Gamify HCI: Device's Human Resolution for Dragging on Touch Screens in a Game with Lab and Crowd Participants

Allan Christensen, Simon André Pedersen, and Hendrik Knoche(✉)

Department of Media Technology, Aalborg University, 9000 Aalborg, Denmark
allanchr@hotmail.com, simon@dk-designer.dk, hk@create.aau.dk

Abstract. We compared a game-based experiment carried out in a lab study to crowdsourced set ups (both uninformed and informed). We investigated the device's human resolution - the minimum size for dragging the finger onto a target on a touch screen. Participants in the lab consistently produced fewer errors than those from the crowd. For lab participants, errors significantly increased between targets of 4 mm and 2 mm in width. The uninformed crowd had too many errors to determine significant differences but the informed crowd yielded useful data and performance declined already for targets between 8 mm and 4 mm width. The smallest selectable target width for dragging for all three groups combined, was between 2 mm and 4 mm on mobile touch devices.

Keywords: Crowdsourcing · Gamification · Device human resolution · Touch-interaction · Dragging · HCI

1 Introduction

Running experiments with human participants drawn from student populations has seen criticism for its poor external validity [5] and crowdsourcing has gained momentum to draw from a wider population that is either monetarily or otherwise incentivized, e.g. by collecting data through games [7] to participate in studies. This paper compares results from crowdsourcing a gamified user study to its lab counterpart with participants from a campus population. One potential strength of crowdscourcing user studies lies in reducing or eliminating acquiescence bias. To this end we compared the performance of naive crowd, informed crowd, and lab participants. The latter two groups knew their game performance was collected for scientific purposes.

2 Background

In this paper, we focus on crowdsourcing with unpaid participants motivated by curiosity or interest in a study, reciprocal altruism towards the experimenter, or motivation to play a game. While improving both population and ecological

A.L. Brooks and E. Brooks (Eds.): ArtsIT/DLI 2016, LNICST 196, pp. 47–54, 2017.
DOI: 10.1007/978-3-319-55834-9_6

validity over lab-based test, crowdsourcing studies raise concerns about internal validity. For example, Henze et al. found implausible results from a game that included rapid touch interactions, which seemed ideal for modelling with Fitts' law [6]. This could have been due to multi-finger entry or other tricks violating how the task was supposed to be carried out.

As a case, we used a study on the unknown limits in precision when dragging a finger onto small targets on a touch screen. We draw on Fitts' Law [4] and the concept of Device Human Resolution (DHR) [2]. Fitts' law predicts the required time for a human to perform a movement over a distance (*amplitude*) from point 'A' to point 'B' with a given a size (*width*). The Index of Difficulty (ID) quantifies the difficulty this task with higher IDs resulting in a harder task and yielding a larger time requirement. We used MacKenzie's extended version of Fitts' ID [9]:

$$Index\ of\ Difficulty\ (ID) = log_2(\frac{amplitude}{width} + 1) \tag{1}$$

Bérard et al. used Fitts' law to determine a Device's Human Resolution (DHR) for mouse, stylus and a free-space device. They defined the DHR as *the smallest target size that a user can acquire with the device, given an ordinary amount of effort*, i.e. without a major decrease in performance in time or accuracy (percentage of successful acquisitions). For mouse input they found a DHR for time (0.036 mm) and error (0.018 mm). Participants were able to maintain a low error rate from 0.036 mm downwards only at the expense of increased time and below 0.018 mm errors increased drastically.

Cockburn et al. compared finger, stylus, and mouse in target acquisition (5, 12.5, and 20 mm width columns) tasks with tapping and dragging [3]. Tapping on 5 mm wide targets with a finger yielded a roughly seven times higher error rate (14%) for acquisition compared to the other devices. Dragging (~0.92 s.) had a significantly higher overall selection time when compared to tapping (~0.57 s.) onto targets mainly attributed to the higher friction when dragging across the screen. But dragging (1% errors) had a significantly higher accuracy than tapping (6.8% errors). The authors attributed this to the offset cursor, which assisted target acquisition while dragging. Tapping had no equivalent feedback on the location of the finger and the 'fat finger' occluded the target. Holz et al. provided two reasons for inaccurate target selections with fingers: (1) users do not know the exact finger surface interaction point - the pixel accurate screen position taken from the skin's contact area with the screen and (2) the imperfect memory of the location of small targets once the finger occludes them [8]. Benko et al. found that users perceive the finger surface interaction point (1) differently [1]. Various design solutions address these problems, e.g. using offsetting the cursor or zooming.

3 Study

The purpose of this study was to compare three different user groups playing a game to investigate DHR for dragging on touch screens. The first consisted of 16 male participants (average age 24, SD = 1.5) from the local university who participated in a lab study including a demographic questionnaire. The uninformed group consisted of 19 participants (crowd), who thought they were merely playing a game and not participating in a study. The third group (crowd-plus) 14 participants (4 female, average age 28, SD = 9.5) knew they were participating in a study. After having completed the game, 86% of them chose to fill in a questionnaire including control variables such as age, environment, and touch device usage. The lab participants used an LG Nexus 4 smart phone running Android 5.1, with a 4.7-inch display and 768 × 1280 resolution, which they held as they pleased.

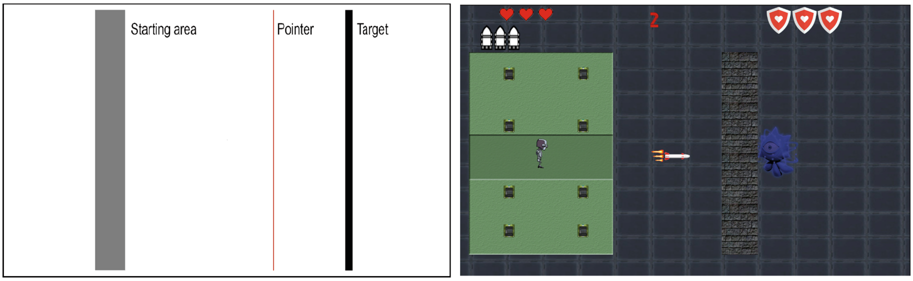


Fig. 1. Design of the standard DHR test (left) and the gamified version (right)

As much as possible we replicated Bérard et al.'s DHR test setup with a game called Wall Destroyer. The user had to tap anywhere within a green starting area (see Fig. 1) and then drag their finger onto a target (wall) that appeared 47 mm away in seven descending widths (32, 24, 16, 8, 4, 2, 1 mm) per round resulting in the following Fitts IDs: 1.32, 1.58, 2, 2.81, 3.7, 4.64, and 5.61. Successful completion of a drag required lifting off the finger when on top of the target. The completion time ran from the touch down event of the dragging finger in the green area to the lift off event on or near the target. Unlike other DHR studies we did not use the second hand to validate target acquisition. The lift-off part of the dragging gesture is essential in understanding the DHR of dragging since the touch area and position at lift off can be different from when the dragged finger comes to a halt on top of the target. If the lift-off occurred on the target a missile appeared and fired as feedback for hits. On misses the missile did not appear. The game provided auditory feedback for both hits and misses but none on the current touch position of the finger input. But given the wall's length the participants were aware of the targets location. To encourage repetitions of these rounds, participants had five lives and a life was only lost on three successive target misses. Even if you missed a target repeatedly you would proceed to the next target. This approach did not enforce an equal number of repetitions, but encouraged most participants to complete multiple game rounds to provide more data.

After the introduction, participants received the smart phone, were prompted to start the game, and watched the ca. 30 s introductory video illustrating how to play and complete the game. The game started on completion of the video. For better between group comparisons, we did not provide any additional assistance to the lab participants in case of questions.

Both crowd groups downloaded the app from the Google Play Store but the crowd-plus participants saw a consent page at start-up. On pressing 'okay' they were redirected to the main menu and from this point on crowd, crowd-plus and lab participants followed an identical procedure. After completion of the game, all saw their own high score. The crowd-plus group further received a pop-up message prompting them to answer a questionnaire.

4 Results

Unless noted otherwise, we used a one-way Analysis of Variance test (ANOVA) with a TukeyHSD as a post-hoc test for analysis. To find differences in slope of all subsets of three successive IDs (e.g. of 1.32, 1.58, and 2) and the overall model slope containing all IDs we ran linear models for each and tested for significant differences between subset model and the overall model. We present the results for the individual groups first and then the overall results with all groups combined, see Fig. 2 as a summary.

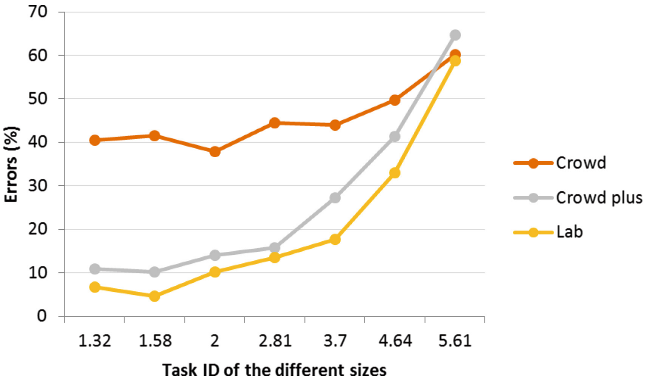


Fig. 2. Target acquisition error rate by target size (in Fitts' ID) for the three groups

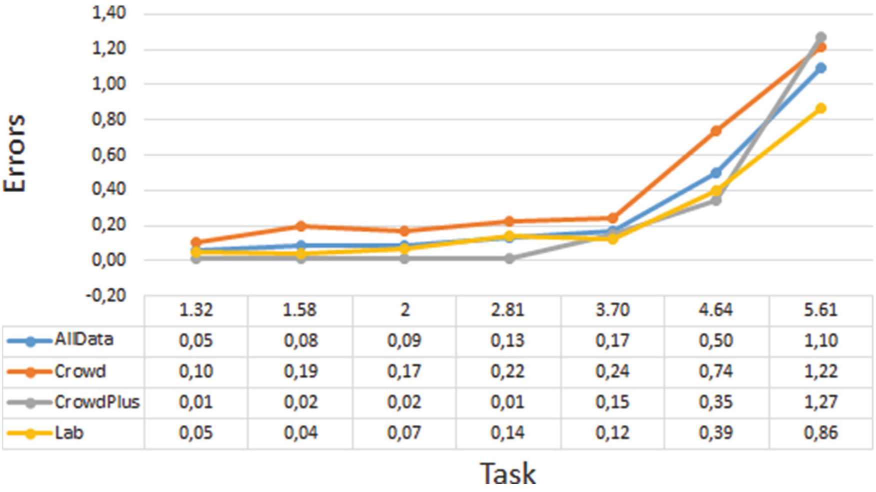
For lab participants we found an effect of size on completion time ($F_{6,1767} = 156.16$, $p \ll 0.001$). However, analyzing the mean slopes for the time data revealed that no subset slope significantly deviated from the overall slope (0.09). For error, a Friedman Ranked Sum test revealed an overall significant difference between the seven sizes ($\chi^2(6) = 63.98$, $p \ll 0.001$) and a post-hoc Friedman test showed significant increases in errors between an ID of 3.7 and 4.64 (4 and 2 mm) and between an ID of 4.64 and 5.61.

For the crowd, we found an overall significant difference for the completion time between the seven sizes ($F_{6,962} = 3.58, p < 0.01$) but no significant deviation from the overall slope (-0.00004). A Friedman Ranked Sum test showed no overall significant difference in errors between the sizes ($\chi^2(6) = 10.80, p = 0.09$). We found a spike between an ID of 2 and 2.81, however, it was not enough of an increase to be significant.

The time data for the crowd-plus showed an overall significant difference between the sizes ($F_{6,840} = 27.46$, $p \ll 0.001$). But the mean slopes for time showed that no subset slope was significantly different from the overall slope (0.09). However, the Friedman Ranked Sum test revealed an overall significant difference for error rates between the sizes ($\chi^2(6) = 23.47$, $p < 0.01$) and the post-hoc Friedman test showed significant increase in errors between target widths of 8 mm and 4 mm (ID of 2.81 and 3.7).

All data. Several participants in the crowd performed very poorly during the experiment. Therefore, we examined the average number rounds the participants in lab (16.6, SD = 10.4), crowd-plus (9.43, SD = 9.34), and crowd (7.7, SD = 9.9) had played. We removed all participants below the average amount of repetitions for each of the participant groups to examine if this change would provide more comparable results. When only including participants performing above average in the number of rounds we retained 7 out of 19 crowd, 6/14 crowd-plus, and 8/16 lab participants.

Group Comparisons. For the time data on the filtered dataset, we found an overall significant difference between the participant groups ($F_{2,2856} = 92.21$, $p \ll 0.001$) and the TukeyHSD found a significant difference between all groups.



	1.32	1.58	2	2.81	3.70	4.64	5.61
AllData	0,05	0,08	0,09	0,13	0,17	0,50	1,10
Crowd	0,10	0,19	0,17	0,22	0,24	0,74	1,22
CrowdPlus	0,01	0,02	0,02	0,01	0,15	0,35	1,27
Lab	0,05	0,04	0,07	0,14	0,12	0,39	0,86

Fig. 3. Error mean per repetition by size (in Fitts' ID) for the three filtered groups and their combined average (AllData)

But when examining the mean slopes no subset slope deviated significantly from the overall slope (0.09) for the time data. We found no significant difference in errors between the participant groups ($F_{2,144} = 0.79$, $p = 0.45$).

Overall DHR for Touch. For the participants with above the average amount of repetitions we found an overall DHR for touch. We found an overall significant difference between the seven tasks ($F_{6,140} = 26.29$, $p \ll 0.001$) in terms of errors. Its TukeyHSD showed a significant difference between an ID of 5.61 and all other tasks. An ID of 4.64 was also significantly different from the other tasks, except for an ID of 1.32. This showed that a significant increase in errors happened between an ID of 3.7 and 4.64 for the overall data. The overall distribution of the error data for each task, after the participants below the average amount of repetitions had been removed, can be seen in Fig. 3 that includes the pooled data from all groups (AllData) after filtering.

5 Discussion

The results confirmed that participants drawn from a campus population playing in a controlled lab environment with no environmental disturbances outperformed crowd participants playing in their own environment for touch tasks. We do not know how performance was affected by the uncontrolled factors: 1. crowd participants' demographics, 2. the environment and setting that they were playing in, 3. differences in task understanding, or 4. a combination of these. But the performance of informed crowd participants who consented to participating in a scientific study was significantly higher than those of naive crowd participants who might have played the game normally with little or no concern regarding their performance. So we could see this as a form of acquiescence bias. The knowledge that their results matter in a scientific study or to a scientist might be motivating to pay more attention and perform better.

Multiple participants did either not understand the dragging task in the Wall Destroyer game, did not want to complete the tasks, and/or performed in general a lot worse compared to others. This was especially the case for the two crowd groups, which had a large spread between the highest and lowest performing participants - much higher than the spread of the lab participants. Furthermore, the repetition data for the three participant groups showed that the lab environment on average completed the game almost twice as many times than the crowd groups. However, crowdsourcing provides access to data at little to no marginal cost. Following Henze et al.'s approach, we removed all the data that was deemed insufficient, i.e. all participants who had below the average amount of task repetitions. In this filtered subset the overall performance between the three groups did not differ significantly.

While measuring the performance differences between groups, we examined the DHR for dragging on a touch screen. The results showed a significant increase in errors between 4 mm and 2 mm target width. This DHR was achieved by the lab participants with all data included, and for the above average performance subset of participants from the two crowd groups. This means that for all participant groups combined, the smallest achievable target width a user can select with a drag on touch screens with little effort is between 2 mm and 4 mm much smaller than average index finger width. We used a target with substantial height which provided cues in terms of the location of the target. Square targets that get completely occluded by the touching finger the DHR might larger in size.

We did not find any differences in completion times. We believe there were two contributing factors. First, the game did not provide any incentives for fast in-game performance. This could be changed in future versions by adding time limits or scores sensitive to time performance, e.g. faster hits yielding higher scores. The average target (wall) acquisition delay across all participants was 0.5 s and the Fitts' law coefficients from MacKenzie's model indicated that most of the movement time was due to the constant ($a = 0.38$) rather than the slope ($b = 0.05$) that depends on the index of difficulty. Furthermore, the fit of Fitts' model even when averaging the time performance of all participants by the different Fitts' IDs was low ($R^2 = 0.23$). We compared this to modeling our data with the ID in Fitts' original model: $log_2(2 \times Distance/Size)$. This approach averaged acquisition delays better ($R^2 = 0.78$) but the coefficients ($a = 0.34, b = 0.07$) were similar to MacKenzie's model. Both MacKenzie's (20.3 bit/s) and Fitts' original model (14.2 bits/s) yielded unrealistically high indices of performance (the inverse of b) that typically lie between 8 and 12 bits/s. In summary, the game in its current design did not yield time performance data that was specific to Fitts' law.

Second, the player did not get any feedback about the actual touch position and could therefore not optimize or correct their finger positions beyond their mental model. This repositioning should yield higher movement times for targets with higher index of difficulty. A setup with positional feedback in a DHR dragging task on touch screens might yield different results in terms of both time and error.

Gamifying existing tests may quickly become tedious for the users, as in our case game elements, scores, lives, animations etc. were insufficient to make the game fun as became clear from our observations during the lab trials and remarks from the lab participants after the experiment. From the lab participants' responses, we believe that adding an overall story for the game in future iterations, may not change the fact that it was neither fun nor very engaging, but rather felt like forcing the participants to do something for an extended period of time, which they did not feel like doing.

6 Conclusion

We found the device's human resolution for dragging with a finger on a touch screen to be between 2 mm and 4 mm. This is comparable to the DHRs for positioning a cursor on targets with in-air interactions (2.4 mm) found by Berard et al. and Bjerre et al. (between 1.2 and 2.4 mm), and much worse than with a mouse (between 0.036 mm for time and 0.018 mm for error). We found significant differences in performance between participants from the crowd (informed and uninformed) and the lab, with lab participants performing best. However, when analysing only participants with above average performance of their respective population, crowd participants performed just as well lab participants. Therefore, using crowdsourcing for HCI user studies should be a feasible solution to get both more participants and acting in their real environment resulting in higher external validity. Informed crowd participants had significantly better performance in the tasks compared to their naive counterparts - another good reason to openly disclose the scientific purpose.

Games or tasks used in crowd studies need to be very easy to understand and engaging. We found high drop-out rates (people ending the game before having lost all their lives) in the crowd compared to the lab, as the participants did not feel as obliged to complete the game. Our response rate in terms of people downloading the game was much lower than what Henze et al. achieved five years earlier. We assume this to be due to a more highly saturated market place for games and entertainment on mobile devices.

References

1. Benko, H., Wilson, A.D., Baudisch, P.: precise selection techniques for multi-touch screens. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 1263–1272 (2006)
2. Bérard, F., Wang, G., Cooperstock, J.R.: On the limits of the human motor control precision: the search for a device's human resolution. In: Campos, P., Graham, N., Jorge, J., Nunes, N., Palanque, P., Winckler, M. (eds.) INTERACT 2011. LNCS, vol. 6947, pp. 107–122. Springer, Heidelberg (2011). doi:10.1007/978-3-642-23771-3_10
3. Cockburn, A., Ahlström, D., Gutwin, C.: Understanding performance in touch selections: tap, drag and radial pointing drag with finger, stylus and mouse. Int. J. Hum.-Comput. Stud. **70**(3), 218–233 (2012)
4. Fitts, P.M.: The information capacity of the human motor system in controlling the amplitude of movement. J. Exp. Psychol. **47**, 381–391 (1954)
5. Henrich, J., Heine, S.J., Norenzayan, A.: The weirdest people in the world? Behav. Brain Sci. **33**, 61–83 (2010)
6. Henze, N., Boll, S.: It does not fitts my data! Analysing large amounts of mobile touch data. In: Campos, P., Graham, N., Jorge, J., Nunes, N., Palanque, P., Winckler, M. (eds.) INTERACT 2011. LNCS, vol. 6949, pp. 564–567. Springer, Heidelberg (2011). doi:10.1007/978-3-642-23768-3_83
7. Henze, N., Rukzio, E., Boll, S.: 100,000,000 taps: analysis and improvement of touch performance in the large. In: Proceedings of the 13th International Conference on Human Computer Interaction with Mobile Devices and Services, pp. 133–142 (2011)
8. Holz, C., Baudisch, P.: The generalized perceived input point model and how to double touch accuracy by extracting fingerprints. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 581–590 (2010)
9. MacKenzie, I.S.: Fitts' law as a research and design tool in human-computer interaction. In: Human-Computer Interaction, pp. 91–139 (1992)