CS3106 - Practical 2

20007413

1 Question 1

1.1 One-Way ANOVA

Results	Table 1 (Gamepad A & B)	Table 2 (Joystick C & D)	Table 3 (Touch Panel E,F & G)
SS_{error}	404	64	440
SS_{total}	596	151.5	$\frac{2258}{3}$
SS_{effect}	192	87.5	$\frac{938}{3}$
n Participants	12	14	21
m Groups	2	2	3
df_{error}	10	12	18
df_{effect}	1	1	2
MS_{error}	40.4	$\frac{16}{3}$	$\frac{220}{9}$
MS_{effect}	192	87.5	$\frac{469}{3}$
F-Distribution	$\frac{480}{101} \approx 4.75248$	$\frac{525}{32} = 16.40625$	$\frac{1407}{220} \approx 6.39545$
Critical value	4.96460	4.74723	3.55456
Significant at $p < 0.05$?	FALSE	TRUE	TRUE
Table 1 Reporting	Gamepad A resulted in fewer average error s than Gamepad B (37 vs. 45 respectively). We assumed these errors were normally distributed. Analysis of variance at significance level of $\theta = 0.05$ showed that this difference was not statistically significant ($F_{1,10} = 4.752, p < 0.05$).		
Table 2 Reporting	Joystick C resulted in fewer average errors than Joystick D (33 vs. 38 respectively). We assumed these errors were normally distributed. Analysis of variance at a significance level of $\theta = 0.05$ showed that this difference was statistically significant ($F_{1,12} = 16.406, p < 0.05$).		
Table 3 Reporting	Touch Panel E, F, and G resulted in different average errors than eachother. We assumed these errors were normally distributed. Analysis of variance at significance level of $\theta = 0.05$ showed that there was a statistically significant difference between at least two of the Touch Panels ($F_{2,18} = 6.395, p < 0.05$).		

Table 1: Results for the one-way ANOVA of three different "between-subjects experiments" on two gamepads A and B, two joysticks C and D, and three touch panel systems E, F, and G.

1.2 Post-Hoc Test

One suitable post-hoc test for the touch panel experiment is Turkey's HSD Test for multiple comparisons to determine exactly which differences in average errors between-groups are statistically significant.

1

2 Question 2

2.1 Non-Hardware Issues to Consider

In developing the TRing interaction technique, there are a number of non-hardware issues to consider.

First, it is important to consider that some may have disabilities which affect their capacity to use a device reliant on deliberate index finger control. This consideration may affect the decision for the sensitivity of activation and control gestures.

Next, it is important to give the user feedback on the system status. This is particularly important in regards to selection of interaction objects, and ending of continuous input.

Another issue to consider is the cultural norms around control gestures. Assumptions about what users will find the most intuitive may be wrong, or differ from culture to culture which would be important if the results of the practicability of the device are to be generalized.

Finally, it is important to consider the comfort of the device. As this type of device is worn on the index finger, it is important to consider the weight and fit of the device, as wearing something not well-fitted may cause injuries.

2.2 Evaluation Experiment

Here we present an potential experimental study for evaluating the effectiveness of the TRing as an alternative input device to a physical remote for interacting with a television.

Experimental Design This experiment will have one independent variable which will be the input device used TRing, Standard Remote. There will also be one dependent variable - the time taken to complete each task given.

Participants 20 participants will be recruited from the university. No pre-requisites will be required for participation.

As there is a risk of asymmetrical skill transfer of completing tasks, counter balancing will be used. Half of the participants will be randomly assigned to use the TRing device in the first half of the experiment, and the physical remote in the second half of the experiment. The rest of the participants will use the physical remote in the first half, and the TRing in the second half.

Apparatus The experiment will be conducted whilst seated on a sofa. In front of the sofa will be a television with a user interface allowing users to see the current program information, channel number. The user interface also allows user to bring up a menu, which allows the user to go to a TV-guide consisting of a list of channels with channel numbers and the title of the current program playing for each channel.

For one half of the experiment, participants will be given a TRing input device, fitted before the start of the trial. The control magnet will be installed into the arm of the sofa, with a labelled control panel. The controls will contain button controls for left, right, up, and down controls; a menu button; a volume slider; and a 9 button numeric keypad.

In the other half, participants will be given a 3D printed intra-red remote control representing the typical TV input method. However, a simplified layout to match that of the TRing control panel. Physical buttons will be used for the left, right, up, down, menu, and 9-digit keypad controls. Instead of having a volume slider, a volume-up and volume-down button will be available.

Material Participants will be given the same sequence of 20 tasks to complete. The task order will be different for the first and second half of the experiment, but will be the same for all participants. This is to prevent skill-transfer effects. The task list will contain tasks common TV control tasks.

Procedure Before the start of the experiment, participants will be taught how to use both input types, and will have 10 minutes to get accustomed to each input device using a demo user interface, different to the trial interface.

Participants will then be asked follow each task in sequence using the first input device assigned. The start time and end time for each task will be recorded, which will be used to calculate the time taken to complete each task. After, the participant will complete the same tasks in a different

2.3 Threats to Experimental Validity

One threat to the validity of the above outlined experiment is participants not being able to complete the tasks given at all, as the reason for incomplete tasks may be due to the difficulty of interacting using the given input device.

Another threat to the validity of the data is experience gained from every-day life. Given the popularity of televisions, most people have experience using physical remotes with buttonc controls to interact with televisions. To mitigate this, we have suggested training users before the start of the experiment in their assigned input device, but there is a risk that this will not overcome their previous experience.

3 Question 3

Online crowdsourcing platforms allow individuals and groups to complete tasks by giving them access to a large, and diverse crowd of workers from around the world, completing micro-

tasks in exchange for monetary compensation.[1]. One type of microtask deployed on these platforms are surveys, which have workers answer a set of questions, usually relating to their personal life or beliefs.[2]

Previous research has suggested that a key drawback from these methods is low-quality responses from workers seeking to maximize their monetary gain through rapid, low-quality completion of microtasks.[3][4]

Crowdsourcing platforms have put into place techniques to curb these malicious workers. One such method is 'gold-standard' questions, which have pre-determined correct answers to detect workers not answering questions accurately, which then tags workers as either 'trustworthy', or 'untrustworthy'.

Authors of the 2015 paper 'Understanding Malicious Behaviour in Crowdsourcing Platforms: The Case of Online Surveys',[2] believed existing techniques were insufficient for curbing malicious activity - especially in regard to surveys, which often contain questions with no real 'correct' reponse, and so measuring their validity is difficult.

The paper identified and categorized common malicious worker behavioural patterns, developing five malicious-worker profiles. The study supports their initial theory that Gold-standard questions are not perfect for preventing malicious behaviours, and so present an alternative technique for more accurately quantifying the maliciousness of workers, with guidance on when to reject worker responses. Finally, the authors outline guidelines for survey creators to prevent malicious workers in the first place.

The experiment consisted of deploying a survey with 34 questions. The question types varied from open-ended (allowing written responses from workers), multiple-choice , and "Likert-type" (attention grabbing question intended to keep workers interested). Two of the questions were Gold-standard, which tagged workers as 'untrustworthy' upon incorrect answers to either question. Experts were then asked to analyse worker responses to the open-ended questions, deeming them either acceptable or unacceptable based on the relevance to the question. A maliciousness score was kept for each worker, incremented for every unacceptable response given. The experts were then asked to find patterns in unacceptable responses, and in the behaviours of workers giving them. An analysis of the gold-standard test compared to expert categorized responses was also done.

The sample consisted of 1000 workers on one of the crowdsourcing platforms. Participants were self-selected, and given a 0.2 $ incentive per-unit completed. The only prerequisite for participants to complete the survey was speaking English (the language in which questions were written) - verified by the platform using location data to check the worker was located in an English-speaking country.

The results concluded that gold-standards tests were generally only effective at detecting one type of malicious worker - the fast deceiver. These workers were the most common type, and tried to bypass the automated input verification to complete the survey as-quickly as possible. However, four other malicious worker profiles were identified. To combat these behaviours, the authors created additional guidelines for survey design.

The next most common malicious behaviour profile were rule-breakers, who did not conform to question rules. Despite not being detected by gold standards questions, the authors note that more stringent input validators would be a practical way to prevent this type of behaviour. They also note that these types of validators for open-ended questions would also help prevent fast-deceivers, who often tried to copy and paste responses for every question.

The malicious behaviour type with the greatest average maliciousness were smart-deceivers, who attempted to more carefully conform to question rules to avoid automated detection. These were very rarely caught by Gold standard tests.

The researchers found that previous malicious responses were a strong predictor for future malicious responses. The authors quantified this by marking the question number at which a worker first gave an unacceptable response, called their 'tipping-point'. They found that workers with earlier tipping-points had a much greater average overall maliciousness. Additionally, workers with a shorter total survey completion time scored, on average, higher in maliciousness. From these findings, the authors concluded that, by verifying the acceptability of a small portion of early trustworthy worker responses to determine tipping points, and by prioritizing verification of workers with short completion times, a large proportion of malicious workers could be detected with little additional cost.

The authors also suggest that this technique could also reduce the impact of the final type of malicious worker profile: gold-standard preys. These workers ranked very low in maliciousness, yet fail the gold-standard questions, and so were incorrectly labelled as untrustworthy. The authors suggest that by verifying early responses from untrustworthy workers, these could be detected.

A potential weakness of this study is in assuming that malicious worker behaviour will not change in response to new detection techniques. The practicability of tipping point detection comes from the relatively few responses that need manual verification. However, given that the results found that some malicious workers already take steps to avoid detection (i.e. smart-deceivers), it is wholly reasonable to assume they would adapt their earlier behaviour to shift their tipping point later. This would then require more effort to verify more questions, which may continue, thus reducing the cost-effectiveness of the technique.

The research methods also note that they did not use all the quality control methods offered by the platform, as they were focusing on malicious worker behaviour. However, the proportion of malicious behaviour types in the results may not be representative of a standard task.

Further research has supported the importance of quality controls for crowdsourcing tasks.[5] However, more systemic changes have been suggested in regard to the reward-driven incentive model of crowdsourcing platforms.[6] In particular, alternate incentives such as social-prestige, gamification, or a sense of common-good have been proposed,[7] which may be a more effective way of preventing malicious behaviours in the first place.

References

[1] Jeff Howe. *Crowdsourcing: How the power of the crowd is driving the future of business*. Random House, 2008.

[2] Ujwal Gadiraju et al. "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*. 2015, pp. 1631–1640. DOI: 10.1145/2702123.2702443.

[3] Yu Zhang and Mihaela Van der Schaar. "Reputation-based incentive protocols in crowdsourcing applications". In: *2012 Proceedings IEEE INFOCOM*. IEEE. 2012, pp. 2140–2148. DOI: 10.1109/INFCOM.2012.6195597.

[4] Winter Mason and Duncan J Watts. “Financial incentives and the" performance of crowds"”. In: *Proceedings of the ACM SIGKDD workshop on human computation*. 2009, pp. 77–85. DOI: 10.1145/1600150.1600175.

[5] Francesco Restuccia et al. “Quality of information in mobile crowdsensing: Survey and research challenges”. In: *ACM Transactions on Sensor Networks (TOSN)* 13.4 (2017), pp. 1–43. DOI: 10.1145/3139256.

[6] Yaron Singer and Manas Mittal. “Pricing mechanisms for crowdsourcing markets”. In: *Proceedings of the 22nd international conference on World Wide Web*. 2013, pp. 1157–1166. DOI: 10.1145/2488388.2488489.

[7] Carsten Eickhoff et al. “Quality through flow and immersion: gamifying crowdsourced relevance assessments”. In: *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*. 2012, pp. 871–880. DOI: 10.1145/2348283.2348400.