Crowdsourcing vs Laboratory-Style Social Acceptability Studies? Examining the Social Acceptability of Spatial User Interactions for Head-Worn Displays

Fouad Alallah[1], Ali Neshati[1], Nima Sheibani[1], Yumiko Sakamoto[1], Andrea Bunt[1], Pourang Irani[1], Khalad Hasan[2]

[1]University of Manitoba
Winnipeg, Canada

[2]University of Waterloo
Waterloo, Canada

{umalallf, neshatia, sheibann, umsakamo, bunt, pourang.irani}@cs.umanitoba.ca,
mkhasan@uwaterloo.ca

ABSTRACT

The use of crowdsourcing platforms for data collection in HCI research is attractive in their ability to provide rapid access to large and diverse participant samples. As a result, several researchers have conducted studies investigating the similarities and differences between data collected through crowdsourcing and more traditional, laboratory-style data collection. We add to this body of research by examining the feasibility of conducting social acceptability studies via crowdsourcing. Social acceptability can be a key determinant for the early adoption of emerging technologies, and as such, we focus our investigation on social acceptability for Head-Worn Display (HWD) input modalities. Our results indicate that data collected via a crowdsourced experiment and a laboratory-style setting did not differ at a statistically significant level. These results provide initial support for crowdsourcing platforms as viable options for conducting social acceptability research.

Author Keywords

Crowdsourcing, Social acceptance, Head-Worn Displays, Input modalities.

ACM Classification Keywords

H5.2. Information interfaces and presentation (e.g. HCI): User Interfaces: Evaluation/Methodology/Input modalities.

INTRODUCTION

Head-Worn Displays (HWDs) offer the opportunity for seamless information access, on-the-go and in settings not possible with traditional computing. To improve the interaction efficiency with current HWDs, researchers have proposed a number of different input modalities [6,10,23,26], which range from highly noticeable and spatial gestures to more subtle ones [3,20]. Extravagant input methods performed in public, however, could naturally trigger unwanted attention from spectators [36] and consequently, make users feel uncomfortable. Social acceptability studies aim to assess these types of non-performance related issues pertaining to novel interaction mechanisms [1,11,27,28]. These studies typically ask the users to perform a set of gestures in one or more public spaces (e.g., in a shopping mall) and then rate their perceived social comfort levels in performing these gestures in a range of potential social contexts (e.g., in front of strangers). A challenge with these studies, however, is that they can be resource intensive and slow to conduct – hardware prototypes may be limited and researchers must accompany participants to public spaces. Consequently, these studies often have relatively small sample sizes [1,28], which can impact the generalizability of the results.

Crowdsourcing platforms have become increasingly attractive as a means of conducting empirical HCI work, due to their rapid and cost-efficient access to large and diverse groups of users. While data quality can be an issue [4,19], a number of studies have shown that with certain precautious outlier removal procedures, crowdsourcing evaluations can lead to the same set of conclusions as laboratory evaluations [19]. When evaluating the social acceptability of novel technologies like HWDs, however, a key distinction is that participants recruited via the crowdsourcing platform will not be able to experience the gestures for themselves, and will have to provide ratings based on watching video clips without the presence of a researcher. Consequently, the extent to which crowdsourcing platforms are viable for this type of research is an open question.

The goal of this paper is to investigate the potential suitability of conducting crowdsourced social acceptability studies of HWD input modalities. To do so, we collected two social acceptability data sets, one using Amazon Mechanical Turk (AMT), where participants watched videos of the gestures being performed, and a traditional laboratory-style dataset, where participants met with a researcher in person, and used the device to perform the gestures in a public space.

CHI 2018, April 21–26, 2018, Montreal, QC, Canada
© 2018 Association for Computing Machinery.
ACM ISBN 978-1-4503-5620-6/18/04...$15.00
https://doi.org/10.1145/3173574.3173884

Our results suggest no statistical differences between the two datasets, providing promising initial support for the feasibility of using crowdsourcing for social acceptability studies of novel input modalities. Our results further provide insight for the social acceptability of the five different forms of input modalities we included in our study (Figure 1).

RELATED WORK

We focus our coverage of related work on social acceptability studies for mobile input technologies and work exploring data collection using crowdsourcing platforms.

Social Acceptability of Novel Mobile Input

The social acceptability of technology is a key determinant in the success or failure of any new technology. Researchers have conducted a number of social acceptance studies examining novel input to mobile devices as well as factors affecting users' social comfort level. Ronkainen et al. [29] first introduced the idea of examining social acceptance of device-based gestures (e.g., swinging a mobile device). Motivated by this work, researchers have studied social acceptance of device-centric gestures and body-centric gestures (e.g., tap on the nose) [1,17,22,27], in-air around-device gestures, and above-device gestures [11]. Their exploration primarily focused on different gesture properties (e.g., gesture size and gesture duration) that influence users' and spectators' willingness to use gestures in different social contexts (e.g., location and audience). Results revealed that the gesture properties affect users' attitude towards using these novel mobile inputs in various social settings.

The above studies have used various methods to collect social acceptability data. One approach has been to provide participants with video clips showing people performing gestures. Participants are then asked to imagine themselves performing these gestures at different locations and in front of different audiences [28,29]. Alternatively, studies also have been conducted in public places (e.g., a shopping mall) to elicit participants' responses to novel mobile inputs in real-world usage situations [1,35,37]. Our study methods are informed by these prior studies.

Crowdsourcing Evaluations

Researchers have shown great interest in crowdsourcing platforms, such as AMT and CrowdFlower, for collecting data (e.g. [7,13,25]) and conducting user studies (e.g., [18,21]). Prior work has also shown, however, that the quality of the collected data can be an issue. The unsupervised setting in such platforms can lead to a range of unwanted participant behaviors such as cheating (i.e. searching for answers or copying them from online sources) [4], not paying attention [12,24], or giving up while performing tasks [30].

Other work has shown that with proper check-and-balances in place, data collected via crowdsourcing can be comparable to data collected in a traditional laboratory setting with supervision. For example, a performance evaluation of two different selection techniques and three different adaptive interfaces found no significant differences between the crowdsourced data and the data collected in the lab [21]. Another study found that for straightforward tasks (i.e. giving examples of sportswear), the crowdsourced participant data were comparable to that from in-lab participants [9]. However, when the complexity of the task increased (i.e. answering questions in detail), differences between crowdsourced and in-lab data emerged [9].

Informed by prior work on crowdsourced data collection, our study examines whether perceived social acceptability of input modalities differ when we compare collected data from crowdsourcing and a laboratory-style setting.

CROWDSOURCED VS. LABORATORY-STYLE SOCIAL ACCEPTABILITY

We aimed to understand the feasibility of using crowdsourcing to conduct social acceptance research of HWD input modalities. Crowdsourcing is considered to be a rapid, cost-effective means of data collection. However, HWDs represent a unique challenge for crowdsourcing research in that many users will not have any experience using these types of devices, and will not get the chance to do so during the study due to its remote nature. We were therefore interested in whether their data would be comparable to data collected via a more traditional laboratory-style evaluation, where participants get to actually experience the gestures with the target hardware.

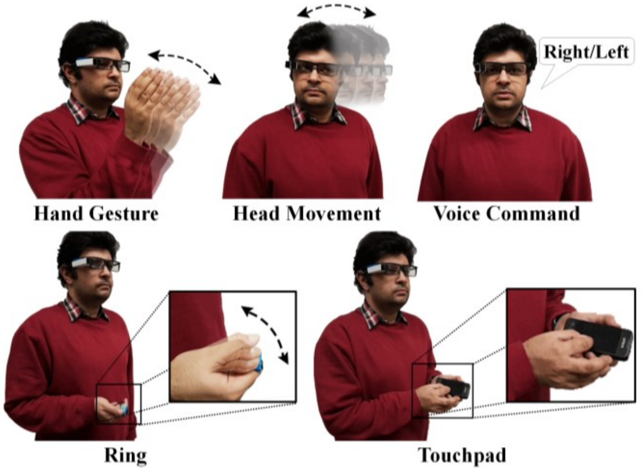


Figure 1. Five input modalities used in the study.

HWDs Input

To inform our choice of specific input modalities, we reviewed prior work on input for HWDs [6,10,23,26,28,31]. We found that touchpads [23], hand gestures [6,26], head movement [16], voice commands [28], and rings [10] are among the most commonly used input modalities for HWDS, and thus we included these five in our study (Figure 1).

Study Design

Our study design and measures are based on multiple prior works on social acceptance [1,27,28,32] and crowdsourcing [7,13,18,21]. As mentioned in the related work section, prior studies on social acceptance have explored novel, yet unavailable input methods (e.g., around-device interaction [14,15], smartphone gestures [31], hand-to-face input [32]) using two prominent approaches: (i) giving participants videos of the input/interaction technology [27,29]; (ii) having participants experience the technology first-hand and asking them to imagine its use in different contexts [1,32]. Our study employs both of these methods. We provide crowdsourced participants with video clips. Laboratory-style participants, on the other hand, are given first-hand technology experience in a single public space. In line with prior social acceptabilities studies [1,27], both our laboratory and crowdsourced data collection involved imaginative contextualization, where participants are asked to imagine themselves in various social contexts.

In our current study, we focused on the following audiences (*alone*, *family members*, *friends*, *colleagues*, and *strangers*) and locations (*at home, public transportation, shopping mall, sidewalk,* and *work*) to understand participants' reactions of using five different input modalities (*hand gesture*, *head movement*, *touchpad*, *ring*, *and voice*).

All participants completed an online questionnaire to collect their social acceptability ratings. We implemented the online questionnaire using PHP and MySQL to collect participants' feedback in all conditions. The questionnaire contains three sections: the first section introduces the study; the second section covers questions about basic participant demographic information. The third section includes video clips showing a co-author using the HWD inputs and questions on using these modalities in different social settings. Participants were asked to rate (from 1 being very socially uncomfortable to 5 being very socially comfortable, Figure 2) the usage of input modalities in front of different audiences and in various locations as performers of the tasks.

In the next sections, we describe the participants in both the laboratory-style study and the crowdsourced study as well as the differences between the two study procedures.

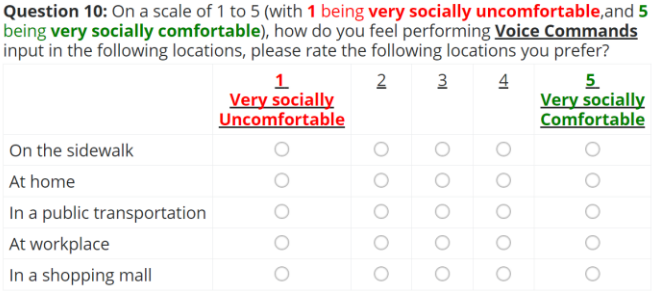
Laboratory-style study

In this study, we recruited 28 participants who received $15 as compensation for their participation in the experiment. The experiment took place in a public space in a crowded university atrium. To help familiarize participants with the gestures, they watched all video clips, each demonstrating one input modality. We then asked them to wear the HWD and to perform tasks that were shown on the clips with each of the five HWD inputs. This step gave them the experience of using the input techniques in a real-world context. A researcher stayed nearby and monitored participants while they were performing the input tasks. Once they completed tasks with an input technique, participants were provided with the online questionnaire to rate the usage of input technique, hypothetically, in front of different audiences and at various locations. Each session lasted around 30 minutes.

Crowdsourced study

We collected data from 106 crowdsourced participants. We posted the link to an online survey in Amazon's Mechanical Turk as a task. We specified the compensation as $1.00 for completing the task, 50% approval rate, and a minimum of 50 completed Human Intelligence Tasks (HITs). We also specified a minimum 50% "Approval Rate", meaning that at least 50% of each participant's prior HITs where considered acceptable by the requestors. Although prior studies have had stricter approval rates [e.g., 21], we chose 50% as our threshold because we had two other data removal steps (i.e., Gold standard questions & 15 secs threshold), which we discuss below.

Our sample sizes (28 laboratory and 106 crowdsourced) are comparable to prior research exploring crowdsourced vs. lab evaluations. For example, Komarov et al. [21] had 96 participants recruited via AMT and 14 local participants. Likewise, Edgar et al. [9] had 1019 participants for a crowdsource study and 44 participants for an in-lab study.



Question 10: On a scale of 1 to 5 (with **1** being **very socially uncomfortable**,and **5** being **very socially comfortable**), how do you feel performing **Voice Commands** input in the following locations, please rate the following locations you prefer?

	1 Very socially Uncomfortable	2	3	4	5 Very socially Comfortable
On the sidewalk	○	○	○	○	○
At home	○	○	○	○	○
In a public transportation	○	○	○	○	○
At workplace	○	○	○	○	○
In a shopping mall	○	○	○	○	○

Figure 2. An example of the online questionnaire.

Crowdsourcing platform participants were presented with the similar online questionnaire used in the laboratory-style study. In contrast, however, they were asked to imagine themselves performing these tasks with the input technique shown in each video without hands-on use. At the beginning of the survey, participants were presented with an introduction to the study and were shown video clips for the HWDs' input modalities.

Since crowdsourcing platforms have a disadvantage regarding the lack of direct supervision of the participants, concerns about the quality of collected data are common. Following guidelines from prior work (e.g., [12,33]), we decided on the criterion for identifying and excluding outliers and/or responses that were not entered in good faith. First, the questionnaire stored participants' IP addresses to be aware of duplicated data [4,34]. Also, the time participants spent on each question was recorded to ensure that they had taken adequate time to read and understand each question before answering [4]. Given prior work indicating that 15 seconds is a lower bound for answering a closed question [2], we used a 15-second threshold to remove data that were entered too quickly, potentially reflecting unreliable responses. We also tested this minimum length in our pilot study. Not removing responses under a certain time limit is considered problematic [2]. Another validation method we used was to add a gold standard, verifiable question [8] to ensure that participants were attending to the video clips. In our case, participants were asked a question about one of the video clip's subtitle captions. Data with incorrect answers were rejected. Finally, we used data only when the participants completed the entire questionnaire. To ensure data quality, we applied these selection criteria both on the laboratory-style and crowdsourcing data.

RESULTS

Disqualified Data

Out of 134 participants (28 in laboratory-style or *Lab,* and 106 AMT crowdsourcing or *AMT*), 44 participants failed to respond correctly to the golden standard question (1 in Lab; 43 in AMT). Next, 44 participants failed to spend adequate time on the questionnaire items (3 in Lab; 41 in AMT). Finally, 11 participants failed to complete the entire questionnaire (only in crowdsourcing). This left us with 24 Lab participants (24/28 = 86% qualified) and 52 AMT participants (52/106 = 49.1% qualified). This quantity of removed crowdsourcing data is consistent with prior work (e.g., ([12] & [33] removed 40% and 70%, respectively).

Data Analysis

Our analysis accounts for our unequal sample sizes and the non-normality distribution of data across conditions by using nonparametric tests. We used Mann-Whitney U Tests with the aggregate of social acceptability across all the five location and audience types, with Bonferroni's adjusted *ps* = .013 (i.e., .05/4). We further report effect sizes as they are independent of sample size issues.

General Demographic Differences

We investigated the demographic differences between in laboratory style and crowdsourcing conditions. Overall, there are more males (n = 50) than females (n = 26) in both conditions (Lab: M = 14, F = 10; AMT: M = 36, F = 16). We also found that the age range was smaller for the laboratory-style condition (18 and 34 yrs) than crowdsourcing counterparts (18 and 74 yrs).

HWDs Input	User Rating -Locations						User Rating -Audiences					
	Lab	ATM	U	z	p	r	Lab	ATM	U	z	p	r
Voice	2.6	2.6	571	-0.6	0.56	0.07	3.6	3.4	563	-0.7	0.49	0.08
Hand Gesture	2.6	2.8	584	-0.4	0.66	0.05	3.4	3.5	600	-0.3	0.79	0.03
Head Movement	2.8	3.6	404	-2.5	0.01	0.28	3.6	3.8	471	-1.7	0.09	0.2
Touchpad	4.9	4.3	483	-1.6	0.11	0.18	4.7	4.6	547	-0.9	0.38	0.1
Ring	4.8	4.6	515	-1.3	0.21	0.14	4.8	4.8	517	-1.2	0.22	0.14

Table 1. Laboratory-style vs. Crowdsourcing: Median Social Acceptability and Confidence Intervals for each modality across all the location and audience types.

Laboratory-style vs. Crowdsourcing across Locations

We first investigated social acceptability levels across all the locations for each input modality, depending on the data-type (i.e., Lab vs. AMT). No significant results were found (Figure 3). Thus, we failed to find any statistically significant differences between laboratory-style and crowdsourcing data when the participants imagined using modalities in various locations. However, these results need to be interpreted with caution as small effects were found (Table 1).

Laboratory-style vs. Crowdsourcing across Audiences

We further investigated social acceptability across audience types for each modality, depending on the data-type. Again, no significant results emerged (Figure 3).

A further exploratory analysis was conducted to investigate overall patterns of the most socially acceptable modality for both locations and audiences. Touchpad and Ring were equally socially acceptable, followed by Head Movement. Voice and Hand Gesture were considered the least socially acceptable modalities: This pattern was consistent across both location types and audience types (Figure 3).

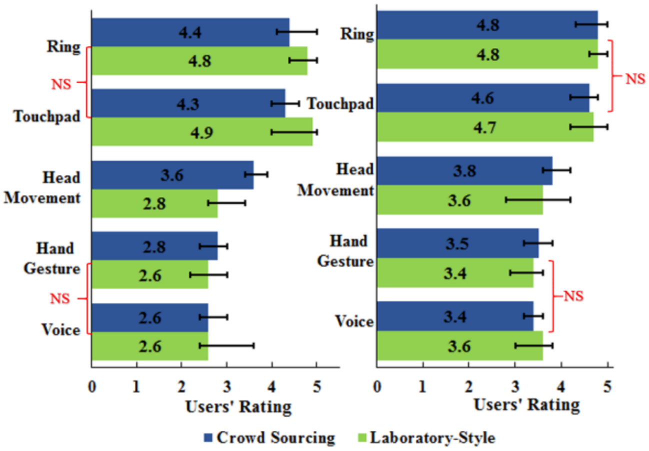


Figure 3. Laboratory-style vs. crowdsourcing: Median social acceptability and Confidence Intervals for each modality across all the location types (Left) and audience types (Right).

Our results indicate that data we collected in the laboratory-style condition and crowdsourcing condition did not differ statistically (*ps* > .013), and the effects were small or marginal (.03 ≤ r ≤ .28). Thus, our results suggest that it is possible to conduct social acceptability research via crowdsourcing. Note, however, that all of our analyses were conducted once all the disqualified data points were removed. Thus, we suggest greater caution is needed when anyone employs crowdsourcing data collection methods.

DISCUSSION, LIMIATIONS, AND FUTURE WORK

Our study compared social acceptability data collected among users in a laboratory-style and that of a crowdsourcing platform. We also examined whether these setups yield different social acceptance results across different hypothetical locations and audiences. Our findings suggest that the results from both groups are statistically similar and potentially comparable. While we saw larger sources of variability with the crowdsourced participants, and we had to remove a number of outliers, the use of crowdsourcing did not impact the overall nature of our findings. These results are consistent with prior work on the feasibility of using crowdsourcing for performance evaluations [21].

By looking at input modalities across audiences and locations, all participants (laboratory-style and crowdsourced) had higher social acceptance ratings for the *ring* and *touchpad* compared to other input modalities. *Touchpad* and *ring* are subtle and less attention seeking, leading to reduced social discomfort. Similar findings were revealed by Ahlström [1] where they suggested that smaller and quicker gestures are less noticeable, and thus more socially acceptable.

The fact that crowdsourcing participants tended to contribute similar data viewing only video clips of the interactions raises a number of interesting avenues for potential research.

Our laboratory-style condition was designed as an initial step toward understanding the social acceptability of input techniques for a HWD in a restricted context. We acknowledge that an in-the-wild study design would provide insight on social acceptability in more realistic usage contexts. Additionally, the evaluation of social acceptability might change based on participants' surrounding social context. Further study conducted in such contexts (e.g., in public/private settings - shopping mall or sidewalk, and in front of different audiences - families, or coworkers) would be needed to validate the generalizability of our results. Furthermore, participants' age as well as their cultural or ethnic background could naturally influence their social acceptability ratings. We unfortunately did not collect participants' ethnicity data in our study. However, in light of these initial feasibility results, further research should consider crowdsourced studies to enrich our understanding of how demographic data such as age range and cultural background might impact social acceptability reactions to novel devices and interactions.

Our study investigated social acceptability of interaction techniques for HWDs in a comparison of laboratory style data and crowdsourcing data. Future research should therefore explore the extent to which remote evaluations are possible with other novel devices and gesture sets. For example, there may be certain interactions that are more difficult to convey via video clips than others, and participants' familiarity with the technology concept might also play a role; while not all the participants in our sample would have had the opportunity to interact with HWDs, some may have seen them depicted in the media. The suitability of crowdsourcing platforms to explore perceptions of more "futuristic" technologies is an interesting area of future research.

SUMMARY

We reported on the results of a study comparing social acceptability ratings of HWDs interaction techniques among users in laboratory-style study and a remote crowdsourcing platform. Our results showed that, overall, data collected from both groups are similar and suggest that there is potential feasibility for running social acceptability studies of new technologies using crowdsourcing platforms if the data is treated carefully.

ACKNOWLEDGEMENT

This research was partially funded by the Natural Sciences and Engineering Research Council of Canada (NSERC).

REFERENCES

1. David Ahlström, Khalad Hasan, and Pourang Irani. 2014. Are You Comfortable Doing That?: Acceptance Studies of Around-Device Gestures in and for Public Settings. *Proceedings of the 16th international conference on Human-computer interaction with mobile devices & services - MobileHCI '14*: 193–202. https://doi.org/10.1145/2628363.2628381
2. Merrill Anderson. 2004. *Bottom-line organization development*. Routledge.
3. Florian Bemmann. 2015. User Preference for Smart Glass Interaction. 3–5.
4. Sabine Buchholz and Javier Latorre. 2011. Crowdsourcing preference tests, and how to detect cheating. In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 3053–3056.
5. Brent C. 2011. How Much Time are Respondents Willing to Spend on Your Survey? Retrieved from https://www.surveymonkey.com/blog/2011/02/14/survey_completion_times/
6. Andrea Colaço, Ahmed Kirmani, Hye Soo Yang, Nan-Wei Gong, Chris Schmandt, and Vivek K Goyal. 2013. Mime: Compact, Low Power 3D Gesture Sensing for Interaction with Head Mounted Displays. In *Proceedings of the 26th Annual ACM Symposium on User Interface Software and Technology* (UIST '13), 227–236. https://doi.org/10.1145/2501988.2502042
7. Kristen Dergousoff and Regan L Mandryk. 2015. Mobile Gamification for Crowdsourcing Data Collection: Leveraging the Freemium Model. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (CHI '15), 1065–1074. https://doi.org/10.1145/2702123.2702296
8. Mira Dontcheva, Robert R. Morris, Joel R. Brandt, and Elizabeth M. Gerber. 2014. Combining crowdsourcing and learning to improve engagement and performance. *Proceedings of the 32nd annual ACM conference on Human factors in computing systems - CHI '14*: 3379–3388. https://doi.org/10.1145/2556288.2557217
9. Jennifer Edgar, Joe Murphy, and Michael Keating. 2016. Comparing Traditional and Crowdsourcing Methods for Pretesting Survey Questions. *SAGE Open* 6, 4: 2158244016671770. https://doi.org/10.1177/2158244016671770
10. Barrett Ens, Ahmad Byagowi, Teng Han, Juan David

Hincapié-Ramos, and Pourang Irani. 2016. Combining Ring Input with Hand Tracking for Precise, Natural Interaction with Spatial Analytic Interfaces. In *Proceedings of the 2016 Symposium on Spatial User Interaction - SUI '16* (SUI '16), 99–102. https://doi.org/10.1145/2983310.2985757

11. Euan Freeman, Stephen Brewster, and Vuokko Lantz. 2014. Towards usable and acceptable above-device interactions. In *MobileHCI 2014 - Proceedings of the 16th ACM International Conference on Human-Computer Interaction with Mobile Devices and Services* (MobileHCI '14), 459–464. https://doi.org/10.1145/2628363.2634215

12. Ujwal Gadiraju, Ricardo Kawase, Stefan Dietze, and Gianluca Demartini. 2015. Understanding Malicious Behavior in Crowdsourcing Platforms: The Case of Online Surveys. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (CHI '15), 1631–1640. https://doi.org/10.1145/2702123.2702443

13. Hervé Goëau, Alexis Joly, Souheil Selmi, Pierre Bonnet, Elise Mouysset, Laurent Joyeux, Jean-François Molino, Philippe Birnbaum, Daniel Bathelemy, and Nozha Boujemaa. 2011. Visual-based Plant Species Identification from Crowdsourced Data. In *Proceedings of the 19th ACM International Conference on Multimedia* (MM '11), 813–814. https://doi.org/10.1145/2072298.2072472

14. Khalad Hasan, David Ahlström, Junhyeok Kim, and Pourang Irani. 2017. AirPanes: Two-Handed Around-Device Interaction for Pane Switching on Smartphones. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (CHI '17)*. ACM, New York, NY, USA, 679-691. https://doi.org/10.1145/3025453.3026029

15. Khalad Hasan, David Ahlström, and Pourang Irani. 2013. Ad-binning: leveraging around device space for storing, browsing and retrieving mobile device content. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '13)*. ACM, New York, NY, USA, 899-908. https://doi.org/10.1145/2470654.2466115

16. Caroline Jay and Roger Hubbold. 2003. Amplifying head movements with head-mounted displays. *Presence: Teleoperators and Virtual Environments* 12, 3: 268–276.

17. Brett Jones, Rajinder Sodhi, David Forsyth, Brian Bailey, and Giuliano Maciocci. 2012. Around Device Interaction for Multiscale Navigation. In *Proceedings of the 14th International Conference on Human-computer Interaction with Mobile Devices and Services* (MobileHCI '12), 83–92. https://doi.org/10.1145/2371574.2371589

18. Aniket Kittur, Ed H Chi, and Bongwon Suh. 2008. Crowdsourcing user studies with Mechanical Turk. *Proceedings of the 2008 SIGCHI Conference on Human Factors in Computing Systems*: 453–456. https://doi.org/10.1145/1357054.1357127

19. Aniket Kittur and Robert E Kraut. 2008. Harnessing the Wisdom of Crowds in Wikipedia: Quality Through Coordination. In *Proceedings of the 2008 ACM Conference on Computer Supported Cooperative Work* (CSCW '08), 37–46. https://doi.org/10.1145/1460563.1460572

20. Barry Kollee, Sven Kratz, and Anthony Dunnigan. 2014. Exploring Gestural Interaction in Smart Spaces Using Head Mounted Devices with Ego-centric Sensing. In *Proceedings of the 2Nd ACM Symposium on Spatial User Interaction* (SUI '14), 40–49. https://doi.org/10.1145/2659766.2659781

21. 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), 207–216. https://doi.org/10.1145/2470654.2470684

22. Sven Kratz, Michael Rohs, Dennis Guse, Jörg Müller, Gilles Bailly, and Michael Nischt. 2012. PalmSpace: Continuous Around-device Gestures vs. Multitouch for 3D Rotation Tasks on Mobile Devices. In *Proceedings of the International Working Conference on Advanced Visual Interfaces* (AVI '12), 181–188. https://doi.org/10.1145/2254556.2254590

23. Meethu Malu and Leah Findlater. 2015. Personalized, Wearable Control of a Head-mounted Display for Users with Upper Body Motor Impairments. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (CHI '15), 221–230. https://doi.org/10.1145/2702123.2702188

24. Winter Mason and Siddharth Suri. 2012. Conducting behavioral research on Amazon's Mechanical Turk. *Behavior Research Methods* 44, 1: 1–23. https://doi.org/10.3758/s13428-011-0124-6

25. Daniel McDuff, Rana el Kaliouby, and Rosalind Picard. 2011. Crowdsourced Data Collection of Facial Responses. In *Proceedings of the 13th International Conference on Multimodal Interfaces* (ICMI '11), 11–18. https://doi.org/10.1145/2070481.2070486

26. Florian Müller, Niloofar Dezfuli, Max Mühlhäuser, Martin Schmitz, and Mohammadreza Khalilbeigi. 2015. Palm-based Interaction with Head-mounted Displays. In *Proceedings of the 17th International Conference on Human-Computer Interaction with Mobile Devices and Services Adjunct* (MobileHCI '15), 963–965. https://doi.org/10.1145/2786567.2794314

27. Julie Rico and Stephen Brewster. 2010. Usable Gestures for Mobile Interfaces: Evaluating Social Acceptability. In *Proceedings of the SIGCHI*

Conference on Human Factors in Computing Systems (CHI ’10), 887–896. https://doi.org/10.1145/1753326.1753458

28. Julie Rico and Stephen Brewster. 2010. Gesture and Voice Prototyping for Early Evaluations of Social Acceptability in Multimodal Interfaces. In *International Conference on Multimodal Interfaces and the Workshop on Machine Learning for Multimodal Interaction* (ICMI-MLMI ’10), 16:1--16:9. https://doi.org/10.1145/1891903.1891925
29. Sami Ronkainen, Jonna Häkkilä, Saana Kaleva, Ashley Colley, and Jukka Linjama. 2007. Tap Input As an Embedded Interaction Method for Mobile Devices. In *Proceedings of the 1st International Conference on Tangible and Embedded Interaction* (TEI ’07), 263–270. https://doi.org/10.1145/1226969.1227023
30. Spencer Rothwell, Steele Carter, Ahmad Elshenawy, and Daniela Braga. 2015. Job Complexity and User Attention in Crowdsourcing Microtasks. *HCOMP 2015 Workshop Papers for “Crowdsourcing Breakthroughs for Language Technology Applications,”* 2008: 20–25.
31. Jaime Ruiz and Yang Li. 2011. DoubleFlip: a motion gesture delimiter for mobile interaction. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '11)*. ACM, New York, NY, USA, 2717-2720. https://doi.org/10.1145/1978942.1979341
32. Marcos Serrano, Barrett M Ens, and Pourang P Irani. 2014. Exploring the Use of Hand-to-face Input for Interacting with Head-worn Displays. In *Proceedings of the 32Nd Annual ACM Conference on Human Factors in Computing Systems* (CHI ’14), 3181–3190. https://doi.org/10.1145/2556288.2556984
33. Mark D Smucker and Chandra Prakash Jethani. 2011. The crowd vs. the lab: A comparison of crowd-sourced and university laboratory participant behavior. In *Proceedings of the SIGIR 2011 Workshop on crowdsourcing for information retrieval*.
34. Shichang Wang, Chu-Ren Huang, Yao Yao, and Angel Chan. 2015. Mechanical Turk-based Experiment vs Laboratory-based Experiment: A Case Study on the Comparison of Semantic Transparency Rating Data. In *PACLIC*.
35. Julie R Wiliamson, Andrew Crossan, and Stephen Brewster. 2011. Multimoda Mobile Interactions: Usability Studies in Real World Settings. In *Proceedings of the 13th International Conference on Multimodal Interfaces* (ICMI ’11), 361–368. https://doi.org/10.1145/2070481.2070551
36. Julie R Williamson. 2012. User experience, performance, and social acceptability: usable multimodal mobile interaction. University of Glasgow.
37. Julie R Williamson, Stephen Brewster, and Rama Vennelakanti. 2013. Mo!Games: Evaluating Mobile Gestures in the Wild. In *Proceedings of the 15th ACM on International Conference on Multimodal Interaction* (ICMI ’13), 173–180. https://doi.org/10.1145/2522848.2522874