Modeling the Perception of User Performance

Max Nicosia
School of Computer Science
University of St Andrews
St Andrews, Fife, UK
ln73@st-andrews.ac.uk

Antti Oulasvirta
Max Planck Institute for
Informatics and Saarland
University, Germany
oantti@mpi-inf.mpg.de

Per Ola Kristensson
School of Computer Science
University of St Andrews
St Andrews, Fife, UK
pok@st-andrews.ac.uk

ABSTRACT

This paper studies how users perceive their own performance in two alternative user interfaces. We extend methodology from psychophysics to the study of interactive performance and conduct two experiments in order to create a model of users' perception of their own performance. In our studies, two interfaces are sequentially used in a pointing task, and users are asked to rate in which interface their performance was higher. We first differentiate the effects of objective performance (speed and accuracy) versus interface qualities (distance between elements and width of elements) on perceived performance. We then derive a model that predicts the amount of change required in an interface for users to reliably detect a difference. The model is useful as a heuristic for predicting if a new interface design is better enough for users to reliably appreciate the obtained gain in user performance. We validate the model via a separate user study, and conclude by discussing how to apply our findings to design problems.

Author Keywords

Perception of user performance; psychophysics

ACM Classification Keywords

H.5.2 Information Interfaces and Presentation: User Interfaces—*Theory and methods*

INTRODUCTION

Present-day computer users are bombarded with possibilities to upgrade, modify, and switch software and hardware. Such changes trigger them to exercise judgment on the interface being presented to them. This could encompass the study of their perceived utility, perceived usability, or perceived user performance. This paper focuses on the perception of performance.

Despite the obvious importance of this topic to human-computer interaction (HCI), the question of how users perceive changes in interactive performance has not received serious attention. Knowledge on this topic would help designers with many tasks where they usually have to run empirical studies. For example, interface designers could focus on usability features where noticeable differences can be gained and marketers could focus advertisements of a new interface on aspects of an interface that users will most likely recognize as improvements. Theories and designers' conceptions of user experience [11, 12] make the point that it is not objective performance that matters but the user's experience and perception of it. However, the link between the perception of user performance and user interface design is not well understood.

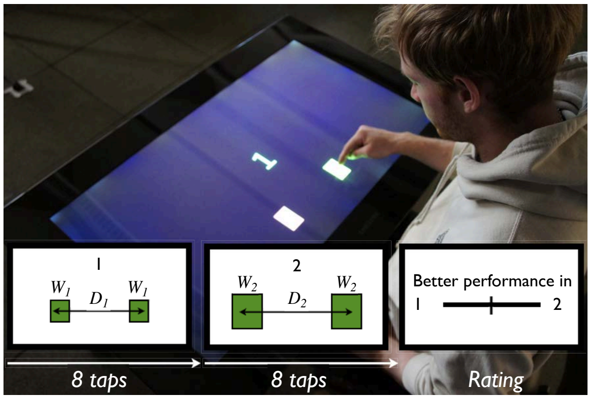

Figure 1. The experimental method in this paper is based on retrospective subjective comparison of two interfaces. Trial 1 and Trial 2 are followed by a rating screen.

This paper builds on *psychophysics methods* to study users' perception of which of two interfaces yields better performance for the same HCI task. In computer science, psychophysical models are used to optimize computer graphics [19], image processing [17], haptics [2], audio [27], and video [13]. In HCI, some psychophysics laws have gained recognition as design guidelines [14]. However, research has been limited to *time perception*, with applications to system response times and progress bars [10, 21].

Our goal is to extend psychophysical methodology to *interactive* tasks. One challenge is that user performance is not passive perception—it involves active engagement of the user over a longer period of time. Another challenge is that the standard psychophysics experiments typically consider relationships between two variables [5, 24]. However, user performance cannot be reduced to a single quality: users can base their judgment on two main categories of cues: *perceived qualities of the interface* and *perceived qualities of the user's own sensorimotor performance*. Important instances of

CHI'14, April 26–May 1, 2014, Toronto, ON, Canada
Copyright © 2014 ACM 978-1-4503-2473-1/14/04...$15.00.
http://dx.doi.org/10.1145/2556288.2557355

the former in HCI are the size, distance, and density of interface elements. Examples of the latter are the speed and number of errors when performing a task, which constitute human performance more generally [20]. Complicating this further is the fact that such variables are necessarily interrelated. For example, a new design that enlarges the sizes of the targets, also affects the time and accuracy of users' target selections. For these reasons, the standard psychophysical methodology has not been readily applicable to interactive tasks.

The method we introduce in this paper includes an *aimed movement* task consecutively carried out with two alternative designs: the standard (the "old" interface), and the alternative (the "new" interface). A participant completes a tapping task with one design (Trial 1), then completes the same task with another design (Trial 2), and is finally asked to rate in which his or her performance was better (Figure 1). This emulates the use of two interfaces that are compared retrospectively. We regard perception of user performance as a second-order representation that is most likely constructed based on other, more directly experienced aspects of interaction. We hypothesize the most important to be: speed and accuracy of aimed movements and visual qualities of the interface, such as the distance to and the width of the target.

The key to our experimental design is that we manipulate *index of difficulty* (ID) (see for example [15, 20]) to exert a predictable effect on user performance. ID is the difficulty of movement required to select an interface element with a distance D and a width W; $ID = \log_2\left(\frac{D}{W}+1\right)$. By manipulating ID, and thus the difficulty of sensorimotor responses, we emulate the fact that the alternative design changes the user performance only indirectly: by affecting the demands placed on sensorimotor control and by changing the perception of the layout.

ID is controlled in two ways in the two experiments reported here. In Experiment 1, we keep the ID of the standard and the alternative design constant, but we change D (distance) and W (width). In Experiment 2, we use a *staircasing method*: the ΔID between the standard and the alternative interface is gradually increased until the user can reliably notice the difference in performance. These two manipulations allow us to address two central research questions:

Experiment 1: What are the criteria users use to judge that performance in one design is better than an alternative design?

Experiment 2: How large does a difference in user performance have to be for users to reliably notice it?

The two experiments map to two common decision contexts in interface design. Experiment 1 addresses what we call a *within-ID design*. In this case, the qualities of interface elements cannot be changed independently of each other. Limited display space would pose such a scenario: the size of the elements cannot be changed independent of inter-element distance. The second experiment addresses the *between-ID* case where the qualities of the new design *can* be changed independently. For example, smartphones with two different screen sizes would have different average IDs.

The results from our experiments expose several previously unreported phenomena on the contribution of different factors and biases. Based on the experimental data we develop a predictive model in the form of a mathematical function that can be used to estimate the percentage of users that will be able to identify a performance change between two interfaces with different IDs. Our model characterizes the probability of a user being able to reliably judge that there is a difference between two designs with different IDs. This model is validated via a separate task, a *Whac-A-Mole* type of game.

In summary, we contribute to the HCI literature by presenting a novel variation of a psychophysical method that we have adapted to interactive tasks. This methodology can be generalized to other studies of user performance in other interactive tasks. The presented model and obtained results are limited to *target acquisition* tasks. While this is a common sub-task in HCI, for example in command selection and typing [15], it leaves room for future work on other task domains. We conclude this paper with a discussion on how to use of this model and deploy the method to other task domains.

RELATED WORK

Psychophysics is the study of a human's perceived ability to distinguish a difference in physical stimuli and events. The classic psychophysics model, the *Weber–Fechner law*, states that equal stimulus ratios produce equal subjective ratios [24]. The *Just Noticeable Difference* (JND) is proportional to the size of the standard stimulus: JND = kS, where S is the size of the standard stimulus and k is a constant (a so-called Weber fraction). k is the proportionate increase the standard stimulus needs to change before it can be reliably discriminated. JND thresholds using the Weber–Fechner law and other models have been charted for perceptual events and for user interface qualities one can expect to be relevant: visual length and area, visual distance, visual velocity, visual flash rate, and duration [24].

Psychophysics research in HCI has focused on time perception, with applications to system response time and progress bars (e.g,. [10, 21]). Recently, cognitive load was found to affect time perception in an HCI task [3]. We are unaware of work in HCI addressing aspects of interactive tasks more broadly.

Evidence from psychology suggests that *interactive performance* can turn out to be special. Studies in psychology have found time perception to be affected by the allocation of attention during the task and the structuring of the stimulus environment [8]. Both attention and the stimulus environment are affected by the interface. More generally, it has been found that judgment tasks that involve *action* by the perceiver at times differ from passive tasks. In such cases, JND functions are not always Weberian and complex interactions emerge. Such tasks include the perception of motion as opposed to stationary stimuli [26] and visually-guided grasping [6]. Moreover, JND thresholds can change during an action. For instance, a moving hand is less sensitive to external stimuli than a static hand [18, 25]. Another complication is posed by multimodal perception. Visual and haptic information can produce superior discrimination when used jointly, compared to using either modality alone [7]. Similar multimodal-unimodal differences have been reported elsewhere (e.g. [4, 22]).

Given this brief overview, we deduce two requirements for our experimental method. First, the Weber–Fechner law, or any other known JND function, cannot be expected to apply. Thus the experimental method should allow a function of any shape to emerge. Second, the experimental method must be able to link perception to both 1) the observable qualities of the interface and 2) the observable qualities of performance.

GENERAL APPROACH

Our research interest is exploratory and encompasses two goals: The first is to understand the factors that affect the perception of user performance and the second is to chart JND thresholds for user performance.

Within a range of commonly studied HCI tasks—such as navigation, command selection and search—we chose *target acquisition* [15] for two reasons. First, visually controlled discrete aimed movements executed with the hand are prevalent in present-day HCI, and thus studies of target acquisition are relevant to a wide range of user interfaces. Second, there exists a well-established predictive model that links interface characteristics with user performance for target acquisition. Fitts' law [15, 23] predicts the movement time (MT) required for a user to hit a target with width W at distance D as:

$$MT = a + bID = a + b\log_2\left(\frac{D}{W} + 1\right), \qquad (1)$$

where a and b are empirically determined parameters[1]. Fitts' law implies that a user's information capacity, as measured by throughput (TP) in bits/s, stays relatively constant as a function of ID [23][2]:

$$TP = \frac{ID}{MT}. \qquad (2)$$

The two experiments in this paper exploit this tendency.

In Experiment 1, we hold ID constant and change D and W within an ID condition. This means that MT should stay at the same level for the the two to-be-compared interfaces, although the interface qualities D and W change. If judgments are based on TP, they should not favor either interface. If this is not the case, we will be able to assess the individual contributions of speed, accuracy, D, and W.

In Experiment 2, we manipulate ID with a so-called staircasing method [5]. We start from a minimum difference between two interfaces and subsequently increase the difference between the IDs (ΔID) with a constant step size until participants are able to reliably tell that their performance has changed. This exploits the prediction of Fitts' law that increasing ΔID increases the difference in MT as well (and possibly inaccuracy). The staircasing method thus allows charting judgment reliability as a function of ΔID.

Knowing that aimed movements at different scales can be associated with different sensorimotor requirements[3], we sample the whole permissible range of IDs afforded by a large interactive surface (a Microsoft PixelSense, model Samsung SUR40). In the two experiments, we study four base-ID conditions ranging from very small ($ID = 1.2$, or "tray icon-sized") to large but still comfortable targets ($ID = 2.4$, similar to the Windows Start Menu). This allows us to examine if the size of the standard ID affects JND thresholds similar to the Weber-Fechner law.

To measure perception we use a rating scale where two sequentially used interfaces are directly compared on a slider ranging from -100 to 0 to $+100$. Participants are instructed to use the slider to express in which interface he or she experienced a "better performance". Performance is explained as the combination of how *quickly* and how *accurately* they were able to do the task [20]. Although rating is a motor task, we do not expect this to affect the reported differences. A measurement via a rating allows participants to use whatever criteria they deem reasonable to make the judgment. Moreover, the scalar value allows them to express their level of certainty. The known drawbacks of the rating scale are nonlinearity of responses, order effects, and scale non-uniformity [5]. We address these in our analysis by making no assumptions of linearity or uniformity, and in the experimental design by randomizing the order of interfaces.

EXPERIMENT 1: CONSTANT INDEX OF DIFFICULTY

Our first experiment investigated if perception of performance is affected by four variables that characterize user performance—perceivable interface qualities: the width W and the distance D between targets; and objective performance qualities: speed and accuracy.

The ID of the two to-be-compared interfaces was kept constant, but W and D were manipulated. In order to examine if sensorimotor demands affect users' perception of performance, we investigated four different IDs.

Method

Participants and Experimental Design

We recruited 18 participants from a university campus. The mean age was 24 years, ranging from 19 to 31 years (*sd* = 3.2). Eight were female, ten were male. Their sight was normal or corrected-to-normal, and they reported no motor or neural disorders.

The experiment was a within-subjects design with one independent variable, ID, with four levels: 1.2, 1.6, 2.0 and 2.4. Sixty trials were carried out in each ID condition by each participant, each trial consisting of two interface designs. Participants performed eight target acquisition tasks in each interface design per trial. A *trial* consisted of two successive sub-trials followed by a rating task in which the participant rated which interface design (the former or the latter, or Trial 1 or Trial 2) resulted in a better perceived performance.

[1] We use the Shannon variant of ID since the values are positive [15].

[2] For a critical view of this standard approach, see Guiard and Olafsdottir [9].

[3] For example, the wrist and the fingers may be used more for targets with small D and W.

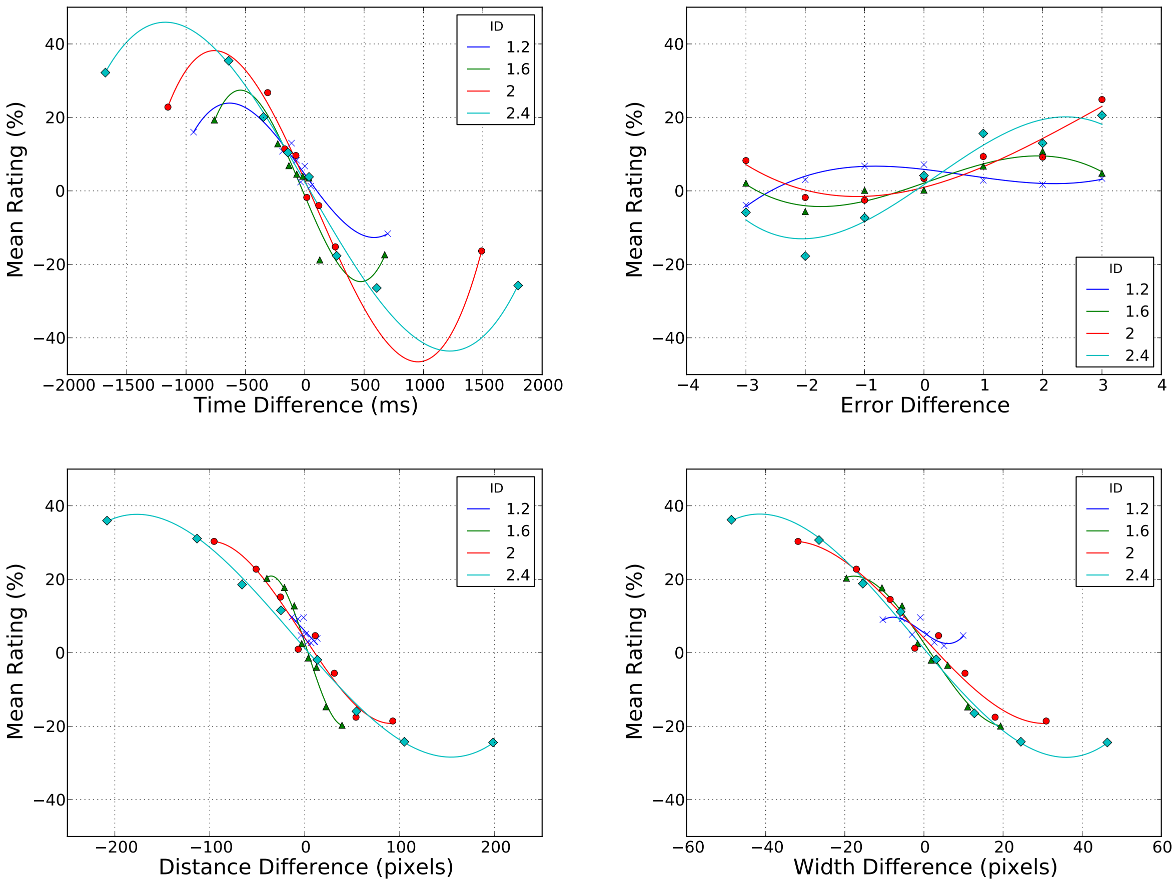

Figure 2. Contributions of *Time*, *Error*, *Distance*, and *Width* on the perception of performance in Experiment 1. Markers denote observed means for each bin, lines are regression models described in Table 1. The x-axes show the differences between interfaces in Trial 1 and Trial 2, the y-axes show the ratings: positive values refer to a preference for Trial 2 (the latter interface), while negative values refer to a preference for Trial 1.

For each trial, D and W were randomly sampled from a uniform distribution covering all permissible combinations within the ID condition. W was restricted to an interval ranging from 30 pixels (13.84 mm) to 173 pixels (79.80 mm) and D to an interval ranging from 39 pixels (17.99 mm) to 740 pixels (341.33 mm). The minimum D was limited by the feasibility of the minimum width for the smallest ID, while the maximum D was based on the screen resolution. ID conditions were balanced using a Latin square design.

Apparatus

The experiment was carried out on a Microsoft PixelSense, model Samsung SUR40. It had a resolution of 1920 × 1080 pixels (885.6 x 498.15 mm). The experimental software was developed in C# and used the Microsoft XNA framework and the Surface SDK. The software was designed to register touches through the surface API. A tolerance threshold was used in order to mitigate tracking errors that may occur on the apparatus for very fast movements. The precision of the timing data was in the order of microseconds.

Task and Procedure

Before starting the experiment, participants were instructed to familiarize themselves with the interface. The experimental task consisted of tapping two targets presented on the surface, sequentially, and as fast as possible without missing the target (see Figure 1). The current target was highlighted with a green color. The target areas were shaped as vertical columns, as in the original Fitts' unidimensional tapping task. Participants were instructed to aim for the middle part of the column that contained the target. However, the whole column was considered the target from the system's point of view. If the participant missed the target but was still within the tolerance range defined by the column, the system accepted the touch.

Each task had two interfaces (Trial 1 and Trial 2) followed by a judging task. Each trial consisted of eight reciprocal target acquisition tasks for a particular interface. Afterwards, the participant made a judgment on the performance of the interfaces by providing a rating on a continuous rating scale shown as a slider on the display (see Figure 1). The numbers "1" and "2" referred to the first and second interface with the corresponding label. Participants were instructed to treat the slider as a continuous rating scale, where the mid point was 0% difference in the interfaces, and each extreme meant a 100% difference in favor of either interface 1 or 2. We instructed participants to think of "performance" as the combination of how quickly and how accurately they were able to do the task.

In total, each participants performed 240 judging tasks (60 trials × 4 *IDs*), with two types of breaks interleaved to prevent fatigue. A 20-second break was scheduled for every 10 pairs of trials and a 1.5 minute break was scheduled for every 50 pairs of trials.

Results

The dependent variable *Rating* was calculated from the position of the slider as positioned by the participant. The left end of the slider was mapped to -100%, i.e. completely better performance in Trial 1 and the right end of the slider was mapped to +100%, i.e. completely better performance in Trial 2.

The predictive variables were calculated as the difference between Trial 1 and 2. *Time* was calculated as the difference in time taken to complete the tapping task in milliseconds, *Error* was calculated as the difference between the number of taps outside the target, *Distance* was calculated as the difference in the distance between the targets, and *Width* was calculated as the difference in the widths of the targets.

We here report general observations from the data and provide models for the effect of *Time*, *Error*, *Distance*, and *Width*. Observations outside three standard deviations were removed. Remaining ratings were grouped into bins per dependent variable, each bin containing roughly the same amount of observations. Binning for *Distance*, *Width* and *Time* was done by allocating 8 bins covering the entire range for each *ID* condition. The sizes of the bins increases with each *ID* as higher *IDs* span over greater ranges. Within each *ID*, bins also increase in width for larger values. For *Error* however, only the bins at either extreme contain more than one error value, all others contain a single error value.

Recency Bias

Overall, the average rating is close to zero, which is expected, because the ID was held constant and the order was randomized in both interfaces. However, participants demonstrated a slight *recency bias*—a bias towards the second and more recent interface they used (the second sub-trial) under all ID conditions. The bias was particularly pronounced in the lowest-ID condition. Calculated biases for *ID*s 1.2, 1.6, 2.0 and 2.4 were 5.80%, 1.35%, 3.94% and 3.82%.

Effects of Time, Error, Distance and Width

Figure 2 shows the effects of *Time*, *Error*, *Distance*, and *Width*, respectively. The markers represent the averaged observations per bin and the curves the fitted models listed in Table 1. Ratings range from -100.00% to 100.00%, where negative values correspond to a perception of better performance in the first interface (Trial 1), and positive values to better performance in the second interface (Trial 2). A rating of 0% expresses no noticeable difference. The following observations were made on the effects of the variables on the perception of user performance:

1. **Time**: *Time* had the strongest effect. Figure 2 shows a strong correlation between the rating and *Time*. The plot also reveals that the Δ*Time* required for participants to adjust the rating was affected by the ID condition. For $ID = 1.2$ the time difference was 500 ms, for 1.6 it was 700 ms and for *IDs* 2.0 and 2.4 they were 1000 ms and 1500 ms respectively. In the case of $ID = 1.2$, participants exhibited a clear recency bias. This may be due to the fact that perceiving a difference in D or W is relatively more difficult with larger and closer targets.

2. **Error**: Errors were not always taken into account when judging performance. $ID = 2.4$ showed a strong correlation between *Error* and rating ($r = 0.90$), while *ID*s 1.6 and 2.0 had a lower correlation ($r = 0.69$). For *ID*s 1.6 and 2.0, participants tended to be biased towards the second interface regardless of the number of errors in each trial. Again, $ID = 1.2$ is an outlier: participants tended to rate the first interface as providing higher performance even if they made more mistakes with it. However, as Figure 2 shows, this effect was negligible.

3. **Distance and width**: Figure 2 shows that *Distance* and *Width* affected judging performance for *ID*s 1.6, 2.0 and 2.4 but, again, *ID* 1.2 did not provide reliable grounds for performance judgments. In this case, the ratings were dominated by the recency bias. This is because the ratings tended to be positive if *any* difference in D or W was present.

Regression Model		
Time	ID	R^2
$4.04 \times 10^{-8}x^3 + 2.99 \times 10^{-6}x^2 - 4.50 \times 10^{-2}x + 4.50$	1.2	0.88
$1.00 \times 10^{-7}x^3 + 1.05 \times 10^{-5}x^2 - 7.67 \times 10^{-2}x - 1.34$	1.6	0.79
$3.36 \times 10^{-8}x^3 - 9.63 \times 10^{-6}x^2 - 7.32 \times 10^{-2}x + 2.88$	2	0.96
$1.30 \times 10^{-8}x^3 - 9.28 \times 10^{-7}x^2 - 5.60 \times 10^{-2}x + 2.50$	2.4	0.98
Error		
$3.47 \times 10^{-1}x^3 - 7.13 \times 10^{-1}x^2 - 1.88x + 5.87$	1.2	0.91
$-5.64 \times 10^{-1}x^3 + 1.55 \times 10^{-1}x^2 + 5.64x + 2.12$	1.6	0.80
$-1.73 \times 10^{-1}x^3 + 1.57x^2 + 4.20x + 9.67 \times 10^{-1}$	2	0.81
$-7.56 \times 10^{-1}x^3 + 3.81 \times 10^{-1}x^2 + 1.11 \times 10^{1}x + 1.68$	2.4	0.85
Distance		
$-2.78 \times 10^{-1}x + 5.94$	1.2	0.51
$1.83 \times 10^{-4}x^3 - 1.50 \times 10^{-3}x^2 - 7.96 \times 10^{-1}x + 2.73$	1.6	0.97
$1.46 \times 10^{-5}x^3 + 1.68 \times 10^{-4}x^2 - 3.92 \times 10^{-1}x + 4.04$	2	0.94
$3.65 \times 10^{-6}x^3 + 1.24 \times 10^{-4}x^2 - 2.98 \times 10^{-1}x + 1.25$	2.4	0.99
Width		
$5.39 \times 10^{-3}x^3 + 1.40 \times 10^{-2}x^2 - 7.65 \times 10^{-1}x + 5.42$	1.2	0.42
$1.47 \times 10^{-3}x^3 - 6.11 \times 10^{-3}x^2 - 1.60x + 2.64$	1.6	0.97
$3.84 \times 10^{-4}x^3 + 1.57 \times 10^{-3}x^2 - 1.17x + 3.96$	2	0.95
$2.85 \times 10^{-4}x^3 + 2.39 \times 10^{-3}x^2 - 1.28x + 1.07$	2.4	0.99

Table 1. Regression models for the judgment of performance separately for the independent variables *Time*, *Error Distance* and *Width* for each of the four base-ID conditions in Experiment 1.

Regression Models

Table 1 lists the regression models that best explained the data. Line plots for the models are shown in Figure 2, together with the observed points. The regression models for D and W can explain 94–95% of the variance for *ID*s 1.6–2.4. Interestingly, these models for *ID* 1.2 can only explain 51% and 42% of the variance respectively. For $ID = 1.2$, the strongest predictor appears to be *Error* ($R^2 = 0.91$) and *Time* ($R^2 = 0.88$). To conclude, except for $ID = 1.2$, the observations reported in the previous subsection could be captured by psychometric functions.

EXPERIMENT 2: STAIRCASING

In Experiment 2, the difference between the two interfaces, or ΔID, was manipulated using the *staircasing* method with

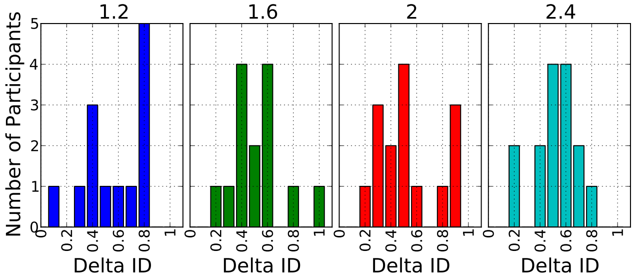

Figure 3. The number of participants that correctly judged 91% (or more) of the trials in Experiment 2 as the ΔID was increased.

a constant increase (0.1) between the steps. We started comparisons from four base-ID conditions (1.2, 1.6, 2.0, 2.4), increasing ΔID of the two to-be-compared interfaces gradually until the user could notice a difference with $>$90% reliability. The term *sub-stair* refers to a trial with a particular ΔID.

Method

We here only report the differences in method compared to Experiment 1.

Participants, Apparatus and Materials

We recruited 16 participants from a university campus. The mean age was 30.5 years, ranging from 22 to 55 years (*sd* = 8.4). Five were female, 11 were male. There sight was normal or corrected-to-normal, and they reported no motor or neural disorders. None of the participants had participated in Experiment 1.

Task and Procedure

Unlike Experiment 1, in Experiment 2 participants were explicitly told that the ID of the two interfaces (Trial 1 and 2) were different. This was necessary because the difference would become salient as the ΔID increased. They were told that there were four different difficulty levels (base-IDs) and that promotion to the next level only occurred when they identified the interface that yielded the highest performance in more than 90% of the trials (10 out of 11 judgments). They were instructed to judge their performance using the same definition of performance provided in Experiment 1. In the case they could not notice any difference, participants were instructed to provide their best guess and to avoid a 0% rating. If a participant was *unable* to judge a difference correctly by the 10th attempt, he or she was promoted to the next base-ID. In the case of the last stair, a *limit* was set by the maximum ID allowed. This *limit* was set because randomly sampling combinations of D and W for IDs higher than 3.2 provided too many distances that are far beyond the physical limits imposed by the resolution of the multitouch display used in the experiment. If participants made too many target selection errors the system beeped and forced the same sub-stair to be repeated.

Results

Our main dependent variable for Experiment 2, *Judgment-Reliability*, was the probability of perceiving an interface with a lower ID to yield a better user performance. It was calculated as the proportion of trials correctly judged out of the total trials for a sub-stair. A participant was considered to judge correctly when he or she rated the sub-trial with the lower ID as providing higher performance, ultimately reducing judgment to a binary decision.

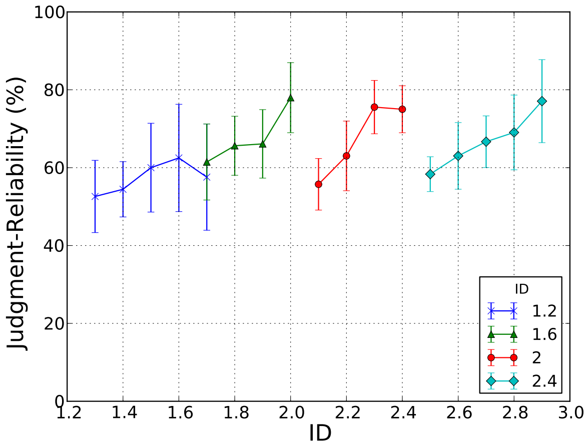

Figure 4. Mean judgment as a function of sub-stair in Experiment 2. Each base-ID is plotted as a separate line. The plot shows only sub-stairs that have at least 2/3 of participants remaining. Error bars show 95% confidence intervals.

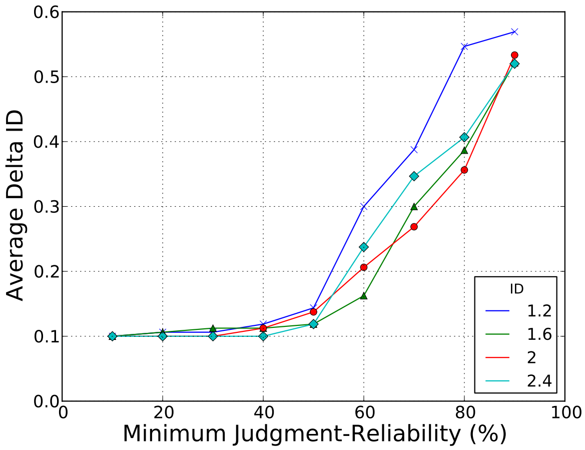

Figure 5. Average ΔID for each minimum Judgment-Reliability level by stair in Experiment 2.

Figure 3 shows a histogram of the number of participants that correctly identified a performance difference on each sub-stair for each stair-ID. There are two key observations. First, no stair showed a 100% completion rate; that is, not all participants were able to achieve a correct judgment above 90% before being pushed onto the next stair. Second, completion rate increased through the stairs. The number of participants that successfully detected a difference was 13, 14, 15 and 15 for IDs 1.2, 1.6, 2 and 2.4 respectively.

Figure 4 shows mean Judgment-Reliability per sub-stair for each of the base-IDs. Only sub-stairs including at least 2/3 of the participants are included in the figure, to ensure representativeness of the whole user group. Figure 4 reveals the

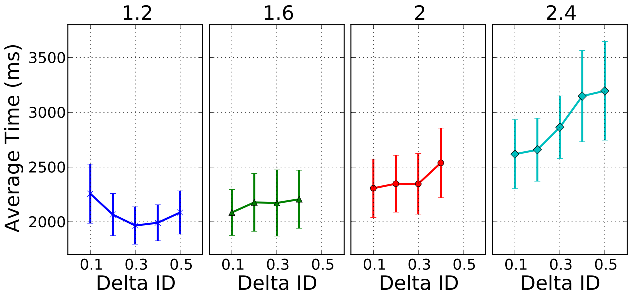

Figure 6. Average time in milliseconds per sub-stair for each stair in Experiment 2. Only sub-stairs containing at least 2/3 of participants are shown.

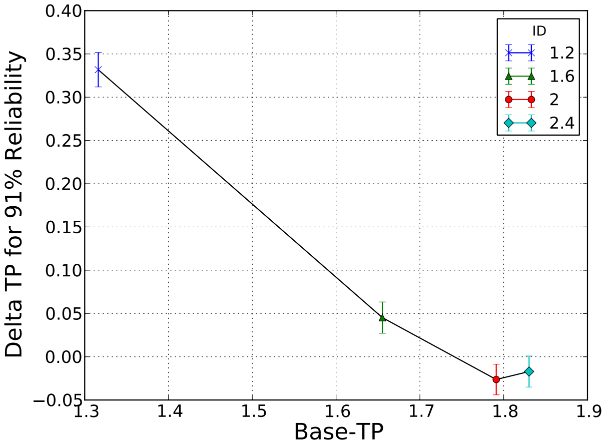

Figure 7. Average ΔTP required for a 91% reliable judgment in performance in Experiment 2, where $TP = ID/MT$. Error bars show 95% confidence intervals.

expected trend that the reliability of perceiving a performance difference increased as participants progressed "up" the sub-stairs, coinciding with increases in completion rates.

Figure 5 shows the average ΔID by minimum Judgment-Reliability. Figure 5 confirms the already mentioned trend that the reliability increased as participants progressed "up" the sub-stairs. Moreover, it also shows that the intuition that can be gained from the histogram in Figure 3 of approximately six sub-stairs being necessary for most participants ($\sim$70%) to reliably notice a difference holds. Base-$ID = 1.2$ is the exception, for reasons observed in Experiment 1: for this base-ID, participants struggled to reliably judge differences due to the small changes in *Distance* and *Width*.

Relationship between ID and Time

Participants tended to take longer to complete the pointing task on later stairs. Figure 6 shows the average *Time* in milliseconds for each sub-stair that still contains at least 2/3 of the participants. The plot shows that *Time* remains fairly constant for the first two stairs, but it increases steadily on the last two stairs, revealing the increased difficulty in tapping the targets. In other words, there is a linear component as presumed by Fitts' law, and some curvature at the extremes due to changes in sensorimotor demands.

Effect of Throughput (TP)

An outstanding question is to what extent it is (TP) that is predictive of Judgment-Reliability, rather than ID.

Figure 7 shows the difference in TP required for a 91% reliable judgment, shown as a function of base-TP. TP was calculated as $TP = ID_e/MT$, where $ID_e = \log_2((D/W_e)+1)$. W_e was calculated as 4.133 multiplied by the standard deviation of all total widths for the condition [15], where a total width was defined as the distance from the center of the target to the x-coordinate where the user touched.

If Interface A is able to provide TP_a bits/s and Interface B TP_b bits/s, how likely is that difference noticeable? We chose the 91 percent cut-off as it corresponds to the stair advance rule used by our method. However, because we did not manipulate throughput (TP) but ID in our experiment, we have a narrower range of TP-differences in the data. This analysis should therefore be regarded as tentative.

The main observation is that *smaller* TP differences are required for larger base-TPs. In other words, when user performance is high, only a small difference is required. However, when it is very high (here, above 1.7 bits/s), TP does not predict the reliability of judgments. Instead, users use some other criteria. In contrast, for low-TP conditions, users require a relatively large (0.35 bits/s) difference. Comparing the curve to our analysis of ID, we conclude that ID is a more powerful predictor for this task.

Individual Differences

To learn about individual differences, we split participants into two groups according to the median of *Judgment-Reliability*. "Better judgers" showed a tendency to stick to two particular strategies when rating their performance: differences in *Width* and *Distance*, or differences in *Time*. For stair-*ID*s 1.2 and 1.6 these participants judged correctly when the *Time* difference was above 500 ms. This practice decreased in the higher stairs as the increase in distance reduced the capacity to estimate *Time*. Participants instead chose shorter *Distances* and smaller *Widths* (depending on the interface quality they looked at) when being unable to estimate *Time*. In contrast, the worse judgers do not reveal an obvious behavioral pattern.

Modeling

Table 2 presents regression models fitted on the data that predict the probability that a participant detected a performance difference for a given ΔID.

To make the models presented in Table 2 more generalizable and reduce the proneness to overfitting we set out to identify a

Regression Model	ID	R^2
$1.38/\left(1+e^{(-0.23-\log x)/0.46}\right)$	1.2	0.96
$0.88/\left(1+e^{(-0.81-\log x)/0.24}\right)$	1.6	0.99
$1.05/\left(1+e^{(-0.73-\log x)/0.39}\right)$	2.0	0.98
$1.08/\left(1+e^{(-0.70-\log x)/0.30}\right)$	2.4	0.97

Table 2. Regression models predicting the probability a participant reliably detected a performance difference as a function of ΔID in Experiment 2.

simplified general predictive model. We found such a model in the form of a variant of the logistic function. This model, referred to as the *generalized model* from now on, predicts the probability $p(x)$ that a user will judge his or her performance as different between two alternative interfaces separated by an ID step-size of x is:

$$p(x) = \frac{1}{1 + e^{-(ax-b)}}, \tag{3}$$

where a and b are model parameters. Based on our data we estimate $a = 8$, $b = 4$. The generalized model is shown with a dashed line in Figure 8 (*GenMod*). We evaluated this model against the actual judging data we collected in Experiment 2 and found that it had good fit for the actual observed frequencies of users being able to reliably judge their performance as being noticeably better for a given ID step-size. The R^2 goodness of fits were 0.71 for $ID = 1.2$, 0.94 for $ID = 1.6$, 0.92 for $ID = 2.0$, 0.98 for $ID = 2.4$, and 0.94 for all IDs.

MODEL VALIDATION STUDY

Having identified the generalized model from data collected in Experiment 2, we validated its predictive power on an alternative task outside of the experimental framework that shaped its functional form. For this purpose we created a variant of the *Whac-A-Mole* game for the Microsoft PixelSense, model Samsung SUR40. In the game, the user has to hit two sets of five circular targets (see Figure 9) in a predetermined order as quickly as possible. Thereafter the user chooses which of the trials resulted in a higher performance.

Method

Each of the circular targets were set with a constant ID. The width of the target was sampled uniformly from an interval between 30 pixels (13.84 mm) and 199 pixels (91.79 mm). The distance to the target was then fully determined by the fixed ID and the width of the target. The configuration of the five targets was randomly generated within the constraints given above. In the game, when the user hit the “Start” button a pre-determined circular target was highlighted. The user was instructed to hit this target as quickly as possible. Thereafter another circular target was highlighted. When the user had hit all the highlighted targets the game was over. The highlighted sequence guided the user through a series of target acquisitions that all had the same ID for an individual game, but all the targets had randomly generated widths and heights within the constraints given above.

We recruited 11 participants from a university campus. The average age was 26.3 years, ranging from 20 to 55 years (*sd* = 11.2). Five were female, six were male. None of the participants had participated in Experiment 1 or 2.

Using the same definition of performance as in Experiment 1 and 2, participants were asked after every two games to choose the game in which they perceived they experienced a better performance. Each pair of games compared a particular base-ID against an alternative ID. To make the results of our validation exercise stronger we took two measures to reduce the possibility of false-positives. First, we took three out of three correct judgments as proof of a participant recognizing a change in performance, thus the probability of choosing the correct ID by chance was only $\frac{1}{8}$. Second, we selected a single base-ID of 0.9. This choice was made based on the observations from Experiment 2, where we noticed that participants had greater difficulty in recognizing differences in low-ID conditions. Coincidentally, this makes the prediction task of our generalized model more challenging as the chosen base-ID for the game is lower than the lowest base-ID of 1.2, which had the worst R^2 fit in Experiment 2.

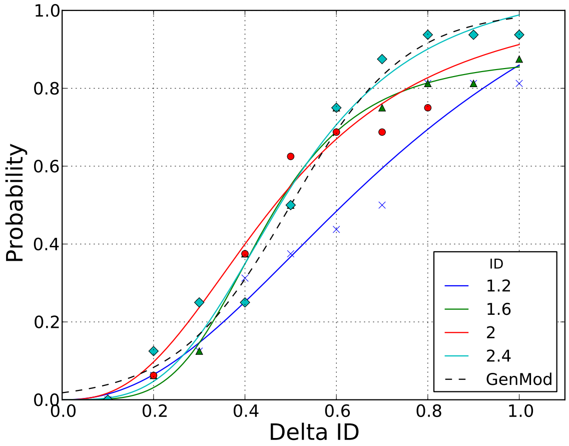

Figure 8. Cumulative probability of reliable judgment in Experiment 2. The models are rendered for base-IDs, and the generalized model (see the text) is shown with a dashed line. The plotted points are the cumulative number of participants that had detected the ID change by that sub-stair.

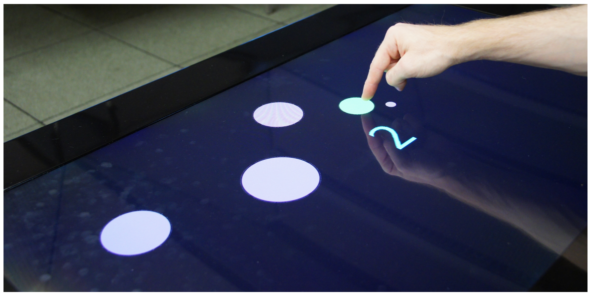

Figure 9. An illustration of the *Whac-a-Mole* game.

In the game we compared nine ΔIDs: $0.1, 0.2, \ldots, 0.9$ for a single base-ID of 0.9. Thus, the ID comparisons ranged from 0.9–1.8. In total, each user played 1 base-ID × 9 ΔIDs × 3 pairs of games per ID comparison = 54 games (27 pairs of games). It took a user circa 7–8 minutes to play all 54 games, which mimics a typical walk-up use-scenario for a tabletop interface.

Results

We calculated the probabilities of a user detecting a change in the base-ID in the data. The procedure was identical to Experiment 2. We then compared the generalized model of perceivable performance gain using the same parameters $a = 8$, $b = 4$ we found in Experiment 2.

We found that the judgments about the ID differences in the game had an R^2 goodness of fit of 0.72. This should be compared against the closest comparable base-ID we used in Experiment 2 (ID = 1.2) and its R^2 model fit of 0.71, which was obtained in a much more controlled setup with less noisy data and more participants. Our good model fit for the game data indicates that our generalized model of perceivable performance gain does generalize beyond the experimental task that shaped its functional form and model parameters.

SUMMARY AND DISCUSSION

Psychophysics models are of paramount importance in graphics, audio, and multimedia, thanks to their ability to directly inform design and engineering. Psychophysics could be a useful asset in the HCI toolbox, too, if it allows us to make reliable quantitative predictions about the effect of a change in an interface on users' perception of interaction. This paper has extended the application of psychophysics in HCI from passive perception to interaction.

To this end, we have presented a method for operationalizing another important psychophysical dependent variable in HCI: a user's *perception* of his or her own performance in a user interface. By definition, user performance is the efficiency of a user carrying out a task [20], which in HCI is measured in terms of speed and accuracy aggregated over acts exhibited in the course of a task. We have presented an adaptation of psychophysics methodology into an interactive task. A central insight is that a change in interface design assumes its potential effect via two interconnected routes: via overt perceivable changes and via changes in a user's objective performance. Changing something in a layout may be noticeable via perception of the elements and/or via how it changes the speed and accuracy of a user's performance. This was operationalized by manipulating the *difficulty* of sensorimotor movement via ID. We registered both interface qualities and objective user performance to predict the user's judgment. A drawback in comparison to more realistic tasks is that the two interfaces in our experiments are used immediately after one another, whereas in actual use they are probably separated more in time.

The results in this paper provide insights into users' perception of interaction. In Experiment 1, we found that:

- Users can indeed reliably judge their performance, but they also exhibit slight biases. Low-ID conditions were markedly different from others: interfaces consisting of large targets close to each other provide no reliable ground for judging one's performance.
- *Width* and *Distance* are not as important cues as *Time*: when interfaces look similar, the differences in the user's own performance dominate the judgment.

In Experiment 2, a staircasing design enabled us to learn what happens when ID changes between two to-be-compared interfaces. We learned that, depending on the base-ID, users' judgment capabilities change. In particular:

- For small base-IDs, users struggle to consistently identify differences.
- Higher base-IDs perform better because participants can obtain more information to make their judgment, such as greater changes in *Distance*, *Width* and *Time*. These differences are needed to improve the reliability of users' judgments.
- Throughput (TP) is not as strong a predictor as ID is. Judgment reliability seems to only be predictable for low-TPs.
- In no case was the user's perception of performance reducible to ID alone, which confirms that the perception of performance is disjoint from actual performance as measured by TP.
- There are inter-subject differences in judgment ability.

Further, we have discovered a mathematical model that predicts the probability that a user will notice a performance difference, as defined by a change in ID. The model was validated in a *Whac-A-Mole* game, which demonstrated that the model's predictions generalized to at least one other HCI task involving discrete aimed movement.

However, we acknowledge that there could be other effects influencing the user, which he or she could not perceive, especially over a prolonged period of time. For example, when the visual layout is more complex, users will have more cues than distance and width of targets that they can use to assess the user interface. Moreover, when the interaction task is a compound task (i.e. it consists of many subtasks), users might base their performance judgments with respect to any of the subtasks. However, we believe that the results in this paper are promising enough to motivate further research in this direction.

Application to Design Problems

Currently, the models derived in this paper can be considered heuristics for interfaces where pointing is important. Such activities involve command selection (menus, hotkeys, etc.), text entry and gaming.

The regression models in Table 1 require knowing the ID of the two user interfaces to be compared, and the difference that is most pronounced between them: this can be the user's speed, accuracy, target distances, or target widths. A model should be selected based on this information.

The regressions models in Table 2 require knowing the IDs of the two user interfaces that are being compared. The lower of these should match or must be matched to the closest base-ID in the table. The output of the model is an estimation of the probability that a user will be able to *reliably* notice the difference in user performance.

It is important to note that the regression models in Table 2 are likely to suffer from overfitting. For general design problems we recommend using the generalized model in Equation 3.

CONCLUSIONS

Psychophysics modeling could have numerous applications in HCI beyond its current limited use. We envision that the ability to better anticipate the subjective perception of interaction might save substantial money in interface development.

The methodology presented in this paper is a starting point for understanding the perception of user performance more fully. With little effort, the paradigm could be extended to continuous aimed movements [1], allowing the study of steering and pursuit tasks common in for example gaming and driving. The methodology can also be extended to tasks involving timing and rhythm, such as music [16]. To extend it beyond aimed movements, the challenge is to find models that allow manipulating user performance with predictable effects.

ACKNOWLEDGEMENTS

This research was supported by the Scottish Informatics and Computer Science Alliance, the Max Planck Centre for Visual Computing and Communication and the Cluster of Excellence for Multimodal Computing and Interaction at Saarland University. Max Nicosia's undergraduate internship in St Andrews was supported by a donation from Nokia.

REFERENCES

1. Accot, J., and Zhai, S. Beyond Fitts' law: models for trajectory-based HCI tasks. In *Proc. CHI 1997*, ACM Press (1997), 295–302.
2. Biggs, S. J., and Srinivasan, M. A. Haptic interfaces. In *Handbook of Virtual Environments*, K. M. Stanney, Ed. Lawrence Earlbaum, 2002, 93–116.
3. Block, F., and Gellersen, H. The impact of cognitive load on the perception of time. In *Proc. NordiCHI 2010*, ACM Press (2010), 607–610.
4. Brewer, B. R., Fagan, M., Klatzky, R. L., and Matsuoka, Y. Perceptual limits for a robotic rehabilitation environment using visual feedback distortion. *IEEE Transactions on Neural Systems and Rehabilitation Engineering 13*, 1 (2005), 1–11.
5. Cunningham, D., and Wallraven, C. *Experimental Design: From User Studies to Psychophysics*. AK Peters, Ltd., 2011.
6. Ganel, T., Chajut, E., and Algom, D. Visual coding for action violates fundamental psychophysical principles. *Current Biology 18*, 14 (2008), R599–R601.
7. Gephstein, S., and Banks, M. S. Viewing geometry determines how vision and haptics combine in size perception. *Current Biology 13*, 6 (2003), 483–488.
8. Grondin, S. Timing and time perception: a review of recent behavioral and neuroscience findings and theoretical directions. *Attention, Perception, & Psychophysics 72*, 3 (2010), 561–582.
9. Guiard, Y., and Olafsdottir, H. B. On the measurement of movement difficulty in the standard approach to Fitts' law. *PloS ONE 6*, 10 (2011), e24389.
10. Hamada, K., Yoshida, K., Ohnishi, K., and Koppen, M. Color effect on subjective perception of progress bar speed. In *Proc. INCoS 2011*, IEEE Press (2011), 863–866.
11. Hassenzahl, M. The thing and I: understanding the relationship between user and product. *Funology* (2005), 31–42.
12. Law, E. L.-C., Roto, V., Hassenzahl, M., Vermeeren, A. P. O. S., and Kort, J. Understanding, scoping and defining user experience: A survey approach. In *Proc. CHI 2009*, ACM Press (2009), 719–728.
13. Le Gall, D. MPEG: A video compression standard for multimedia applications. *Communications of the ACM 34*, 4 (1991), 46–58.
14. Lidwell, W., Holden, K., and Butler, J. *Universal Principles of Design*. Rockport Publishers, 2010.
15. MacKenzie, I. S. Fitts' law as a research and design tool in human-computer interaction. *Human-Computer Interaction 7*, 1 (1992), 91–139.
16. Maury, S., Athénes, S., and Chatty, S. Rhythmic menus: toward interaction based on rhythm. In *Ext. Abstr. CHI 1999*, ACM Press (1999), 254–255.
17. Nadenau, M. J., Winkler, S., Alleysson, D., and Kunt, M. Human vision models for perceptually optimized image processing—a review, 2000. Unpublished manuscript.
18. Rank, M., Shi, Z., J. Müller, H., and Hirche, S. Perception of delay in haptic telepresence systems. *Presence: Teleoperators and Virtual Environments 19*, 5 (2010), 389–399.
19. Reddy, M. Perceptually optimized 3D graphics. *IEEE Computer Graphics and Applications 21*, 5 (2001), 68–75.
20. Sanders, A. F. *Elements of Human Performance: Reaction Processes and Attention in Human Skill*. Lawrence Erlbaum, 1998.
21. Seow, S. C. *Designing and Engineering Time: The Psychology of Time Perception in Software*. Addison-Wesley, 2008.
22. Shi, Z., Hirche, S., Schneider, W. X., and Muller, H. Influence of visuomotor action on visual-haptic simultaneous perception: A psychophysical study. In *Proc. Haptics 2008*, IEEE Press (2008), 65–70.
23. Soukoreff, R. W., and MacKenzie, I. S. 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.
24. Stevens, S. S. On the psychophysical law. *Psychological Review 64*, 3 (1957), 153–181.
25. Yang, X. D., Bischof, W. F., and Boulanger, P. Perception of haptic force magnitude during hand movements. In *Proc. ICRA 2008*, IEEE Press (2008), 2061–2066.
26. Zanker, J. M. Does motion perception follow Weber's law? *Perception 24* (1995), 363–372.
27. Zwicker, E., and Zwicker, U. T. Audio engineering and psychoacoustics: Matching signals to the final receiver, the human auditory system. *Journal of the Audio Engineering Society 39*, 3 (1991), 115–126.