The Perils of Confounding Factors: How Fitts’ Law Experiments can Lead to False Conclusions
Julien Gori, Olivier Rioul, Yves Guiard, Michel Beaudouin-Lafon

To cite this version:
Julien Gori, Olivier Rioul, Yves Guiard, Michel Beaudouin-Lafon. The Perils of Confounding Factors: How Fitts’ Law Experiments can Lead to False Conclusions. CHI’18, Apr 2018, Montréal, Canada. 10.1145/3173574.3173770. hal-01690122

HAL Id: hal-01690122
https://hal.archives-ouvertes.fr/hal-01690122
Submitted on 22 Jan 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
The Perils of Confounding Factors: How Fitts’ Law Experiments can Lead to False Conclusions

Julien Gori\textsuperscript{1} Olivier Rioul\textsuperscript{1} Yves Guiard\textsuperscript{2,1} Michel Beaudouin-Lafon\textsuperscript{2}

\textsuperscript{1}LTCI, Telecom ParisTech
Université Paris-Saclay
F-75013, Paris, France

\textsuperscript{2}LRI, Univ. Paris-Sud, CNRS,
Inria, Université Paris-Saclay
F-91400, Orsay, France

\{julien.gori,olivier.rioul,yves.guiard\}@telecom-paristech.fr, mbl@lri.fr

ABSTRACT
The design of Fitts’ historical reciprocal tapping experiment gravely confounds index of difficulty ID with target distance D: Summary statistics for the candidate Fitts model and a competing model may appear identical, and the validity of Fitts’ model for some tasks can be legitimately questioned. We show that the contamination of ID by either target distance D or width W is due to the common practices of pooling and averaging data belonging to different distance-width (D,W) pairs for the same ID, and taking a geometric progression for values of D and W. We analyze a case study of the validation of Fitts’ law in eye-gaze movements, where an unfortunate experimental design has misled researchers into believing that eye-gaze movements are not ballistic. We then provide simple guidelines to prevent confounds: Practitioners should carefully design the experimental conditions of (D,W), fully distinguish data acquired for different conditions, and put less emphasis on $r^2$ scores. We also recommend investigating the use of stochastic sampling for D and W.

ACM Classification Keywords
H.5.2. Information Interfaces and Presentation (e.g. HCI): User Interfaces

Author Keywords
Fitts’ law; factor confounds; models

INTRODUCTION
Fitts’ law is a well known rule for human-aimed movement that predicts the movement time (MT) it takes to reach a target of width W located at a distance D:

$$MT = a + b \log_2(1 + D/W) = a + b \text{ID},$$

(1)

where ID is the index of difficulty\textsuperscript{1} and $a$ and $b$ are estimated empirically.

Fitts’ law gained importance in the HCI community after the seminal study by Card et al. [4] that measured the performance of four pointing devices on desktop computers. The law plays a prominent role in HCI, where it is heavily used to predict and evaluate the performance of input techniques. It has been shown to handle many conditions and tasks quite well, resulting in an incredibly wide spectrum of applications. Studies include reciprocal tapping between two targets using different limbs and body parts, such as tapping pedals [9] and manipulating a cursor [15] with the feet, controlling a stylus attached to the chin with head movements [2], and moving a cursor by rolling the head [16]. More unexpected applications include pointing and dragging sequences [12], rapid elbow flexion [6], eye-gaze movements [25, 21], and a study of patients with cerebral palsy [3].

Fitts’ pointing paradigm for the reciprocal tapping experiment [10, Experiment I] has proven to be very influential. Fitts’ apparatus was comprised of two plates of width W, separated by distance D. Distance and width were systematically varied: The four different conditions of D (2,4,8,16 (in.)) were crossed with four conditions for W (1/4,1/2,1,2 (in.)), resulting in 16 conditions. MT was then evaluated for each one of the 16 conditions. Today’s version of a generic 1-D Fitts’ law experiment is in many cases a simple adaptation of this protocol, where physical plates are replaced by targets on computer screens, and reciprocal tapping is sometimes replaced by discrete tapping, a cleaner version of the protocol due to Fitts & Peterson [11]. While the ISO standard [24] for the evaluation of pointing performance recommends a 2-D multi-directional tapping task, composed of circular targets of diameter W arranged in a circle of diameter D to control the effect of direction, many studies, old or recent, are conducted with the simple 1-D task.

Notations
Throughout the paper we use the following notations:

- **Factors** are noted in roman capital letters: distance D, width W, index of difficulty ID, and movement time MT.

\textsuperscript{1}There are several formulations for ID, the one given here, known as the Shannon formulation [18, 19] being the most widely used in HCI.

\textsuperscript{1}There are several formulations for ID, the one given here, known as the Shannon formulation [18, 19] being the most widely used in HCI.
• These factors take values in sets, denoted by their corresponding calligraphic capital letters: $D = \{d_1, d_2, \ldots, d_n\}$, $W = \{w_1, w_2, \ldots, w_m\}$, $\mathbb{D} = \{d_1d_2, \ldots, d_1d_6\}$.
• When two factors, e.g., $D$ and $W$, are fully crossed, each element of $\mathbb{D}$ is paired with each element of $\mathcal{W}$, resulting in $n \times m$ pairs. We use the condensed notation $\mathbb{D} \times \mathcal{W}$ to denote all these pairs.
• Several $(D, W)$ pairs may result in the same $D/W$ ratio. For example if $D = \{10, 20, 50\}$ and $\mathcal{W} = \{1, 2, 4\}$, pairs $(10,1)$ and $(20,2)$ have the same $D/W$ ratio of 10. We call $\mathbb{D}$ the variable corresponding to the average distance computed over equal ratios $D/W$.
• The list of averaged values $\overline{D}$ is noted $\mathbb{D}$. With $\mathbb{D}$ and $\mathcal{W}$ as above, the set of ratios is $\{2.5, 5, 10, 12.5, 20, 25, 50\}$ and the corresponding list of values of $\overline{D}$ is $\mathbb{D} = \{10, 15, 15, 50, 20, 50, 50\}$.
• Similarly, we note $\overline{MT}$ and $\overline{W}$ the variables corresponding to the averages of $MT$ and $W$ computed over equal ratios $D/W$.

Flaws in Fitts’ Literature

While Fitts’ law has proven to be incredibly robust and useful, a number of potential flaws in the experimental designs used in the literature have been identified.

Guiard [14] showed that the design of Fitts’ original tapping experiment correlates ID with $D$: “The dependence between $D$ and ID is strong and systematic: on average target distance is raised by 11.7 cm for each extra bit of information” [14]. Hence Fitts’ design makes $D$ a confounding factor: a factor that has an effect on both a dependent and an independent variable in a controlled experiment.

It may then appear that manipulating the independent variable leads to variations in the dependent variable, when in fact both variations are due to the confounding factor. Thus, within Fitts’ design, manipulating the ID affects movement time, but in fact both ID and MT are affected by the confounding factor $D$, making it impossible to disentangle the effects of $D$ and ID on MT.

The correlations between $D$ and ID are even stronger if one “builds an average on execution times for one ID first and calculate the correlation afterwards” [7]. This procedure, commonly used by HCI researchers, considers all blocks corresponding to the same ID as equivalent and computes a single average per value of ID, leading to what we have noted $\overline{D}$, $\overline{W}$ and $\overline{MT}$. Not only does this operation pre-supposes the validity of Fitts’ law [7, 8], it also strongly correlates $\overline{D}$ and ID [14, 7, 8]. As we shall illustrate, it then becomes impossible do distinguish if the effects on $\overline{MT}$ are due to $\overline{D}$ or ID.

The implications of these confounds are wide ranging. According to MacKenzie [20], referring to Glencross and Barrett [13], “It has been suggested that the model [Fitts’ law] would hold for the mouth or any other organ for which the necessary degrees of freedom exist and for which a suitable motor task could be devised”. If $D$ is indeed a confounding factor for ID, one can then wonder about the validity of Fitts’ law in some of its applications.

Positioning and Goals of The Paper

One solution proposed by Guiard [14] to disentangle $D$ and ID is what he called the complete form $x$ scale design. Although this solution does guarantee the decorrelation of $D$ and ID, it is impractical for pointing experiments in that it strongly restrains the range of variation of ID. The ID cannot be raised over 6 bits or so because that would require, at lower scale levels, impractically small values of $W$; nor can the ID be lowered below 4 bits or so because that would require, at higher scale levels, impractically large values of $D$ [14].

While the form $x$ scale design ensures a totally independent variation of factors $D$ and ID, in “real-world” HCI, the decision about the validity of Fitts’ law is usually based on the informal criterion that $r^2$ between ID and MT should be high enough (e.g., $r^2 > 0.9$ in [24]). What matters then, as we will show, is not that the confound be totally removed, but rather that its “strength” be reduced. We define $X$ and $Y$ as strongly confounded factors if $r^2(X, Y) > 0.9$, otherwise they are weakly confounded. In contrast to Guiard [14], the focus of this paper is to find designs that prevent strong confounds.

The goal of this paper is threefold: First, we identify the objective conditions that lead to strong confounds in Fitts’ law experimental designs. Such conditions can be evaluated on past and future experiments to quickly determine if the design is at risk of strong confounds. Second, once these conditions are clearly identified, we show a documented case in eye-gaze pointing where the validation of Fitts’ model is the result of a strong confound between $D$ and ID. Third, we provide recommendations to protect the experimenter from confounding factors when designing experiments for Fitts’ law. These recommendations may also prove useful for experimenters in HCI in general.

PRACTICAL EFFECTS OF STRONG CONFOUNDS

Guiard [14] and Drewes [7, 8] have illustrated the strong confound between $D$ and ID that occurs in Fitts’ original design and raised ensuing qualitative issues. The goal of this section is to show a numerically worked out example of strong confounds between $D$ and ID, on a different design than Fitts’.

A Surprising Simulation

Let us consider two potential generative processes for MT:

Law A is Fitts’ law, where MT is given by Eq. (1). Values for $a$ and $b$ are taken respectively as $a = -1047$ ms and $b = 391$ ms/bit.

\[ MT = a' + b' D, \quad (2) \]

where $a' = -251$ ms and $b = 1.956$ ms/mm.

The explanation for the specific values of $a$, $a'$, $b$ and $b'$ will appear below. We now consider the following experimental

\[ 3 \text{As noted in [14], a full form} \times \text{scale design on Fitts’ design for Experiment I [10] requires, e.g., } W = 0.02 \text{ cm.} \]

\[ 3 \text{The operational value of } r^2 \text{ above which a confound is deemed strong or not depends of course on what is expected by the experimenter. The value .9 is common in the Fitts’ law literature.} \]
We run a simulation experiment where trials for each process are generated for the \( D \times W \) conditions by adding a random noise sample to the underlying generative process. MT, for each trial \( i \), is then given by,

\[
\begin{align*}
\text{MT}_i & = -1047 + 391 \log_2(1 + D/W) + z_i \quad \text{(law A)} \\
\text{MT}_i & = -251 + 1.956 D + z'_i \quad \text{(law B)}
\end{align*}
\]

where \( z_i \) and \( z'_i \) are the outcomes of a centered random process, which we took equally distributed according to a zero-mean Gaussian law with standard deviation \( \sigma = 300 \) ms, meaning that approximately \( 96\% \) of the samples for MT are located within a 1200 ms interval. As we shall see later, the shape of the actual distribution is of little impact as long as it is centered.

We generated 200 trials per condition; the two datasets are represented in a MT vs. ID plane in Fig. 1. Five vertical scatters are obtained, one for each different D/W ratio. To get rid of the strong variability, we consider MT, the averages of MT per ID. After averaging, we end up with the blue dots and orange diamond markers in Fig. 1. The surprising result is that the summary \( \bar{MT} \) for law A is indistinguishable from that of law B. Both datasets, after averaging, are extremely well fitted with Fitts’ law: The \( r^2 \)-squared between \( \bar{MT} \) and ID is \( r^2 = 0.9995 \) for law A and \( r^2 = 0.9921 \) for law B. This is unexpected for the second dataset since it was generated with law B, not Fitts’ law.

### A Scenario for an Experimenter

The scenario for the above simulation is very close to a real experiment: Let us consider an experimenter who wants to investigate two tasks A and B and tries to find reasonable models to explain MT in both tasks. Let us assume that MT is actually governed by law A (Fitts’ law) for task A and by law B for task B. The experimenter does not know this but has the intuition that both tasks should be reasonably well modeled by Fitts’ law (law A).

After setting up the experimental design described above, he runs the experiment for both tasks. The resulting datasets are probably very similar to those in Fig. 1, with 5 vertical scatters for each task. After averaging over ID, he ends up with the large markers (blue disks and orange diamonds) of Fig. 1. This is precisely the averaging procedure discussed by Drewes [7] and mentioned in the introduction. After averaging, the experimenter then computes a fit using linear regression, and finds very high \( r^2 \)-s for both tasks. He can thus conclude that his intuition was right and that Fitts’ law is a good model to describe MT for both tasks. Unfortunately, as we have shown, in the case of task B, this is an artifact due to the experimental design.

### Decoding the Simulation

We now explain the results of the simulation.

#### Law A

For a given ID, the average of MT is equal to Fitts’ law evaluated at this given ID, plus the average value of the \( z_i \)’s. Since the noise is centered, the law of large numbers implies that the average value of the noise is close to 0. While we used a Gaussian distribution, this holds for any shape of the noise distribution. We could also have used an asymmetric distribution to guarantee positive movement times, this would not have changed the values obtained after averaging. A similar observation can be made about the standard deviation. By virtue of the central limit theorem, a larger standard deviation would only require more trials for the sample mean to be close enough to the statistical mean.

Thus, the fact that we get an \( r^2 \) between \( \bar{MT} \) and ID close to 1 for law A is not surprising. As expected, we find that the linear regression between \( \bar{MT} \) and ID is very close to the one used for generating the data: \( \bar{MT} = 386 \text{ ID} - 1040 \) for the simulation results vs. \( \bar{MT} = 391 \text{ ID} - 1047 \) for the formula generating the data.

#### Law B

The dataset generated with law B has the same summary statistic as the one generated with law A. Although surprising, this can be explained as follows. In the MT versus ID plane, there can be more than one \((D, W)\) condition leading to the same D/W ratio, e.g. for \( D/W = 32 \), there are three different values for \( D: D = 1024 \text{ (W = 32)}, D = 512 \text{ (W = 16)}, \) and \( D = 256 \text{ (W = 8)}. \) Therefore while it appears that there is only one vertical scatter plot per ID, there are in fact three overlapping ones. As the contribution of the noise will again be close to zero by the law of large numbers, average MT can thus be evaluated by inputting the average D = \( 1/3 \times (256 + 512 + 1024) = 600 \) into
D and ID for di MT. According to the D and ID. D and ID so strongly. In this section, D.

MT is given by D and ID is very high (r^2 = 0.99), see [14, 8]). We have established that the design of the goal passing task becomes a simple distance covering task. Assuming a constant maximum speed c, movement time would simply be given by

\[ MT = t_0 + 1/c \times (D - d_0), \]  

where \( t_0 \) is the time needed to reach a speed of c, and \( d_0 \) the distance traversed until c is reached. This is a linear model as in Eq. (2) with slope \( 1/c \) and intercept \( t_0 - d_0/c \).

Conditions for Strong Confounds
Fitts’ design for the tapping experiment [10, Experiment I] strongly confounds \( D \) with ID (\( r^2 \) between \( D \) and ID above .99), see [14, 8]). We have established that the design of the goal passing task [1, Experiment I] suffers from a similar confound (\( r^2 \) between \( D \) and ID above .99). In this section, we investigate the reason for these strong confounds.

Fitts-Like Designs
We first define a class of experimental designs, which we call Fitts-like designs, characterized by experimental conditions of the following general form:
1. \( D = \{d_i \} \), where \( 0 \leq i \leq N - 1 \) and \( d \) is fixed.
2. \( W = \{w_j \} \), where \( 0 \leq j \leq M - 1 \) and \( w \) is fixed.
3. \( D \) and \( W \) are fully crossed.

Conditions 1 and 2 state that the values of \( D \) and \( W \) follow a geometric progression. Both Accot & Zhai’s goal passing task and Fitts’ tapping experiment are Fitts-like designs.

Table 1 shows the \( r^2 \) values between \( D \) and ID for four Fitts-like designs: “Tapping”, “Disc Transfer” and “Pin Transfer” refer to the experiments conducted by Fitts [10]; “Goal Passing” refers to the study conducted by Accot & Zhai [1, Experiment I]. All four experiments lead to strong confounds between \( D \) and ID.

We also compute the \( r^2 \) values between \( D \) and ID for different sizes of \( D \) and \( W \). Fig. 3 shows that any Fitts-like design strongly confounds \( D \) with ID if \( N \) and \( M \) are small enough. Practical considerations often limit the values of \( N \) and \( M \) to
D and ID. As W grows at the same rate
$W$ and $D$ to be strongly confounded with ID, i.e. to be linearly de-
confound between $D$ and $ID$. As $W$ grows at the same rate
Geometric Progression of D&W Causes Strong Confound
The strong confound between
Table 1. Characteristics of four Fitts-like experiments. Tapping, Disc
Transfer and Pin Transfer by Fitts (1954) and Goal Passing by Ac-
W have $d$ and $w$ corresponding
We can then always choose the values of $D$ and compute the
The combination of a small number of conditions, the fact that
range of IDs used in practice is small. This explains why $r^2$ decreases
Care must also be taken with designs that are not Fitts-like.
the range of IDs being investigated is small. This explains why $r^2$ decreases
2. the range of IDs used in practice is small.
Other Strong Confounds
We have just identified the reasons that make Fitts-like designs
strongly confound $D$ with $ID$. As $W$ grows at the same rate
Care must also be taken with designs that are not Fitts-like.
the combination of a small number of conditions, the fact that
they are usually chosen according to some structure (such as a
linear or geometric progression), and the smoothing effect of
averaging make it very likely that $D$ or $W$ can be approached
by a simple function of $ID$.
We illustrate this with an example. Consider a candidate model for the dependent variable \( Y \) as a function of the independent variable \( X \):

\[
Y = f(X).
\] \hspace{1cm} (9)

Now consider a third variable \( Z \), that can be expressed as a function of \( X \):

\[
Z = g(X).
\] \hspace{1cm} (10)

We will call \( g \) the confusion function. A competing model of the form

\[
Y = f(g^{-1}(Z))
\] \hspace{1cm} (11)

will inevitably be indistinguishable from the candidate model, since plugging (10) into (11) gives:

\[
Y = f(g^{-1}(g(X))) = f(X)
\] \hspace{1cm} (12)

which is equivalent to Eq. (9).

The variable \( Z \) used here is very general and can represent any factor. In the example developed in the previous section (“Thought experiment”), \( Z \) was a square root of \( Y \). We used the procedure described in Fig. 4 with the following summary statistics of the two laws almost identical. More high for Law B:

\[
D = a \sqrt{ID} + b
\] \hspace{1cm} (13)

Then, \( g^{-1}(x) = (x/a) - b' \), and \( f(g^{-1}(x)) = a - bb' + b(x/a)^2 \), so that a quadratic model for MT cannot be distinguished from Fitts’ model, as MT is also linearly related to ID. If we then take a model of the form

\[
\bar{MT} = a - bb' + b(\bar{D}/d')^2
\] \hspace{1cm} (14)

plugging Eq. (13) into Eq. (14) gives

\[
\bar{MT} = a + bID.
\] \hspace{1cm} (15)

This is a simple linear function between \( \bar{MT} \) and ID, thereby showing that a quadratic law for movement time can be indistinguishable from Fitts’ law in some designs.

While in general, a weak confound between \( D \) and ID implies that any function of \( D \) is (weakly) confounded with ID, this is not true anymore for strong confounds. We have shown here that most designs of Fitts’ law experiments are likely to have a strong confound between ID and some function of \( \bar{D} \) and \( \bar{W} \). If one knows precisely this strong confound, e.g. Eq. (13), it is easy to determine which model for MT will have almost the same summary statistic \( \bar{MT} \) and \( r^2 \) as Fitts’ model. For example, with Eq. (13), consider the model of Eq. (14).

Creating Strong Confounds Between Any Two Factors

In the previous subsections, we explained how strong confounds between ID and functions of \( D \) or \( W \) could make two models indistinguishable from each other. We now give a general method to create strong confounds between ID and almost any function of \( D \) or \( W \), as a constructive illustration.

---

In Fitts’ law, \( Y \equiv \bar{MT}, X \equiv ID, \) and \( f \) is a linear function. \( X, Y \) and \( Z \) may also represent averaged quantities, e.g. \( \bar{D} \) or \( \bar{MT} \).
In summary, we first demonstrated that Fitts-like designs create strong confounds between $D$ and ID. We then showed that the issue is much deeper than Fitts-Like designs: With any design, there exists a possibility that some function of $D$ or $W$ is strongly confounded with ID.

**EYE-GAZE EXPERIMENTS**

We discuss two studies conducted on eye-gaze movements using the results of the previous section. These demonstrate that the issues we have underlined until now are not only theoretical constructs, but do appear in real-world scenarios. The first study, by Miniotas [21], validated Fitts’ model for movement time in a pointing task using eye-gaze. Drewes [8] discussed this study and pointed out that outside HCI, researchers would generally use Carpenter’s formula [5] (a formula not dependent on $W$) to model eye-gaze data. We compare this first study to another study by Miniotas et al. [22], as yet uncommented, and show that the results of these two studies on the validity of Fitts’ model for eye-gaze data are strikingly inconsistent. We will explain the difference by formally applying Carpenter’s formula to Miniotas’ paradigm and using the previous results on Fitts-like designs.

**Fitts’ Law for Eye-Gaze Interaction**

The first study [21] was conducted to validate Fitts’ law for modeling eye-gaze interactions. According to Miniotas [21], prior work by Ware and Mikelian [25] suggested that Fitts’ law would be an adequate model for movement time in eye-gaze interaction, but the range of D explored then was very narrow (less than 3 bit wide), and the width $W$ had been kept constant. The motivation was thus to create an empirical dataset that was more thorough.

In line with Fitts’ paradigm, the task was to move the cursor to a target of width $W$ located at a distance $D$ by using their eyes instead of a stylus. An eye tracker was used to control the cursor. The control variables were $D$ ($D = \{26, 52, 104, 208\}$ (mm)) and $W$ ($W = \{13, 26\}$ (mm)). It is not clear how the $r^2$ between $MT$ and ID was computed, but Miniotas reported $r^2 = 0.982$. He concluded that Fitts’ law was a good fit for his dataset, a valuable result for designers.

However, it turns out that the design of the experiment is “Fitts-like” according to our definition: First, both $D$ and $W$ have a geometric progression; there are 8 conditions, yet only 5 different $D/W$ ratios. Second, the range of IDs, $[1, 4.1]$, is quite small. Accordingly, the correlation between $D$ and ID is very strong: $r^2 = 0.974$, leading to a strong confound. As a consequence, a competing model such as Eq. (2) could equally well explain the summary of the gathered data. Note that Ware & Mikelian [25], recognizing that eye-saccades are ballistic, stated that they had used Fitts’ law “only as a convenient way of summarizing the results, not because [they] wish[ed] to make any theoretical claims”.

**Carpenter’s Formula**

A reliable relation known as Carpenter’s formula [5] relates $MT$ and angular amplitude $\alpha$ (Eq. (17)) for eye-gaze movements:

$$MT = a + b \alpha, \quad (17)$$

where $MT$ is the movement time needed to cover the angular amplitude $\alpha$. Carpenter’s formula can be applied to Miniotas’ paradigm as illustrated in Fig.6, where $L$ is fixed and $S$ corresponds to $D$ in Fitts’ paradigm. In line with the ballistic nature of eye movements, $W$ does not appear in Carpenter’s formula. The angle $\alpha$ can be expressed in terms of the available parameters as

$$\alpha = 2 \arctan \left( \frac{S}{2L} \right) \approx \frac{S}{L} = \frac{D}{L} \quad \text{when} \quad \alpha \quad \text{is small enough.} \quad (18)$$

If we input this equation into formula Eq. (17), we obtain

$$MT = a + b \alpha \approx a + b' D, \quad \text{where} \quad b' = b/L. \quad (19)$$

Therefore, Carpenter’s formula, the leading explanatory model, predicts that movement time is linearly related to $D$.

Thanks to our previous analysis, we can now safely assert that since the design of Miniotas’ experiment [21] is Fitts-like, Fitts’ model is almost equivalent in terms of fitting $MT$ to the leading explanatory model derived from Carpenter’s formula. Therefore one cannot conclude from that experiment that eye-gaze follows Fitts’ law.

**Eye-Gaze Interaction with Expanding Targets**

Miniotas et al. [22] conducted a second study on eye-gaze interaction with expanding static targets. Expanding static targets are targets whose appearance does not change for the user, but to which the interface responds as if it were larger. The expansion is predetermined, hence the term “static”.

Although the goal was not to validate Fitts’ law, Miniotas et al. did check for goodness of fit of the Fitts model and found $r^2 = 0.69$. The experiment is very similar to that described in the first study [21], with a notable exception: the progression for $W$ is not geometric, but linear: $D = \{128, 256, 512\}$ (pix.) and $W = \{12, 24, 36\}$ (pix.). Because this design is not Fitts-like, we expect $r^2$ between $D$ and ID to be weaker than for Fitts-like designs, and indeed we find $r^2 = 0.74$.

Miniotas et al. attributed the decrease in correlation between MT and ID from 0.98 in the 2000 study [21] to 0.69 in the 2004 study [22] to the presence of a visible cursor in the 2000 study.

<table>
<thead>
<tr>
<th>XP</th>
<th>Fitts-like</th>
<th>$r^2(D, id)$</th>
<th>$r^2(MT, id)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Miniotas 2000</td>
<td>yes</td>
<td>0.99</td>
<td>0.98</td>
</tr>
<tr>
<td>Miniotas et al. 2004</td>
<td>no</td>
<td>0.74</td>
<td>0.69</td>
</tr>
</tbody>
</table>

Table 2. Table summarizing the relevant $r^2$ in the Miniotas (2000) and Miniotas et al. (2004) studies. $r^2(x,y)$ is the coefficient of determination between $x$ and $y$. 

[Figure 6. Carpenter’s Formula applied to Fitts’ pointing paradigm. $L$ is the fixed distance between the user and the screen, $S$ is the on-screen distance that the cursor must travel, i.e. $D$ with the notations of Eq. (1).]
whereas there was no visual feedback in the 2004 study. This is because they assumed Fitts’ law to be a valid model. However, the comparison between the two studies tells a different story, as shown in Table 2. The design used in the 2000 study [21] was Fitts-like, making Fitts’ model almost equivalent to the one derived from Carpenter’s formula, whereas the design used in the 2004 study [22] was not Fitts-like. As a result the confound between D and ID is not as strong (r² between D and ID of .69). Not surprisingly, r² between MT and ID drops with a similar magnitude (r² between MT and ID of .69). We conclude that Fitts’ model’s apparent validity in [21] is an artifact caused by the experimental conditions.

COMBATING STRONG CONFOUNDS
In this section we give four recommendations to protect experiment designers from strong confounds.

Do Not Trust a Good r²
The two datasets generated from law A and law B in the Thought Experiment had the same MT and r². From this summary alone, they were indistinguishable. If one had used different summaries, such as those from Jude et al. [17], striking differences would have emerged. For example, the variance of the two datasets are very different.

The evaluation of Fitts’ model in the HCI community relies almost exclusively on high r² values. As emphasized by Roberts & Pashler [23] however, a good fit reveals nothing about the flexibility and variability of the data (what the model can and cannot fit), nor the likelihood of other models. Indeed, we have shown that within a Fitts-like design, two different models could fit the same summary data.

Do Not Average or Pool Data From Different Conditions
We have seen (Sect. 3) that the strong confound between D and ID was made possible because of the averaging procedure. For example, in Fitts’ tapping experiment, the r² between D and ID is .48; after averaging the r² between D and ID is .99. It is unfortunately common for experimenters to pool data and average movement times that correspond to the same D/W ratio but that do not come from the same D × W condition (see, e.g., Drewes [7, 8]). This practice involves a confirmation bias: The logic behind averaging movement times that correspond to the same D/W ratio but come from different pairs (D,W) is to believe that because data was acquired under the same ratio and that movement time is supposedly dependent on the ratio only, the two conditions would essentially be the same. But this is only true if indeed, the ratio explains all the variability of MT, which is precisely what we want to test when trying to validate Fitts’ law in the first place.

Averaging before knowing the validity of Fitts’ model may then result in a premonitory experiment where Fitts’ law can be validated simply because it was pre-supposed to hold, as shown in the section analyzing Miniotas’ experiment [21].

Note that for experimenters using the effective index of difficulty, IDₑ [24], a different value of IDₑ is calculated for each block based on the participants variability in that specific block. Therefore different conditions will almost always result in different values of IDₑ, even if they correspond to the same D/W ratio. The net result is that this procedure eliminates the risk of averaging across (D, W) conditions.

Consider Competing Models
We have shown that Fitts-like designs create strong confounds between ID and both D and W. We further showed that we could construct a design that strongly confounds ID with almost any simple function of D or W. It is then important, when evaluating whether Fitts’ model is a good fit for a task, to also consider competing models. If there are any, the experimenter should make sure that the design does not strongly confound factors of both models. For example, in the eye-gaze study it would have been safer to also evaluate Carpenter’s formula (Eq. 17) on the experimental data.

Notice that once a competing model is identified, it is easy to verify the risk of strong confounds among factors by checking the correlations between them.

Use Stochastic Conditions
In a Fitts’ law experiment, D and W are varied and MT is measured. The average MT of each block represents one sample in the (D,W) space. Experimental data can thus be visualized as a set of samples in the (D,W) space⁶. Fitts [10] showed that this representation could be summarized by a simple formula – now known as Fitts’ law. We have shown that some sampling strategies such as Fitts’, i.e., geometric progressions and orthogonal sampling in the (D,W) space, may lead to strong confounds between factors. A different sampling strategy is Guiard’s [14] orthogonal sampling in the form×scale space, which provides a theoretical solution to the issue, but can lead to physical values of D and W that are hard to implement in practice.

We have shown that sampling issues leading to strong confounds occur under very specific conditions, i.e., when the conditions are generated by some rule. For example, a Fitts-like design is characterized by a geometric progression for D and W. Therefore, we believe that a simple solution to get rid of potential sampling artifacts is to adopt stochastic conditions for D and W, possibly with some constraints. For example, we could divide the (D,W) space into a grid and choose experimental conditions by drawing a point uniformly within each rectangle defined by the grid, resulting in pairs of (D,W) values.

CONCLUSION AND PERSPECTIVES
In the experimental testing of any mathematical model, we may distinguish two steps:

1. Sampling the factor space, e.g. Fitts’ traditional (D,W) space or Guiard’s form×scale space, thus defining a set of experimental conditions for data collection;

2. Processing the data by applying operations that yield a score. In Fitts’ law studies, this traditionally involves computing means of movement times and r² values between ID and MT.

⁶Incidentally, this is precisely how Fitts summarized his data in his historical study [10, Fig. 4]
We have shown that a Fitts-like sampling of the \((D, W)\) space solely associated with the computation of \(r^2\) between ID and MT creates strong confounds between D and ID. We attributed this to the geometric progression of D and W. A simple workaround would seem to be to avoid such a sampling. However, using a constructive approach, we devised a sampling strategy that strongly confounds ID with any simple function of D and W. Avoiding Fitts-like designs is thus insufficient to avoid strong confounds.

Based on these new results, we analyzed an apparent contradiction between the results of two eye-gaze pointing experiments using. We resolved the contradiction by noting that Carpenter’s formula is a widely accepted model for eye-gaze data and by showing that in one of the experiments, Fitts’ model was indistinguishable from Carpenter’s model due to the use of a Fitts-like design.

Finally, we provide guidelines to avoid strong confounds between factors. We believe that Fitts’ law studies place too much emphasis on high \(r^2\) values. It is crucial to introduce other considerations when validating a model, such as the flexibility of the evaluated model, the variability of the dataset and the possibility of competing models. Working with block averages or, worse, averages computed for equal values of ID, such as MT, dramatically decreases the number of points to be fitted, thereby mechanically increasing \(r^2\) values.

An interesting and simple way to prevent strong confounds is to use stochastic sampling of the \((D, W)\) space. Stochastic sampling is a promising perspective, especially when considering the replication of studies, as a different but equivalent design can be ensured with each replication. However, more conceptual work is needed to support the idea of using random conditions in a controlled experiment.

**ACKNOWLEDGMENTS**

This research was partially funded by Labex DigiCosme (ANR-11-LABEX-0045-DIGICOSME), operated by the French Agence Nationale de la Recherche (ANR) as part of the program “Investissement d’Avenir” Idex Paris-Saclay (ANR-11-IDEX-0003-02), and by ERC European Research Council (ERC) grant 695464 “ONE: Unified Principles of Interaction”.

**REFERENCES**


14. Yves Guiard. 2009. The problem of consistency in the Fitts’ formula is a widely accepted model for eye-gaze data and by showing that in one of the experiments, Fitts’ model was indistinguishable from Carpenter’s model due to the use of a Fitts-like design.


