Content Is King: Impact of Task Design for Eliciting Participant Agreement in Crowdsourcing for HRI

Alisha Bevins[1], Nina McPhaul[2], and Brittany A. Duncan[1(✉)]

[1] University of Nebraska-Lincoln, Lincoln, NE 68588, USA
{abevins,bduncan}@cse.unl.edu
[2] Howard University, Washington, D.C. 20059, USA
nina.mcphaul@bison.howard.edu

Abstract. This work investigates how the design of crowdsourced tasks can influence responses. As a formative line of inquiry, this study sought to understand how users would respond either through movement, response, or shift of focus to varying flight paths from a drone. When designing an experiment, running several proto-studies can help with generating a dataset that is actionable, but it has been unclear how differences in things such as phrasing or pre- and post-surveys can impact the results. Leveraging methods from psychology, computer-supported cooperative work, and the human-robot interaction communities this work explored the best practices and lessons learned for crowdsourcing to reduce time to actionable data for defining new communication paradigms. The lessons learned in this work will be applicable broadly within the human-robot interaction community, even outside those who are interested in defining flight paths, because they provide a scaffold on which to build future experiments seeking to communicate using non-anthropomorphic robots. Important results and recommendations include: increased negative affect with increased question quantity, completion time being relatively consistent based on total number of responses rather than number of videos, responses being more related to the video than the question, and necessity of varying question lengths to maintain engagement.

Keywords: Crowdsourced · Gesture · Aerial vehicle

1 Introduction

This work seeks to inform future researchers on lessons learned and recommendations for conducting a crowdsourced study to elicit participant responses to non-anthropomorphic robots that produce high rater agreement. To explore this we embarked on an exploratory study, which included exploring question types

Supported by NSF-IIS-1925148, NSF-IIS-1750750, NSF-CNS-1757908, and NSF-IIS-1638099.

A. R. Wagner et al. (Eds.): ICSR 2020, LNAI 12483, pp. 640–651, 2020.
https://doi.org/10.1007/978-3-030-62056-1_53

to elicit desired responses, selection of content for labeling, and gaining insight into the differences in the presentation of questions. These design methods, in combination with trying to minimize issues of participant fatigue, negative affect, and lack of attention to prompts, led to researcher reflection and lessons learned which could be valuable to others seeking to leverage these methods in future experiment design. Additional information for designers of aerial vehicle flight paths will be provided, but is presented as a case study to exemplify the impact of the question types, repetitions, and length of tasks on the participant responses.

The research questions in this work are twofold:

1. When designing an open-ended crowdsourced study, how do various parameters impact participants and their responses?
2. Does the design of the questions or content of videos impact consistency of participant responses?

The findings and their relation to the impact of study parameters, including: the number of questions and/or videos, length of tasks, and forms of questions, on participants and their responses should inform future researchers on how to best structure their studies for success. The second question is generalized to understand the impact of the content and questions and will be examined here within the context of the case study. Success for that study was defined as eliciting responses to understand what the participants perceived the robot was requesting of them or how they were being directed and how they would respond to those requests, referred to as an action based response.

Lessons learned from this work include:

1. Participants quickly assume questions are the same if they are of similar length
2. Asking additional questions has a similar impact on increasing timing whether they are presented individually or as multiple associated with the same content
3. Questions which elicited more positive participant affect seemed to indicate higher agreement

These lessons led to recommendations that researchers focus most on the content presented rather than on their underlying questions to produce the most actionable responses, cycle through similar questions of different lengths if participant attention is key to their responses, and gather the data that is anticipated to be helpful because reducing questionnaires does not appear to have a meaningful impact on time to completion.

2 Related Work

2.1 Question Design

Best practices for creating questions are that they should be: concise, easily interpreted, and use accessible language in order to appeal to the diversity of participants likely to be recruited in crowd sourcing studies [2,6].

Previous works with anthropomorphic robots have shown that free responses yield the most diverse or creative results [3,14]. The non-anthropomorphic nature of many robots can lead to participants simply describing the motion of the vehicle, rather than inferring requests or deriving information from the actions. In order to elicit more humanlike responses to the sUAS small Unmanned Aerial System), questions can be worded to request more human-like descriptions, as seen in anthropomorphic studies. Anthropomorphic studies tend to include questions that imply that the robot had intention and was intelligent [1,3,5,14] which increased participants' confidence in the robot [14].

2.2 Crowdsourced Data

Positive Aspects of Crowdsourcing. A major problem with performing only in-person studies is the lack of diversity in participants (and difficulty recruiting in general). In-person studies typically result in testing a small subset of people living within a short distance from the researchers and with limited diversity in culture and/or age range. Conducting these experiments online, when possible, allows answers from around the country or world. Crowdsourcing is also a convenient solution for obtaining large amounts of responses to quick answer questions [13,15].

When comparing crowdsourced results to in-person, researchers have seen minimal to no difference in their results between the participants who came in person and those who completed tasks online [4,16].

Microtask Design in Crowdsourcing. It should be noted the difference between the tasks presented here and micro-tasks. Micro-tasks, another common usage of crowdsourcing, has the survey-taker complete very short tasks (each requiring no more than a few seconds) typically in large quantities. A significant amount of research has been completed about how to run these properly. Gadiraju et al. [9] researched the impact of malicious behavior in these platforms by exploring the concept of "Untrustworthy workers". These are workers who provide wrong answers in response to straightforward attention-check questions. In this work, we'll refer to them as "Non-Responses". Gadiraju, Yang, and Bozzon [10] also explored how to design instructions and task titles to explain how tasks can be crafted to provide clearer instructions to workers.

3 Methods

3.1 Participants

In total there were 80 Amazon mTurk (46 male, 33 female, 1 no answer) participants between the ages of 24–68 years old ($M = 38.58$ $SD = 10.66$). With an education level ranging from high school to graduate degree, breaking down into 4 Graduate School, 35 Bachelor's, 11 Associate's, 17 Some College, 12 High

school, and 1 No Answer. A majority identified as American (76), 3 identified as Indian, and 1 identified as Chinese.

All participants were required to be considered an mTurk Master, as determined by Amazon through analyzing worker performance over time. They must continue to pass the statistical monitoring in place to retain that qualification. Each participant was paid 4 dollars and Amazon was paid 1 dollar for recruitment.

3.2 Design and Method

Participants selected the study from mTurk and then were taken to a webpage where they were asked to complete a consent form followed by a demographic questionnaire, the Positive and Negative Affect Scale (PANAS) (based on their condition), and the Negative Attitude towards Robots Scale (NARS). PANAS was used to assess how participant affect changed throughout the study to understand the impact of manipulation, and NARS was used due to the findings of Riek [12] that found people with high NARS had difficulty in recognizing robot motions when interacting with a humanoid robot. Following these tests they were then redirected to a Google Form where they were asked to watch 16 unique 30-second videos of a drone flying in a pattern, either once or twice depending on their condition, followed by 1 or 2 questions about each video. Although the participants were requested to watch the entire video, they did have the capability of answering the question and proceeding on to the next question before the end of the video. After the videos, they all completed a post-survey questionnaire consisting of a few questions about the study. If they completed PANAS prior to the videos, they were asked to complete PANAS again at this time. The participants were allotted 1 h to complete the tasks and averaged 30.7 min overall.

4 Approach

4.1 Question Variants

Three variations of question types were investigated to elicit a variety of responses. The groupings were based on whether they were expected to elicit a replication description, speech, or physical response from the participant. Two questions were available for each of the three question types. A full listing of the conditions used, the questions, question character length, whether that test used PANAS, and how many participants were in each of the conditions can be seen in Table 1.

Gesture-based questions are meant to elicit a response regarding how the participant may relate the action of the drone to an action they are familiar with seeing in other people (of similar culture/area). Speech based questions were asked to see how participants may assign verbal communication to the drone's actions. One question from the Speech type and one question from the

Table 1. Study Conditions. L is the character length of the question, N is the number of participants in that condition

Condition	Question(s)	L	PANAS	N
1 Speech	If you saw this drone in real life, what would it say to you?	61	Yes	8
			No	8
2 Speech	If you saw this drone in real life, what would it say to you?	61	Yes	8
	If this drone could speak what would it tell you to do?	55		
1 Gesture	What human gesture does this remind you of?	43	Yes	8
			No	8
2 Gesture	What human gesture does this remind you of?	43	Yes	8
	If you had to replicate this movement with your head and/or body, what would you do?	84		
1 Speech 1 Gesture	If you saw this drone in real life, what would it say to you?	61	Yes	8
	What human gesture does this remind you of?	43	No	8
1 Physical	If you were in the room with the robot, what would you do immediately following the robot's action?	99	Yes	8
1 Physical	If you were in the room with the robot, how would you respond immediately following the robot's action?	103	Yes	8

Gesture type were selected to run together in order to see if people would give complementary responses across both types and whether these responses would give greater insight into their responses. A set of Takayama's [14] questions were reformatted to ideally capture both the speech and gesture question types, while allowing the participant to answer in either way or with a more physical response.

4.2 Length of Tasks

Reduce Questionnaires. We observed that the participant responses seemed to indicate less engagement, through becoming either less informative or more hostile, towards the end of the tasks. We wanted to see if we could minimize these types of responses by reducing the amount of requested tasks, to elicit more engagement or variability within the answers compared to those participants who had the full length survey. To test this we reduced the questionnaires by removing PANAS for three of the conditions.

Additional Videos. When considering the number of videos that were presented to participants, it was necessary to repeat the set of videos when asking two questions from the same category (two speech or two gestures), as they would appear on separate pages. This presentation was chosen to allow us to see whether participants answered consistently throughout the condition and whether slight differences in wording elicited more informative responses.

4.3 Video Content

Sixteen videos were created to include the motions from [7], as well as additional videos that corresponded to both the taxonomy and the most popular flight paths from [8]. Each video was 30 s in length with repetitions of the flight added to reach the desired length of the video, as necessary. Flight paths were held constant for speed and distance covered as much as possible.

5 Participant Responses

As a formative line of inquiry, this study sought to classify how users would respond either through movement, response, or a shift of focus to varying flight paths from a drone. While lines of inquiry for human speech or gesture were investigated, the underlying goal was to understand what the participants perceived was being requested of them or how they were being directed by the vehicle and how they would respond to those requests.

5.1 Category Definition

Two raters were obtained to independently label responses across three classification categories and directions were provided to give context for categorization without guiding the raters to any responses. The raters were given 3 questions asking them if the response indicated an intention to or request for: (1) participant movement, (2) a verbal or physical response to the drone, or (3) a shift of focus, chosen from an initial high-level categorization of responses.

5.2 Findings by Question

After their independent assessments, the raters' results were compared in order to calculate Cohen's Kappa for their agreement according to [11]. When considering the raters' responses based on the questions, the agreement scores had higher variability than those by video. Thirteen out of eighteen categories had a Kappa of at least .61 indicating "Substantial" or .81 indicating "Near Perfect" agreement, with only one category having less than .41 "Moderate" agreement.

When considering answers chosen more frequently by raters, those with chance assignment outside the average plus or minus one standard deviation, there were three questions that were relatively successful at prompting the types of responses we were requesting. "If this drone could speak what would it tell you to do?" was likely to elicit responses requesting participant movement with

Moderate agreement (.60). "If you saw this drone in real life, what would it say to you?" was likely to elicit a verbal or physical response with higher than average probability and Substantial agreement (.71). "If you were in the room with the robot, how would you respond immediately following the robot's action?" was likely to elicit a shift of focus with Substantial agreement (.75). Finally, when combining the two speech questions, there was a likely participant movement request or intention with Near Perfect agreement (.81).

5.3 Findings by Video

When labeling the results by video, the raters had "Substantial" or "Almost Perfect" agreement on all videos across all three categories, since all Kappa values were above 0.61.

This agreement indicates a higher likelihood of participant responses indicating an intention to or request for participant movement when responding to the videos for Back and Forth, Descend and Shift, and Diagonal Descend. On the contrary, participant responses were unlikely to indicate an intention to or request for participant movement when viewing Hover, "U", or "X" Shape.

When considering intention to or request for verbal or physical response to a drone, only Horizontal Figure 8 was likely to elicit a positive response. Conversely, "U" Shape and Yaw were unlikely to elicit intention to or request for verbal or physical response.

Finally, a shift of focus seemed indicated by Yaw, Hover, and Diagonal Descend while Spiral and "U" Shape were unlikely to elicit a shift of focus.

6 Results on Crowdsourcing Methods

Important results from this study point to some key insights when designing crowdsourced studies to elicit the responses that researchers are seeking without reducing participant engagement. We considered the participant responses holistically, looking at responses to the PANAS, number of rejections per HIT, and ultimately question success in eliciting consistent, actionable responses as assessed by the raters.

6.1 PANAS

Multiple conditions were run with 56 participants taking PANAS and 24 participants not taking PANAS to understand the impact of question type on participant affect, but also consider whether reducing the amount of tasks improved participant response quality. When considering the length of time on tasks, it was relatively stable across the conditions with two questions and was 20% shorter when participants answered only one question per video.

If participants were becoming fatigued by the number of questions, then we would expect their PANAS scores to be significantly impacted in the two question conditions when compared to the one question conditions. While we did have five participants (out of 24) with overwhelmingly negative affect in the two question

conditions, there were three out of 32 with similarly negative feelings in the single question conditions. There were also four (out of 32) participants with overwhelmingly positive affect after completing the single question conditions, compared to none in the two question conditions. This is very preliminary data, but does indicate that participants are showing a lack of positive feelings, which could be due to fatigue or frustration.

Interestingly, the overwhelmingly positive responses were related to the speech and physical questions, indicating that these questions might be more approachable or interesting to the participants. These questions also produced a high level of rater agreement regarding verbal or physical response to the drone and shift of focus, respectively.

6.2 Rejections

Occasionally a participant would complete the study, but be rejected later due to failing an attention check or lack of responsiveness. The first and main reason was that an attention check was performed within the set of questions to make sure the participants were reading each question and watching each video. We placed a word in the middle of one video and asked them to type out the word they saw on the screen. Lack of reporting this word alone caused 15 Rejections. The second reason for a rejection was if the participant clearly did not complete the study appropriately. This method of rejection included if they used vulgar language or had over 50% non-responses. Non-responses were classified as repeated "Nothing", "Not sure", or repeating the same identical answer in multiple boxes. 6 people were rejected for non-responses. In total there were 21 Rejections (who were not paid) out of 101 participants who attempted the mTurk HITs. Additionally, there were 4 people who were paid for their work on the HITs, but provided at least 25% (and less than 50%) answers that were considered non-responses.

The number of rejections for lack of attention was highest both numerically and by percent for the single speech and gesture questions (13 rejections over 32 requested participants, resulting in 45 HIT attempts) compared to a single rejection for the 24 participants in the three double question categories (two speech, two gesture, and one speech/one gesture). The much higher rate for the single speech and gesture questions might be related to the length of the question (43 or 61 characters, both fitting on a single line) being similar to the attention check (46 characters) and the participants not paying attention to the questions since the questions did not change between pages in those conditions (as opposed to the two question conditions). The question length consideration is reinforced when comparing to the single rejection for 16 participants in the physical question categories (99 or 103 characters).

6.3 Time Reduction

As described earlier, participant fatigue was an open question for this work and steps were taken to reduce fatigue by eliminating the PANAS questionnaire on some conditions. Unfortunately, this reduction (removing the three sets of 20

Table 2. Significant kappa and chance values by question

Participant question	Kappa	Chance	Significant rater question
If this drone could speak what would it tell you to do?	0.603	52.779	Participant Movement
If you saw this drone in real life, what would it say to you?	0.708	116.459	Response to Drone
If you were in the room with the robot, how would you respond immediately following the robot's action?	0.748	12.750	Shift of Focus
What human gesture does this remind you of?	0.965	0.536	Participant Movement
If you saw this drone in real life, what would it say to you? If this drone could speak what would it tell you to do?	0.808	42.055	Participant Movement
	0.636	93.507	Response to Drone
	0.665	3.209	Attention Shift

PANAS questions, or 60 total) only reduced time from 31.5 to 29.1 min. Given that this reduction only saved on average 2.4 min per participant, the additional information was likely more valuable than the fatigue that was produced from the surveys.

Another consideration was whether the addition of two questions per video would impact the amount of time participants spent on the overall tasks. For this we considered one question per video (28.1 min), two questions per video (35.0), and two questions with each video shown twice (35.8 min). It is interesting to note that doubling the questions resulted in basically the same amount of additional time (about 7 min) whether the participants were explicitly asked to watch the videos again or not.

Finally, the paired questions were run as a test to see if they helped with triangulating participant responses. Categorization from two raters showed that participants gave similar answers for two questions asked about the same subject around 40% of the time overall. When participants were asked the two speech questions, found in Table 1, only 24.2% of the responses were categorized as similar with a Kappa of 0.71. This supports the idea that people are providing complimentary information, rather than the same answer reworded for speech questions. This also confirms the findings of Table 2, which says the two speech questions elicit different types of responses, one being more informational and the other seeking a command. On the other hand the gesture questions when asked together provide an agreement 57.8% of the time (Kappa 0.69), showing that participants interpreted the questions in similar ways, and thus asking the two gesture questions was less beneficial for this complimentary information, but better when seeking consistent responses. The presence of two questions with substantially different foci, one speech and one gesture, produced converging ideas around 40% of the time.

7 Lessons Learned

As referenced in the title, the most important lesson from this study is the content (or in this case videos) being tested has the biggest impact on participant responses. While we hope that the different questions will result in interesting taxonomic differences, the current finding is that the videos generated more consistent action-based responses than the questions. This finding is promising when considering the ability to elicit consistent responses from novices, but troubling when considering how to shape participant responses.

The other key lessons learned from this work involve the ease with which participants are lulled into not reading questions, the impact of multiple questions, and the relationship between participant affect and agreement of responses.

7.1 Participant Attention

The key lesson from the attention check is that the participants stopped reading the questions if they look (even at a very high level) to be the same. Most participants who were rejected for not completing the attention check responded that the word never appeared in any video they watched and only later found the attention check after carefully reviewing their HIT. Most of the rejections answered reasonably about the condition question in reference to the attention check video and simply missed the change in question and word in the middle of the video. This also indicates that many participants were not watching all iterations of the flight paths while considering their answers to the questions.

This raises some questions about how quickly participants acclimate to consistency in questions and how to change questions between tasks in order to maintain some level of engagement with questions that may not all be consistent.

7.2 Multiple Questions

We found that asking multiple questions was not highly predictive of lack of engagement (as reflected by number of rejections), so this is a positive finding for researchers moving forward to continue asking triangulating questions. However, care should be taken to consider that asking multiple questions significantly slows participant responses by about the amount of time allocated per task (in this case 7 min) whether or not they are explicitly repeated on different pages, or asked similar questions on the same page. This is important for task design because it indicates that even if responses are expected to be highly correlated, participants are still seemingly considering them independently. The increase in time also indicates that, contrary to the findings on the attention check, participants do appear to be paying attention to what is asked and considering their responses.

8 Discussion

8.1 Limitations

A taxonomy of participant responses would be a valuable contribution to understand the emotions elicited, the target of communications, and other common themes from the responses, but is outside the scope of this paper. As a standalone guide for eliciting action-based responses, this paper provides insights into the perception of the questions posed to participants and the impact of the task design on the attitudes and responses provided, which was thought to be a valuable contribution in its own right.

8.2 Recommendations

When developing the content to test, it is important to understand the underlying responses that are sought and to ensure that the prompts are appropriate for the responses. The video content was more important because the questions were relatively consistent and focused on the participant perceptions or responses to the content. This finding is supported by the high Kappa values for each of the videos (all with substantial or near perfect agreement) compared to the variable Kappas across the questions (only 13/18 categories across all 6 questions had substantial or near perfect agreement) with a subset shown in Table 2.

Additionally, when examining the responses and considering the questions that are asked, it would be a good practice to cycle through multiple forms of questions to keep participants engaged in reading and responding to the prompts. This is supported by the fact that tests which had similar question length (43 or 61 character questions) to the attention check question length (46 characters) had 3–4 more rejections on all four of the tests than the other 6 tests which are visually different (either having two questions reduced to one or having a 99 or 103 character question). This is also supported by the convergent answers to questions with similar requests related to gestures and divergent answers for questions with complementary requests related to speech; similar questions presented to the same participants resulted in additional information depending on the content of the question.

One surprising finding was the relatively stable amount of time it took for participants to complete the HIT regardless of the length of the questionnaires and the relationship between the number of questions (rather than videos) and the time to completion. Results showed that participants completed the tasks in about 28 min with only one question at a time and about 35 min with two questions, showing minor fluctuations of 1–2 min with other edits to length of task, such as removal of PANAS or requesting the video be watched twice. This indicates to us that we should continue to collect information that might complement the participant responses and continue to test multiple questions in the way that best makes sense. An incorrect perception we had was that asking participants to watch the videos again would take too long, but the reality is that developing the responses was the time consuming part of the task.

References

1. Bartneck, C., Kanda, T., Mubin, O., Al Mahmud, A.: Does the design of a robot influence its animacy and perceived intelligence? Int. J. Social Robot. **1**(2), 195–204 (2009)
2. Brosnan, K., Babakhani, N., Dolnicar, S.: "i know what you're going to ask me" why respondents don't read survey questions. Int. J. Market Res., 1470785318 821025 (2019)
3. Cañamero, L., Fredslund, J.: I show you how i like you-can you read it in my face? [robotics]. IEEE Trans. Syst. Man Cybern.-Part A: Syst. Hum. **31**(5), 454–459 (2001)
4. Casler, K.: Separate but equal? A comparison of participants and data gathered via Amazon's MTurk, social media, and face-to-face behavioral testing. https://doi.org/10.1016/j.chb.2013.05.009
5. Chaminade, T., et al.: Brain response to a humanoid robot in areas implicated in the perception of human emotional gestures. PLoS ONE **5**(7), e11577 (2010)
6. Christensen, L.B., Johnson, B., Turner, L.A., Christensen, L.B.: Research methods, design, and analysis (2011)
7. Duncan, B.A., Beachly, E., Bevins, A., Elbaum, S., Detweiler, C.: Investigation of communicative flight paths for small unmanned aerial systems* this work was supported by NSF NRI 1638099. In: 2018 IEEE International Conference on Robotics and Automation (ICRA), pp. 602–609. IEEE (2018)
8. Firestone, J.W., Quiñones, R., Duncan, B.A.: Learning from users: an elicitation study and taxonomy for communicating small unmanned aerial system states through gestures. In: 2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI), pp. 163–171. IEEE (2019)
9. Gadiraju, U., Kawase, R., Dietze, S., Demartini, G.: 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, pp. 1631–1640. ACM (2015)
10. Gadiraju, U., Yang, J., Bozzon, A.: Clarity is a worthwhile quality: on the role of task clarity in microtask crowdsourcing. In: Proceedings of the 28th ACM Conference on Hypertext and Social Media, pp. 5–14. ACM (2017)
11. Landis, J.R., Koch, G.G.: The measurement of observer agreement for categorical data. Biometrics, 159–174 (1977)
12. Riek, L.D., Rabinowitch, T.C., Bremner, P., Pipe, A.G., Fraser, M., Robinson, P.: Cooperative gestures: effective signaling for humanoid robots. In: Proceedings of the 5th ACM/IEEE International Conference on Human-Robot Interaction, pp. 61–68. IEEE Press (2010)
13. Sorokin, A., Berenson, D., Srinivasa, S.S., Hebert, M.: People helping robots helping people: crowdsourcing for grasping novel objects. In: 2010 IEEE/RSJ International Conference on Intelligent Robots and Systems, pp. 2117–2122, October 2010. https://doi.org/10.1109/IROS.2010.5650464
14. Takayama, L., Dooley, D., Ju, W.: Expressing thought: improving robot readability with animation principles. In: 2011 6th ACM/IEEE International Conference on Human-Robot Interaction (HRI), pp. 69–76. IEEE (2011)
15. Tellex, S., et al.: Understanding Natural Language Commands for Robotic Navigation and Mobile Manipulation, p. 8
16. Toris, R., Kent, D., Chernova, S.: The robot management system: a framework for conducting human-robot interaction studies through crowdsourcing. J, Hum.-Robot Interact. **3**(2), 25 (2014)