Performance Evaluation of Crowdsourced HCI User Studies

Allan Christensen
Aalborg University
Denmark
akch10@student.aau.dk

Simon André Pedersen
Aalborg University
Denmark
sape09@student.aau.dk

ABSTRACT

In this study, we examined the viability for HCI user studies to use crowdsourcing as a participant group. Potentially it could yield higher attendance for the studies and studies would not rely on subjects classified as WEIRD. We conducted a preliminary study to determine if touch or tilt controlled a game better within a lab environment. We found that touch outperformed tilt. Following this study, we examined if an informed crowd (informed about being in an experiment) and uninformed crowd could perform equivalent to participants in a controlled lab environment. The study showed that in our first level, a touch controlled game, the lab environment outperformed both of the crowds, while the informed crowd performed better than the uninformed. The second level featured a device human resolution experiment through Fitts' law, to determine the smallest selectable target with little effort. The data revealed that the lab consistently produced fewer errors and we saw a significant increase in errors between a Fitts' ID of 3.70 and 4.64. For the informed crowd we saw a spike in errors for a Fitts' ID between 2.81 and 3.70. The uninformed crowd had generally too many errors to determine a significant increase in errors. The smallest selectable target for all three groups combined, was between 2 mm and 4 mm for touch devices.

Author Keywords

Crowdsourcing; gamification; device human resolution; touch-interaction; human computer interaction; user studies;

ACM Classification Keywords

H.5.2 Information Interfaces and Presentation: User Interfaces

1. INTRODUCTION

Human-Computer Interaction (HCI) studies are conducted to increase usability of products, and the effectiveness of HCI. These studies typically focuses on analyzing user behavior when interacting with computers, through question sheets, observations, and experiments on user groups of various sizes, and demographics.

Due to the difficult nature of measuring human behavior, these methods of analysis often lead to incomplete, partially biased, or wrong results. One of the issues often seen with the analyses is the controlled testing environment, where the users know that their results matter to the people conducting the tests. This can cause an increase in e.g. user performance, and cause response bias, as users might give the technique a higher rating. Furthermore, the testing setups can influence the data for the technique, as these testing environments do not suffer from e.g. noise in the real environment. Thus, raising the question if the ecological validity of the data is actually comparable to the environment that the product subsequently will be used in.

Another issue related to HCI experiments is the demography of users. Henrich et al. [11] brand the participants from social science studies as WEIRD, which is an acronym for Western, Educated, Industrialized, Rich, Democratic. This category of people is the most used in HCI experiments and studies, and shows that 96% of test participants in behavioral science research come from western industrialized countries [1], with a majority of undergraduates; 67% of American samples and 80% in other countries. Henrich et al. also state that the WEIRD subjects often occupies the extreme ends of the behavioral science distribution and will therefore often provide very different results than other subject groups. Furthermore, Gächter [9] points out that some of the issues are not only related to the WEIRD subjects, but rather the way that the subjects are used. Currently, the subjects are used for all kinds of tests and experiments. Therefore, Gächter suggests that WEIRD subjects instead could be used as a benchmark or starting point for investigating generalizability to other social groups. Thus concluding that the WEIRD subjects could be used as a baseline, and then later expand the experiments to other subject groups, if it is relevant. A method of addressing these issues is to utilize crowdsourcing of HCI experiments.

This paper investigates the effects of crowdsourcing of existing HCI user studies compared to an experimental setup using WEIRD subjects. The goal is to investigate if crowdsourcing can be used to generalize results for HCI user studies and if information regarding the experimental aspect of an application has any implications on the outcome. First, we examine related work on crowdsourcing, Fitts' law and Device Human Resolution, and user's touch interaction. Further, the effectiveness of crowdsourcing is analyzed by comparing crowdsourcing of applications versus testing applications in a closed experimental environment. The primary goal of gamifying existing HCI tests is to eliminate user bias of the experimental part of the application by hiding the fact that the users are participating in an experiment in which their achievable score, and performance matters to the creators. To examine the method of gamification of tests, the results of crowdsourced participants who are aware that they are part of an experiment versus participants who are unaware of this, are separated into two groups.

2. RELATED WORK

2.1 Crowdsourcing

In this study, we define crowdsourcing as: "recruitment of nonpaid global workforces that are voluntarily working on a specifically defined task or set of tasks". The key features in this definition are (1) nonpaid worker, (2) works voluntarily and (3) is completing a task or set of tasks. This means that there is no increase in motivation to volunteer in the study because of payment, as often seen in traditional research studies. The motivation can instead be increased if the participants find the game interesting and immersing. However, using crowdsourcing as the main data distributor raises several concerns that need investigation to ensure valid and comparable data.

2.1.1 Population size and Market Analysis

Empirical research in laboratories (lab) use between 14 and 30 participants [10, 16]. Participants are often recruited from local universities and classified as WEIRD [11].

When examining the global smartphone market, at least 1.3 billion people currently have a smartphone [26]. In the Google Play Store, 1.4 million apps are available [25]. Releasing experiments as an app to the Google Play Store will potentially increase the amount of participants in experiments. For example, Henze et al. have 108.000 installations of their app, over 72 days from release [14].

However, this also raises some concerns, such as, who participates in the experiments. When recruiting participants, researchers can validate that participants fulfill the requirements of the study, e.g. reject a participant for being too young, in order to fit a certain target group. For a crowdsourcing experiment, it is not possible in the same way to exclude participants, as the researcher will never interact directly with the participants [19]. Henze et al., did not include demographic questions within their app; therefore it was impossible to exclude participants because of e.g. being underage [14]. Henze et al. collected data from 99.749 installations, but had to remove $\sim$10% of the data as it was deemed insufficient or meaningless, due to multiple installations that shared the same identifier [14].

2.1.2 Validity

When conducting experimental research using crowdsourcing, it raises questions about the validity of the experiment. This is mostly concerned with the reliability of the data.

2.1.2.1 Population Validity

Population validity describes if the sample population (sample), resembles the entire population. Bracht and Glass [5] describes it as the accessible population versus the target population (target). The accessible population is the population the researcher has available for the study, while the target is the total group of subjects the researcher tries fit his research to and use the accessible population to resemble the target. An example is the WEIRD subject population which is a sample that often does not resemble the target very well. Henze et al. has a very large sample of smartphone users for their study [14]. The population validity is therefore high, because their sample resembles a large part of the target, which is all smartphone users.

2.1.2.2 Ecological Validity

Ecological validity describes the testing environment surrounding the technology. The testing environment shall resemble the real world environment. If the usage of the technology resembles a real world scenario, the ecological validity is high [23]. Again examining the experiment by Henze et al., we see that they tested touch points on mobile devices. The app is only released on touch devices and the devices are tested in people's everyday environment. Therefore, they have a high ecological validity as well [5, 14]. However, they also experience huge variances in the data and implausible results for the Fitts' law part of their experiment [12].

2.1.3 User Consent

When conducting experiments in a lab, the researcher will most often brief the participants about the experiment before starting, and participants usually fill out a consent form, which allows the researcher to use the data for later analysis [19]. When working with crowdsourcing there are no definitive guidelines or rules about informing users about data collection [13]. Henze et al. examine five different ways to inform users. They state that the two methods that provides the highest amount of data is: not asking users (83.68%) and showing a consent sheet with no option to opt-out (81.31%) [13]. Due to concerns of having tested the above on different apps they followed-up with examining four ways for users to accept sending anonymous data [21]. They state that a single 'okay' button with no opt-out feature provides the highest acceptance level (87.6%) from the participants, however, this raises concerns about being unethical. Therefore, they recommend using a 'yes / no' button (67.4%), which forces participants to decide if they want to participate in the study [13, 21].

2.2 Target Acquisition

In this study we mainly focus on the differences between participant groups in existing HCI experiments. As touch interfaces in recent years have become a more common technology, there is a need to know how precise users are able to select targets. The following section investigates Fitts' Law, Device Human Resolution (DHR), and the differences between pointing and dragging on touch.

2.2.1 Fitts' Law

Fitts' law [8] predicts the time required for a human to perform a movement from point 'A' to point 'B'. In order for Fitts to predict the time requirements, he introduced an Index of Difficulty (ID) that describes how difficult it is to hit a target. The formula is a function between the travel distance and the target size. A higher Fitts' ID results in a harder task for humans to perform and therefore a larger time requirement.

MacKenzie [17] further extended the calculation of Fitts' ID to:

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

Several studies have examined a variety of devices for their performance using Fitts' law. A study by Cockburn et al. compare finger input, a stylus, and a mouse in target acquisition tasks. The results shows a higher error rate (14%) for a width of 5 mm for finger dragging compared to the other devices. However, they only examine widths of 5, 12.5, and 20 mm [6]. Furthermore, Sasangohar et al. experience high error rates (9.8%) for touch input, with 5 mm targets, compared to a mouse (2.1%) [22]. For smaller target acquisition tasks Sears and Shneiderman [24] state that using a stabilized touch screen for targets of a single pixel, users have a significantly higher error rate (64%) compared to the mouse (20%).

2.2.2 Device Human Resolution

Bérard et al. examine the human performance on several devices using a theory they call Device Human Resolution (DHR) [3]. This theory uses Fitts' law to determine the smallest target that is achievable by humans with little effort. Standard Fitts' law experiments examine the overall performance of a human using a specific input device, while DHR examines the smallest possible target for the human to select without a decrease in performance. Bérard et al. examines three different input devices. Their results for mouse shows that a participant's performance decreases below 0.036 mm targets; they are able to maintain a low error rate at this size but at the expense of time. This means the DHR for a mouse is 0.036 mm (for time) and below 0.018 mm (for error). They further test a free space device and a stylus, which results in a DHR of 2.4 mm and 0.23 mm, respectively [3]. Furthermore, Bjerre et al. [4] investigate the DHR for in-air interactions and finds the DHR to be between 1.2 and 2.4 mm, when using a Leap Motion.

2.2.3 Pointing and Dragging

The experiments mentioned in the previous sections all focus on pointing tasks. MacKenzie et al. examines the performance differences between pointing and dragging for Fitts' law tasks [18]. A pointing acquisition task in Fitts' law includes a movement and a click on a marked target as fast and precise as possible. When clicking the target the participant has to point and click a new target. Dragging tasks are similar, but instead the participant clicks and then drags the cursor to another target location.

Cockburn et al., examines an offset cursor in a tapping task versus a dragging task using a finger, a mouse, and a stylus as input methods. They state that there are no difference in the mean selection time between the mouse and stylus. However, the finger has a significantly higher overall selection time when dragging (∼922 ms) compared to tapping (∼572 ms) mainly explained by the higher friction when dragging across a screen. They further experience that tapping (6.8% errors) with the finger has a significantly lower accuracy compared to dragging (1% errors). The target sizes used in the experiment are 5, 12.5 and 20 mm. Two reasons explained this: (1) there was no system feedback on the location of the finger prior to completing selections by tapping the screen [6] and (2) the 'fat finger' problem, meant that using a finger was a large and relatively crude pointing method for small targets [6].

2.3 Finger Size

When designing touch user interfaces, we had to consider button sizes for the users to click with their fingers. Dandekar et al. examines tactile sensing with the index finger of adult users and find that the average width of the index finger is between 16 - 20 mm [7]. Several studies have experienced that users tends to have 'fat fingers' [2, 6, 15]. Holz et al. mentions two common known reasons for inaccurate target selections. (1) Users do not know the exact interaction point of their finger, meaning the position of the skin that is initially in contact with the screen. (2) At the same time, the finger will occlude the target the user is trying to select when targets get too small [15]. Holz et al. extends (1) by describing that users also perceive the interaction point differently. Various design solutions tries to overcome the fat finger problem, e.g. offsetting the cursor or zooming [2]. However, one question that still needs answer is what the smallest achievable target size, which humans are actually able to select, on touch devices is.

2.4 Summary

The benefit associated with the use of crowdsourcing is that it is possible to get a large amount of participants, which yield high external validity. However, some of the drawbacks are that it is hard to control the user's environments compared to a lab environment. There are no clear guidelines to ask about consent, however, letting the users know without the possibility to opt-out gives the highest amount of data. Furthermore, exclusion of data might be necessary, because post validation of participants is nearly impossible when using crowdsourcing.

We will draw upon the concept of gamification to create an experiment where we examine the DHR for touch applications through Fitts' law. The expectations from the related work assumes that users will experience occlusion problems with targets below 5 mm [6, 22]. Furthermore, we found that using dragging instead of tapping when interacting with touch yields a higher selection time, but fewer errors.

3. METHODS AND MATERIALS

In this section, we explored the differences between crowdsourcing and lab testing through two user studies. The first was a preliminary study, where we examined user preference and performance for two different control schemes. Furthermore, the study also explored the overall technical aspect of the application (app) and how it could be improved in terms of user interaction. The findings from the preliminary study was used to inform the main user study about the highest performance control scheme, so that only one of them had to be used in the main user study, as well as technical improvements to the app, before it could be released to the crowd and tested in the lab.

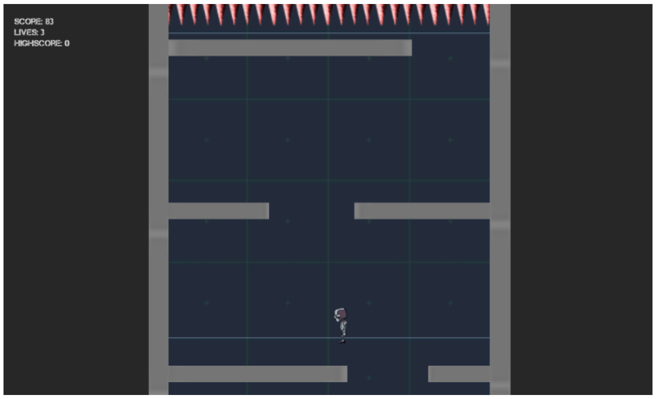


Figure 1. A screen-shot of the first level Drop as seen by participants in the game. The goal of the game was to survive for as long as possible by moving the character to the holes with the bottom of the map moving upwards at an increasing pace.

The second was the main user study. As described in the introduction in Section 1, the overall goal for the project, and therefore the main user study was to determine if there were significant differences in performance between crowdsourcing an app or testing it in a lab environment on WEIRD subjects. Furthermore, we wanted to investigate if there were performance differences between informing participants that they are participating in an experiment and not giving them any information about the testing aspect of the game. Lastly, we wanted to investigate the DHR for touch applications in order to make comparisons to other devices.

3.1 Game Design

Before moving into the two user studies we want to shortly present our design of the app. The app was a small game that consisted of two levels, where each level represents a different HCI user study. The first level was called Drop, and was based on an existing Google Play app with the same name [20]. The goal was to control a character from side to side to get through the map as can be seen in Figure 1. The two control schemes consisted of touch and tilt, where touch controlled the character on screen by pressing on either side of the screen, which made him move towards that direction. The other control scheme was tilt, where the user had to tilt the touch device left or right and the accelerometer in the device would detect the tilt and the character would move in that direction.

The second level was based on the DHR test by Bérard et at. [3], which we called Wall Destroyer. The original test functioned as a Fitts' law experiment to find the resolution for different devices that humans were capable of interacting with. In the test the user had to click within a starting area and then click seven targets of different sizes always presented in descending order, where an example of the interface used can be seen in Figure 2. Information about the seven targets can be seen in Table 1.

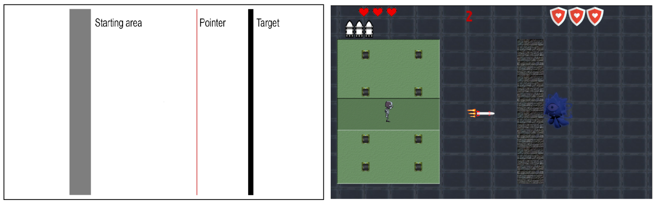


Figure 2. Screen-shot of the interface used in the original DHR experiment by Bérard et al. [3](left) and our gamified version (right).

Task	1	2	3	4	5	6	7
Width (mm)	32	24	16	8	4	2	1
Fitts' ID	1.32	1.58	2	2.81	3.70	4.64	5.61

Table 1. Table showing the task number, width, and Fitts' ID of the seven tasks used in both user studies. For all the tasks a distance of 48 mm between the starting point and the center of the target was used.

Our level was a gamified version of the test where we added game elements to the different tasks, as well as a score that made the user continue to play the game. An example of the interface presented to the users in the game can be seen in Figure 2. Furthermore, in the original experiment all participants completed 20 repetitions of the seven target sizes. In order to force repetitions we gave each participant five lives, and a life was only lost if they missed a target three times in a row. This did not provide a constant amount of repetitions, but forced most participants to complete the game multiple times.

3.2 Preliminary Study

After creating the app, we conducted a preliminary study that focused on the control methods for the first level Drop in order to examine which control method would yield the best performance and had the highest preference for the users. Second, we explored the interaction for the second level Wall Destroyer, where we examined if there were any problems associated with the way the level had been implemented.

3.2.1 Participants and Apparatus

Six voluntary participants were recruited from the university campus. The participants were all male students in the age range of 24 to 27 years old (M = 24.83, SD = 1.17). All participants reported to be daily users of touch devices.

The test was conducted on a LG Nexus 4 smartphone, with a 4.7 inch display and a 768x1280 resolution. Participants were allowed to hold the phone as they pleased to achieve the most comfortable interaction pose.

3.2.2 Procedure

Before the test, we gave each participant a short demographic questionnaire (age, profession, touch interaction, and gaming habits) followed by a short introduction to the test. After the introduction the participants were presented with the first level of the app, namely Drop. The order of the controls for Drop were counterbalanced, meaning that three of the participants would start with tilt, and the other three would start with touch. All participants understood that the goal was to stay alive for as long as possible. When a participant had completed both controls for Drop, they moved on to the second level, Wall Destroyer, where the goal was to destroy the seven targets as many times as possible. After completion of the DHR level, we conducted a short semi-structured interview in order to further explore the participants' opinions on the game related parts of the application.

For Drop, we measured the survival time for both control schemes. For Wall Destroyer, we measured the time it took to destroy each wall, as well as the amount of errors that users made for each wall, during the level. The entirety of the test took approximately 15 minutes.

3.2.3 Expectations

For Drop, we expected the survival time to be similar for both control schemes, however, we expected touch to outperform tilt, as users in general had more experience using touch than tilt. Furthermore, the control scheme for touch enabled usage in every setting, whereas tilt needed more special conditions to function optimally. An example of this could be playing with the phone lying on a table. You could still play the touch version, however, in order to play the tilt version you had to hold the phone in free space, thereby limiting the control scheme. Because of this, we also expected that the users would prefer the touch control over the tilt.

For Wall Destroyer, we expected the results to be too good to be comparable to the original DHR experiment. The reason for this was that users were able to use two hands while interacting with the game, which made Fitts' law obsolete, similar to the results found by Henze and Boll [12]. This meant that the results from the DHR part of the test were of less importance, and the part was mainly included in the test to reveal improvements to the level before we released the app.

3.2.4 Results

In this section we present the results from the preliminary study. We have split the results into two sections covering each level individually.

3.2.4.1 Level 1 – Drop

For the first level Drop, we investigated the survival time for each of the control schemes for the six participants. This meant the higher the time the better. Each participant had three lives, which meant that they had to repeat the game six times in order to move onto the second level.

Through a dependent t-test we found that the control scheme significantly affected the survival time ($t(17) = 3.10$, $p < 0.01$). Examining the mean values and the distribution of the data as seen in Figure 3, we found that touch (*M* = 126, *SD* = 11.55) had a significantly higher survival time than tilt (*M* = 118.5, *SD* = 6.37).

The results from the semi-structured interview, conducted after the test was completed showed that, consistent with the results of the data, 5 out of 6 participants (83%) preferred to use touch over tilt when controlling the character during the game.

3.2.4.2 Level 2 – Wall Destroyer

For the second level which focused on the DHR of the participants, we investigated the time between clicking within the starting area and hitting the different target sizes. Furthermore, we examined the error rate when users tried to hit the targets. Users could make three errors before they were advanced onto the next target, which was equivalent to the original experiment. As mentioned in the expectations in Section 3.2.3, the data could not be compared to the original test as there was no actual travel distance because the users used two hands (one for the starting area and one for the target). This was especially the case for the time data.

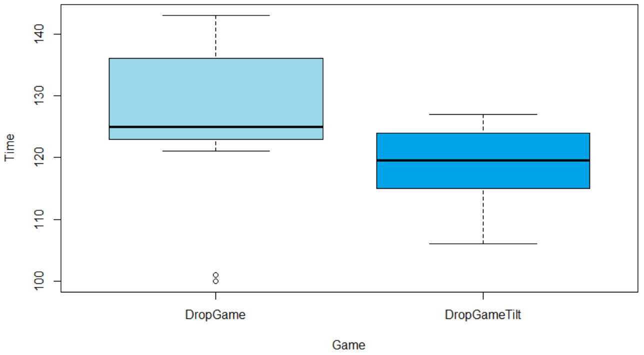


Figure 3. Image showing the distribution of survival time for the first level Drop, with Game type (control scheme) on the x-axis and the time on the y-axis.

Using an ANOVA test on the time data we found that there was an overall significant difference between the tasks ($F_{6,953} = 17.01$, $p \ll 0.001$). Using a TukeyHSD post-hoc test we found that there was a significant difference between an ID of 5.61 and the rest, except for an ID of 1.32. Furthermore the time for ID of 1.32 was also significantly different from the rest, except for an ID of 4.64. Examining the mean slope for the data we found that no slope deviated significantly from the overall slope (0.06).

Using a Friedman Ranked Sum test on the error data we found that there was an overall significant difference between the tasks in terms of errors ($\chi^2(6) = 30.96$, $p \ll 0.001$). The post-hoc of the Friedman test is shown in Table 2.

Task	Task 7	Task 6	Task 5	Task 1	Task 4	Task 2	Task 3
Fitts' ID	5.61	4.64	3.70	1.32	2.81	1.58	2
Rank	A	B	C	CD	CD	D	D

Table 2. The results from the Friedman post-hoc test for the error data in the preliminary study. The tasks were significantly different if they were assigned a new letter.

The error percentages for each task can be seen in Figure 4. Examining the graph we found that a spike in errors seemed to occur between an ID of 3.70 and 4.64. This was consistent with the data from the Friedman test. Furthermore, there was another spike in errors between an ID of 4.64 and 5.61, which again was consistent with the results above.

Observations during the test showed that all users used two hand/finger interaction during the DHR part of the test, which again showed that Fitts' law was not applicable for the preliminary study.

3.2.5 Partial Conclusion

The preliminary study showed various things that had to be improved before the main user study could be conducted.

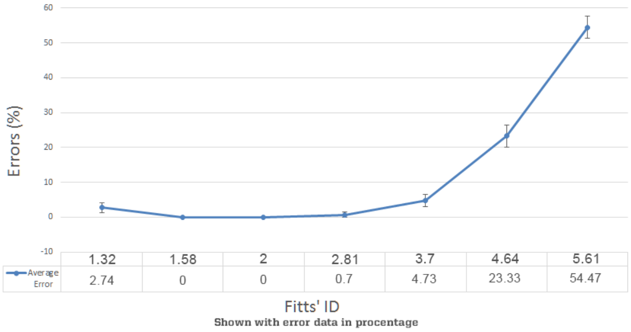


Figure 4. Image showing the distribution of error percentages for the second level Wall Destroyer in the preliminary study, with Fitts' ID (Task) on the x-axis and the error percentages on the y-axis. The error data was not precise, as all participants used two hands for the interaction. The bars show the standard errors for each task.

First, for the main user study we have implemented a drag for the DHR part of the experiment. This ensures that users are only using one hand to interact with the device which makes Fitts' law applicable; therefore, the time results will be more representative in the main user study. However, as described in Section 2.2.3 this also creates some issues, as there are differences in performance between pointing and dragging on touch. In order to examine if this has an influence on the results, we will compare the error results from DHR between the preliminary study and the main user study.

The errors for DHR showed that users were able to reach a low amount of errors up until an ID of 3.70. The errors also showed that for an ID of 5.61, users had a very high error percentage which was consistent with the findings by Sears and Shneiderman [24].

Furthermore, the results from Drop showed that touch had a significantly higher performance than tilt, as well as preferred by most users, and therefore only touch was implemented for the main user study.

3.3 Main User Study

The purpose of this study was to analyze and compare results from three different user groups. The three groups would go through two levels. These two levels were the same as the ones presented in the preliminary study. The first was touch control from the Android game Drop, and the second was the DHR experiment to test the participants' performance using a touch device.

3.3.1 Participants and Apparatus

This study was divided into three groups as mentioned above. The first group resembled a standard lab environment. We recruited 16 participants from the local university who could be categorized as being part of the WEIRD population. All participants were male subjects and had an age range of 22 - 26 years old (M = 24, SD = 1.5).

The second group consisted of an uninformed crowd (crowd), which meant that they were not presented with consent and informed that they were participating in a study. The crowd consisted of 19 participants. We did not ask this crowd for any demographic questions, and therefore we did not know age, environment, or touch device usage for this group.

The third group resembled the informed crowd and was called 'crowd-plus'. This crowd knew that they were participating in an experiment. The group consisted of 14 participants and after they played the game, they were asked to fill out a questionnaire, which 12 of the participant decided to answer (85.71%). It revealed an age range of 22 to 57 years old (M = 28.16, SD = 9.52) and four of these were female.

For the lab environment, the participants used a LG Nexus 4 smartphone running Android 5.1, with a 4.7-inch display and a 768x1280 resolution. We ensured comparability with the two crowds by letting the participants hold the device as they pleased.

3.3.2 Procedure

The lab participants went through the same procedure as in the preliminary study, described in Section 3.2.2. First, they filled a short demographics questionnaire and received a short introduction. It was important that they did not receive additional information about the actual test compared to the crowd, as it could influence the overall experiment. After the introduction, the participants received the smartphone and was prompted to start the game and watch the introductory video. For each of the levels an approximately 30 second video showed the participant how to complete the level. After the video, the Drop level started. When the participants had completed the Drop level, the introductory video for Wall Destroyer started and the level started instantaneously after the video. After the second level was over, the participants were asked to fill out a questionnaire.

Both crowd groups had to download the app from the Google Play Store. When they started the game, crowd-plus would be presented with a consent page. At this point we assumed that they understood that they were participating in an experiment, as described in Section 2.1.3. When pressing 'okay' they were redirected to the main menu. From this point on crowd, crowd-plus and lab were identical, and followed the same procedure as the lab participants. After crowd completed the game, we presented them with their own high score. When crowd-plus completed the game, we showed a pop-up message prompting them to answer a questionnaire. If they touched the screen they were redirected to the questionnaire.

3.3.3 Expectations

The overall goal for this experiment was to examine if there was a significant performance difference between a controlled environment and an uncontrolled environment. For the uncontrolled environment, we further examined if there was a difference between users knowing and not knowing about their participation in an experiment.

For the Drop level we expected that the users from the lab environment had a significantly higher survival time compared to both of the crowds. This was mainly explained by the hypothesis which stated that data from a controlled lab environment would outperform public users' environment. However, we expected that crowd-plus participants would survive

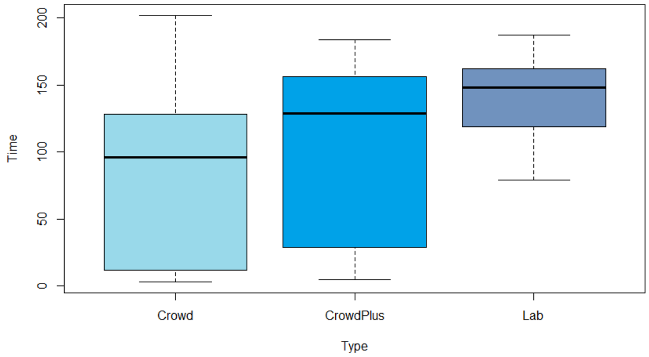


Figure 5. Image showing a box-plot for the first level Drop, with group type shown on the x-axis and the survival time on the y-axis.

longer compared to the crowd group, because we expected them to try harder because of the awareness of being test participants.

For the Wall Destroyer level, the overall expectation was similar to that of Drop. The lab participants would perform better in both time and errors compared to both crowds, while crowd-plus would perform better than crowd. This should reveal the best DHR for lab, then crowd-plus, and worst for crowd. However, we did expect the three DHRs to be similar and the overall DHR (the three participant groups combined) to be comparable to the original study by Bérard et al. [3].

3.3.4 Results

For the results of the main user study we used a significance level of $\alpha = 0.05$. If nothing else is noted, we used a one-way Analysis of Variance test (ANOVA) with a TukeyHSD as post-hoc to analyze the results. Furthermore, the results are divided into individual sections for the two levels.

3.3.4.1 Drop Level

In this experiment the participants had to stay alive for as long as possible, which meant that a high survival time indicated a better performance. This study was mainly used to validate if the results were comparable between all participant groups. For the three groups we found the survival time to be significantly affected by the groups ($F_{2,286} = 23.48$, $p \ll 0.001$, $\eta^2 = 0.14$). A TukeyHSD post-hoc test showed that lab was significantly better than both crowds, and crowd-plus was significantly better than the crowd. When we examined the results, we discovered that the data revealed a high amount of data entries for both crowds, which had very low survival times compared to the lab. A box-plot illustrating the distribution of the time data can be seen in Figure 5.

Due to the high amount of low survival times for both crowds, we examined the data by only including the highest survival time for each participant. The data showed that there was no overall significant difference between the groups ($F_{2,38} = 0.205$, $p = 0.82$, $\eta^2 = 0.01$).

3.3.4.2 Wall Destroyer Level

As mentioned above, we divided the test results into the three groups to compare them individually. After the results for the individual groups, we present the overall results with all groups combined.

Lab

We examined the completion time between the seven tasks and found an overall significant difference ($F_{6,1767} = 156.16$, $p \ll 0.001$, $\eta^2 = 0.35$). Analyzing the mean slopes for the time data revealed that no subset slope significantly deviated from the overall slope (0.09).

For the error data a Friedman Ranked Sum test revealed an overall significant difference between the seven tasks ($\chi^2(6) = 63.98$, $p \ll 0.001$). The error distribution of the data can be seen in Figure 6.

Examining the post-hoc of the Friedman test we found that there was a significant increase in errors between an ID of 3.70 and 4.64. There was also a significant increase between an ID of 4.64 and 5.61.

An overview of the data for the lab group can be seen in Table 3. This table includes both time, error-rate and mean slopes for each task with the corresponding Fitts' ID.

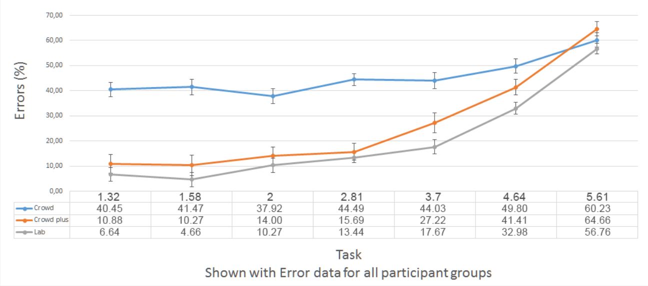
Fitts' ID	Mean Time (sec)	Failure Rate (percent)	Mean Slope
1.32	0.55	6.64	
1.58	0.40	4.66	0.10
2	0.44	10.27	0.11
2.81	0.55	13.44	0.12
3.70	0.74	17.67	0.09
4.64	0.83	32.98	-0.13
5.61	0.99	56.76	

Table 3. Table showing the data for lab. There was no mean slope value for the first and last task / ID because the calculation of the mean slope calculates the mean between the previous and following target sizes.

Crowd

For crowd, we found that there was an overall significant difference for the completion time between the seven tasks ($F_{6,962} = 3.58$, $p < 0.01$, $\eta^2 = 0.03$). For the slope data, none of the subset slopes significantly deviated from the overall slope (-0.00004). Therefore, we again examined the error data for the crowd participant group. The error distribution for crowd can be seen in Figure 6.

A Friedman Ranked Sum test showed no overall significant difference in errors between the tasks ($\chi^2(6) = 10.80$, $p =$



	1.32	1.58	2	2.81	3.7	4.64	5.61
Crowd	40.45	41.47	37.92	44.49	44.03	49.80	60.23
Crowd plus	10.88	10.27	14.00	15.69	27.22	41.41	64.66
Lab	6.64	4.66	10.27	13.44	17.67	32.98	56.76

Figure 6. The distribution of data for the three groups in the main user study. The seven tasks are shown on the x-axis with the errors in percent on the y-axis. Each line represents a participant group and the bars show the standard errors for each task.

0.09). Examining the post-hoc we found a spike between an ID of 2 and 2.81, however, it was not enough of an increase to be significant.

An overview of the data for the crowd group can be seen in Table 4. This table includes time, error-rate and mean slopes for each task with the corresponding Fitts' ID.

Fitts' ID	Mean Time (sec)	Failure Rate (percent)	Mean Slope
1.32	0.61	40.45	
1.58	0.44	41.47	0.03
2	0.44	37.92	0.04
2.81	0.45	44.49	0.02
3.70	0.47	44.03	0.01
4.64	0.52	49.80	-0.24
5.61	0.52	60.23	

Table 4. Table showing the data for crowd. There was no mean slope value for the first and last task / ID because the calculation of the mean slope calculates the mean between the previous and following target sizes.

Crowd-plus

The time data for the crowd-plus showed an overall significant difference between the tasks ($F_{6,840} = 27.46$, $p \ll 0.001$, $\eta^2 = 0.16$). Examining the mean slopes for the time data again revealed that no subset slope was significantly different from the overall slope (0.09).

For the crowd-plus error data, a Friedman Ranked Sum test revealed an overall significant difference between the tasks ($\chi^2(6) = 23.47$, $p < 0.01$). The post-hoc of the Friedman test revealed that there was a significant increase in errors between an ID of 2.81 and 3.70. The error distribution for the crowd-plus participant group can be seen in Figure 6.

An overview of the data for the crowd-plus group can be seen in Table 5. This table includes time, error-rate and mean slopes for each task with the corresponding Fitts' ID.

Fitts' ID	Mean Time (sec)	Failure Rate (percent)	Mean Slope
1.32	0.62	10.89	
1.58	0.48	10.27	0.13
2	0.54	14.00	0.11
2.81	0.62	15.69	0.10
3.70	0.71	27.22	0.11
4.64	0.82	41.41	-0.09
5.61	0.95	64.66	

Table 5. Table showing the data for crowd-plus. There was no mean slope value for the first and last task / ID because the calculation of the mean slope calculates the mean between the previous and following target sizes.

All data

In the following section we describe the overall comparisons between the three participant groups. However, examining the results, we found some tendencies towards the data being skewered for the DHR level. The data showed that several participants in the crowd performed very poorly during the experiment. Therefore, we examined the amount of repetitions, as they varied a lot between participants. The participants who did not complete a high amount of repetitions, were considered to have performed worse at the tasks than participants with a high amount of repetitions. Therefore, we tried to remove all participants below the average amount of repetitions for each of the participant groups to examine if this change would provide better and more comparable results. When only including the data from participants performing above the average amount of repetitions, we ensured that participants had understood the exercise task. For crowd, the average amount of repetitions was 7.68, which meant that only 7 out of 19 participants remained. For crowd-plus the average amount of repetitions was 9.43, and therefore only 6 out of 14 participants remained. Lastly, for lab the average amount of repetitions was 16.56, which meant that only 8 out of 16 participants remained.

Group Comparisons

For the time data on the remaining dataset, we found an overall significant difference between the participant groups ($F_{2,2856} = 92.21$, $p \ll 0.001$, $\eta^2 = 0.06$). Using a TukeyHSD post-hoc test, we found a significant difference between all groups. Examining the mean slopes we saw that no subset slope deviated significantly from the overall slope (0.09) for the time data.

The error data showed that there was no overall significant difference between the participant groups ($F_{2,144} = 0.79$, $p = 0.45$, $\eta^2 = 0.01$). Examining the TukeyHSD post-hoc we found that crowd was not significantly different from crowd-plus and lab. Lab was also not significantly different from crowd-plus.

Overall DHR for Touch

For the dataset where the participants below the average amount of repetitions was removed, we found the overall DHR for touch. Examining the error data, we found an overall significant difference between the seven tasks ($F_{6,140} = 26.29$, $p \ll 0.001$, $\eta^2 = 0.53$). A TukeyHSD post-hoc test showed that there was 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 hap-

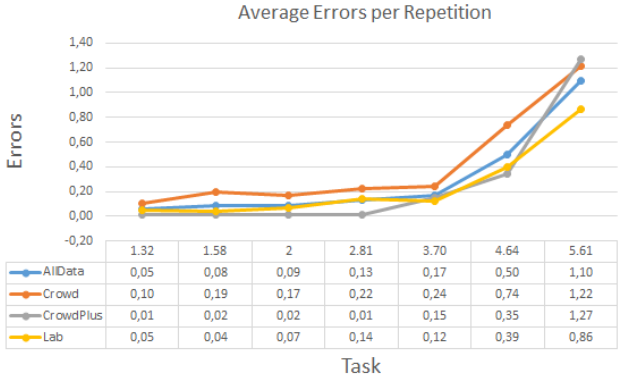


Figure 7. Graph illustrating the mean error rate for each task per repetition for the three participant groups as well as the overall data. The x-axis shows the Fitts' IDs and the y-axis shows the amount of errors.

pened between an ID of 3.70 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 Figure 7, where the data for all participant groups have been included.

An overview of the gathered data for all participant groups can be seen in Table 6. This table includes time, error-rate and mean slopes for each task with the corresponding Fitts' ID.

Fitts' ID	Mean Time (sec)	Failure Rate (percent)	Mean Slope
1.32	0.53	11.88	
1.58	0.41	10.85	0.10
2	0.44	9.70	0.12
2.81	0.51	10.78	0.12
3.70	0.61	10.21	0.09
4.64	0.74	27.35	-0.12
5.61	0.84	52.08	

Table 6. Table showing the data for the gathered participant data. There is no mean slope value for the first and last task / ID because the calculation of the mean slope calculates the mean between the previous and following target sizes.

4. DISCUSSION

The results from the main user study confirm our initial expectations that the lab setup will outperform the crowd setups for touch tasks. An issue when using crowdsourcing is that we cannot with guarantee state if the differences in performance is due to the participants, the environment that they are playing in, or a combination of both. However, through the conducted studies we have found that both of these factors can influence the touch performance, and that WEIRD subjects (and specifically in this case the western university students) in a controlled lab environment, with no environmental disturbances, will in most cases provide higher performance results than the subjects playing in their own environment. A high amount of the results gathered from the crowd participants can, however, skewer the data in a negative direction.

First, we found that for the Drop level, the lab participants performed significantly better than both crowds ($p \ll 0.001$). However, as we examined the data more closely we found that the data for both crowd groups was skewered in a negative direction. Therefore, by examining only the highest survival time for each participant in the Drop level (they provided at least three due to three repetitions of the level) we found that there was no significant difference between the crowd-plus and lab environment. This shows that even for very simple touch assignments users will provide a better overall performance in a lab environment, however, the users that know they are participating in a study (crowd-plus group) are able to provide performance results that are significantly equivalent to the lab environment.

For the Wall Destroyer level (DHR), we found that multiple participants did either not understand the tasks, did not want to complete the tasks, and/or were in general a lot worse compared to other participants. This was especially the case for the two crowd environments, where we saw that the difference between the highest and lowest performing participants was much higher than in the lab environment. Furthermore, the repetition data for the three participant groups showed that the lab environment on average completed the level almost twice as many times as the crowdsourced groups. Following the example of Henze et al. [14], we removed all the data that was deemed insufficient (meaning all participants below the average amount of repetitions). The new data showed that there was no overall significant difference between the three groups. This shows similar results to those found in the Drop level, namely that when all data is included, the lab participants will have a higher performance than the crowdsourced participants, however, when only including the highest survival time for each participant we see that the crowds are able to perform at the same level as lab. This shows that the results from the lab are more consistent than the crowds, but that all participant groups can perform at the same level.

While measuring the performance differences between groups, we also examined the DHR for touch applications. The results showed that a significant increase in errors happens around a Fitts' ID between 3.70 and 4.64, and using even smaller target size (Fitts' ID of 5.61), another significant increase in errors can be expected. This DHR was achieved for the dataset where all the data from the test was included, and for the dataset where the participants below the average amount of repetitions had been removed. This means that for all participant groups combined, the smallest achievable target a user can select on touch, with little effort, is between ∼2 and ∼4 mm.

By investigating the issue on consent and informing users that they were participating in a study, we saw that for both levels there was a significant increase in performance for the users from the crowd-plus participant group. A reason for this could be the same reason as to why participants in the lab environment are also performing better, namely that when knowing that their results matter and have an influence on the results from a test, the users will in general try their very best and in most cases increase their performance while interacting with the application.

In Section 2.2.3 we examined the difference between using pointing and dragging for selection tasks. Contrary to the results by Cockburn et al. [6], we found that completing the DHR tasks in the preliminary study using tapping, participants made an average of 0.23 errors per repetition (16.77%) compared to an average amount of 0.30 errors per repetition for dragging in the main user study (21.46%). It should again be noted that the results from the preliminary study is not bound by Fitts' law, as all participants used two hands to do the interaction, and thereby having a distance of 0 between the targets.

Lastly, we investigated if there were any issues related to gamifying existing HCI user studies in order to make users in the crowd group believe that they were only playing a game and thereby hiding the fact that they were participating in a user study. We cannot with absolute certainty state that users thought they were only playing a game, however, as stated above the users who were informed about the study (crowd-plus) had a significantly better performance than crowd, which showed that the participants in the crowd played the game as normal, but might have focused less on their own performance compared to the two other groups. We also found that gamifying existing tests may quickly become tedious for the users, as game elements, scores, lives etc. are simply not enough to make a game fun. With the user responses from the test, we can also see that even adding an overall story for the game in future iterations, may not change the fact that the DHR level of the game is simply not fun and engaging, but rather feels like forcing the users to do something for an extended period of time, which they do not feel like doing.

5. CONCLUSION AND FUTURE WORK

In this study, we evaluated the effect of crowdsourcing existing HCI user studies. We examined if there were significant differences between using crowdsourcing or a lab environment in a gamified Device Human Resolution experiment on touch devices. The test consisted of two crowdsourced groups; one group informed about their participation in a study, and one uninformed group. We compared our crowdsourced groups, in uncontrolled environments, with a laboratory group in a controlled environment. From the related work, we expected that using a lab environment would reveal better results than the crowdsourced groups, in terms of time and error-rate performance.

The initial results showed that there were significant differences in performance between crowdsourcing and lab testing, with lab being superior. However, the findings indicate that when taking the high variance of the crowd into consideration, the crowd will be able to perform equivalent to the lab environment. Therefore, using crowdsourcing for HCI user studies is a feasible solution to get more participants from the real environment. This will result in a higher external validity for HCI user studies. However, studies needs to ensure that the experiment is easily understandable and engaging, as it can influence the drop-out rate by the crowd compared to lab, as the participants do not feel as obliged to complete the experiment. Furthermore, we examined the differences between informing participants about their participation in a study and not informing them. This yielded significantly better results for the informed participants compared to the uninformed participants.

We also examined the overall DHR for touch. The findings showed that the DHR is between $\sim$2 mm and $\sim$4 mm for all data entries combined. Comparing the results to our own previous work [4] and the original work by Bérard et al., [3] we see that the DHR performance of touch input is comparable to in-air interactions, however, the mouse, as an input device, is still unbeatable for small target selection tasks.

In a future iteration of this study, further iteration of the DHR level, and a possible redesign of the level may be necessary. We observed some issues with the understanding of the level, which especially influenced the results from the crowds. Furthermore, the feedback on gamification we received from the participants, displayed issues. Such as a non-immersive gaming factor resulting in early drop-outs in the crowd experimental group. This needs to be addressed in future versions, especially to ensure that the app is efficient enough for conducting HCI user studies. For the future development of the app, we will examine other HCI studies for touch and implement them as new levels.

6. REFERENCES

1. Arnett, J. The neglected 95%: Why American psychology needs to become less American. *Behavioral and brain sciences 63* (2008), 602–614.
2. Benko, H., Wilson, A. D., and Baudisch, P. Precise Selection Techniques for Multi-Touch Screens. *Proceedings of the SIGCHI conference on Human Factors in computing systems* (2006), 1263–1272.
3. Bérard, F., Wang, G., and Cooperstock, J. On the limits of the human motor control precision: the search for a device's human resolution. *Proceedings of the 13th IFIP TC 13 International Conference on Human-Computer Interaction 6947* (2011), 107–122.
4. Bjerre, P., Christensen, A., Pedersen, A. K., and Pedersen, S. A. Transition times for manipulation tasks in in-air user interfaces. `http://projekter.aau.dk/projekter/files/207459117/Paper_MTA14930.pdf`. Accessed 19. May 2015.
5. Bracht, G. H., and Glass, G. V. The external validity of experiments. *American educational research journal* (1968), 437–474.
6. Cockburn, A., Ahlström, D., and Gutwin, C. Understanding performance in touch selections: Tap, drag and radial pointing drag with finger, stylus and mouse. *International Journal of Human-Computer Studies 70*, 3 (2012), 218–233.
7. Dandekar, K., and B. I. Raju, a. M. A. S. 3-D finite-element models of human and monkey fingertips to investigate the mechanics of tactile sense. *Journal of Biomechanical Engineering 125* (2003), 682–691.
8. Fitts, P. M. The information capacity of the human motor system in controlling the amplitude of movement. *Journal of Experimental Psychology 47* (1954), 381–391.
9. Gächter, S. (Dis)advantages of student subjects: What is your research question? *Behavioral and brain sciences 33* (2010).
10. Heer, J., and Bostock, M. Crowdsourcing Graphical Perception: Using Mechanical Turk to Assess Visualization Design. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (2010), 203–212.
11. Henrich, J., Heine, S. J., and Norenzayan, A. The weirdest people in the world? *Behavioral and brain sciences 33* (2010).
12. Henze, N., and Boll, S. It does not Fitts my data! Analysing large amounts of mobile touch data. *Human-Computer Interaction – INTERACT 2011* (2011), 564–567.

13. Henze, N., Boll, S., Pielot, M., Poppinga, B., and Schinke, T. My App is an Experiment: Experience from User Studies in Mobile App Stores. *International Journal of Mobile Human Computer Interaction* (2011), 71–91.

14. Henze, N., Rukzio, E., and Boll, S. 100,000,000 Taps: Analysis and Improvement of Touch Performance in the Large. *Proceedings of the 13th International Conference on Human Computer Interaction with Mobile Devices and Services* (2011), 133–142.

15. Holz, C., and Baudisch, P. The Generalized Perceived Input Point Model and How to Double Touch Accuracy by Extracting Fingerprints. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (2010), 581–590.

16. Kolly, S. M., Wattenhofer, R., and Welten, S. A Personal Touch: Recognizing Users Based on Touch Screen Behavior. *Proceedings of the Third International Workshop on Sensing Applications on Mobile Phones* (2012), 1–5.

17. MacKenzie, I. S. Fitts' law as a research and design tool in human-computer interaction. *Human-Computer Interaction* (1992), 91–139.

18. MacKenzie, I. S., Sellen, A., and Buxton, W. A. S. A comparison of input devices in element pointing and dragging tasks. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (1991), 161–166.

19. Nosek, B. A., Banaji, M. R., and Greenwald, A. G. E-Research: Ethics, Security, Design, and Control in Psychological Research on the Internet. *Journal of Social Issues 58* (2002), 161–176.

20. Out of Pixels. Drop. `https://play.google.com/store/apps/details?id=com.infraredpixel.drop`. Accessed 17th March 2015.

21. Pielot, M., Henze, N., and Boll, S. Experiments in App Stores - How to Ask Users for their Consent? *CHI 2011* (2011), 1–4.

22. Sasangohar, F., MacKenzie, I. S., and Scott, S. D. Evaluation of mouse and touch input for a tabletop display using Fitts' reciprocal tapping task. *Proceedings of the Human Factors and Ergonomics Society 53rd Annual Meeting* (2009), 839–843.

23. Schmuckler, M. A. What Is Ecological Validity? A Dimensional Analysis. *Infancy 2* (2001), 419–436.

24. Sears, A., and Shneiderman, B. High precision touchscreens: design strategies and comparisons with a mouse. *International Journal of Man-Machine Studies 34* (1991), 593–613.

25. Statista Inc. Number of available applications in the Google Play Store from December 2009 to February 2015, 2015. `http://www.statista.com/statistics/266210/number-of-available-applications-in-the-\google-play-store/`. Accessed 10. May 2015.

26. Statista Inc. Number of smartphone users* worldwide from 2012 to 2018, 2015. `http://www.statista.com/statistics/330695/number-of-smartphone-users-worldwide/`. Accessed 10. May 2015.