EAI Endorsed Transactions
on Creative Technologies

Research Article **EAI.EU**

A Comparison of Gamified HCI Studies with Lab and Crowd Participants

Hendrik Knoche,[*] Allan Christensen, Simon André Pedersen

Department of Media Technology, Aalborg University, Rendsburggade 14, 9000 Aalborg, DK

Abstract

We compared a game-based experiment carried out in the lab with crowdsourced set ups (informed and uninformed participants) on the device's human resolution (DHR) - the minimum size for dragging the finger onto a target on a touch screen. Lab participants produced fewer errors than the crowd. From lab participants, we found the smallest selectable target width for dragging onto non-occluded targets with visual target position feedback, was between 2mm and 4mm on mobile touch devices. Performance data on *error* not *time* allowed for drawing this conclusion as participants from all groups did not take enough care and time to acquire the targets. The bi-modal performance distributions of crowd participants required filtering data.

Received on 18 November 2016; accepted on 04 February 2017; published on 05 April 2017

Keywords: HCI, Crowdsourcing, gamification, device human resolution, DHR, touch interaction, touch screen, dragging, acquiescence bias

doi:10.4108/eai.5-9-2017.153058

1. Introduction

Running experiments with human participants drawn from student populations has seen criticism for its poor external validity [17] and crowdsourcing studies has gained momentum to draw from a wider population. Studies either monetarily or otherwise incentivize participation], e.g. through games [20], or run covert with participants being unaware of their involvement in a scientific knowledge creation endeavour [36].

We compare results from a gamified study in the lab with a campus to a crowd population. One potential strength of crowdscourced studies lies in eliminating acquiescence bias. To this end, we compared the performance of lab participants with an *informed* and a *naive* crowd. The former two groups knew their game performance was collected for scientific purposes. We found that data analysis benefited from removing outliers in all three participant groups and that in general lab participants performed better than both crowd groups. Removing below absolute median performance participants in a game requiring *accuracy* and using only a participants best performance in a reaction *time* game eradicated the differences between all groups.

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.

2.1. Crowdsourcing and Game-based Studies

Crowdsourcing studies addresses criticism about poor population validity of social science studies, which draw on WEIRD (Western, Educated, Industrialized, Rich, Democratic) participants [17]. In published behavioral science research 96% of test participants came from western industrialized countries [1], with a majority of undergraduates (67% in American samples and 80% in other countries). WEIRD subjects often occupy the extreme ends of the behavioral science distribution and provide different results than other subject groups [17]. Gächter suggested that WEIRD subjects could be used as a benchmark or starting point for investigating generalizability to other social groups [15].

[*] Corresponding author. Email: hk@create.aau.dk

Amazon's Mechanical Turk (mturk) is one of the most common solutions for crowdsourcing research. Investigators typically upload small tasks that mturk workers (the crowd) finish in exchange for monetary remuneration [26, 37]. In lab studies, remuneration for lab participants is generally higher than for mturk workers and running a large scale crowdsourced study requires less time per participant [25].

Games or gamifying tasks are another way of crowdsourcing studies that draw on people's intrinsic motivation to master or play a game or a gamified task for fun [14, 39] rather than an extrinsic motivation of monetary gain. However, such games require careful and genuine design. Deterding et al.'s criticized the current simplistic state of gamification, which uses points, badges and leaderboards on everything to make it seem like a game [10, 11].

While improving both population and ecological 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 [18]. This could have been due to violations of how the task was supposed to be carried out but that was not enforced or controlled for in the logged data.

Obtaining informed consent represents standard practice in lab settings. After having shown up, few sign-ups decide to not participate when asked to read and sign an informed consent sheet. But in crowdsourced studies users often did not carry through to play a game when this entailed the sending of data. Henze et al. examined five different ways to inform users and found the highest carry through by not asking users (83.7%) and showing a consent form with no opt-out option (81.3%) [19]. Pielot looked at four ways for users to accept sending anonymous data [33] and found a single 'okay' button with no opt-out feature provides the highest carry through (87.6%). For ethical reasons they recommended using a forced choice 'yes/no' button (67.4%) for participation in the study [19, 33]. Internet based study participants are more likely to voluntarily opt-out before completing the required tasks of a study than in a lab-based environment in which they come in face-to-face contact with the person running the study [20, 26, 30]. For a recent discussion of and rethinking of informed consent in Internet based studies see Brown et al.'s work [6].

2.2. Device Human Resolution (DHR)

As a case, we used a study on the unknown limits in precision when dragging a finger onto small-width targets on a touch screen. We draw on Fitts' Law [12], the concept of Device Human Resolution (DHR) [3], and the literature on touch accuracy. Fitts' law predicts the required time for a human to perform a movement over a distance (*amplitude*) from from an origin to target with a given a size (*width*). The Index of Difficulty (ID) quantifies the difficulty of the task with higher IDs resulting in a harder task and requiring more time. Error probabilities increased with rising IDs in Fitts' original experiments but did not go above 4% [12]. Due to the widespread use in the HCI community and the recent disputes about its validity [22] we use both MacKenzie's Shannon version of Fitts' ID [29]:

$$ID = log_2\left(\frac{amplitude}{width} + 1\right) \tag{1}$$

and Fitts and Peterson's [13] original version:

$$ID = log_2\left(\frac{2 \times amplitude}{width}\right) \tag{2}$$

Studies use the empirically obtained target acquisition times and regress them onto Fitts ID using linear models yielding an intercept and a slope component. 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). Their analysis used the participants' regressed slopes from subsets of the ID range and tested for significant differences to the slope of the participants' whole ID range used in the study. For mouse input they found a DHR for *time* (0.036mm) and *error* (0.018mm). Participants could maintain a low error rate from 0.036mm downwards only at the expense of increased time and below 0.018mm errors increased drastically.

2.3. Touch Performance

Factors moderating accuracy and time of finger pointing on touch interfaces include: target size [9] and shape [5], the pointing device occluding the target's position and having larger size than the target (fat finger problem) [23], acquisition time limits, target location in relation to screen borders [5, 32] and to other targets [24], finger orientation (roll, pitch, yaw) [23], posture of the user, e.g. sitting vs standing [16, 34], the vertical [5] and horizontal [27] tilt of the screen, parallax [4], and feedback on whether a target was successfully touched [5].

Holz et al. provided two reasons for inaccurate target selections with fingers on touch screens: (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 [23]. Benko et al. found that users perceived the finger surface interaction point (1) differently [2]. Various design solutions have addressed these problems, e.g. by offsetting the cursor or zooming.

Cockburn et al. compared finger, stylus, and mouse in target acquisition (5, 12.5, and 20mm width columns) tasks with tapping and dragging [8]. Tapping on 5mm wide targets with a finger yielded a roughly seven times higher error rate (14%) for acquisition compared to the other devices. Dragging (~0.92 sec.) had a significantly higher overall selection time when compared to tapping (~0.57 sec.) 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 acquisitions during dragging. Tapping had no equivalent feedback on the location of the finger position in relation to the target that the finger occluded.

In summary, the smallest size of targets for dragging on touch screens with no additional feedback is currently unknown and a large number of factors many of which cannot be controlled for in crowdsourced studies affect time and accuracy of touch interactions.

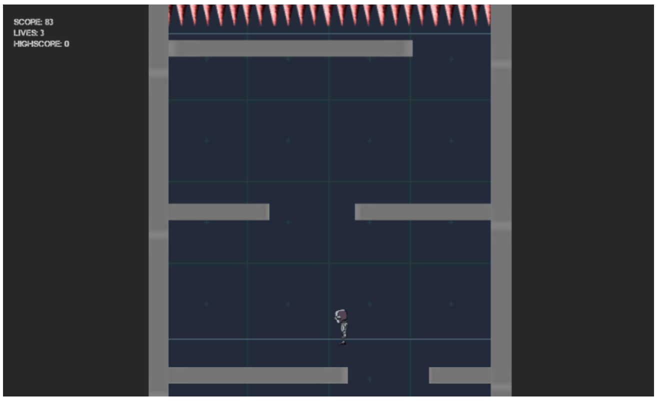
3. Study Context and Participants

Our application included two levels that were distinctly different games but both had the same robotic looking, main character appear in them (*c.f.* Figure 1 and Figure 4).

Three different participant groups played the games. 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 14 participants (crowd), who thought they were merely playing a game and not participating in a study. The third group (crowd-plus) included 11 participants (4 female, average age 28, $SD = 9.5$) who knew they were participating in a study.

The lab participants used an LG Nexus 4 smart phone running Android 5.1, with a 4.7-inch display and 768x1280 resolution, which they held as they pleased. After an introduction, the lab participants received the smart phone and were prompted to start the application.

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. For each game they watched a ca. 30 second introductory video illustrating how to play and complete the game. The game started after the video had finished. For better comparison between groups, we did not provide any additional assistance beyond the introductory videos to the lab participants in case of questions. After completion of each game, all participants saw their own high score. The crowd-plus group further received a pop-up message prompting them to answer a questionnaire. A large majority (86%) of the crowd-plus group chose to fill in the questionnaire, which included control variables such as age, environment played in, and touch device usage.



SCORE: 83
LIVES: 3
HIGHSCORE: 0

Figure 1. The first level of Drop. The goal of the game was to survive by not hitting the walls for as long as possible by moving the character left and right through the holes with the bottom of the screen moving upwards at an increasing pace.

4. Study 1 – Drop Game

The game of the first level was inspired by Drop [31] an Android game. In our version the user had to control a robotic character using either the gyroscope, thereby tilting the device from left to right in order to make the character move in either direction, or creating a touch-point by pressing somewhere on the screen and having the character move towards that point (see Figure 1 and 3). The floor then moved upwards at an increasing pace throughout the game and the user had to stay alive for as long as possible. Higher survival times indicated better performance. The survival time in seconds was equal to the points awarded. Players could gain additional points by collecting stars throughout the game. After the player had lost three lives the game ended and proceeded to the second game (Wall Destroyer) described in section 5. Repeated play of Drop required playing through Wall Destroyer unless the player shut down the application on the device, which we did not log.

This study allowed for comparing results between the participant groups with a task that was based on reaction time and felt more like a game than the second level.

4.1. Results

The lab group played Drop once, but on average the crowd (4.2) and crowd-plus (1.9) groups played more often. The crowd group that had no commitment to the scientific goal of the study but potentially the highest intrinsic motivation for engaging with the game played most often.

We found a significant difference in average survival times between the *groups* ($F_{2,38}$=10.13, p<0.001). A TukeyHSD post-hoc test showed that lab participants performed significantly better than both the crowd and the crowd-plus group. Both the crowd and crowd-plus participants had a high number of very low survival times compared to the lab. Figure 2 (top) summarizes the estimated densities of survival times of the three groups. It includes each participant's minimum and maximum survival time and shows clear bi-modal distributions for the crowd and crowd-plus groups.

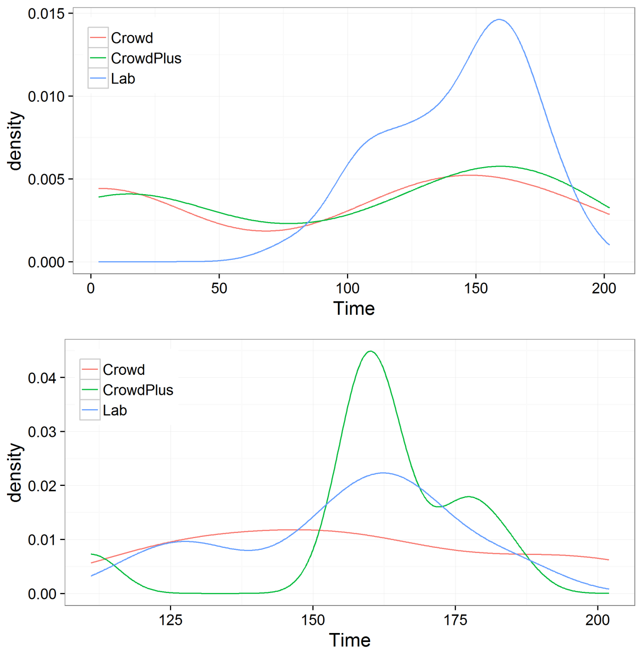


Figure 2. Estimated density of survival times using each participants minimum and maximum survival time (top) and only the maximum survival time (bottom)

Due to the high amount of low survival times for both crowd groups, we analysed the data that included only the highest survival time for each participant. In this data set, while the distributions are not very similar as illustrated in Figure 2 (bottom), we found no significant difference in survival times between the groups ($F_{2,38}$=0.205, p=0.81).

Observations

The majority of the lab participants decided to hold the phone in a firm grip with both hands and used both thumbs for interaction, see Figure 3. They did not have to be precise and therefore they focused on fast movement. Several tried putting the phone on the table but eventually picked it up and held it in the firm grip. A few left the phone on the table for the whole duration of the test (see Figure 3).

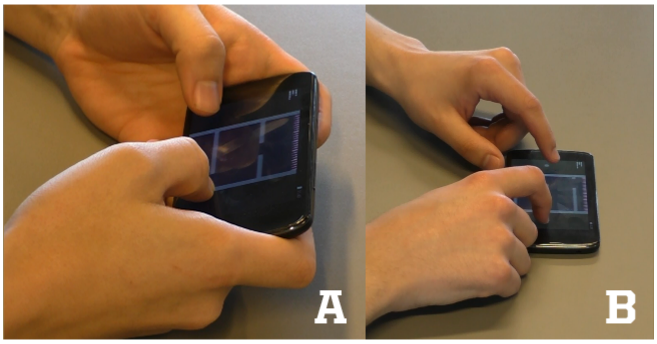


Figure 3. Different grip types while playing Drop in the lab: *Firm grip* input with two thumbs (A) and *table rested* input with index finger (B).

5. Study 2 – Wall Destroyer Game

The purpose of this study was to compare the three different user groups playing a game to investigate the DHR for dragging on touch screens. As much as possible we replicated Bérard et al's setup [3] with a game called Wall Destroyer.

The user had to tap anywhere within a green starting area (see Figure 4) to make the wall appear in front of the monster and then drag their finger onto the wall. The wall appeared always 47mm away from the finger's touch down location in seven descending widths (32, 24, 16, 8, 4, 2, 1mm) per round always starting with the largest. This resulted in the following Fitts IDs for the target (wall): 1.32, 1.58, 2, 2.81, 3.7, 4.64, and 5.61.

Unlike other DHR studies we did not use the second hand to validate target acquisition. Successful acquisition of a drag required lifting off the finger when on top of the target. 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 neither auditory nor visual feedback on the current touch position of the finger input. But given the wall's length the participants were aware of the targets location. The acquisition 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.

To encourage repetitions of these rounds, participants had five lives. Players lost a life only after three successive misses on the same target size after which the player advanced to the next target size. This approach did not enforce an equal number of repetitions of all sizes, but encouraged most participants to complete multiple game rounds to provide more data.

5.1. Results

The combined data from the three participant groups amounted to 5404 wall acquisitions. For each participant we removed data from incomplete repetitions (a round which did not complete ID 5.61) and from the resulting set we removed all participants who did not complete at least one round (of all sizes/IDs).

Since we were only interested in systematic increases we conservatively removed acquisition time outliers that were more than four absolute deviations from the median [28] for each Fitts ID. This removed a total of 137 or 2.5% of the acquisitions (59 from crowd participants, 37 crowd-plus, and 41 lab) and across all wall sizes (the number of removed acquisitions from largest to smallest wall size were: 32, 18, 17, 7, 14, 17, and 32). Unless noted otherwise, we used one-way Analysis of Variance tests (ANOVA) with a TukeyHSD for post-hoc tests or in case of violations of model assumptions a Friedmann tests. This was typically the case when analysing acquisition errors. We ran these as a first pass test since *size* should have an effect on acquisition time and/or error. Subsequently, we tested whether *time* or *error* increased disproportionally at a given size. To this end, we used linear models to determine each participant's slope for each subset of three successive IDs (e.g. the first subset contained the IDs 1.32, 1.58, and 2) and the overall linear model slope containing all IDs of all participants. For each subset of three IDs, we used t-tests to determine whether the subset slopes were significantly larger than the overall model slope.

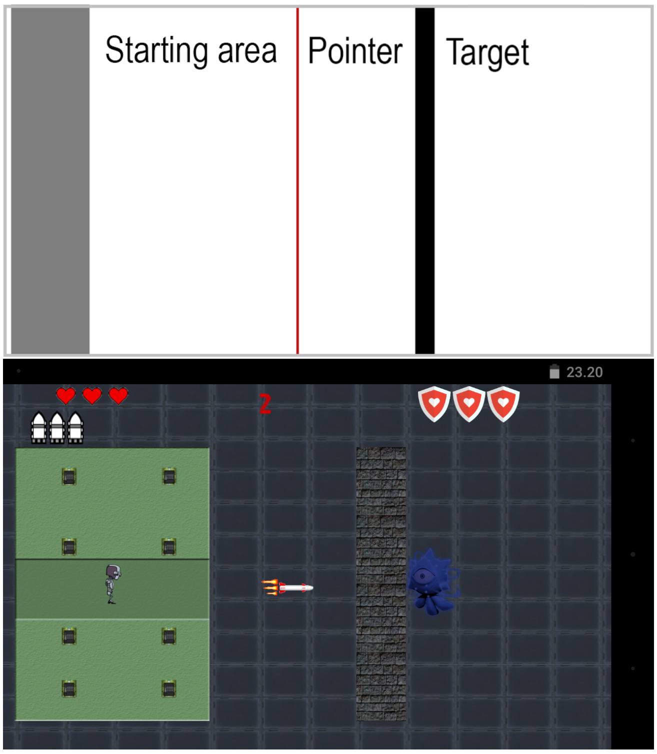


Figure 4. Design of the standard DHR test (top) and the gamified version (bottom) with the starting area in green, no indication of the pointer but the missile as feedback on successful drag actions on the targets (wall)

We present an overall comparison across groups first and then the results for the individual groups, see Figure 5 and Figure 6 as summaries of time and error performance.

Group comparison. A mixed ANOVA on acquisition *time* checked *group*, *size*, and their interaction for effects. We found no effect for *group* ($F_{2,33}$=0.80, p=0.46, η_p^2=0.01) and after necessary sphericity corrections none for *size* ($F_{6,198}$=3.6, ϵ_{GG}= 0.2, p=0.057, η_p^2=0.07) or interaction

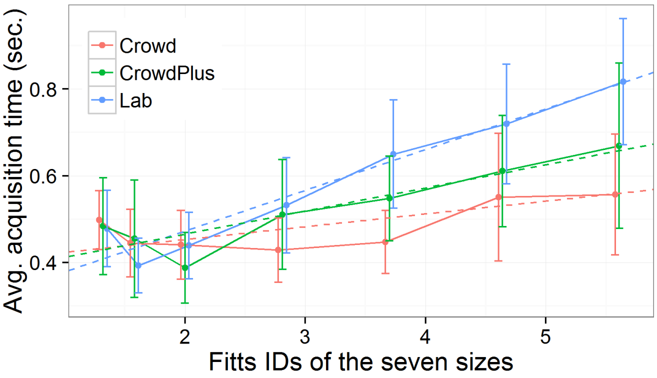


Figure 5. Target acquisition time by target size (in Fitts' IDs) and group including 0.95 confidence interval error bars

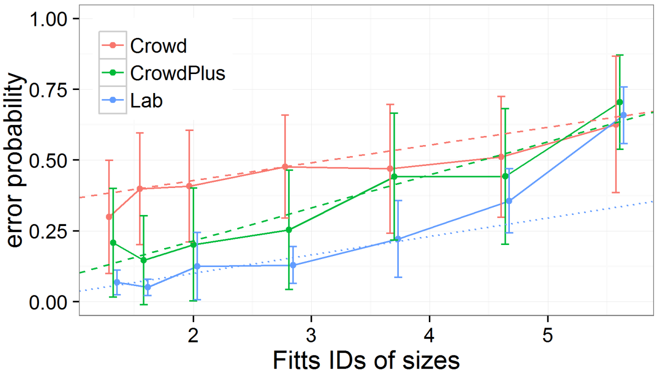


Figure 6. Target acquisition error rate by target size (in Fitts' IDs) and group including 0.95 confidence interval error bars. Dashed slope lines include all IDs, the dotted line for the Lab had the two highest IDs excluded based on subset slope deviations.

of the two ($F_{12,198}$=1.81, ϵ_{GG}= 0.2, p=0.17, η_p^2=0.07) either - see Figure 5 for an overview.

Figure 6 summarizes the results on *error* probabilities at the different IDs for the groups. The lab participants had the lowest average *error* probability followed by crowd-plus and then crowd particpants. For the highest ID the *error* probabilities of all three groups converged. Given the non-normal distribution of the *error* probabilities we used a Kruskal-Wallis test to compare the groups and an aligned rank transform test [21] for an interaction (*group×size*). We found only a significant difference ($\chi^2(2)$=6.48, p=0.04) between *groups*. A posthoc test using Tukey and Kramer (Nemenyi) test showed that this was due to the performance of the lab participants being significantly better than the crowd group. Given the convergence of error probabilities and the sub-slope analysis results presented in the following sections we tried whether removing the highest ID from the *error* analysis would yield significant differences between the groups, but this was not the case.

Lab. For the 16 lab participants, the ANOVA showed an effect of *size* on *acquisition time* ($F_{6,90}$=18.0, ϵ_{GG}= 0.27, p≪0.001, η_p^2=0.25) but for none of the subsets were the participants' individual slopes significantly higher than the overall acquisition time slope (0.09, *c.f.* Figure 5). Surprisingly, the largest size (smallest ID) that came first during each repetition in the game had a higher acquisition time average than the next smaller size.

For the error data, a Friedman Ranked Sum test showed an overall significant difference in errors for *size* ($\chi^2(6)$=185.2, p≪0.001). The Friedman post-hoc showed that all sizes had significantly different error probabilities apart from the two largest sizes. The fourth Fitts ID (2.81) had a lower probability than the third ID (2).

The sub-slope analysis showed a significant difference between the subset from IDs 3.7 to 5.61 to the overall slope (0.12) of error probability. We would iteratively remove the highest IDs from the overall slope regression and again for another test of all subset slopes against it if we found significant differences between subset slopes and overall slope. Specifically, we excluded the highest ID (5.61), remodelled the overall slope of the data, and re-ran the sub-slope analysis. Based on this data, the participants' slopes from the second highest subset (IDs 2.81 to 4.64) were significantly higher than the revised overall slope (0.08), too. Again we excluded the highest ID of this set (4.64) and obtained a new overall slope (0.06). With this last overall slope we found no further significant differences from sub-set slopes to it. We included this slope (0.06) in Figure 6 as a dashed line.

Crowd. For the nine crowd participants, we found no significant effect of *size* on *acquisition time* ($F_{6,48}$=2.53, ϵ_{GG}=0.29, p=0.12, η_p^2=0.16) and no significant increase in the participants' slope subsets from the overall slope (0.03, *c.f.* Figure 5).

A Friedman Ranked Sum test showed an overall significant difference in errors between the *sizes* ($\chi^2(6)$=185.5, p≪0.017) with all but the two smallest IDs being signficantly different from one another. But none of the subset slopes were significantly different from the overall slope (0.07, *c.f.* Figure 6). Given the subset slope results in the lab group, we tried removing the highest ID (5.61) from modelling the overall slope of errors and repeated the test of the subset slopes. Even with the highest ID removed, the subset slopes did not differ significantly from the overall slope of errors.

Crowd-plus. For the 11 crowd-plus participants, we found no significant effect of *size* on *acquisition time* ($F_{6,60}$=1.69, p=0.14, η_p^2=0.11) and none of the participants' slope subsets were significantly higher than the overall slope (0.05, *c.f.* Figure 5).

The Friedman Ranked Sum test revealed an overall significant difference in error probabilities between the sizes ($\chi^2(6) = 198.2$, $p \ll 0.001$) but as for the crowd group the subset slope analysis showed no significant deviation from the overall slope (0.11, *c.f.* Figure 6). Again we removed the highest ID from the overall slope calculation but this did not change the outcome - none of the subset slopes were significantly different from the overall slope.

Filtered data. Several participants in the crowd performed very poorly during the experiment as evident from the large error probabilities on the target sizes. Studies have shown that many of the sizes the Wall Destroyer study employed (32-16mm) were unproblematic in terms of acquisition [20, 24]) and average fingers (e.g. 16mm for index fingers [35]) could not have occluded the two largest targets. Therefore we cannot attribute the comparatively low acquisition performance to the fat finger problem but assume motivational and task differences. Fitts law studies often instruct their participants to carry out the task "*as fast as possible and as accurately as possible*" resulting in rather small error probabilities of around 4% [38]. The game did not prompt the participants to do this and their hit probability was much lower. Although acquisition times of all groups increase with larger IDs the time taken was not sufficient to arrive at error rates typical for Fitts ID tasks. Most notably, the lab participants who on average took more time than both crowd and crowd-plus had the same high error probability for the highest ID as the crowd groups who used the least time of all groups on this wall size. Clearly the game did not encourage the participants enough to take sufficient care to acquire the targets. We therefore turned to participants that took more care than the others as evidenced through higher repetitions of game rounds. We consulted the median number of repetitions (completed rounds) the lab (10), crowd-plus (13), and crowd (11) participants had played (see Figure 7). We removed all participants with below their group's median number of repetitions. Only including participants with above median repetitions we retained 5 out of 9 crowd, 6/11 crowd-plus, and 8/16 lab participants.

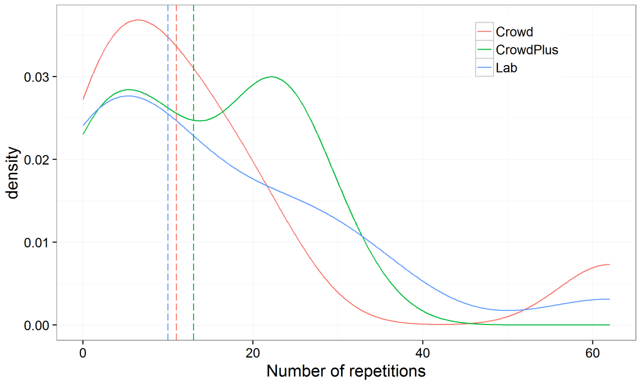


Figure 7. Estimated density of repetitions of complete rounds played by group along with group medians as vertical dashed lines

Group Comparisons. For acquisition time in the filtered dataset, a mixed ANOVA yielded significant differences between the *sizes* ($F_{6,102}$=43.3, ϵ_{GG}=0.32, $p \ll 0.001$, η_p^2=0.32) but none for participant *groups* ($F_{2,17}$=1.31, p=0.34, η_p^2=0.10) or the interaction between *size* and *groups* ($F_{12,102}$=1.47, p=0.145, η_p^2=0.03). Figure 8 illustrates the results. But when examining the mean slopes for acquisition time no subset slope deviated significantly from the overall slope in any of the groups.

A Kruskal-Wallis test showed that there was no statistically significant difference between the error probabilities for *group* ($\chi^2(2)$=3.31, p=0.19) in the filtered dataset. But only the lab participants had a significantly higher slope for the highest subset of IDs than the overall slope of all IDs. When the highest ID was removed from the overall slope calculation the slopes of the highest subset for both *lab* and *crowd-plus* were significantly higher than the overall slope. Figure 9 includes both of these overall slopes in which the highest IDs were removed as dotted lines.

Table 1 provides an overview of the outcomes of the statistical and sub-sloope tests on the two dependent variables *time* and *error* in both the overall and filtered data set.

Observations. The lab participants referred to Wall Destroyer as a precision game and therefore the majority used their index finger. Several participants started with a firm grip (see Figure 3, A) but ended up either using one hand (see A, B, E in Figure 10) or placing the phone on the table (see C, D in Figure 10). When going from firm grip to the other grip types meant going from thumb to index finger.

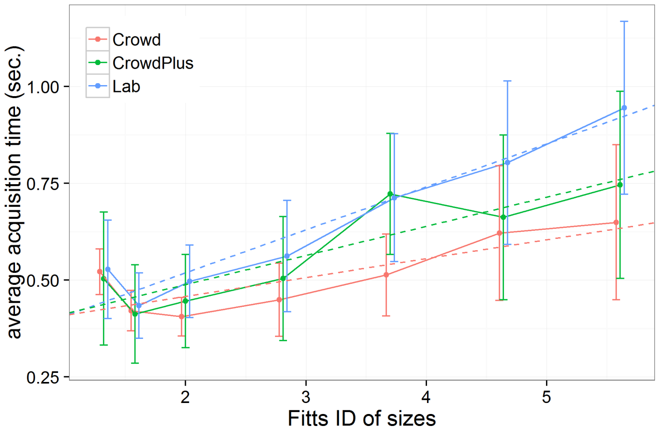


Figure 8. Average acquisition time by size (in Fitts' ID) including 0.95 confidence interval error bars for the three filtered groups

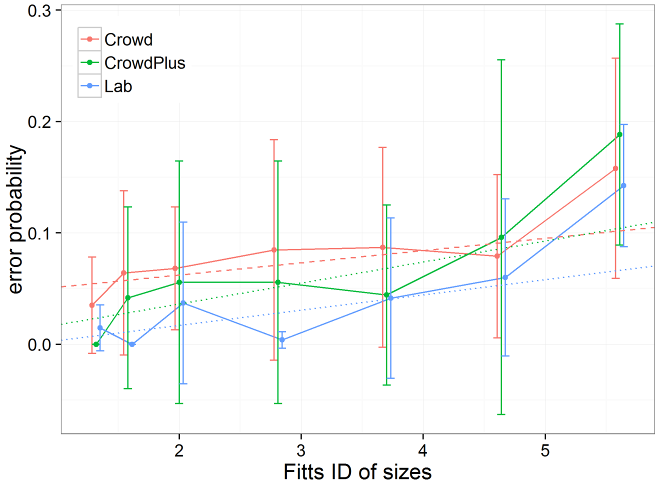


Figure 9. Error mean per repetition by size (in Fitts' ID) including 0.95 confidence interval error bars for the three filtered groups.

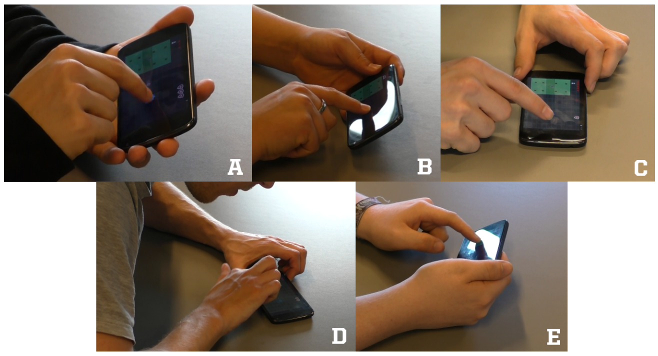


Figure 10. Different grip types of lab participants while dragging with their index finger in Wall Destroyer

	time				*error*			
	ANOVA		sub-slopes		non-param.		sub-slopes	
groups	all	filt.	all	filt.	all	filt.	all	filt.
crowd	-	*size*	-	-	*size*	-	-	-
crowd-plus	-	*size*	-	-	*size*	*size*	-	>4.64
lab	*size*	*size*	-	-	*size*	*size*	>3.70	>4.64
all groups combined (group as factor)	-	*size*	-	-	*group*	*size*	>4.64	>4.64

Table 1. Overview of test results from ANOVA, non-parametric tests, and the subset slopes analyses for *time* and *error* for both the overall (all) and above median filtered (filt.) dataset. Dashes (-) denote no effects; Sub-slopes entries denote IDs that as part of a subset had significantly higher slopes than the the overall slope. See the results section of the lab participants as an example.

Some participants mentioned that their thumb was “too big” for the task and they had to use their index finger, which some found too large for the task as well. Participants tried during this level only to use the tip of the index finger to get the highest precision, therefore no participant used a flat finger on the screen in line with findings by Holz & Baudisch [23].

5.2. Discussion

Given that Cockburn et al’s study had shown that Fitts IDs moderated acquisition times of finger dragging tasks on a touch screen it came as a surprise that acquisition times in the overall cross-group ANOVA and in the crowd and crowd-plus groups did not yield significant differences for *sizes* in the Wall Destroyer game. Cockburn et al’s lab study used a Fitts task similar to what Figure 4 depicts, an ID range overlapping with our study (from 2.3 to 5.4), and a small target (5mm) within that range. Failure to find significant differences in acquisition times from sizes could be due to a lack of statistical power as evident from the high variance in the data as apparent from the large confidence interval error bars in Figure 8 and that *size* was a significant factor in the ANOVA of the lab participants’ *acquisition times*. In our game the players additionally did not take enough care or time in acquiring the targets as evident from the very high error probabilities that dwarf those of typical Fitts law tasks and the below analysis of the empirical coefficients when modelling the acquisition times with Fitts law.

With the initial acquisition time outliers removed the fit of time performance of all participants averaged by the different Fitts’ IDs the fit with MacKenzie’s model (R^2=0.9, a=0.29, b=0.08) and Fitts and Peterson’s original model [13] was good (R^2=0.88, a=0.34, b=0.07). Especially when considering that this included the anomaly (c.f. Figure 5) of the smallest ID that took participants on average longer than the second one. When the smallest ID was removed the fit was almost perfect for both models (R^2=0.99). On the same data set but with no outliers removed Christensen et al. [7] had found poorer fits of Fitts law (R^2=0.23 for MacKenzie’s and R^2=0.78 for Fitts’ and Peterson’s model). However, the average target (wall) acquisition delay across all participants was 0.57 seconds and the Fitts’ law coefficients from MacKenzie’s and Fitts and Peterson’s model indicated that most of the movement time was due to the constant (a) rather than the slope (b), which depends on the index of difficulty. The model attributed more time to initiating the movement and lifting off the finger than the actual movement. Furthermore, both MacKenzie’s (12.5bit/second) and Fitts’ original model (14bits/second) yielded rather high indices of performance (the inverse of b) that usually lie between 8 and 12 bits/second. Given that Cockburn had shown dragging to be significantly slower than tapping for target selection make our results with high indices of performance at the limit of human motor control very unlikely to model the data correctly.

In summary, the game in its current design with the very high error probabilities and the way touch events were logged (from touch down to lift-off) did not yield time performance data that was very specific to Fitts’ law.

The game made it difficult for players to be very precise in the acquisition task since it provided no feedback about the actual touch position and players could therefore not optimize or correct their finger positions beyond their mental model. 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.

The Wall Destroyer game did not provide incentives for fast performance either. We did not control for lingering times in the start area before initiating the movement towards the wall, which might yield more representative results and better modelling with Fitts law. Future versions could a) log the touch events with higher granularity, e.g. the beginning and end of movement on the target, b) add time limits or scores sensitive to time performance. This could include faster hits yielding higher scores, missing the target incurring higher penalties, and providing rewards proportional to ID.

Gamifying existing tests may 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. However, our participants in the crowd groups did on average play the games repeatedly without further coercion. The naive crowd group played the most.

While lab participants on average took more time and had lower acquisition error rates than the crowd and crowd-plus participants these differences were not significant and the median of completed games similar between groups. Earlier analysis of the data set [7] found larger spreads between the highest and lowest performance in acquisition time for both crowd groups when comparing them to the lab participants. But after removal of outliers as outlined in this paper this was not the case.

We followed Henze et al.'s approach by excluding data from participants with low or insufficient performance, i.e. in our case those who had below the median amount of task repetitions to see whether this would yield different results. This sub-analysis of only above median performers resulted in less conservative results for the sub-slope analysis (see Table 1 and the next section for details) with the few participants that remained in the data set.

DHR. Target acquisition *times* did not provide any insight into a DHR limit (c.f. Table 1) but acquisition *errors* did. While acquisition error probabilities were unusually high in all groups in comparison to standard Fitts' law experiments (typically up to 4%), sizes smaller than 4mm (IDs larger than 3.7) yielded disproportionally and significantly higher acquisition error probabilities for the lab participants through the sub-slope analysis. This suggests that the DHR for dragging onto non-occluded targets with no visual feedback on touch position lies between 4 and 2mm - much smaller than average index finger width. This is comparable to the DHRs for positioning a cursor on targets with in-air interactions (2.4mm) found by Bérard et al. and Bjerre et al. (between 1.2 and 2.4mm), and much worse than with a mouse (between 0.036mm for *time* and 0.018mm for *error*). We used a target with substantial height which provided cues in terms of the location of the target. For targets that get completely occluded by the touching finger the DHR might be larger in size.

In the above median performer subset error probabilities were an order of magnitude smaller than in the overall data set (c.f. Figure 6 and Figure 9). In this case both the lab and the crowd-plus participants had significantly higher slopes than the overall slope - but only in the highest subset of IDs (from 3.7 to 5.61). However, we only had few participants in this filtered subset and statistical power might have been too low to show differences for smaller IDs.

6. Overall Discussion

The results from both games confirmed that participants drawn from a campus population playing in a controlled lab environment with no environmental disturbances outperformed both naive and informed crowd participants playing in their own environment in touch tasks embedded in games when speed (Drop) or accuracy (Wall Destroyer) were required. 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 (e.g. posture and support while holding the device), 3. differences in task understanding, 4. motivational differences (naive crowd participants played more than the informed crowd-plus ones), or 5. a combination of these. 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 most likely did not feel as obliged to complete the game.

While Christensen et al in an earlier analysis of the dataset at hand [7] found a significant difference between the crowd and crowd-plus groups in their error probability in Wall Destroyer we did not find significant differences neither in the complete nor in the filtered data set. However, on average *error* probabilities were higher for the crowd than the crowd-plus participants and the statistical might have been insufficient to detect the differences given the large variance of the data. This was further supported by the crowd-plus group in the filtered data set having both a significant effect of target *size* on *errors* and a deviation of a sub-set slope from the overall slope which the crowd group did not yield. Future research needs to investigate these differences with more participants.

Crowd participants both naive and informed did not perform as consistently as the lab participants in Drop, which required fast reactions. Earlier work [7] resorted to removing entire participants (performing below group average) our outlier removal (absolute deviations from the median) in Wall Destroyer allowed for retaining data from participants in all groups. After the removal the performance in acquisition *time* subsequently did not differ significantly between groups. The groups still differed in their error probabilities, which vanished upon removal of participants below each group's median performance level. Future research needs to investigate how to best filter data coming from crowd participants.

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.

7. Conclusion

Absolute median filtering of outliers can be used to improve analysis of Fitts law data. Bèrard et al's approach of DHR through sub-slope analysis allowed for finding the limit for dragging on small touch screen devices from error probabilities but not from acquisition times. The device's human resolution for dragging with a finger onto non-occluded targets with no feedback of touch position on touch screens lies between 2mm and 4mm.

We found significant differences in performance between participants from the crowd (informed and naive) and the lab, with lab participants performing better than the crowd both in game that required fast reactions and one that required accuracy. However, when analysing only participants with above median performance of their respective population in a game requiring accuracy or each participant's best score in a game requiring quick reactions, both crowd groups performed just as well as lab participants.

Although we did not find significant differences in a game requiring accuracy the sub-analysis of informed crowd participants yielded more useful and expected statistical results than that of the naive crowd participants. Potentially, this was due to insufficient statistical power and large scale studies are required for verification of these results. However, researchers should be aware of potential higher performance from informed crowd participants when deciding whether to disclose the scientific purpose of a study or running it covert [36].

To reduce variability in performance or control for it crowdsourced studies ideally need to log a large amount of contextual data e.g. a device's placement and holding posture, finger angle, environmental disturbances e.g. when being on a bus from noise and shake. The effort required in adding these logging details need to be considered vis-a-vis a lab study in which researchers can control these factors. While crowdsourcing might reach a demographically different participants our informed crowd group was demographically not that dissimilar from our lab group.

References

[1] Arnett, J. (2008) The neglected 95%: Why American psychology needs to become less American. *Behavioral and brain sciences* **63**: 602–614.

[2] Benko, H., Wilson, A.D. and Baudisch, P. (2006) Precise Selection Techniques for Multi-Touch Screens. *Proc. of CHI'06* : 1263–1272.

[3] Bérard, F., Wang, G. and Cooperstock, J. (2011) On the limits of the human motor control precision: the search for a device's human resolution. *Proc. of INTERACT'11* **6947**: 107–122.

[4] Beringer, D.B. (1990) Target size, location, sampling point and instructional set: more effects on touch panel operation. In *Proc.of the Human Factors and Ergonomics Society Annual Meeting* (SAGE), **34**: 375–379.

[5] Beringer, D.B. and Peterson, J.G. (1985) Underlying behavioral parameters of the operation of touch-input devices: Biases, models, and feedback. *Human Factors: The Journal of the Human Factors and Ergonomics Society* **27**(4): 445–458.

[6] Brown, B., Weilenmann, A., McMillan, D. and Lampinen, A. (2016) Five Provocations for Ethical HCI Research. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16 (New York, NY, USA: ACM): 852–863. doi:10.1145/2858036.2858313.

[7] Christensen, A., Pedersen, S.A. and Knoche, H.O. (2016) Gamify HCI: Device's Human Resolution for Dragging on Touch Screens in a Game with Lab and Crowd Participants. In *Proc. of ArtsIT'16* (Esbjerg, Denmark: Springer).

[8] Cockburn, A., Ahlström, D. and Gutwin, C. (2012) Understanding performance in touch selections: Tap, drag and radial pointing drag with finger, stylus and mouse. *International Journal of Human-Computer Studies* **70**(3): 218–233.

[9] Colle, H.A. and Hiszem, K.J. (2004) Standing at a kiosk: Effects of key size and spacing on touch screen numeric keypad performance and user preference. *Ergonomics* **47**(13): 1406–1423. doi:10.1080/00140130410001724228.

[10] Deterding, S., Dixon, D., Khaled, R. and Nacke, L. (2011) From Game Design Elements to Gamefulness: Defining "Gamification". *Proc. of MindTrek'11* : 9–15.

[11] Fabian, G. (2012) Gamification: State of the art definition and utilization. *Research Trends in Media Informatics* **39**: 39–46.

[12] Fitts, P.M. (1954) The information capacity of the human motor system in controlling the amplitude of movement. *Journal of Experimental Psychology* **47**: 381–391.

[13] Fitts, P.M. and Peterson, J.R. (1964) Information capacity of discrete motor responses. *Journal of experimental psychology* **67**(2): 103.

[14] Flatla, D.R., Gutwin, C., Backe, L.E., Bateman, S. and Mandryk, R.L. (2011) Calibrating Games: Making Calibration Tasks Enjoyable by Adding Motivating Game Elements. *Proc. of UIST'11* : 403–412.

[15] Gächter, S. (2010) (Dis)advantages of student subjects: What is your research question? *Behavioral and brain sciences* **33**: 92–93.

[16] Hall, A.D., Cunningham, J.B., Roache, R.P. and Cox, J.W. (1988) Factors affecting performance using touch-entry systems: Tactual recognition fields and system accuracy. *Journal of applied psychology* **73**(4): 711–720.

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

[18] Henze, N. and Boll, S. (2011) It does not Fitts my data! Analysing large amounts of mobile touch data. *Proc. of INTERACT'11* : 564–567.

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

[20] Henze, N., Rukzio, E. and Boll, S. (2011) 100,000,000 Taps: Analysis and Improvement of Touch Performance in the Large. *Proc. of Mobile HCI'11* : 133–142.

[21] Higgins, J.J., Blair, R.C. and Tashtoush, S. (1990) The aligned rank transform procedure. In *Proc. of Applied Statistics in Agriculture*. URL http://newprairiepress.org/agstatconference/1990/proceedings/18/.

[22] Hoffmann, E.R. (2013) Which Version/Variation of Fitts' Law? A Critique of Information-Theory Models. *Journal of Motor Behavior* **45**(3): 205–215. doi:10.1080/00222895.2013.778815.

[23] Holz, C. and Baudisch, P. (2010) The generalized perceived input point model and how to double touch accuracy by extracting fingerprints. In *Proc. of CHI'10* (ACM): 581–590.

[24] Hwangbo, H., Yoon, S.H., Jin, B.S., Han, Y.S. and Ji, Y.G. (2013) A study of pointing performance of elderly users on smartphones. *Int'l J. of Human-Computer Interaction* **29**(9): 604–618.

[25] Kittur, A., Chi, E.H. and Suh, B. (2008) Crowdsourcing User Studies with Mechanical Turk. *Proc. of CHI'08* : 453–456.

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

[27] Leahy, M. and Hix, D. (1990) Effect of touch screen target location on user accuracy. In *Proc. of The Human Factors and Ergonomics Society Annual Meeting* (SAGE), **34**: 370–374.

[28] Leys, C., Ley, C., Klein, O., Bernard, P. and Licata, L. (2013) Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. *Journal of Experimental Social Psychology* **49**(4): 764–766. doi:10.1016/j.jesp.2013.03.013.

[29] MacKenzie, I.S. (1992) Fitts' law as a research and design tool in human-computer interaction. *Human-Computer Interaction Journal* **7**: 91–139.

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

[31] Out of Pixels (2015) Drop. https://play.google.com/store/apps/details?id=com.infraredpixel.drop. Accessed 17th March 2015.

[32] Perry, K.B. and Hourcade, J.P. (2008) Evaluating one handed thumb tapping on mobile touchscreen devices. In *Proc. of Graphics interface'08* (Canadian Information Processing Society): 57–64.

[33] Pielot, M., Henze, N. and Boll, S. (2011) Experiments in App Stores - How to Ask Users for their Consent? *Proc. of CHI'11* : 1–4.

[34] Sears, A. (1991) Improving touchscreen keyboards: design issues and a comparison with other devices. *Interacting with computers* **3**(3): 253–269.

[35] Wang, F. and Ren, X. (2009) Empirical Evaluation for Finger Input Properties in Multi-touch Interaction. In *Proc of CHI'09*, CHI '09 (New York, NY, USA: ACM): 1063–1072.

[36] Williamson, J.R. and Sundén, D. (2015) Deep Cover HCI: A Case for Covert Research in HCI. In *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems*, CHI EA '15 (New York, NY, USA: ACM): 543–554. doi:10.1145/2702613.2732500.

[37] Williamson, V. (2014) On the Ethics of Crowd-Sourced Research. http://scholar.harvard.edu/files/williamson/files/mturk_ps_081014.pdf. Accessed 9th March 2015.

[38] Zhai, S., Kong, J. and Ren, X. (2004) Speed–accuracy tradeoff in Fitts' law tasks—on the equivalency of actual and nominal pointing precision. *International Journal of Human-Computer Studies* **61**(6): 823–856. doi:10.1016/j.ijhcs.2004.09.007.

[39] Zichermann, G. and Cunningham, C. (2011) *Gamification by design: Implementing game mechanics in web and mobile apps* (O'Reilly Media, Inc.).