# Evaluating Fitts' Law Performance With a Non-ISO Task

Maria Francesca Roig-Maimó
University of Balearic Islands
Dept. of Mathematics and Computer Science
Palma, Spain
xisca.roig@uib.es

I. Scott MacKenzie
York University
Dept. of Electrical Engineering and Computer Science
Toronto, Canada
mack@cse.yorku.ca

Cristina Manresa-Yee
University of Balearic Islands
Dept. of Mathematics and Computer Science
Palma, Spain
cristina.manresa@uib.es

Javier Varona
University of Balearic Islands
Dept. of Mathematics and Computer Science
Palma, Spain
xavi.varona@uib.es

## ABSTRACT

We used a target-selection task to evaluate head-tracking as an input method on a mobile device. The procedure used a non-ISO Fitts' law task since targets were randomly positioned from trial to trial. Due to a non-constant amplitude within each sequence of trials, throughput was calculated using two methods of data aggregation: by sequence of trials using the mean amplitude and by common *A-W* conditions. For each data set, we used four methods for calculating throughput. The grand mean for throughput calculated using the division of means and the adjustment for accuracy was 0.74 bps, which is 45% lower than the value obtained using an ISO task. We recommend calculating throughput using the division of means (and not the slope reciprocal from the regression model) and with the adjustment for accuracy. We present design recommendation for non-ISO tasks: Keep amplitude and target width constant within each sequence of trials and use strategies to avoid or remove reaction time.

## CCS CONCEPTS

• **Human-centered computing → User studies**; *HCI theory, concepts and models*;

## KEYWORDS

Fitts' law, throughput, ISO 9241-411, mobile HCI, head-tracking

**ACM Reference format:**
Maria Francesca Roig-Maimó, I. Scott MacKenzie, Cristina Manresa-Yee, and Javier Varona. 2017. Evaluating Fitts' Law Performance With a Non-ISO Task. In *Proceedings of Interacción '17, Cancun, Mexico, September 25–27, 2017,* 8 pages.
DOI: 10.1145/3123818.3123827


*Interacción '17, Cancun, Mexico*

DOI: 10.1145/3123818.3123827

## 1 INTRODUCTION

Fitts' law [3] describes the relationship between movement time, movement distance, and selection accuracy for people engaged in rapid-aimed movements. In Human-Computer Interaction (HCI), Fitts' law applies to pointing and dragging using input devices. Since Fitts' original publication, the law has proven one of the most robust, highly cited, and widely adopted models to emerge from experimental psychology [8].

Since 1954, when Fitts' law was originally introduced, the model has been refined in primarily three ways: in its mathematical formulation, in the accommodation of the distribution of movement end-points, and in the means to calculate input device *throughput*. However, these refinements were not universally adopted and as a result multiple variations of the law have been applied [13]. Inconsistency is common and this weakens between-study comparisons based on throughput.

As an effort to bring consistency, the ISO 9241-9 standard was published in 2002 [6]. ISO 9241-9 describes performance tests for evaluating human performance of non-keyboard input devices. The standard was updated in 2012 as ISO 9241-411 [7].[1] Despite standardization efforts, inconsistency in the calculation of throughput is still common in the HCI literature.

In the field of mobile devices, this inconsistency is even more problematic. Due to the particularities of small displays or to the motivation of the research, it may not be suitable to use the tests described in the ISO standard. For that reason, researchers sometimes design a custom Fitts' law task. However, every custom task brings its own challenges for calculating throughput, which may again lead to inconsistent values of throughput.

We present the Face Me experiment, which was not initially conceived for Fitts' law analysis. The experiment used a non-ISO task specifically designed according to the motivation of the study: testing target selection over the entire display surface using a novel head-tracking method. However, as the study involved point-select movements over a range of amplitudes and target widths, it may be possible to calculate troughput with the Face Me experiment data. We detail the Face Me task and its challenges for calculating throughput, jointly with approaches to overcome these challenges.

[1] With respect to performance evaluation, the two versions of the standard are the same.

The contribution of the present research is to identify situations where an empirical evaluation involving point-select tasks uses, by necessity and by design, tasks that do not conform to those described in ISO 9241-411. We demonstrate how to calculate Fitts' throughput — the dependent measure defined in ISO 9241-411 — and identify potential problems that arise due to the task properties. This is followed with recommendations for designing non-ISO tasks to obtain valid and consistent throughput values.

## 2 FITTS' LAW AND THE CALCULATION OF THROUGHPUT

In the field of Human-Computer Interaction (HCI), Fitts' law has been applied in mainly two ways, as a predictive model, and as a means to derive the dependent measure throughput (Fitts' *index of performance*) as part of the comparison and evaluation of pointing devices.

The calculation of throughput (*TP*) is performed over a range of movement amplitudes or distances and with a set of target widths involving tasks for which computing devices are used. The primary tests in ISO 9241-411 involve point-select tasks using either a one-dimensional (1D) or multi-directional (2D) task (see Figure 1). Although user performance is typically evaluated with multiple dependent variables (e.g., speed, accuracy), the only performance measure stipulated in the ISO standard is throughput.

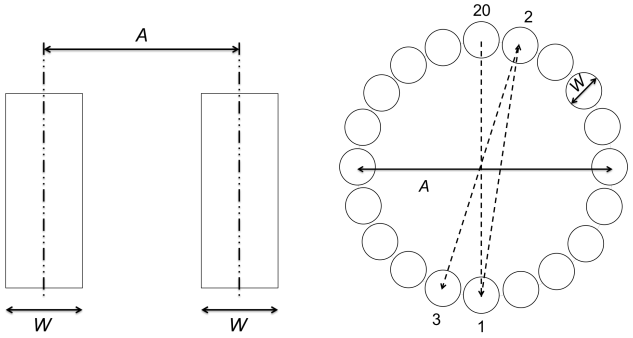


**Figure 1: ISO tests for pointing evaluation: (a) one-directional point-select test (b) multi-directional point-select test.**

Throughput (*TP*) is calculated as follows:

$$TP = \frac{\text{Effective index of difficulty}}{\text{Movement time}} = \frac{ID_e}{MT}, \quad (1)$$

where $ID_e$ is computed from the movement amplitude ($A$) and target width ($W$) and $MT$ is the per-trial movement time averaged over a sequence of trials. Because $ID_e$ has units "bits" and $MT$ has units "seconds", the units for throughput are "bits per second" or "bps".

The effective index of difficulty ($ID_e$) is a measure of the difficulty and user precision in accomplishing a task:

$$ID_e = log_2\left(\frac{A}{W_e} + 1\right), \quad (2)$$

where $W_e$ is the effective target width, calculated from the width of the distribution of selection coordinates made by a participant over a sequence of trials. The effective target width is calculated as follows:

$$W_e = 4.133 \cdot S_x, \quad (3)$$

where $S_x$ is the standard deviation of the selection coordinates in the direction that movement proceeds. Alternately, if the standard deviation of the selection coordinates is unavailable, the error rate may be used to approximate the effective value,

$$W_e = \begin{cases} W \cdot \frac{2.066}{z(1-error/2)} & \text{if } error > 0.0049\% \\ W \cdot 0.5089 & \text{otherwise} \end{cases} \quad (4)$$

where *error* is the error rate in the sequence of trials, and $z(x)$ is the inverse of the standard normal cumulative distribution; that is, the $z$-score corresponding to the point where the area under the normal curve is $x\%$ of 1. Assuming a normal distribution in the selection coordinates, $W_e < W$ if the error rate < 4% and $W_e > W$ if the error rate > 4%.

The effective value is used to include spatial variability in the calculation. The effective amplitude ($A_e$) can also be used if there is an overall tendency to overshoot or undershoot [13].

Using the effective values, throughput is a single human performance measure that embeds both the speed and accuracy in human responses. The index of difficulty calculated without the effective values (*ID*) quantifies the movement task that the experimenter wants subjects to perform; however, subjects often do not actually perform at this index of difficulty. The difference between *ID* and $ID_e$ is a natural consequence of motivated subjects' desire to perform well. However, a large discrepancy may indicate that the movement tasks were extremely easy, extremely hard, or simply ill-suited to the conditions under investigation [13].

Despite standardization efforts, inconsistency in the calculation of throughput remains common in the HCI literature and this weakens between-study comparisons. A point of particular contention is the interpretation of throughput as the slope reciprocal ($1/b$) from the regression equation. Although $1/b$ has units "bits per second", this term cannot be used as a dependent variable in experimental research because of the wavering influence of the intercept, $a$, which is absent in $1/b$. Throughput calculated as $1/b$ will be similar to the value calculated via Equation 1 only if the intercept $a$ is 0, or close to 0. It is worth noting that Fitts originally defined throughput not as the slope reciprocal but as a division of means: "The average rate of information generated by a series of movements is the average information per movement divided by the time per movement" [3, p. 390].

A detailed description of the calculation of throughput is found in other sources [9, 13].

## 3 RELATED WORK

Below we provide two detailed examples of the inconsistency in the calculation of throughput. We focus specifically on experiment design issues impacting the calculation of throughput.

Perry and Hourcade [10] analyzed one-handed thumb tapping on mobile touchscreen devices. The study included five different target sizes and 25 unique positions that intersect a four-by-four grid

which divided the screen of the mobile device. They used a custom task for the experiment where every combination of position (25 unique positions) and size (five different target sizes) appeared randomly for each sub-block (two sub-blocks). The start target was positioned below the targets unless the destination target was in the bottom two rows of the four-by-four grid; in this case, the start target was above the destination target. The start target was always positioned 27.9 mm away from the destination target. With this custom task, the amplitude is constant over all the trials ($A$ = 27.9 mm) and the target width ($W$) varies from trial to trial within each sequence of trials. They calculated the index of performance (throughput) using the index of difficulty ($ID$) instead of the effective index of difficulty ($ID_e$) applying the formula $TP = \frac{ID}{MT}$. Thus, the measure does not embed the accuracy of users' responses. They reported a throughput around 4 bps.

Henze and Boll [5] analyzed the touch behavior for smartphones using the game *Hit It!*, published on the Android Play Store. They designed part of the game as a Fitts' law task. Multiple circles of different sizes were displayed and the player sequentially touched the targets. As soon as a target was successfully touched, it disappeared. The player must touch all targets in a certain timeframe. The game was lost if the timeframe expired or if the player missed three targets. When the player touched the first circle, the touched position and the current time were used as the start for the first task. Amplitude ($A$) was the distance from the position of the first touch to the center of the second circle and width ($W$) was the diameter of the second circle. With this custom task, $A$ and $W$ are not constant within each sequence of trials. To apply Fitts' law, they computed the least-squares prediction equation as $MT = a + b \cdot log_2\left(\frac{A}{W} + 1\right)$ and calculated throughput as $1/b$. They reported an implausibly high value for throughput in the range of 16 to 25 bps. As the prediction equation had a low correlation, they argued that the high throughput suggests that the task might be different from what is commonly used as a Fitts' law task (recommended by ISO 9241-411) and the low correlation shows that Fitts' law is not a good model for the tasks employed.

The two studies just cited are examples of the disparate ways in which throughput is calculated in the Fitts' law literature. Although numerous other examples exist, space precludes a detailed review. In the following section, we present an example of a user study where a non-ISO point-select task was used. We then examine issues concerning the calculation of throughput.

## 4 THE FACE ME EXPERIMENT

The Face Me experiment used a mobile head-tracking interface to investigate the effect of device orientation (portrait, landscape), gain (1.0, 1.5), and target width (88 pixels, 176 px, 212 px). In the initial poster [12], the evaluation was limited to selection accuracy, cursor velocity, and selection errors.

### 4.1 Participants

Nineteen unpaid participants (four females) were recruited from the local town and university campus from an age group of 23 to 69. The average age was 38.2 years ($SD$ = 14.1). None of the participants had previous experience with head-tracking interfaces.

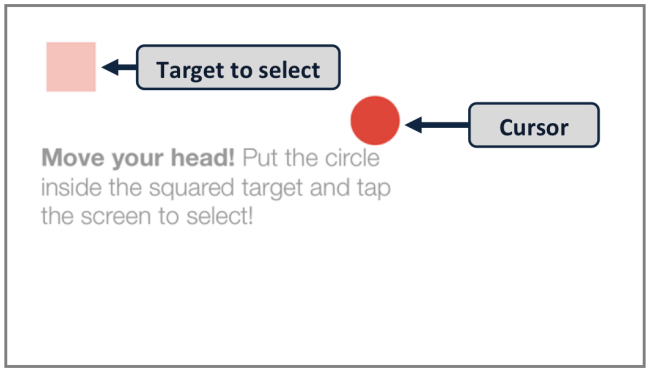


**Figure 2: Example of a target condition ($W$ = 88 px, orientation = landscape) with annotations and the procedure instructions (task belonging to the practice sequence).**

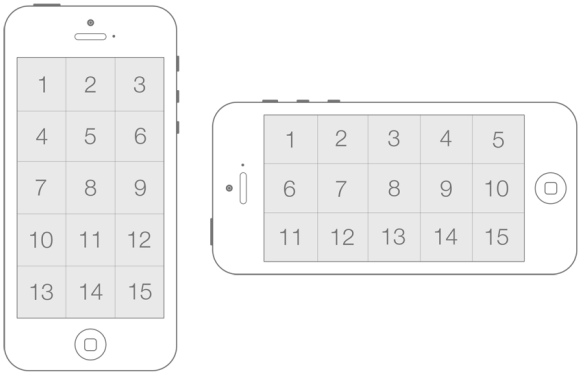


**Figure 3: Regions of the display surface in portrait and landscape orientation. (Numbers added to identify regions.)**

### 4.2 Apparatus and Experiment Task

The experiment was conducted on an Apple *iPhone 5* with a resolution of 640 × 1136 px and a pixel density of 326 ppi. This corresponds to a resolution of 320 × 568 Apple points.[2]

The experiment involved a point-select task that required positioning a circle cursor inside a square target (see Figure 2). User input combined mobile head-tracking for pointing and touch for selection. Selection occurred by tapping anywhere on the display surface with a thumb. The target was highlighted in green when the center of the cursor was inside the target; a selection performed in these conditions was considered successful.

The experiment sought to determine if all regions of the device screen were accessible for users. For this purpose, the display surface was divided into 15 regions of approximately 213 × 227 px (see Figure 3). The targets were centered inside the regions.

The task was implemented in both portrait and landscape (right) orientations.

[2]Apple's point (pt) is an abstract unit that covers two pixels on retina devices. On the *iPhone 5*, one point equals 1/163 inch (Note: 1 mm ≈ 6 pt).

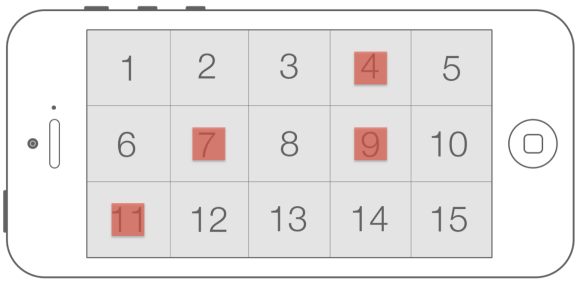


**Figure 4: Random example of the first four trials of a possible sequence of trials with targets placed inside their regions (selection order: 7, 9, 4, 11) of a target condition ($W$ = 88 px, orientation = landscape). Note that only one target is visible for each trial (numbered regions and already selected targets added for clarification purposes).**

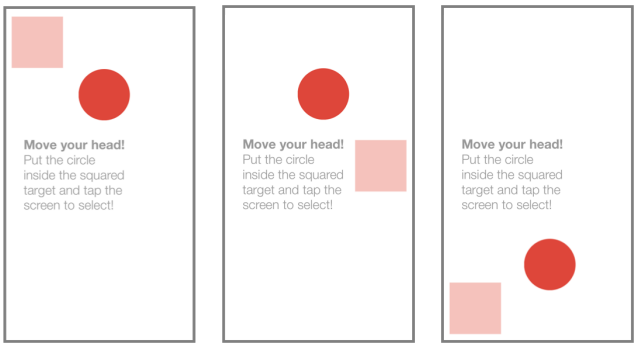


**Figure 5: Example of the practice sequence for the target condition.**

A sequence of trials consisted of 15 target selections, one for each of the 15 regions of the display, presented randomly and without replacement. Upon selection, a new target appeared centered inside one of the remaining regions. Selections proceeded until all 15 regions of the display were used as target centers (see Figure 4).

According to the iOS Human Interface Guidelines [1], the optimal size of a tappable UI element on the *iPhone* is 44 × 44 pt, which is equivalent to the minimim level chosen for target width: 88 × 88 px. The maximum level for target width was limited by the size of the screen regions (approximately 213 × 227 px).

For each user session, all twelve conditions were used and presented in random order until all trials were completed.

To adapt to each new condition, participants were required to correctly select a practice series of three targets to start the sequence of trials (see Figure 5). The practice series was not registered as experiment data.

(a) Portrait orientation (b) Landscape orientation

**Figure 6: Participants performing the experiment: holding the device in (a) portrait orientation and in (b) landscape orientation. Moving the cursor by moving the head and selection by tapping anywhere on the display surface with a thumb.**

### 4.3 Procedure

After signing a consent form, participants were briefed on the goals of the experiment and were instructed to sit and hold the device (in the orientation indicated by the software) in a comfortable position (see Figure 6). The only requirement was that their entire face was visible by the front camera of the device.

The experiment task was demonstrated to participants, after which they did a few practice sequences. They were instructed to move the cursor by holding the device still and moving their head. Participants were asked to select targets *as quickly and as close to the center as possible*. They were allowed to rest as needed between sequences. Testing lasted about 15 minutes per participant.

### 4.4 Design

The experiment was fully within-subjects with the following independent variables and levels:

- Orientation: portrait, landscape
- Gain: 1.0, 1.5
- Width: 88, 176, 212 px

The total number of trials was 19 Participants × 2 Orientations × 2 Gains × 3 Widths × 15 Trials = 3420.

A detailed description of the Face Me experiment and the results obtained is found elsewhere [12].

## 5 FITTS' LAW ANALYSIS (FACE ME EXPERIMENT DATA)

The Face Me experiment was not initially conceived for Fitts' law analysis. But, as the study involved point-select movements over

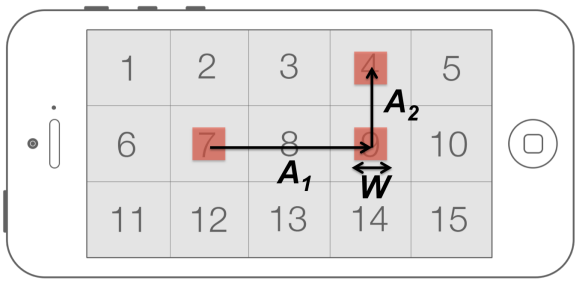


**Figure 7: Definition of *A* and *W* for a random example of the first three trials of a possible sequence of trials (selection order: 7, 9, 4) for a target condition (*W* = 88 px, orientation = landscape).**

a range of amplitudes and target widths, it may be possible to calculate throughput with the experiment data. This is examined below.

A sequence of trials was formed by 15 target selections, with all targets of the same width but presented in a random region of the displace surface. Therefore, target width (*W*) was constant within each sequence of trials, but amplitude (*A*) varied.

To apply Fitts' law to the Face Me task, we defined for each trial *W* as the width of the current target, *MT* as the time between selection of the previous target and selection of the current one, and *A* as the distance from the center of the previous target to the center of the current target (see Figure 7). We considered the time when the first target was selected as the beginning of the sequence; therefore, the first target of every sequence was discarded for the analysis.

The sequential appearance of targets in random regions creates two main issues for the calculation of throughput:

(1) Movement time might include a reaction time, and
(2) Amplitude is not constant from trial to trial within a sequence.

Methods to accommodate these issues are explored below.

## 5.1 Reaction Time

Fitts' law is a model for point-select movement tasks. The task should not involve a reaction time or, if it does, the reaction time should be excluded from the task completion time [4]. For the Face Me experiment task, each target was randomly positioned and appeared only when the preceding target was selected. Thus, the task likely included a reaction time component. Arguably, this should be removed from the task completion time.

It is known that that time to react to a visual stimulus is about 200 ms [2, p. 41]. Thus, one way to remove reaction time for the Face Me task is to reduce the movement time for each trial by about 200 ms. The effect of doing so is included in our discussions of issue #2.

**Table 1: Aggregation by sequence of trials: comparison of throughputs**

| Condition | Throughput (bps) $ID/MT$ | $ID_e/MT$ |
|---|---|---|
| Portrait (gain = 1.0) | 0.90 | 0.82 |
| Portrait (gain = 1.5) | 0.89 | 0.77 |
| Landscape (gain = 1.0) | 0.89 | 0.82 |
| Landscape (gain = 1.5) | 0.92 | 0.73 |
| *Mean* | 0.90 | 0.78 |

## 5.2 Non-constant Amplitude

Since targets were randomly positioned for each trial in the Face Me task, the amplitude of movements varied from trial to trial. The amplitude might be small (e.g., neighboring targets) or large (e.g., targets on opposite corners of the display surface). To overcome the non-constant amplitude issue, we explored two methods of trial aggregation: (1) aggregation by trial sequence with *A* equal to the mean amplitude of movement over the trials in the sequence, and (2) aggregation by common *A*-*W* conditions. Each of these methods is examined below.

*5.2.1 Aggregation by Sequence.* Using this approach, we calculated throughput within each sequence of trials, defining *A* as the average amplitude for all trials in the sequence.

After discarding the first trial, a sequence of trials was formed by 14 target selections. For each of the 19 participants, we had 12 sequences of trials (one sequence per condition). Therefore, we had 228 sequences of trials (19 Participants × 2 Orientations × 2 Gains × 3 Widths). Although we had potentially 228 different, but very similar, amplitudes, it makes no sense to construct a Fitts' law regression model, as the index of difficulties are the same, or similar.

Since the standard deviations of the selection coordinates were available, we calculated $W_e$ using the standard deviation method (Equation 3).

The grand mean for throughput using $ID_e$ was 0.78 bps (see Table 1). By orientation, the means were 0.79 bps (portrait) and 0.77 bps (landscape). The difference was not statistically significant ($F_{1,18} = 0.43$, ns). By gain, the means were 0.82 bps (1.0) and 0.75 bps (1.5). The difference was statistically significant ($F_{1,18} = 9.19$, $p < .01$).

The grand mean for throughput using *ID* was 0.90 bps. For both independent variables (orientation, gain), the difference between the means was not statistically significant (ns).

If we reduce the movement time by 200 ms for each trial, to remove reaction time (see Section 5.1), throughput increases by about 10%, to 0.87 bps (or to 1.00 bps without using effective values).

Regression models were not built using the aggregation by sequence approach. The reason, as noted above, is that the *ID*s were all the same, or similar, since the amplitude for each sequence was the mean amplitude computed across the trials in the sequence.

*5.2.2 Aggregation by A-W Condition.* Using this approach, we calculated throughput across sequences of trials by aggregating on equivalent *A*-*W* conditions.

**Table 2: Examples of amplitudes and adjusted amplitudes (see Figure 3 for region identification)**

| Orientation | Target region | | Difference | Amplitude (px) | |
|---|---|---|---|---|---|
| | Start | End | in regions | Original | Adjusted |
| Landscape | 8 | 3 | 1 | 213 | 220 |
| Landscape | 11 | 12 | 1 | 227 | 220 |
| Landscape | 7 | 9 | 2 | 454 | 440 |
| Landscape | 12 | 2 | 2 | 426 | 440 |

The randomization of target positions yielded a large number of different amplitudes, which was worsened by the slightly off-square size of the original display regions (213 × 227 px).

To achieve a practicle number of index of difficulty conditions, and because the aspect ratio of the display regions was close to 1, we adjusted the task amplitudes according to the number of regions between the start and end targets instead of the difference in coordinates. Therefore, we considered an adjusted squared region size of 220 × 220 px (220 px = average size of the two dimensions of the original region), so the adjusted amplitude between the start and end targets separated by the same number of regions was equivalent. As an example and with reference to Figure 3, a task between between regions 8 and 3 in landscape orientation (vertical movement, difference in regions = 1) had an original amplitude of 213 px and an adjusted amplitude of 220 px (*new region dimension* × 1). For a task between regions 11 and 12 in landscape orientation (horizontal movement, difference in regions = 1), the original amplitude was 227 px, which is different from the amplitude of the vertical movement across region; but, with the adjusted amplitude (220 px) both tasks had the same amplitude value (see Table 2).

With the adjustment, there were 11 different amplitudes. Therefore, the number of index of difficulty conditions was 33 (11 Amplitudes × 3 Widths). Consequently, aggregation by *A*-*W* condition produced 33 *A*-*W* conditions, each with a different number of trials and from different participants.

As trials for each *A*-*W* condition were formed by tasks from different participants, it makes no sense to calculate $W_e$ using the standard deviation method. Instead, $W_e$ was calculated using the discrete error method. See Equation 4.

The grand mean for throughput was 0.70 bps using $ID_e$ and 0.87 bps using *ID* (see Table 3). In both cases and for both independent variables (orientation, gain), the difference between the means was not statistically significant.

We constructed two Fitts' law regression models for each of the four experiment conditions (2 Orientations × 2 Gains) by using the 33 combinations of movement amplitudes and target widths. One model used the effective index of difficulty ($ID_e$) and the other used the presented index of difficulty (*ID*). As a matter of demonstration, throughput was also calculated as $1/b$ from the regression equations. See Table 3. Note that for all eight models throughput calculated using $1/b$ (columns 8 and 9) is higher than the corresponding throughput calculated using the division of means (columns 10 and 11). This is due to the presence of a positive intercept in the models. A positive intercept lowers the slope coefficient ($b$) and, hence, increases $1/b$. Note also in right two columns (10 and 11)

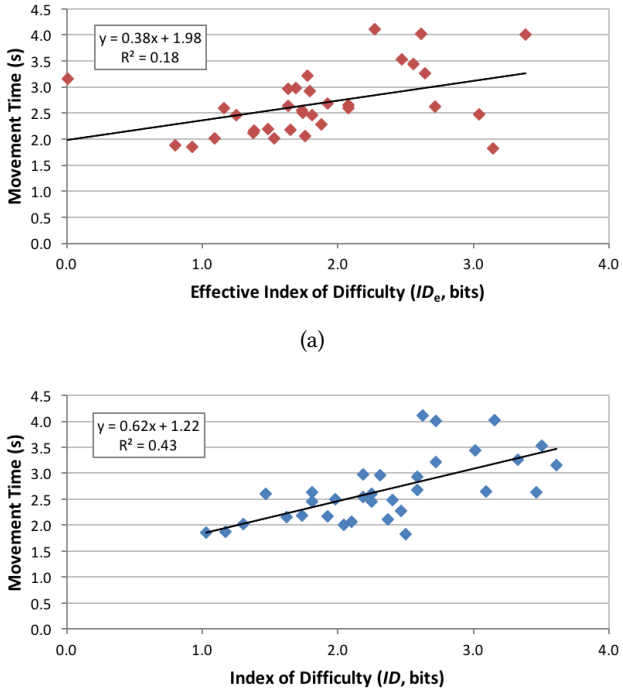


**Figure 8: Regression models for the experiment condition *orientation* = landscape, *gain* = 1. (a) Using $ID_e$, and (b) using *ID*. See text for discussion on the left-most point in (a).**

that throughput is lower when computed using $ID_e$ versus *ID*. This occurs because the $ID_e$ values overall were lower than the *ID* values (see Figure 8). With a lower numerator, the quotient (throughput) is naturally lower.

As mentioned above, each *A*-*W* condition had a different number of trials. The variation in the number of trials could produce a bias. In particular, sequences with a low number of trials have a disproportionately high influence, both in calculating the mean over *A*-*W* conditions and in building a regression model. As an example, Figure 8 shows the two scatter plots and regression models for the experiment condition *orientation* = landscape, *gain* = 1.0. Besides the low correlation in both models, an extreme case appears for the condition $A$ = 492.8 px and $W$ = 44 px. Only two trials had this combination of $A$ and $W$. Furthermore, both trials resulted in errors; thus, $ID_e$ was close to 0. This is seen at the left-most point in Figure 8a.

As in the previous approach, the values of throughput could be recalculated by removing the reaction time from the movement time. This has a similar effect of increasing the throughput values by about 10%.

## 6 DISCUSSION

In the previous sections, two methods of aggregation were applied to overcome the non-constant amplitude for calculating throughput:

**Table 3: Aggregation by *A-W* condition: comparison of regression models and throughputs**

| Condition | *MT-ID* regression | | | *MT-ID*$_e$ regression | | | Throughput (bps) | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | *a* (s) | *b* (s/bit) | $R^2$ | *a* (s) | *b* (s/bit) | $R^2$ | 1/*b* (*MT-ID*) | 1/*b* (*MT-ID*$_e$) | *ID*/*MT* | *ID*$_e$/*MT* |
| Portrait (gain = 1.0) | 0.52 | 0.97 | 0.74 | 1.37 | 0.72 | 0.50 | 1.03 | 1.39 | 0.85 | 0.71 |
| Portrait (gain = 1.5) | 0.04 | 1.18 | 0.70 | 1.65 | 0.64 | 0.20 | 0.85 | 1.56 | 0.86 | 0.68 |
| Landscape (gain = 1.0) | 1.22 | 0.62 | 0.43 | 1.98 | 0.38 | 0.18 | 1.61 | 2.63 | 0.88 | 0.70 |
| Landscape (gain = 1.5) | 0.68 | 0.86 | 0.50 | 1.94 | 0.41 | 0.17 | 1.16 | 2.44 | 0.89 | 0.70 |
| | | | | | | *Mean* | 1.16 | 2.00 | 0.87 | 0.70 |

**Table 4: Summary of four methods for calculating throughput (bps) by aggregation method**

| Method | 1/*b* | | Division of means | |
|---|---|---|---|---|
| | *ID* | *ID*$_e$ | *ID* | *ID*$_e$ |
| Aggregation by sequence | - | - | 0.90 | 0.78 |
| Aggregation by *A-W* | 1.16 | 2.00 | 0.87 | 0.70 |
| | | *Mean* | 0.89 | 0.74 |

aggregation by sequence and aggregation by *A-W* condition. In the case of aggregation by sequence (see Table 4, first row), throughput was calculated two ways:

(1) using the division of means using *ID* (column 4), and
(2) using the division of means using *ID*$_e$ (column 5).

In the case of aggregation by *A-W* condition (see Table 4, second row), throughput was calculated four ways:

(1) using the slope reciprocal using *ID* (column 2),
(2) using the slope reciprocal using *ID*$_e$ (column 3),
(3) using the division of means using *ID* (column 4), and
(4) using the division of means using *ID*$_e$ (column 5).

Although these different calculations (and others) are all represented in the Fitts' law literature, the values in the right-hand column in Table 1, Table 3, and Table 4 are the most realistic. They were computed both using the division of means and using the effective value for the index of difficulty (to include accuracy). This is the method of calculation stipulated in ISO 9241-411 (see also [13] for a further discussion on the calculation of throughput).

As noted earlier, the interpretation of throughput as the slope reciprocal should be avoided because it does not comport with the original definition given by Fitts and also because of the wavering influence of the intercept. In Table 3, the effect of the intercept (*a*) on throughput is apparent when using the division of means and when using the slope reciprocal. In cases where the intercept is close to 0, both values are similar (see Table 3, row: portrait (gain = 1.5), columns: 2, 8, and 10). As the intercept increases, the difference between values also increases (see Table 3, row: landscape (gain = 1.0), columns: 2, 8, and 10). Therefore, henceforth we focus on throughput calculated using the division of means.

For both methods of aggregation, the throughput values were higher when calculated using *ID* compared to *ID*$_e$. The increases were 15% (by sequence) and 25% (by *A-W* condition). This occurred because participants overall made > 4% errors (with a corresponding increase in $SD_x$). Thus, the *ID*$_e$ values were slightly lower than the *ID* values and this tends to lower throughput when computed using the division of means. In summary, throughput calculated including the adjustment for accuracy is the most realistic approach because it embeds both the speed and accuracy in human responses. Therefore, the calculation of throughput should use the effective values whenever possible.

Since the Face Me experiment evaluated a novel head-tracking interaction method and there was a specific motivation to test target selection over the entire display surface, the Face Me experiment did *not* use tasks following ISO 9241-411. The grand mean for throughput using the non-ISO task was 0.74 bps (or around 0.81 bps if throughput is recalculated by removing the reaction time from the movement time). This value is low compared to the value of throughput of 1.42 bps obtained for the same head-tracking interaction method using the ISO task [11]. The increase of throughput using the ISO task, together with the low correlation values obtained in the regression models, may suggest that the Face Me task is ill-suited for the evaluation of Fitts' law performance.

The Face Me task presented two main issues for the evaluation of Fitts' law performance: (1) reaction time was included in the movement time, and (2) the amplitude varied within each sequence of trials . These two issues resulted from the sequential appearance of targets in random regions. We now present design recommendations for non-ISO tasks used in the evaluation of Fitts' law performance.

(1) Amplitude (*A*) and target width (*W*) should be constant within each sequence of trials.
(2) Each *A-W* condition (sequence of trials) should be presented in a randomized order and with as many repetitions as the experimental procedure allows.
(3) Throughput should be calculated on each sequence of trials (and not on a single trial or on trials aggregated across sequences).
(4) Movement time should not include reaction time (nor dwell time nor homing time). A good practice to avoid reaction time might be to dimly highlight the target to select following the current target (see Figure 9).

## 7 CONCLUSION

We presented the Face Me experiment, which evaluated a novel head-tracking interaction method with a specific motivation to test target selection over the entire display surface. Therefore, the Face Me experiment did *not* use tasks following ISO 9241-411. Despite using a non-ISO task, we evaluated Fitts' law performance using

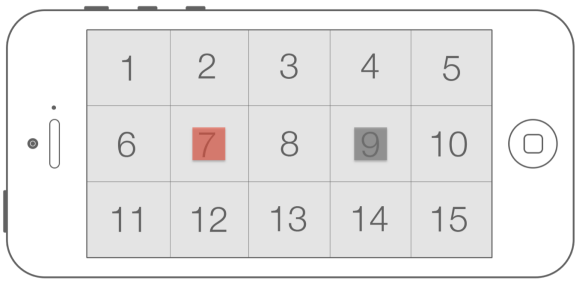


**Figure 9: Example of dimly highlighting the target to select following the current target to avoid reaction time. Target to select = 7, next target to select = 9 (highlighted in gray).**

the Face Me experiment data, trying to overcome the challenges brought forth by the custom task.

We recommend calculating throughput by the division of means and with the adjustment for accuracy, with the latter grounded on the benefits of including spatial variability in human responses. The grand mean for throughput was 0.74 bps, which is 45% lower than the value obtained using an ISO task. We conclude that the low value of throughput, together with the low correlation values obtained in the regression models, may suggest that the Face Me task was ill-suited for the evaluation of Fitts' law performance.

Hence, we presented design recommendations for non-ISO tasks: Keep amplitude and target width constant within each sequence of trials and use strategies to avoid or remove reaction time.

## ACKNOWLEDGMENTS

This work has been partially supported by the project TIN2016-81143-R (AEI/FEDER, UE), grant BES-2013-064652 (FPI) and mobility grant EEBB-I-15-10293. Finally, we also acknowledge the support of the "Visiting lecturers on short visits" programme by the OSR (University of Balearic Islands).

## REFERENCES

[1] Apple Inc. 2014. iOS human interface guidelines: Designing for iOS. (Nov. 2014). Retrieved November 11, 2014 from https://developer.apple.com/library/ios/documentation/userexperience/conceptual/mobilehig/

[2] R. W. Bailey. 1996. *Human performance engineering: Designing high quality professional user interfaces for computer products, applications and systems* (3 ed.). Prentice Hall, Upper Saddle River, NJ.

[3] P. M. Fitts. 1954. The information capacity of the human motor system in controlling the amplitude of movement. *Journal of Experimental Psychology* 47, 6 (1954), 381.

[4] P. M. Fitts and J. R. Peterson. 1964. Information capacity of discrete motor responses. *Journal of Experimental Psychology* 67, 2 (1964), 103–112.

[5] N. Henze and S. Boll. 2011. It does not Fitts my data! Analysing large amounts of mobile touch data. In *Proceedings of the IFIP Conference on Human-Computer Interaction - INTERACT 2011*. Springer, Berlin, 564–567.

[6] ISO. 2002. 9241–9. 2000. Ergonomics requirements for office work with visual display terminals (VDTs) – Part 9: Requirements for non-keyboard input devices. *International Organization for Standardization* (2002).

[7] ISO. 2012. 9241–411. 2012. Ergonomics of human-system interaction – Part 411: Evaluation methods for the design of physical input devices. *International Organization for Standardization* (2012).

[8] I. S. MacKenzie. 1992. Fitts' law as a research and design tool in human-computer interaction. *Human-Computer Interaction* 7, 1 (1992), 91–139.

[9] I. S. MacKenzie. 2015. Fitts' throughput and the remarkable case of touch-based target selection. In *Proceedings of the 17th International Conference on Human-Computer Interaction - HCII 2015*. Springer, Switzerland, 238–249.

[10] K. B. Perry and J. P. Hourcade. 2008. Evaluating one handed thumb tapping on mobile touchscreen devices. In *Proceedings of the Graphics Interface 2008 - GI 2008*. Canadian Information Processing Society, Toronto, 57–64.

[11] M. F. Roig-Maimó, I. S. MacKenzie, C. Manresa-Yee, and J. Varona. Submitted. Head-tracking interfaces on mobile devices: Evaluation using Fitts' law and a new multi-directional corner task for small displays. *International Journal of Human-Computer Studies* (Submitted).

[12] M. F. Roig-Maimó, J. Varona Gómez, and C. Manresa-Yee. 2015. Face Me! Head-tracker interface evaluation on mobile devices. In *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems - CHI 2015*. ACM, New York, 1573–1578.

[13] R. W. Soukoreff and I. S. MacKenzie. 2004. Towards a standard for pointing device evaluation: Perspectives on 27 years of Fitts' law research in HCI. *International Journal of Human-Computer Studies* 61, 6 (2004), 751–789.