Crowdsourcing Culture in HRI: Lessons Learned From Quantitative and Qualitative Data Collections

Michiel Joosse
Human Media Interaction
University of Twente
Enschede, the Netherlands
m.p.joosse@utwente.nl

Manja Lohse
Human Media Interaction
University of Twente
Enschede, the Netherlands
m.lohse@utwente.nl

Vanessa Evers
Human Media Interaction
University of Twente
Enschede, the Netherlands
v.evers@utwente.nl

ABSTRACT
In recent years crowdsourcing started to become a methodology for gathering data in academic research. In this paper we present two studies in which we collected data by harnessing the "wisdom of the crowd" in order to elicit people's preferences for robot behaviors. Our main goal was to investigate cultural differences between national cultures, which led to specific choices in our methodology. We collected both quantitative (N=181) and qualitative (N=118) data. This paper presents how we ensured data quality, and which advantages and challenges exist when using crowdsourcing when researching cultural robotics.

Categories and Subject Descriptors
J.4 [**Computer Applications**]: Social and Behavioral Sciences

Keywords
Human-Robot Interaction, cultural differences, crowdsourcing, methodology

1. INTRODUCTION
In order to develop and deploy robots in different areas around the world, insight is needed in people's preferences for specific robot behaviors. Previous research in Human-Robot Interaction (HRI) indicates that cultural differences exist regarding people's attitudes towards robots [3], mental models of robots [12] and preferences for communication styles [17]. This implies that insight is needed into potential cultural differences regarding people's preferences for robot behavior. As we are aware that different cultures exist at different levels (such as nations and organizations), we will limit ourselves in this paper to "national cultures" when using the word "culture".

As part of the EU FP7-project SPENCER we aim to develop a demonstrator robot to guide people at airports, an environment with people having diverse cultural backgrounds.

Figure 1: Example illustration of a robot and a group as used in study 2

Our specific task involves conducting user studies to investigate which motion behaviors are deemed appropriate by passengers from different countries. We are especially interested in those countries where more and more people get the opportunity to fly to other continents, thus being inexperienced with flying, which provides an use case for the SPENCER project. One such country is China.

User studies in general, but especially in HRI, are resource-intense, and when investigating cultural differences with real robots, one either has to bring the robot to the participants (such as [17]) or bring the participants to the robot, for example by recruiting participants with different cultural backgrounds. A disadvantage of the latter method could be that ecological validity is compromised as at least a part of the sample is not observed in their natural culture [2]. This is a disadvantage as a cultural identity not only consists of one's social norms and ideas (subjective culture) but also the way one lives, including a culture's architecture, dress and food (material culture) [15]. A solution to this problem is to "bring a robot" to participants without using an actually physical embodied robot. An example of such a solution is the video-based HRI method employed by Walters et al. [16] and Woods et al. [18] where participants watched videos of a robot interacting with an actor and subsequently rated extrovert and introvert robot behaviors. We believe that these kinds of methods are specifically useful for research on cultural differences as it allows researchers to provide the stimuli in one's native culture.

In this paper we will address the potential of "crowdsourcing" as a methodology for cultural robotics, as this method allows us to recruit participants within their natural culture. We will first provide an overview of related work in HRI. Based upon two studies we conducted we will discuss the opportunities and challenges we encountered when using crowdsourcing in HRI. We will conclude this paper with a set of recommendations for researchers.

2. CROWDSOURCING IN HRI

Webster [1] defines crowdsourcing as *"the practice of obtaining needed services, ideas, or content by soliciting contributions from a large group of people and especially from the online community rather than from traditional employees or suppliers"*. For researchers, crowdsourcing presents an opportunity to run online studies with people from all over the world. This is an advantage as there is a tendency to recruit participants who can be described as belonging to Western, Educated, Industrialized, Rich, and Democratic (WEIRD) countries [7]. Crowdsourcing could be a way of providing researchers with more diverse samples.

In HRI crowdsourcing has been used as a means of gathering data. One example by Read & Belpaeme includes having people rate videos with robot behaviors [13]: participants were shown videos of a robot being subject to various actions, and responding with a positive or negative non-linguistic utterance, or not at all. For each video participants provided a valence rating based upon how they thought the robot felt. Another example of crowdsourcing in HRI is using data recorded online to train behavioral models, also called demonstration from learning. Breazeal et al. [4] developed a game called *Mars Escape* in which two players took the roles of a robot and an astronaut, and had to solve a collaborative task together. These behavior models were later implemented on a robot in a real-life version of the game.

Above work differs from our own crowdsourcing studies in which we specifically recruited and compared participants from various national cultures [11]. In particular, we were interested to determine whether people from different cultures have different proxemics expectations of robots. In the next section we will introduce the methods we used (Section 3). Subsequently we will discuss the lessons we learned (Section 4).

3. CASE STUDIES

In this section we will describe two crowdsourcing studies we conducted through the CrowdFlower platform[1]. The first study was of a quantitative nature (Section 3.1), the second of a qualitative nature (Section 3.2). We introduce these two different studies as the type of data influences the methods (to be) employed to ensure data quality, both prior- and post data collection. In this paper we will discuss the methodological challenges; the results of both studies are reported elsewhere.

[1]http://www.crowdflower.com

Figure 2: In study 1 we collected quantitative data by asking participants to rate the appropriateness of robot positioning

We used the CrowdFlower platform to host our studies as this platform provides advantages when investigating cultural differences. CrowdFlower allows researchers to specifically target countries and/or language spoken by people. While the CrowdFlower platform can be used to create surveys, it is not designed for that. Therefore, we chose to recruit and pay participants through CrowdFlower, but host the survey itself on the SurveyMonkey platform[2] which allows for a broader variety of question types and logic. At the end of each survey, a code was provided which participants had to input on the CrowdFlower platform, in order to get paid.

To maximize the quality of the data a priori, we limited both surveys in that only CrowdFlower workers with the highest level of accuracy were allowed to participate[3].

3.1 Study 1. Quantitative data

The goal of study 1 was to investigate whether people of different national cultures have different proxemics expectations from a robot that approaches a small family. The results of this study have been reported in Joosse et al. [11]. In this study we recruited participants from the United States, China and Argentina, as these three countries have previously been found to belong to different (societal) cultural clusters [6, 8]. These three different cultures are relevant for the SPENCER project, which, as a EU project, will be deployed at an European airport.

We asked participants to look at images containing a small family group and a robot, and to indicate on a 7-point Likert scale "how appropriate they believed the position of the robot was". The survey was distributed to participants with three different nationalities (between-subjects), and every survey contained 18 different images (such as Figure 2), each image was shown twice. Participants thus provided 36 rat-

[2]http://www.surveymonkey.com

[3]According to CrowdFlower, level 3 contributors "account for 7% of the monthly judgments and maintain the highest level of accuracy across an even larger spectrum of CrowdFlower jobs"

ings. Additionally, two manipulation checks were included, with answer options randomized in the survey:

1. Participants had to indicate whether the robot approached (a) from the same direction, (b) in between the same two persons, or (c) from different directions. The correct answer was (c).
2. Participants had to indicate whether the robot generally stopped (a) at the same distance, or (b) at different distances from the group. The correct answer was (b).

Out of the 281 participants who opened the survey, 244 completed the entire survey. We controlled for the quality of our data in a three-step procedure in which we excluded 63 participants (26%):

1. We first excluded 28 participants who failed to correctly answer the two manipulation check questions.
2. Participants rated each of the 18 robot-scene situations twice. Participants who rated four or more situations with a difference of 3 or more points were also excluded from the survey. This led to the exclusion of 32 additional participants.
3. We looked for abnormal patterns in the data: 3 participants who indicated each of the 36 robot-scene situations were exactly equally (in)appropriate were excluded.

After applying the exclusion criteria, the total sample contained 181 participants. Each of the 244 participants who completed the survey was paid $1. On average (mean), participants took 49 minutes to complete the survey, which implies we should have payed more. We noticed that the standard deviation was very high; over 3 hours. Analysis of the data revealed that four participants took very long to complete the survey (12, 23, 23 and 29 hours), thereby increasing both the average time to completion and standard deviation. If we look at the median, however, we find that people took on average 18 minutes to complete the survey. Hence, $1 was an appropriate payment.

3.2 Study 2. Qualitative data

Our second crowdsourcing study was aimed at collecting qualitative data from participants from the United States and China. We did not include Argentina (or another South American country) as we did not find cultural differences between the United States and Argentina in the previous study [11]. We asked people to write sentences to describe how they thought a robot should react to everyday situations at an airport, ranging from a passenger who stopped to tie his shoelace to passengers needing to go to the bathroom.

Participants were instructed that they were on their way to a connecting flight with a guide robot (Figure 1). We manipulated the instructions we provided to the participants:

1. Participants were part of a small or larger group[4],
2. This group was either in a hurry, or had more than enough time to wander around
3. The situation applied to either the minority or majority of the group.

We therefore had (2x2x2=)8 conditions. In the survey we provided participants with 19 situations which either concerned themselves, or only other members of the group. An example of such a situation is:

> *You and one other group member can't find the robot guide and other group members anymore. Please describe shortly how you think the robot should react to this situation (2-3 sentences).*

One manipulation check was included, which asked how many people were in the group. The size of the group was depicted on pictures and text on every previous page. Participants did not have to provide the exact answer; for the small group conditions reported size between 2 and 4 were deemed correct, in the larger group conditions sizes between 8 and 12. We included this error margin as we not so much interested in knowing if participants remembered the exact text, but rather if they were part of a small or big group.

From the 332 participants who opened the survey, 225 completed it. We excluded 107 participants (47%) who satisfied one or more of the following three exclusion criteria:

1. Participants whose answers did not consist of text (words, sentences), but rather random numbers or strings of characters.
2. Participants who did not provide complete sentences, or the same answer for all 19 situations)
3. Participants who failed to answer the manipulation check question.

A pilot study revealed it would take about 30 minutes to complete the survey. Based upon our experiences with the previous survey, we decided to pay participants $3. As participants had to input text data which we could not validate directly (if the provided answer was a real sentence, or just some random keystrokes), we payed participants $1 directly after completing the survey, and the remaining $2 after we had validated whether or not text was provided. 118 participants (52.4% of the 225 who completed the survey) were included in the data analysis. Mean completion time was 33 minutes (sd=56 minutes), as with study 1 there was an outlier which caused this high standard deviation.

4. CROWDSOURCING CULTURAL ROBOTICS

Based upon the insights gained from the two studies described in the previous sections we will describe the advantages and challenges we foresee for cultural robotics when using crowdsourcing to collect data.

[4]Based upon research of [10] we defined a small group containing 3 people, and a larger group 10.

Table 1: Self-reported area of employment and demographics in study 1 and 2

	Study 1	Study 2
Employment area		
Retired	5.5% (10)	3.4% (4)
Student	14.4% (26)	16.9% (20)
Unemployed	23.2% (42)	11% (13)
Arts or Entertainment	5.5% (10)	0.8% (1)
Broadcasting	1.1% (2)	0% (0)
Education	7.7% (14)	11.0% (13)
Construction	3.3% (6)	5.1% (6)
Finance and Insurance	4.4% (8)	5.1% (6)
Health Care	2.2% (4)	5.9% (7)
Hotel and Food Services	0.6% (1)	2.5% (3)
Information Services	3.3% (6)	8.5% (10)
Processing	0.6% (1)	0.8% (1)
Legal Services	1.7% (3)	0% (0)
Manufacturing	3.9% (7)	5.1% (6)
Public Administration	5% (9)	0.8% (1)
Publishing	0.6% (1)	0.8% (1)
Real Estate	0.6% (1)	1.7% (2)
Research	0% (0)	0.8% (1)
Retail	5.5% (10)	5.1% (6)
Software	5% (9)	6.8% (8)
Telecommunications	1.1% (2)	1.7% (2)
Technical Services	2.2% (4)	3.4% (4)
Utilities	1.7% (3)	1.7% (2)
Other	1.1% (2)	0.8% (1)
Age	M=37.0 (sd=12.5)	M=32.33 (sd=11.98)
Age range	16-73	17-66
Male	51.4% (93)	61% (72)
Female	48.6% (88)	39% (45)
(missing)		0.8% (1)

4.1 Crowdsourcing provides a culturally diverse sample

The first advantage of using crowdsourcing for cultural robotics is that a culturally diverse sample is there. We have successfully employed crowdsourcing to gather data both from the United States and China.

The data we gathered is not only relevant for cultural robotics, as the sample was also more diverse in terms of educational background. In both studies we asked participants what their primarily area of employment was. Only 14% and 16% respectively indicated they were students. The employment area of the other participants has been reported in Table 1. As can be seen in both studies, more so in study 1, the level of participants who indicated they were unemployed was quite high. It could very well be that these people use crowdsourcing as means of employment, and in retrospect it might have been useful to include this category in the list of occupations. In general, the diverse areas of employment, along with the higher age range of respondents leads us to conclude that the sample is quite varied in terms of age and occupation (Table 1).

Lesson 1 Crowdsourcing has the potential to provide a more (culturally) diverse sample as compared with lab studies.

4.2 Crowdsourcing potentially allows for quick data collection

The second advantage of using crowdsourcing is that it is possible to gather data relatively quick. Looking back we noticed in both studies that data collection for the United States sample went quite fast, especially compared with the other countries investigated. We have plotted the time to complete the survey in Figure 3. As can be seen the data for the United States sample was in within 6 hours, in stark contrast to the data from China which took 4 weeks to collect. Therefore, this can only be an advantage when cultural robotics researchers sample from countries which are represented enough on the crowdsourcing platform of choice. We will reflect more on this further down this section.

Lesson 2 Crowdsourcing can be a method to collect data quickly.

We used CrowdFlower to host our experiment, and while there are multiple crowdsourcing platforms[5], in academia Amazon Mechanical Turk is the most frequently used platform. In a survey of crowdsourcing use in Information Systems (IS) research, Zhao & Zhu [19] found that nearly half of the surveyed papers used AMT as platform.

When looking at the demographics of AMT, a report by Ross et al. [14] showed that even though the percentage of U.S.-based respondents decreased, in February 2010 still 85% of the respondents originated either from the U.S. or India. A survey conducted by CrowdFlower, which we used for our research, found that indeed a lot of their participants originated from the United States[6], though these data do not tell us anything about generalizability to the entire CrowdFlower workforce. In our research we also had this impression, i.e. because it was quicker to collect data from the U.S. While we could still collect enough data for our purposes, depending on the countries, or cultures of interest it might be necessary to investigate the usage of local crowdsourcing platforms. An alternative would be to settle for responses from countries sharing common values, thereby clustering based upon societal values. An example of such clustering is the GLOBE project [8].

Lesson 3 Crowdsourcing allows researchers access to different countries, though researchers should be aware of the number of available participants in each country.

4.3 Quality control requires manual labor

One of the major disadvantages, or challenges, with using crowdsourcing for cultural robotics is the effort required to ensure data of high enough quality to be used for academic research. In study 1 we collected quantitative data, and even though the data was only comprised of numbers it took manual labor to assess its quality; we applied the rule-based exclusion criteria, and afterwards we manually went through all remaining data to see which participants filled out the same answer to each question. We excluded a lot of participants with this method, and therefore, we can recommend

[5]Such as Amazon Mechanical Turk (AMT), CrowdFlower and SocialSci

[6]https://www.statwing.com/open/datasets/558c919142209fdc7d88c75aa18b43e926f90c45# workspaces/10722

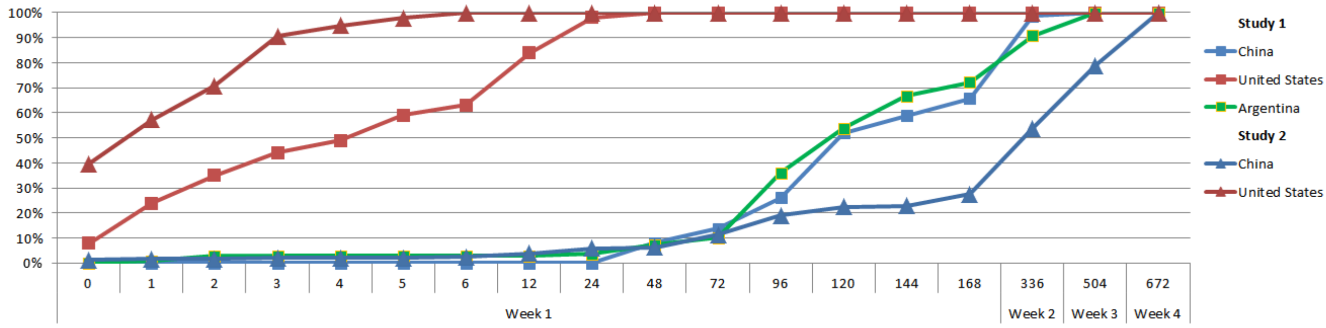

Figure 3: Time to completion in hours, measured from survey launch. Square markers denote study 1; triangular markers study 2

this control method to other researchers collecting qualitative data.

In the second study we decided to only pay a percentage of the reimbursement after completing the survey, as we were aware of the potential pitfalls of using crowdsourcing through our earlier experiences. The quality control procedures here were more time-consuming, in part because we had to import the data, assess the quality and manually searching for the CrowdFlower ID to administer a bonus. This was different from study 1 where we could do the quality control once at the end; in study 2 we repeated this process daily in order to pay participants their bonus in time. Though the quality control took some time in terms of man-hours manually checking the data and administrating the payment, we found it quite beneficial as we got a feel for the data during the collection, and could where necessary increase the number of required participants by reopening the system.

In both studies we employed different methods to assess validity of the data. We found that a simple repeated-measure is effective to check qualitative data, whereas a check of the content is an effective way in case of qualitative data.

Lesson 4 Effective quality control requires manual labor, the appropriate way to check for validity of the data is dependent on the type of data collected.

4.4 Crowdsourcing requires thought about compensation

Especially when using crowdsourcing we believe that it is important to pay attention to compensating participants for their efforts. We assume that payment pays a bigger role in participants' motivation to complete the survey compared with lab studies. In lab studies people might be more motivated, as they are not anonymous. It could also be that people complete the survey as they do not want to disappoint the experimenter. In a way, this might be seen as an example of the Hawthorne effect [5][7]. With crowdsourcing, we believe that there are only two drivers behind completing the survey: intrinsic motivation to contribute to scientific progress and payment.

[7]"a type of reactivity in which individuals modify or improve an aspect of their behavior in response to their awareness of being observed" [5]

For future work we are not sure whether it would make sense to increase or decrease the amount we payed participants. Compared with other crowdfunding studies we believe that the sum we paid was a bit above average, especially in the second study we payed an average hourly wage of \$5.45. This is slightly above the hourly wage of \$4.8 as estimated by Ipeirotis [9]. When designing the study we reasoned that we would also have payed locally-recruited students if they completed the questionnaire, and more importantly: we wanted to have good quality data. We wanted our respondents to take our survey seriously; both contained many questions and required thinking. Therefore, we believed it was only logical to take the respondents serious as well, and start with a serious compensation for their efforts.

Lesson 5 Crowdsourcing requires thought about compensation.

In study 2 we learned that an effective way is to pay only a percentage after completing the survey, and only paying the remainder after assessing validity of the data. Therefore, not only the amount one pay's seems important, also when one does.

Lesson 6 Compensating participants partly after checking validity of the data is effective, especially when collecting qualitative data.

4.5 Robots are not part of everyday life

A final challenge for cultural robotics is the fact that technology is not spread even among cultures. A technology which could be part of everyday life in one part of the continent may not be used at all in another part. Especially with robots, a topic people generally have stereotypical expectations of, a researcher cannot be sure if the participants and researchers are on the same line regarding the notion of what constitutes a robot and what does not.

Lesson 7 As robots are not part of everyday life it is important to find out how much experience a participant has with robots.

5. DISCUSSION & CONCLUSION

In this paper we presented two crowdsourcing studies we conducted and the advantages and disadvantages we encountered while conducting these studies. For cultural robotics crowdsourcing has several advantages. The first advantage is the ease of which participants from a culturally diverse pool can be recruited. The second advantage of crowdsourcing is that it can be relatively quick, though we also noted that it took quite long for the Chinese sample to come in.

Challenges for crowdsourcing (HRI) research in general include both the availability of participants from a specific culture, which was in our case a national culture. The second challenge we encountered was the amount of manual labor involved in quality control, and the associated number of participants a researcher might have to exclude. This strengthens the argument for carefully considering the compensation participants receive for completing a survey (or experiment). What we found especially challenging from cultural robotics research is that social robots are a new technology, and with crowdsourcing it was hard to assess how much experience people had with technology, and whether they could envision the capabilities and limitations of the technology at hand.

Currently we do not know yet how reliable the results of crowdsourcing for cultural robotics are. However, if we look at related work in HRI, such as the work of Walters et al. [16] and Woods et al. [18] we can assume that the results gathered through crowdsourcing are an accurate representation of people's feelings of (in our case) robot behaviors.

Looking back at our studies we are still convinced that crowdsourcing is a very useful research instrument for HRI researchers interested in cross-cultural research. From our experience it was more straightforward to assess the quality of the responses in the second study, as plain text is (for us, people) easier interpretable than a string of numbers, without running comparisons first. We would therefore recommend researchers to incorporate at least some open questions, such as manipulation checks. By doing this, validity can be assessed in a more objective way.

To conclude: while crowdsourcing research involves certain challenges, based upon our research we believe that this could provide opportunities for researchers in cultural robotics to conduct cross-cultural HRI research.

Acknowledgements
This research has been partly supported by the European Commission under contract number FP7-ICT-600877 (SPENCER).

6. REFERENCES
[1] Merriam-webster. http://www.merriam-webster.com/dictionary/crowdsourcing. Accessed: 2015-09-03.

[2] M. Baldassare and S. Feller. Cultural variations in personal space. *Ethos*, 3(4):481–503, 1975.

[3] C. Bartneck, T. Suzuki, T. Kanda, and T. Nomura. The influence of people's culture and prior experiences with aibo on their attitude towards robots. *Ai & Society*, 21(1-2):217–230, 2007.

[4] C. Breazeal, N. DePalma, J. Orkin, S. Chernova, and M. Jung. Crowdsourcing human-robot interaction: New methods and system evaluation in a public environment. *Journal of Human-Robot Interaction*, 2(1):82–111, 2013.

[5] R. H. Franke and J. D. Kaul. The hawthorne experiments: First statistical interpretation. *American sociological review*, pages 623–643, 1978.

[6] V. Gupta, P. J. Hanges, and P. Dorfman. Cultural clusters: Methodology and findings. *Journal of world business*, 37(1):11–15, 2002.

[7] J. Henrich, S. J. Heine, and A. Norenzayan. The weirdest people in the world? *Behavioral and brain sciences*, 33(2-3):61–83, 2010.

[8] R. J. House, P. J. Hanges, M. Javidan, P. W. Dorfman, and V. Gupta. *Culture, leadership, and organizations: The GLOBE study of 62 societies*. Sage publications, 2004.

[9] P. G. Ipeirotis. Analyzing the amazon mechanical turk marketplace. *XRDS: Crossroads, The ACM Magazine for Students*, 17(2):16–21, 2010.

[10] J. James. The distribution of free-forming small group size. *American Sociological Review*, 1953.

[11] M. P. Joosse, R. W. Poppe, M. Lohse, and V. Evers. Cultural differences in how an engagement-seeking robot should approach a group of people. In *Procs of the 5th ACM Int. Conf. on Collaboration across boundaries: culture, distance & technology*, pages 121–130. ACM, 2014.

[12] S.-l. Lee, I. Y.-m. Lau, S. Kiesler, and C.-Y. Chiu. Human mental models of humanoid robots. In *Procs of the 2005 IEEE Int. Conf. on Robotics and Automation (ICRA)*, pages 2767–2772. IEEE, 2005.

[13] R. Read and T. Belpaeme. Situational context directs how people affectively interpret robotic non-linguistic utterances. In *Procs of the 2014 ACM/IEEE Int. Conf. on Human-robot interaction*, pages 41–48. ACM, 2014.

[14] J. Ross, L. Irani, M. Silberman, A. Zaldivar, and B. Tomlinson. Who are the crowdworkers?: shifting demographics in mechanical turk. In *CHI'10 Extended Abstracts*, pages 2863–2872. ACM, 2010.

[15] H. C. Triandis. *The analysis of subjective culture*. Wiley, New York, 1972.

[16] M. L. Walters, M. Lohse, M. Hanheide, B. Wrede, D. S. Syrdal, K. L. Koay, A. Green, H. Hüttenrauch, K. Dautenhahn, G. Sagerer, et al. Evaluating the robot personality and verbal behavior of domestic robots using video-based studies. *Advanced Robotics*, 25(18):2233–2254, 2011.

[17] L. Wang, P.-L. P. Rau, V. Evers, B. K. Robinson, and P. Hinds. When in rome: the role of culture & context in adherence to robot recommendations. In *Procs of the 5th ACM/IEEE Int. Conf. on Human-robot interaction*, pages 359–366. IEEE Press, 2010.

[18] S. Woods, M. Walters, K. L. Koay, and K. Dautenhahn. Comparing human robot interaction scenarios using live and video based methods: towards a novel methodological approach. In *Advanced Motion Control, 2006. 9th IEEE International Workshop on*, pages 750–755. IEEE, 2006.

[19] Y. Zhao and Q. Zhu. Evaluation on crowdsourcing research: Current status and future direction. *Information Systems Frontiers*, 16(3):417–434, 2014.