes are voice commands.

Symbol	Original Referent	Accuracy	New Referent	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%

For Morris's set of 43 symbols, participants were able to correctly identify the referents for 38 symbols and assigned new referents for five symbols. Table 3 lists the five symbols with new referents from Morris's study, the original referents, and the percentage of proposals of the

original referent—we refer to it here as "accuracy." The table also lists the newly assigned referents, and their percentage of the total number of proposed referents.

5.2.3 Comparability. We compared the agreement scores of Morris's symbols to Crowdlicit's symbols 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.) We found that the symbols' referent-agreement scores resulting from the Crowdlicit elicitation study were significantly higher than Morris's symbols ($t(37) = 2.99$, $p < .005$). In addition to higher agreement scores, the fact that all of Crowdlicit's symbols were correctly identified as a result of the identification study leads us to believe that the symbol set resulting from the Crowdlicit study was more identifiable than Morris's.

5.3 Interaction Habits and 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 3 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 symbols resulting from the Crowdlicit elicitation study is made up entirely of voice commands.

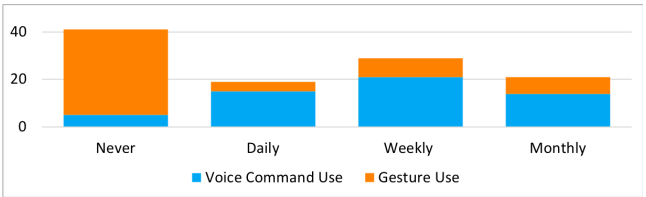

Figure 3. The frequency of using voice *vs.* mid-air gestures to interact with technology from 55 participants: 33 from the elicitation study and 22 from the identification study.

Participants generally had a positive experience interacting with Crowdlicit itself (Table 4). 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 symbols after submitting—a feature we are including as future work. 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."*

Table 4. 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.

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)

6 DISCUSSION

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

6.1 Crowdlicit Benefits

We demonstrated that running an elicitation study using Crowdlicit is not only possible but yields more identifiable symbols 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 our first study, an elicitation study based on Morris's "Web on the Wall" study [26], it took six hours to recruit and collect data from 78 participants. Fifty of our 78 participants completed the entire study, twice the number of participants as Morris's study [26], which had 25. Twenty-eight of our participants gave partial answers. Morris's study took about 12 hours to run (12 groups of participants × 1 hour per session as reported in [26]). Crowdlicit allowed us 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 [28].

Our 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 we only asked people who were already participating in our online study. However, participants' willingness to partake in online studies over lab-based ones complements findings by Zyskowski *et al.* [42] 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

We collected a number of spam answers in our elicitation study. The first task in the study (open browser) had 58 unusable symbols out of 224 total elicited symbols.

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. We attempted to limit spam answers by assigning a minimum time threshold of 10 seconds to answers, 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. We intend further measures to limit spam in future work (*e.g.*, by adding the option to randomize the order of tasks in the Task Manager page). By limiting spam, researchers can collect data efficiently from a large number of remote participants.

Our participants reported that they used voice-commands more frequently than gestures in their daily lives, which could explain why the set of symbols we gathered in our Crowdlicit elicitation study mostly comprised voice-commands. Another contributing factor for the popularity of voice-commands in our 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.

In our work, we relied on the imagination of our remote participants, which provided sound results. However, Crowdlicit and online elicitation studies might not be well-suited to designing interactions for novel technologies that require users to imagine unfamiliar environments *e.g.*, virtual reality or entirely new platforms.

6.3 Identification Studies

We formalized the end-user identification study method to evaluate the symbols from two elicitation studies. We calculated referent agreement using a form of Wobbrock *et al.'s* [40,41] original symbol agreement equation (Eq. 1). Agreement can also be measured using one of the other agreement measures proposed by the extensive prior work published on elicitation studies [12,36,38,39].

Our *comparability* analysis can be used to compare two sets of symbols meant to invoke the same set of referents. 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.

We showed in this work that voice commands were more identifiable than gestures for this particular use case. In this case, we 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 our participants had more experience interacting with voice-enabled technologies (see Figure 3) than devices that accept mid-air gestures like the Microsoft Kinect, which was used in Morris's study [26]. This preference may indicate the influence of legacy bias on users' preferences.

Crowdlicit's set of symbols was more identifiable than Morris's as participants were able to correctly identify all 15 referents for the Crowdlicit symbol set. For Morris's symbol set, the identification study participants were able to correctly identify 38 referents for the 43 symbols and assigned new referents to five symbols. In addition to the correct identification of all referents for the Crowdlicit symbol set, the agreement rates for the Crowdlicit set were significantly higher than for Morris's set. We attribute Crowdlicit's symbol set's high identifiability to the larger pool of participants proposing the symbols—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, we decided to run our elicitation study unsupervised; the lack of supervision gave us the advantage of collecting a large amount of data in a short amount of time. On the other hand, we did not have any study notes or added insight into the participants' answers besides the symbols they proposed, some basic demographic information, and the feedback they gave us on Crowdlicit.

The majority of our participants were college-educated and from the United States. We intend to follow up this work with studies with more diverse populations. For example, running studies with people from different parts of the world and comparing their results, or running studies with people with disabilities to create inclusive technologies, would be useful ways to exercise and extend Crowdlicit. Given that the work we present in this paper is the first attempt at conducting distributed elicitation studies at all, additional studies exploring different interaction modalities would only enhance our understanding of the strengths and limitations of our approach and the Crowdlicit system. Also, we wish to

replicate published elicitation studies with more participants to help generalize (or refute) their findings; examples are [9,27,33]. Crowdlicit makes running iterative elicitation studies practical—an important aspect of its contribution.

Understanding the limits of online elicitation studies for novel modalities, such as VR, is also an important area for future work. However, for elicitation studies that require special equipment or environments, Crowdlicit can still offer benefits such as: (1) use participants' imaginations, like we demonstrated in our study, to get fast preliminary results that inform more in-depth lab-based elicitation; (2) organize, store, and export the collected data from in-lab studies for efficient analysis using the Crowdsensus tool [2]; and (3) evaluate the results of a lab-based elicitation study using online identification studies.

The identification study method we put forth in this work measures the identifiability of symbols by having participants propose referents—the reverse of how an elicitation study works. We imagine future work extending this method to have participants match symbols to referents from a list rather than freely proposing referents from an "infinite" set of possibilities. We would also extend this method to add new measures, *e.g.*, those capturing preferences, ease of use, and aesthetic appeal, to name a few.

7 CONCLUSION

This paper reports on *Crowdlicit*, a system for conducting distributed elicitation and identification studies. We also introduced *end-user identification studies*, which are the reverse of elicitation studies, as studies that evaluate the identifiability of interface actions and symbols and how well they map to intended referents. Our 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 our hope that researchers, designers, and developers will use Crowdlicit to efficiently run crowd-powered end-user elicitation studies, gaining quality data in little time.

ACKNOWLEDGMENTS

This work was supported in part by funding from Microsoft Research, the Mani Charitable Foundation, and the National Science Foundation under grant IIS-1702751. Any opinions, findings, conclusions or recommendations expressed in our work are those of the authors and do not necessarily reflect those of any supporter.

REFERENCES

[1] Luis von Ahn, and Laura Dabbish. (2004). Labeling images with a computer game. *Proceedings of the ACM conference on Human factors in computing systems (CHI'04)*. Vienna, Austria. New York: ACM Press, pp. 319–326.

[2] Abdullah X. Ali, Meredith R. Morris, and Jacob O. Wobbrock. (2018). Crowdsourcing Similarity Judgments for Agreement Analysis in End-User Elicitation Studies. *Proceedings of the ACM Symposium on User Interface Software and Technology (UIST '18)*. Berlin, Germany. New York: ACM Press, pp. 177-188.

[3] Michael S. Bernstein, Greg Little, Robert C. Miller, Björn Hartmann, Mark S. Ackerman, David R. Karger, David Crowell, and Katrina Panovich. (2010). Soylent: a word processor with a crowd inside. *Proceedings of the 23nd annual ACM symposium on User interface software and technology (UIST'10)*. New York, NY. New York: ACM Press, pp. 313–322.

[4] Jeffrey P. Bigham, Chandrika Jayant, Hanjie Ji, Greg Little, Andrew Miller, Robert C. Miller, Aubrey Tatarowicz, Brandyn White, Samual White, and Tom Yeh. (2010). VizWiz: nearly real-time answers to visual questions. *Proceedings of the ACM symposium on User interface software and technology (UIST '10)*. New York, NY. ACM, New York, NY, USA, pp. 333-342.

[5] Maria Claudia Buzzi, Marina Buzzi, Barbara Leporini, and Amaury Trujillo. (2015). Exploring Visually Impaired People's Gesture Preferences for Smartphones. *Proceedings of the 11th Biannual Conference on Italian SIGCHI Chapter (CHItaly'15)*. Rom, Italy. New York: ACM Press, pp. 94–101.

[6] Krista Casler, Lydia Bickel, and Elizabeth Hackett. (2013). Separate but equal? A comparison of participants and data gathered via Amazon's MTurk, social media, and face-to-face behavioral testing. *Computers in Human Behavior* 29, (6) pp. 2156–2160.

[7] Jessica R. Cauchard, Jane L. E., Kevin Y. Zhai, and James A. Landay. (2015). Drone & me: an exploration into natural human-drone interaction. *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp'15)*. Osaka, Japan. New York: ACM Press, pp. 361–365.

[8] Joel Chan, Steven Dang, and Steven P. Dow. (2016). Improving crowd innovation with expert facilitation. *Proceedings of the ACM Conference on Computer-Supported Cooperative Work & Social Computing* (CSCW'16). San Francisco, CA. New York: ACM Press, pp. 1223-1235.

[9] 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. *Proceedings of the 12th International Conference on Interaction Design and Children (IDC'13)*. New York, NY. New York: ACM Press, pp. 277–280.

[10] Giuseppe Desolda, Carmelo Ardito, and Maristella Matera. (2017). Empowering End Users to Customize their Smart Environments. *ACM Transactions on Computer-Human Interaction* 24, (2) pp. 1–52.

[11] Nem Khan Dim, Chaklam Silpasuwanchai, Sayan Sarcar, and Xiangshi Ren. (2016). Designing Mid-Air TV Gestures for Blind People Using User- and Choice-Based Elicitation Approaches. *Proceedings of the 2016 ACM Conference on Designing Interactive Systems (DIS'16)*. Brisbane, Australia. New York: ACM Press, pp. 204–214.

[12] Leah Findlater, Ben Lee, and Jacob O. Wobbrock. (2012). Beyond QWERTY: augmenting touch screen keyboards with multi-touch gestures for non-alphanumeric input. *Proceedings of the ACM conference on Human Factors in Computing Systems (CHI'12)*. Austin, Texas. New York: ACM Press, pp. 2679–2682.

[13] Kraig Finstad. (2013). Response to commentaries on "The usability metric for user experience." *Interacting with Computers* 25, (4) pp. 327–330.

[14] 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) pp. 1032–1043.

[15] Lynn Hoff, Eva Hornecker, Sven Bertel. (2016). Modifying Gesture Elicitation : Do Kinaesthetic Priming and Increased Production Reduce Legacy Bias ? *Proceedings of the Tenth International Conference on Tangible, Embedded, and Embodied Interaction (TEI'16)*. Eindhoven, Netherlands. New York: ACM Press, pp. 86–91.

[16] Shaun K. Kane, Jacob O. Wobbrock, and Richard E. Ladner. (2011). Usable gestures for blind people. *Proceedings of the 2011 annual conference on Human factors in computing systems (CHI'11)*. Vancouver, Canada. New York: ACM Press, pp. 413-422.

[17] Aniket Kittur, Ed H. Chi, Bongwon and Suh. (2008). Crowdsourcing user studies with Mechanical Turk. In *Proceeding of the twenty-sixth annual conference on Human factors in computing systems (CHI'08)*. Florence, Italy. New York: ACM Press, pp. 453-456.

[18] Anne Köpsel, Nikola and Bubalo. (2015). Benefiting from legacy bias. *Interactions* 22, (5) pp. 44–47.

[19] Christine Kühnel, Tilo Westermann, Fabian Hemmert, Sven Kratz, Alexander Müller, and Sebastian Möller. (2011). Im home: Defining and evaluating a

gesture set for smart-home control. *International Journal of Human Computer Studies* 69, (11) pp. 693–704.

[20] 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)*. Quebec City , Canada. New York: ACM Press, pp. 817–828.

[21] Hoo Yong Leng. (2017). A User-Defined Gesture Set for Music Interaction in Immersive Virtual Environment. (2) pp. 44–51.

[22] Hoo Yong Leng, Noris Mohd Norowi, and Azrul Hazri Jantan. (2017). A User-Defined Gesture Set for Music Interaction in Immersive Virtual Environment. *Proceedings of the 3rd International Conference on Human-Computer Interaction and User Experience in Indonesia (CHIuXiD'17,)* Jakarta, Indonesia. New York: ACM Press, pp. 44-51.

[23] 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)*. Oldenburg, Germany. New York: ACM Press, pp. 74–83.

[24] Erin Mcaweeney,Haihua Zhang, Michael Nebeling, M. (2018). User-Driven Design Principles for Gesture Representations. *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI'18)*.Montreal, Canada. New York: ACM Press, pp. 1–13.

[25] David McGookin, Stephen Brewster, and WeiWei Jiang. (2008). Investigating touchscreen accessibility for people with visual impairments. *Proceedings of the 5th Nordic conference on Human-computer interaction building bridges (NordiCHI'08)*. Lund, Sweden. New York: ACM Press, pp. 298-307.

[26] Meredith R. Morris. (2012). Web on the Wall: Insights from a Multimodal Interaction Elicitation Study. *Proceedings of the 2012 ACM international conference on Interactive tabletops and surfaces (ITS'12)*. Cambridge, MA. New York: ACM Press, pp. 95–104.

[27] Meredith R. Morris, Andreea Danielescu, Steven Drucker, Danyel Fisher, Bongshin Lee, m. c. Schraefel, and Jacob O. Wobbrock. (2014). Reducing legacy bias in gesture elicitation studies. *Interactions* 21, (3) pp. 40–45.

[28] Meredith R. Morris, Jacob O. Wobbrock, and Andrew D. Wilson. (2010). Understanding users' preferences for surface gestures. *Proceedings of Graphics Interface (GI'10)*. Ottawa, CA. New York: ACM Press, pp. 261–268.

[29] Miguel A. Nacenta, Yemliha Kamber, Yizhou Qiang, and Per Ola Kristensson, P. O. (2013). Memorability of pre-designed and user-defined gesture sets. *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI'13)*. Paris, France. New York: ACM Press, pp. 1099-1108.

[30] 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. *Proceedings of the 9th ACM International Conference on Interactive Tabletops and Surfaces (ITS'14)*. Dresden, Germany. New York: ACM Press, pp. 15–24.

[31] Michael Nebeling. (2017). XDBrowser 2 . 0: Semi-Automatic Generation of Cross-Device Interfaces. *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI'17)*. Denver, Colorado. New York: ACM Press, pp. 4574–4584.

[32] Michael Nebeling, David Ott, and Moira C. Norrie. (2015). Kinect Analysis: A System for Recording, Analysing and Sharing Multimodal Interaction Elicitation Studies. *Proceedings of the ACM SIGCHI Symposium on Engineering Interactive Computing Systems (EICS'15)*. Duisburg, Germany. New York: ACM Press, pp. 142-151.

[33] Mohammad Obaid, Markus Häring, Felix Kistler, René Bühling, and Elisabeth André. (2012). User-defined body gestures for navigational control of a humanoid robot. *Proceedings of the International Conference on Social Robotics*. Berlin: Springer, pp. 367-377.

[34] Thammathip Piumsomboon, Adrian Clark, Mark Billinghurst, and Andy Cockburn. (2013). User-defined gestures for augmented reality. *Proceedings of INTERACT 2013*. Berlin: Springer, pp. 282-299.

[35] Katharina Reinecke, and Krzysztof Z. Gajos. (2015). LabintheWild: Conducting Large-Scale Online Experiments With Uncompensated Samples. *Proceedings of the ACM Conference on Computer Supported Cooperative Work & Social Computing (CSCW'15)*. Vancouver, Canada. New York: ACM Press, pp. 1364–1378.

[36] Theophanis Tsandilas. (2018). Fallacies of Agreement: A Critical Review of Consensus Assessment Methods for Gesture Elicitation. *ACM Transactions on Computer-Human Interaction* 25, (3) pp. 1–49.

[37] Radu-Daniel Vatavu. (2012). User-defined gestures for free-hand TV control. In *Proceedings of the 10th European conference on Interactive tv and video (EuroiTV'12)*. Berlin, Germany. New York: ACM Press, pp. 45-48.

[38] Radu-Daniel Vatavu and Jacob O. Wobbrock. (2015). Formalizing Agreement Analysis for Elicitation Studies. *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI'15)*. Seoul, Republic of Korea. New York: ACM Press, pp. 1325–1334.

[39] Radu-Daniel Vatavu and Jacob O. Wobbrock. (2016). Between-Subjects Elicitation Studies. *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI'16)*. San Jose, California, New York: ACM Press, pp. 3390–3402.

[40] Jacob O. Wobbrock, Htet Htet Aung, Brandon Rothrock, and Brad A, Myers. (2005). Maximizing the guessability of symbolic input. *Proceedings of the ACM* confrence *on Human factors in computing systems (CHI'05)*. Portland, OR. New York: ACM Press, pp. 1869-1872.

[41] Jacob O. Wobbrock, Meredith R. Morris, and Andrew D. Wilson. (2009). User-defined gestures for surface computing. *Proceedings of the ACM conference on Human factors in computing systems (CHI'09)*. Boston, MA. New York: ACM Press, pp. 1083 -1092.

[42] Kathryn Zyskowski, Meredith R. 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. *Proceedings of the ACM Conference on Computer Supported Cooperative Work & Social Computing (CSCW'15)*. Vancouver, Canada, New York: ACM Press, pp. 1682–1693.