Exploring Dwell-time from Human Cognitive Processes for Dwell Selection

TOSHIYA ISOMOTO, University of Tsukuba, JAPAN
SHOTA YAMANAKA, Yahoo Japan Corporation, JAPAN
BUNTAROU SHIZUKI, University of Tsukuba, JAPAN

In order to develop future implicit interactions, it is important to understand the duration a user needs to recognize a visual object. By providing interactions that are triggered after a user recognizes an object, confusion resulting from the discrepancy between completing a cognitive process, which we define as the process from perceiving a visual stimulus to determining a selection, and triggering an interaction can be reduced. To understand this duration, we developed a model to derive dwell-times, allowing dwell selection to be performed after completing a cognitive process based on the Model Human Processor and the number of fixations. Our model revealed a minimum dwell-time of 174.2 ms for a colored target selection task. For an image selection task, the minimum dwell-time was 272.5 ms, which increased to 835.8 ms when a participant had not previously fixated on the object.

CCS Concepts: • **Human-centered computing → HCI theory, concepts and models**; *User models*; Interaction techniques.

Additional Key Words and Phrases: gaze interface, model human processor, user modeling, perceptual behavior, implicit interaction

ACM Reference Format:
Toshiya Isomoto, Shota Yamanaka, and Buntarou Shizuki. 2023. Exploring Dwell-time from Human Cognitive Processes for Dwell Selection. *Proc. ACM Hum.-Comput. Interact.* 7, ETRA, Article 159 (May 2023), 15 pages. https://doi.org/10.1145/3591128

1 INTRODUCTION

Implicit interactions have been gaining attention in the human-computer interaction (HCI) field as future interactions, moving instead of the current explicit interactions, such as using a mouse or touch panel. Explicit interactions rely on a user's deliberate actions, allowing them to trigger the interaction at will. In contrast, implicit interactions are triggered in conjunction with the detected user's intent, based on human behavior. One of the advantages of implicit interaction is that it can be performed before or simultaneously as a user decides to perform it. Although this advantage holds promise for faster and more intuitive interactions, there has been limited investigation into the appropriate timing for triggering such interactions from the perspective of human cognitive processes, which we define as the processes occurring between a human perceiving a visual stimulus and determining a selection.

For example, dwell selection, which enables a user to select a target by merely looking at it for a certain duration [15–17], relies on a time threshold called dwell-time to detect the user's intent to select the target. From the HCI perspective that faster interactions are preferable, most dwell selection research has focused on using shorter dwell-times. However, previous studies have reported that a small dwell-time (e.g., <200,ms) decreases usability, leading to user confusion [9,

Authors' addresses: Toshiya Isomoto, isomoto@iplab.cs.tsukuba.ac.jp, University of Tsukuba, Tsukuba, Ibaraki, JAPAN; Shota Yamanaka, syamanak@yahoo-corp.jp, Yahoo Japan Corporation, Chiyoda, Tokyo, JAPAN; Buntarou Shizuki, shizuki@cs.tsukuba.ac.jp, University of Tsukuba, Tsukuba, Ibaraki, JAPAN.

12, 34]. Thus, from the viewpoint of usability in gaze-based interaction, a small dwell-time may not always be the optimal solution, contradicting the HCI perspective. Since "looking" is a human behavior closely connected to cognitive processes, a user may feel confused if an interaction, such as target selection, is performed before they complete a cognitive process in dwell selection, regardless of whether the interaction is successfully executed. While determining appropriate dwell-times is crucial for faster, more accurate, and more usable dwell selection, there is no systematic approach to determine dwell-time. Consequently, various dwell-times have been used across different studies and tasks.

If we derive the duration a user requires to complete a cognitive process, it can be helpful for future interactions, including gaze-based interaction. We developed a model that derives the dwell-time systematically, enabling dwell selection after a user completes the cognitive process based on their behavior. We first devised three hypotheses regarding the relations between the information of fixation and cognitive processes. Referring to the findings of a previous study [13] involving the model human processor (MHP) [8], we then developed our model to derive the dwell-time using a number of fixations ($N_{fixation}$) and the duration of fixations ($D_{fixation}$) that a user performs for a target. The limitations of that study are that the $N_{fixation}$ required for recognizing a target should be predicted beforehand, but no prediction method was presented, and that the applicable situation is only for image selection [13]. In contrast, we use $N_{fixation}$ determined from user behavior. Using this $N_{fixation}$, our model derives the dwell-time that fits the relation between $N_{fixation}$ and cognitive processes. Since cognitive processes differ depending on the task, we conducted five selection tasks of different cognitive levels to evaluate our hypotheses and develop our model. We summarize the contributions of this paper as follows.

- We devised three hypotheses about fixation during a selection and validated them through an experiment involving five tasks having different cognitive levels.
- We developed a model that derives dwell-time systematically from the perspective of human cognitive processes.
- We showed how our model derives dynamically changing dwell-times based on user behavior, especially $N_{fixation}$.

2 RELATED WORK

We describe how dwell-time has been determined in research on dwell selection. We then describe how the MHP interprets the user's cognitive processes during a target selection and how dwell-time is determined using it.

2.1 Dwell-time in Dwell Selection

In dwell selection, solving the Midas-touch, which is a user's unwanted selection, with a small dwell-time has been investigated. Researchers on dwell typing (i.e., dwell selection on a key) used 180–600 ms as dwell-times from two perspectives: user preference and robustness against the Midas-touch [19, 22, 28, 29, 32]. To select a key that is likely to be selected, a small dwell-time prevents the Midas-touch and enables a faster selection, while using a large dwell-time prevents the Midas-touch for a key that is unlikely to be selected. Dwell-times dynamically decrease/increase along with the previously typed keys and the probability of the next key typed.

Researchers reported that the dwell-time that can prevent the Midas-touch differs depending on the task. For example, 650 ms is adequate for a low-cognitive (selecting two digits) task, 1,100 ms is not adequate for a high-cognitive task (selecting two-words) [34], and 400 ms is adequate for selecting a colored circle [22, 33]. This is because a selection task includes visual searching. Visual-search time varies with the size, arrangement, and color of a target or its content [25, 31]. Since

"looking" and "searching" are human behaviors, a large dwell-time should be used to separate those behaviors and dwelling and to prevent the Midas-touch. For searching tasks in a shopping situation, [26] reported that 1,000 ms is not adequate and 2,000 ms should be used.

These dwell-times are determined to prevent the Midas-touch as well as provide a faster interaction. However, since gaze-based interaction is closely connected to cognitive processes, dwell-times should also be derived from these processes, not only considering the Midas-touch and faster interaction. With a dwell-time less than 100 ms, a user would feel confused when the target is selected before they look at it [12]. This suggests the necessity of investigating a dwell-time that enables users to select a target just after finishing their cognitive processes.

Since cognitive processes differ depending on the user's previous behavior, we assumed that dwell-time could be dynamically changed along with the user's behavior, especially with the user's fixation before a selection. This dynamic change of dwell-time is the same concept as dwell typing. For example, if a user knows about an object or has looked at it many times, they can quickly recognize it. Thus, our model derives a small dwell-time systematically by taking into account cognitive processes and potentially reduces user confusion.

2.2 Model Human Processor [8]

The MHP demonstrates human perceptual behavior in response to visual (and auditory) stimulus by dividing the information-processing system into three subsystems: perception, cognition, and motor. The perception subsystem completes *perceiving* a visual stimulus and encoding into a visual code within τ_p=100 [50–200] ms. Each range indicates that the Fastman (e.g., an expert) takes the minimum time and the Slowman (e.g., a novice) takes the maximum time. The cognition subsystem completes *recognizing* the visual code, *classifying* the recognized code into a meaning, *matching* the meaning and instruction loaded on the working memory beforehand, and *requesting* to press a button. The required cognitive processes differ among tasks. The time taken for one cognitive process (τ_c) is 70 [25–170] ms. The motor subsystem completes *pushing* a button along with the request from the cognition subsystem within τ_m=70 [30–100] ms.

From the MHP, we can interpret a human's perceptual behavior on performing an interaction, especially selection with button pressing. While the cognitive processes in the perception and motor subsystems are the same regardless of the task, those in the cognitive subsystem differ depending on the task. For a simple reaction task, the cognitive subsystem requires one process (*requesting*). For a class-match task, however, the subsystem requires four processes (*recognizing*, *classifying*, *matching*, and *requesting*).

A previous study using the MHP explored a user's preferred dwell-time in the dwell selection of an image referring to the class-match task [13]; the preferred dwell-time is derived from the model as follows:

$$\tau_p + (3\mathrm{N}_{\mathrm{fixation}} + 1)\tau_c \ [13]. \tag{1}$$

In Equation 1, one cognitive-cycle, which consists of *recognizing*, *classifying*, and *matching* processes, is done during one fixation, and after a certain $\mathrm{N}_{\mathrm{fixation}}$, *requesting* is processed. Thus, $3\mathrm{N}_{\mathrm{fixation}}$+1 of τ_c is required for completing a cognitive process. Because dwell selection does not require pressing a button, τ_m is not counted. For example, if we know that $\mathrm{N}_{\mathrm{fixation}}$ required to complete an image selection is three, the user's preferred dwell-time may be 800 ms ($100 + (3 \times 3 + 1) \times 70$). The limitations of that previous study [13] are that the method of determining $\mathrm{N}_{\mathrm{fixation}}$ is unclear since they reported that predicting $\mathrm{N}_{\mathrm{fixation}}$ beforehand is challenging and that the applicable task is only for image selection. We improved upon their model using a user's previous selection behavior, especially using $\mathrm{N}_{\mathrm{fixation}}$ and $\mathrm{D}_{\mathrm{fixation}}$ before a selection. We developed our model on the basis of the results of five different selection tasks to make it more applicable to a wider range of tasks.

2.3 Models of Human Cognitive Processes and Behavior for Visual Search Tasks

In many studies, human behavior and cognitive processes were modeled, similarly to the MHP. For instance, the Adaptive Control of Thought–Rational (ACT-R) model [2, 3] is a representative model of cognitive processes, including visual attention. The ACT-R model reports that it takes a human 186 ms to shift attention, with or without eye movement. In a visual search task, three processes repeatedly occur: 1) responding "yes" (i.e., a looking candidate is a target), taking a "base" time of 208 ms, 2) shifting attention, taking a "shift" time of 186 ms, and 3) responding "no" (i.e., there is no target after searching all candidates), taking a "base" and "neg" time of 133 ms[1]. Another representative model is Fitts' law [11, 18], which is for pointing behavior. The time for pointing is expressed as $a \times \log_2(A/W + 1) + b$, where A is the distance between the position of a cursor and target, D is the target size, a can be interpreted as the time required for the motor process (e.g., moving a hand for a mouse-based interaction), and b can be interpreted as the time required for decision-making and triggering action. Numerous models have been proposed for GUIs (e.g., [5, 10, 27]).

These models provide a precise representation of human cognitive processes and behaviors, including the time required for each, making them useful for systematically determining dwell-time. However, we use the MHP for the following reasons. The MHP describes the human cognitive process through three systems, each having a required time (τ_p for the perception system, τ_c for the cognition system, and τ_m for the motor system). The MHP describes processes required for completing visual search tasks, as outlined in Section 2.2.

3 HYPOTHESES

We devised the following three hypotheses.

- **H1.** *N_{fixation} required for selecting a target increases along with the cognitive level of a task.* We assume that the user needs to fixate on the target many times to recognize a more complex target before selecting it.
- **H2.** *D_{fixation} of the last fixation before a selection decreases along with increasing total N_{fixation}.* We assume that the user can select the target by looking at it for a shorter period by fixating on it frequently and recognizing it beforehand.
- **H3.** *D_{fixation} for large N_{fixation} converges to one duration regardless of the task.* We assume that if a user has already recognized a target, they can make a decision to select the target with a duration regardless of the target.

4 EXPERIMENT

We used five selection tasks with different cognitive levels to verify the hypotheses and determine the N_{fixation} required for a selection and D_{fixation}.

4.1 Participants and Apparatus

We recruited 20 university students (one female and 19 males, all Japanese) aged 20–26 (M = 22.9). They use GUI-based interfaces on a daily basis. Fifteen previously participated in an experiment using an eye-tracker. Each received JPY 2,500 (~USD 18).

We used the Tobii Pro Spectrum, which samples gaze data at 1,200 Hz (0.833 ms/sample). The eye tracker was attached to the bottom of a 24-inch (1980×1080 pixels) non-glare display. The participants' heads were positioned approximately 65 cm from the display. The participants used

[1]These values depend on the difference in the types of targets and distractors (e.g., letters versus numbers) and the number of candidates present in a visual search task.

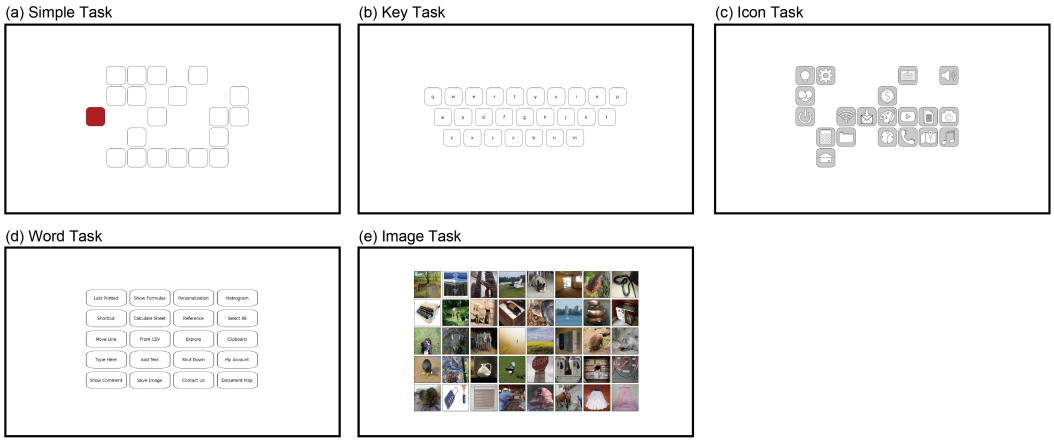

Fig. 1. Displays used for Experiment 1.

a wire-connected keyboard to control the task. The experiment was conducted in a room with fluorescent light at approximately 810 lux.

4.2 Selection Method

We used *gaze-button* selection, which is performed on the gaze coordinate when pushing the 'Enter' key of the keyboard. Systematically, the selection is allowed when the gaze coordinate is inside a target; otherwise, no selection is performed even if the participants push the key.

4.3 Selection Task and Interface

For gaze-button selection, we asked the participants to complete five selection tasks: *simple, key, word, icon*, and *image*. One trial involved completing a selection. Each task consisted of 51 trials. We used this number by taking into account the concentration and fatigue of the participants and used the first trial as a training trial (not used for our analysis). The order of the tasks was randomized among the participants. Before beginning the experiment, we calibrated the eye tracker with Tobii's 9-point calibration for each participant. The task began with the instruction display, which gave instructions to the participants for each task. The participants read the instructions then pushed the space key to move on. The task display (Fig. 1) was then shown, and the participants were asked to select a target with gaze-button selection. Between the tasks, we asked the participants to take rest for at least one minute. The experiment took approximately 25 min.

The task display consisted of candidates specific to the task and one target. We did not give the participants visual feedback for all tasks to eliminate any potential side effects. We determined the target size at which the eye-tracking performance (i.e., the offset and precision) did not affect the selection, as described in Sections 4.3.1–4.3.5. Although gaze data should be collected from various tasks, it is difficult to conduct an experiment with such diverse tasks. Thus, we used these five tasks representing daily interaction situations in reference to a previous study [14]. We list the relationship between tasks and cognitive levels in Table 1.

4.3.1 Simple task. The simple task involves selecting a red circular target (Fig. 1a). We instructed the participants to "select a red object." This task is similar to the *simple reaction* task in which a participant pushes a button after the visual stimulus is displayed [8]. The selection time shows the minimum duration necessary for a human to perceive the visual stimulus. We displayed one red

Table 1. Our definition of tasks and their cognitive levels. "Known candidates" means whether or not a participant knew which keys/icons/words/images were shown in candidates before a task began. We assigned "Cognitive levels" in accordance with the row of "Similar task in MHP" that each task requires a different number of required cognitive processes. The difference in "Known candidates" between the key and icon tasks results in a different cognitive level even though the task in the MHP is the same.

Tasks	Target type	Known candidates	Similar task in MHP	Cognitive level
Simple	colored object	beforehand	simple reaction	1 (minimum)
Key	key	beforehand	physical match	2
Icon	desktop icon	beforehand	physical match	3
Word	menu item	depend on candidate	name match	4
Image	image	none	class match	5 (max)

target and 19 white candidates in a random position in an 8×5 grid. The size of each target was 2.5°×2.5°.

Since there is one red target and the others are white, the participants would not need to search for it and would know all candidates before the task. This selection corresponds to a real situation of a preprogrammed selection that can be done with less confirmation. For example, a close button of the web browser could be selected with less confirmation. Such buttons are positioned at the top corner of the browser[2], and the user knows the content before looking at it. A situation of selecting the most frequently selected targets is another example. Because these situations would be the easiest interaction situations, we defined the cognitive level of the simple task as the lowest among the tasks.

4.3.2 Key task. The key task involves selecting a key (Fig. 1b). For example, we instructed the participants to "select [a] key". This task is similar to the *physical-match* task in which a participant pushes a button if a visual shape of a candidate and an instruction are the same [8]. We displayed 26 keys in a qwerty alignment. One key among the 26 was randomly chosen as a target. The size of each target was 2.5°×2.5°.

Since we used a qwerty alignment, which the participants were familiar with, they could recognize the position of all candidates and the content (i.e., a key) with less effort. However, more recognition is needed to confirm a target than in the simple task. This task corresponds to a real situation of a key selection and a selection of a radio button with a character.

4.3.3 Icon task. The icon task involves selecting a target that resembles a desktop icon (Fig. 1c). For example, we instructed the participants to "select a [call] icon". This task is similar to the *name-match* task in which a participant pushes a button if the meaning of a candidate and instruction are the same [8]. We used an icon set consisting of 20 icons that resemble desktop icons. As opposed to the key task, the instruction and target differ (i.e., verbal instruction and visual target) while their meanings are the same. We displayed one target and 19 candidates in a random position in an 8×5 grid. The size of each target was 2.5°×2.5°.

Before beginning the task, we asked the participants to memorize the correspondence between the images and instructions to eliminate preconceived notions on the basis of previous interaction experience. The participants needed to recognize the visual stimulus then match the meanings of the stimulus and instruction before pushing the button. This selection task corresponds to the real situation of a relatively simple image selection. For example, the desktop icons and tab-icons of the web browser that a user already knows before looking at them.

[2]top-right in Microsoft Edge and top-left in Safari

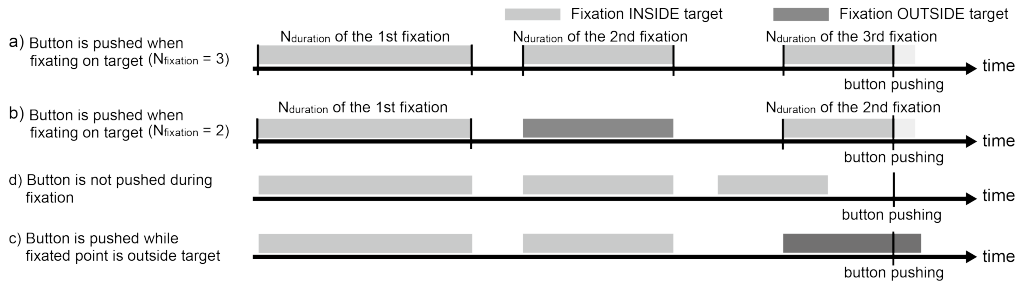

Fig. 2. Fixations we used (a and b) and did not use for later analysis (c and d).

4.3.4 Word task. The word task involves selecting a one- or two-word target consisting of at least eight characters (Fig. 1d). For example, we instructed to the participants to "select a [copy text]". We created a word set consisting of 20 words extracted from text- and image-editing interfaces such as Microsoft Word and Adobe PhotoShop. Similar to the key task, both the instruction and target are verbal. The difference is the character length: one character vs. at least eight characters. We randomly selected one target and 19 candidates from the word set in a random position in a 4×5 grid. The size of each target was 5.5°×2.5°.

Unfortunately, there is no similar task in the MHP [8]; however, the word task was used as a higher cognition level than the one-character task [34]. For the above reason, the word task required a higher cognitive level of the participants than the simple, key, and icon tasks.

4.3.5 Image task. The image task involves selecting an image target (Fig. 1e). For example, we instructed to the participants to "select a [dog] icon". We used the image-set extracted from Visual Genome[3]. As opposed to the icon task, we did not show all images to participants beforehand. While the icon and image tasks are both selection tasks against nonverbal candidates, there is a difference in whether the participants knew or did not know which images/icons were shown in candidates before a task has begun. This task is similar to the *class-match* task in which a participant pushes a button if a class (e.g., a letter or digit) of a candidate and instruction is the same [8]. We displayed one target and 39 candidates randomly selected from the image set in a random position in an 8×5 grid. The size of each target was 3.5°×3.5°.

The participants needed to recognize the visual stimulus, classify the stimulus into an image type (e.g., an image of a dog), then match the classes of the stimulus and instruction before pushing the button. This selection task corresponds to a real situation of relatively more complex image selection than that in the icon task. For example, images in an image-search result and an image that a user rarely sees. For the above reason, the image task requires the highest cognitive level among the tasks.

4.4 Results

We measured $N_{fixation}$ and $D_{fixation}$ that the participant performed on a target before pushing the button. Using them, we first validated our hypotheses then developed our model that derives the dwell-time that allows dwell selection to be performed after a user completes a cognitive process on the basis of their behavior. Although a previous study used a duration in which gaze coordinates are inside a target [13], we used $D_{fixation}$ because it contains more meaningful information than gaze coordinates.

[3]https://visualgenome.org/, licensed under CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/)

Table 2. $N_{fixation}$ required for completing each task. The number in the brackets is that of the participants. For example, twelve participants required two fixations in 20 trials to complete the key task.

Task \ $N_{fixation}$	1	2	3	4	5	6	7
simple	987 (20)	3 (3)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
key	946 (19)	20 (12)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
icon	865 (20)	87 (20)	6 (6)	5 (5)	0 (0)	0 (0)	0 (0)
word	768 (20)	172 (20)	23 (14)	3 (3)	1 (1)	0 (0)	0 (0)
image	548 (20)	308 (20)	79 (20)	3 (3)	2 (2)	2 (2)	0 (0)

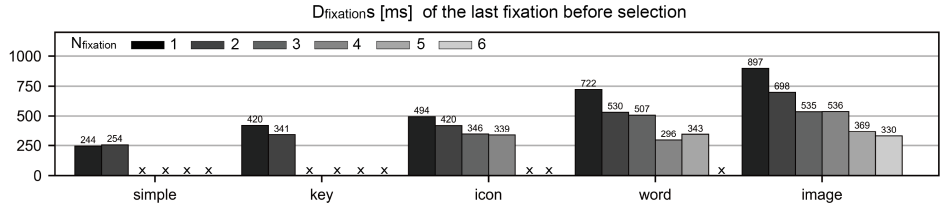

Fig. 3. $D_{fixation}$ of the last fixation before selection for each task and each $N_{fixation}$.

We discarded the first trial of each task as practice, thus used 1,000 (= (51−1) trials×20 participants) trials for each task. Before detecting a fixation, we first excluded eye-tracking noise by applying the median filter with a window size of six samples, which is equal to 5 ms with 1,200 Hz of the eye-tracker. We then applied the I-DT algorithm [30] with a dispersion threshold of 30° and used 100 ms as the minimum duration of the fixation. Thus, we defined the gaze coordinates in which the velocity of gaze movement is below 30°/sec over 100 ms as one fixation. We used the fixations detected with the algorithm and parameters in which the fixation point (i.e., the average gaze coordinates) was inside the target. For later analysis, we used trials in which the participant pushed a button and succeeded in a selection during fixation so that the trials would be consistent with the analysis (Fig. 2a and b). We did not use the trials in which selection was not done during a fixation (Fig. 2c) and the fixation was outside a target (Fig. 2d).

4.4.1 Number of fixations. We detected fixations for 4,828 trials. We could not detect fixations in 172 trials (3.4% of all trials) with the algorithm and parameters due to the eye-tracking noise and our definition of fixation. For example, some noise may have remained and been affected by the algorithm. Since we did not use the trials in which selection was not made during a fixation, these trials were determined as errors, although the participant successfully selected a target. Thus, we removed them as outliers.

We show the $N_{fixation}$ the participants required for completing each task in Table 2. We did not instruct participants on the selection strategy (e.g., select a target with a small $N_{fixation}$), and all selections could be made by looking at a target at least once to observe participants' selection behavior. Although the participants did not frequently require a large $N_{fixation}$, they seemed to require it (e.g., $N_{fixation} \geq 3$) for completing tasks with a higher cognitive level (i.e., icon, word, and image tasks). From these results, we concluded that this result verifies **H1** that $N_{fixation}$ *required for selecting a target increases along with the cognitive level of a task.*

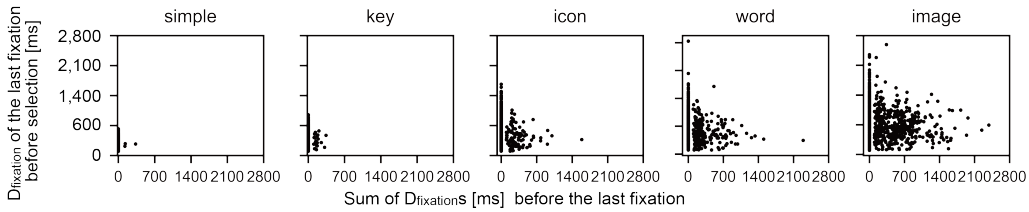

Fig. 4. $D_{fixation}$ of the last fixation against the sum of $D_{fixation}$ before the last fixation.

Table 3. Models and regression results for each task.

Equ.	Model	Simple	Key	Icon	Word	Image
(2)	$a + b\times(N_{fixation}\text{-}1)$	R^2 =1.0 a =244.2 b =10.2 AIC =-111.8	R^2=1.0 a = 420.4 b = -79.4 AIC = -110.5	R^2=0.918 a = 480.5 b = -53.9 AIC = 38.5	R^2=0.860 a = 677.7 b = -99.0 AIC = 58.5	R^2=0.935 a = 834.1 b = -109.2 AIC = 67.8
(3)	$a + b\times\log_2(N_{fixation})$	R^2=1.0 a=244.2 b=10.2 AIC= -111.8	R^2=1.0 a = 420.4 b = -79.4 AIC = -110.5	R^2=0.972 a = 494.3 b = -82.6 AIC = 34.1	R^2=0.907 a = 721.9 b = -175.4 AIC = 56.5	R^2=0.972 a = 905.8 b = -217.9 AIC = 62.8

4.4.2 Duration of fixation. We first measured the $D_{fixation}$ of the last fixation before the participant pushed a button. Because the last fixation included the participant's button pushing in our analysis, we used the $D_{fixation}$ of the last fixation as the duration required for recognizing the target then making a decision. The average $D_{fixation}$ has a tendency to decrease along with increasing $N_{fixation}$, as shown in Fig. 3. When $N_{fixation}$ was one (i.e., the participant fixated a target once), $D_{fixation}$ increased along with the increasing cognitive level of the tasks. We then investigated the relation between the $D_{fixation}$ of the last fixation and the sum of $D_{fixation}$s before the last fixation (Fig. 4). For example, if $N_{fixation}$ is three, we calculate the sum of the first two $D_{fixation}$s, and if $N_{fixation}$ is one, the sum becomes zero. The relation indicates that the $D_{fixation}$ of the last fixation decreases along with increasing sum. That is, when the participant fixated on a target for a long time, they could make a decision in a small duration. Although certain $D_{fixation}$s did not decrease along with the increasing cognitive level and the sum of $D_{fixation}$s before the last fixation, these results may verify **H2** *$D_{fixation}$ of the last fixation before a selection decreases along with increasing total $N_{fixation}$.*

5 MODEL SYSTEMATICALLY DERIVING TIME REQUIRED FOR COMPLETING HUMAN COGNITIVE PROCESS ON SELECTION TASK

Using the results of the experiment and the MHP, we developed our model. In contrast to a previous study [13], we used $N_{fixation}$, which is counted from the user's behavior.

5.1 Equations of Model

To evaluate our model, we first evaluated how $N_{fixation}$ linearly affects the duration as a common model of analysis with the following equation:

$$y = a + b \times (N_{fixation} - 1). \tag{2}$$

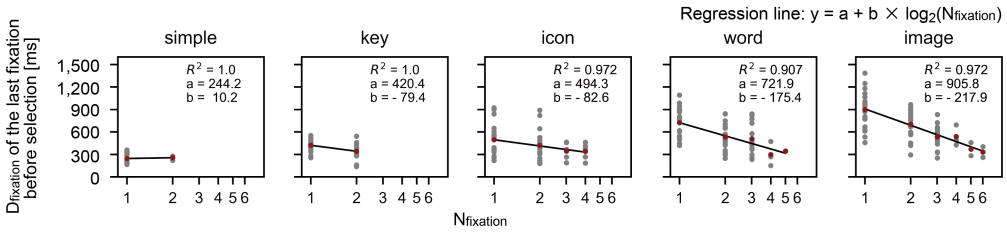

Fig. 5. Regression results with average $D_{fixation}$ and equation $\log_2(N_{fixation})$ for each task. Gray plots are the average $D_{fixation}$ for each participant. Red plots are the average for each $N_{fixation}$.

We then explored the following equation as a more efficient model in which $N_{fixation}$ logarithmically affects the duration:

$$y = a + b \times \log_2(N_{fixation}). \tag{3}$$

In these two models, y indicates the duration in a certain $N_{fixation}$, a indicates the duration when $N_{fixation}$ is one, and b indicates a change in the $D_{fixation}$ of the last fixation against increasing $N_{fixation}$. We show the regression results of both models in Table 3 and Fig. 5. The R^2 of Equation 3 was higher than that of Equation 2 for the icon, word, and image tasks. Because the maximum $N_{fixation}$ was two in the simple and key tasks, R^2 was 1.0. We assumed that the reason for achieving a higher R^2 in Equation 3 as the $D_{fixation}$ of the last fixation logarithmically decreases, not linearly, because a human can remember a stimulus (visual image in this paper) and proceed with the cognitive process by referring it.

To determine a better formulation in a statistical manner, we compared the *AIC* values [1] of the two models. As a brief guideline, a model with a lower *AIC* is better, and a model with $AIC \leq (AIC_{minimum} + 2)$ is probably comparable with better models [6]. Thus, we used $\log_2(N_{fixation})$ as an independent variable of the expression in our model.

5.2 Decrease in $D_{fixation}$ along with Increasing $N_{fixation}$s

To interpret a decrease in $D_{fixation}$ along with increasing $N_{fixation}$, we used the principle that a human completes one cognitive cycle, which consists of *recognition, classification*, and *matching* during one fixation [13]. Since the slopes of the equations (i.e., b in Equation 3) represent a downward trend from 10.2 (simple task) to −217.9 (image task) as the cognitive level increases, we clarify why these slopes are derived with the above principle by showing the relation between the slope and estimated slope by using the MHP (i.e., the estimated time required for one cognitive cycle) in Table 4. The differences between the slope with our model and the estimated slope are under 35.4 ms. Since the original τ_c also ranged from 25 ms to 170 ms, this difference could be considered covered by the range.

5.3 Minimum $D_{fixation}$ for each task

There is approximately a 90 ms difference in $D_{fixation}$ when $N_{fixation}$ is the largest between the simple task and other tasks. This difference is due to the simple task requiring only *requesting*, while the others require at least two cognitive processes. Moreover, 90 ms is within the range of τ_c (25–170 ms). Thus, we assumed that the difference in the $D_{fixation}$ of the last fixation between the simple task and other tasks could be interpreted due to the difference in the cognitive processes. Since users can generally select the target in a well-familiarized interface without careful fixation, even if the target is a key, icon, word, or image, they can select the target without a cognitive process, in other words, that they can select the target as equal to the simple task. For example, selecting a close button

Table 4. The relation between the slope and estimated slope by using the MHP. The units of all digits are in milliseconds. With respect to the principle of a previous study [13], the τ_c, *requesting* is not included in the duration required for one cognitive cycle. Thus, the estimated slope was calculated with the number of required cognitive processes except *requesting*. For example, in the image task, the cognitive processes are *recognizing*, *classifying*, *matching*, and *requesting*, which require $4\tau_c$. As the τ_c for *requesting* is not included for one cognitive cycle, the estimated slope with the MHP is similar to $3\tau_c$= 210 ms.

Task	Required cognitive process				Slope (ours)	Slope (MHP)	Diff.
Simple				*requesting*	10.2	0.0 ($0\tau_c$)	10.2
Key			*matching*	*requesting*	−79.4	−70 ($1\tau_c$)	9.4
Icon			*matching*	*requesting*	−82.6	−70 ($1\tau_c$)	12.6
Word	*recognizing*		*matching*	*requesting*	−175.4	−140 ($2\tau_c$)	35.4
Image	*recognizing*	*classifying*	*matching*	*requesting*	−217.9	−210 ($3\tau_c$)	7.9

Table 5. Summary of regression results for each task.

Task	Equation	Max $N_{fixation}$	Smallest dwell-time
Simple	174.2 + 10.2 × $\log_2(N_{fixation})$	2	174.2
Key	350.4 - 79.4 × $\log_2(N_{fixation})$	2	271.0
Icon	424.3 - 82.6 × $\log_2(N_{fixation})$	4	259.3
Word	651.9 - 175.4 × $\log_2(N_{fixation})$	5	244.6
Image	835.8 - 217.9 × $\log_2(N_{fixation})$	6	272.5

in a web browser that is often located at the top corner of the browser could be selected without careful fixation (most are designed to be done so). Thus, we concluded one minimum $D_{fixation}$ exists regardless of the task and $D_{fixation}$ converges to one in the simple task (i.e., 244 ms), which verifies **H3** *$D_{fixation}$ for large $N_{fixation}$ converges to one duration regardless of the task.*

5.4 Range of $D_{fixation}$

In addition to the above analysis focusing on average values, we analyzed how the $D_{fixation}$ in each $N_{fixation}$ varied among participants (Fig. 5). These ranges may be due to the same reason as in the MHP, that is, the Fastman can complete a task with minimum duration, and the Slowman requires maximum duration. Since we did not instruct participants on the selection strategy, $D_{fixation}$ also varied for each participant and each selection. Personality and background may have also affected the results. For example, a user carefully searching for a target requires a large τ_c, and a user familiar with a target (e.g., a user has used the menu item in the word task) requires a small τ_c. Thus, using average values is generally one simple solution to reflect the duration that a human requires to finish a cognitive process. However, using a calibrated τ_c for users is the better solution to estimate a more precise duration.

6 USE OF OUR MODEL FOR DWELL SELECTION

We describe how our model can be applied to dwell selection. Since no action of pushing a button is needed for dwell selection, we first subtract a duration of τ_m=70 ms from the model. Using Equation 3 and regression results, we then define the adapted model for each task. We summarize the equations and dwell-times derived with our model for each task in Table 5, which suggest that we can dynamically change dwell-times with our model. For example, in an image-selection task,

if a user fixates on a target three times beforehand, we can use 490.4 ms as the dwell-time; if six times, we can use 272.5 ms.

We consider a span that keeps counting $N_{fixation}$. For example, a system keeps counting $N_{fixation}$ and discards those over 609, 996, 2,455, 3,620, and 23,565 ms for the simple, key, icon, word, and image tasks, respectively. We determined these spans from average durations taken for the trial (i.e., from displaying a target to finishing a selection) in the experiment. We did not consider $N_{fixation}$ more than those observed in our experiment (more than max $N_{fixation}$ in Table 5) and determined the minimum $N_{fixation}$ for each task. However, as described in Section 5.3, the minimum $N_{fixation}$ may become one for the simple task (i.e., 174.2 ms). Of course, if users prefer a faster interaction, they can use under 174.2 ms at will. Such a smaller dwell-time can be considered when users are familiar with the situation.

In research on preventing the Midas-touch, a faster and more accurate dwell selection has been developed (i.e., the best solution was been regarded as 0 ms of dwell-time and zero Midas-touches); however, this seems to be difficult. A minimum dwell-time that does not decrease usability may be derived with our model, and the researcher can aim to prevent the Midas-touch with the minimum dwell-time. Thus, solving the Midas-touch will be more realistic than aiming for 0 ms as dwell-time. As a dynamic adjustment of dwell-time for improving usability (e.g., [19, 22]), dwell-time can be adjusted from the context of human cognitive processes by using our model.

Although we have described the use of our model in a real interaction, we can not strongly conclude that it is useful due to the limitations of our experimental conditions and results. Further investigation with an application adopting dwell selection with our model should be conducted.

7 LIMITATIONS AND FUTURE WORK

Our findings are limited by the experimental conditions. It is unclear whether our findings, i.e., the duration that a human requires to finish a cognitive process, would hold under other conditions. Regarding selection tasks, there are numerous situations of real interactions, for example, selecting a sentence and thumbnail within an interface consisting of various types of targets. Since τ_p, τ_c, and τ_m were derived from certain user attributes [8], our model using the MHP may be suitable for users whose attributes differ from those of the participants in this experiment (e.g., different ages, experience with computer interaction, and experience with gaze-based interaction). However, this is only a hypothesis, and we could not conclude anything from our current results, so further research should be conducted with a large number of participants and more diverse participants. Although we concluded that our model based on Equation 3 could effectively derive duration, it is necessary to evaluate the model under other experimental conditions.

We developed our model from the perspectives of linear- and logarithmic-based equations and the MHP [8]. Similar to Fitts' Law [11] and ACT-R [2, 3], which has numerous variations of a model regarding the context, we assumed that we could investigate a variation of our model for a specific context or user attributes. For example, the keystroke-level model [7] indicates that the time to complete a typing (i.e., key selection) task varies depending on the context and the user's typing skill. As with previous studies on adjusting dwell-time (e.g., [20]), our model can be improved using the keystroke-level model. Our model is not the only one model and further development combined with the above studies is also potentially helpful to understand dwell-time (or other "time" parameters for visual search) from human cognitive processes. Of course, our model should also be investigated in a real interaction situation, e.g., a dynamic adaption of dwell-time for dwell selection.

As with the other adaptions, our model has the possibility of improving the usability of implicit interaction, especially, an interaction driven by an intent prediction. Current interactions, especially using GUIs (e.g., for a mouse [4, 21, 23]) will enable a selection that is done just before a user

performs a specific action. Research has shown that an interaction system automatically corrects an error input through intent prediction using gaze data [24]. Unfortunately, the time when the error correction should be done has not been investigated in detail. In these scenarios, the same as with dwell-time, we can adapt the duration that a human requires to finish a cognitive process to improve usability. These are, of course, speculations; thus, it is necessary to verify them.

8 CONCLUSION

We developed a model that derives the dwell-time systematically, enabling dwell selection after a user completes the cognitive process on the basis of their behavior. We first conducted an experiment involving five tasks of different cognitive levels to measure the number of fixations and their duration through a user's selection behavior. We then validated our three hypotheses related to fixations and developed our model using the fixations and durations by referring to the MHP. We then showed how our model can be used for dwell selection and discussed the use of our findings for future implicit interactions.

ACKNOWLEDGMENTS

This work is supported by JSPS, KAKENHI Grant Number 21J10301, Tateisi Science and Technology Foundation.

REFERENCES

[1] Hirotugu Akaike. 1974. A new look at the statistical model identification. *IEEE Trans. Automat. Control* 19, 6 (1974), 716–723.

[2] John R. Anderson, Michael Matessa, and Scott A. Douglass. 1995. The ACT-R Theory and Visual Attention. In *Proceedings of the Seventeenth Annual Conference of the Cognitive Science Society*. Hillsdale, NJ: Lawrence Erlbaum Associates, 61–65.

[3] John R. Anderson, Michael Matessa, and Christian Lebiere. 1997. ACT-R: A Theory of Higher Level Cognition and its Relation to Visual Atention. *Human-Computer Interaction* 12, 4 (1997), 439–462. https://doi.org/10.1207/s15327051hci1204_5

[4] Takeshi Asano, Ehud Sharlin, Yoshifumi Kitamura, Kazuki Takashima, and Fumio Kishino. 2005. Predictive Interaction Using the Delphian Desktop. In *Proceedings of the 18th Annual ACM Symposium on User Interface Software and Technology* (Seattle, WA, USA) *(UIST '05)*. Association for Computing Machinery, New York, NY, USA, 133–141. https://doi.org/10.1145/1095034.1095058

[5] Gilles Bailly, Antti Oulasvirta, Duncan P. Brumby, and Andrew Howes. 2014. Model of Visual Search and Selection Time in Linear Menus. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Toronto, Ontario, Canada) *(CHI '14)*. Association for Computing Machinery, New York, NY, USA, 3865–3874. https://doi.org/10.1145/2556288.2557093

[6] Kenneth P Burnham and David R Anderson. 2003. *Model selection and multimodel inference: a practical information-theoretic approach*. Springer Science & Business Media, Heidelberg, Germany.

[7] Stuart K. Card, Thomas P. Moran, and Allen Newell. 1980. The Keystroke-Level Model for User Performance Time with Interactive Systems. *Communications of the ACM* 23, 7 (July 1980), 396–410. https://doi.org/10.1145/358886.358895

[8] Stuart K. Card, Allen Newell, and Thomas P. Moran. 1983. *The Psychology of Human-Computer Interaction*. L. Erlbaum Associates Inc., USA.

[9] Myungguen Choi, Daisuke Sakamoto, and Tetsuo Ono. 2022. Kuiper Belt: Utilizing the "Out-of-Natural Angle" Region in the Eye-Gaze Interaction for Virtual Reality. In *CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 357, 17 pages. https://doi.org/10.1145/3491102.3517725

[10] Andy Cockburn, Carl Gutwin, and Saul Greenberg. 2007. A Predictive Model of Menu Performance. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '07)*. Association for Computing Machinery, New York, NY, USA, 627–636. https://doi.org/10.1145/1240624.1240723

[11] P. M. Fitts. 1954. The Information Capacity of the Human Motor System in Controlling the Amplitude of Movement. *Journal of Experimental Psychology* 74 (1954), 381–391.

[12] Toshiya Isomoto, Toshiyuki Ando, Buntarou Shizuki, and Shin Takahashi. 2018. Dwell Time Reduction Technique Using Fitts' Law for Gaze-Based Target Acquisition. In *Proceedings of the 2018 ACM Symposium on Eye Tracking*

Research & Applications (Warsaw, Poland) *(ETRA '18)*. Association for Computing Machinery, New York, NY, USA, 26:1–26:7. https://doi.org/10.1145/3204493.3204532

[13] Toshiya Isomoto, Shota Yamanaka, and Buntarou Shizuki. 2021. Relationship between Dwell-Time and Model Human Processor for Dwell-based Image Selection. In *Proceedings of the 2021 ACM Symposium on Applied Perception* (virtual) *(SAP '21)*. Association for Computing Machinery, Article 6, 5 pages. https://doi.org/10.1145/3474451.3476240

[14] Toshiya Isomoto, Shota Yamanaka, and Buntarou Shizuki. 2022. Dwell Selection with ML-Based Intent Prediction Using Only Gaze Data. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, 3, Article 120 (Sep. 2022), 21 pages. https://doi.org/10.1145/3550301

[15] Robert J. K. Jacob. 1990. What You Look at is What You Get: Eye Movement-based Interaction Techniques. In *Proceedings of the 1990 CHI Conference on Human Factors in Computing Systems* (Seattle, Washington, USA) *(CHI '90)*. Association for Computing Machinery, New York, NY, USA, 11–18. https://doi.org/10.1145/97243.97246

[16] Robert J. K. Jacob. 1991. The Use of Eye Movements in Human-computer Interaction Techniques: What You Look at is What You Get. *ACM Transaction on Information Systems* 9, 2 (1991), 152–169.

[17] Robert. J. K. Jacob. 1993. Eye Movement-Based Human-Computer Interaction Techniques: Toward Non-Command Interfaces. *Advances in Human-Computer Interaction* 4 (1993), 151–190.

[18] Ian Scott Mackenzie. 1991. *Fitts' Law As a Performance Model in Human-computer Interaction.* Ph. D. Dissertation. Toronto, Ont., Canada, Canada.

[19] Päivi Majaranta, Ulla-Kaija Ahola, and Oleg Špakov. 2009. Fast Gaze Typing with an Adjustable Dwell Time. In *Proceedings of the 2009 CHI Conference on Human Factors in Computing Systems* (Boston, MA, USA) *(CHI '09)*. Association for Computing Machinery, New York, NY, USA, 357–360. https://doi.org/10.1145/1518701.1518758

[20] Martez E. Mott, Shane Williams, Jacob O. Wobbrock, and Meredith Ringel Morris. 2017. Improving Dwell-Based Gaze Typing with Dynamic, Cascading Dwell Times. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI '17)*. Association for Computing Machinery, New York, NY, USA, 2558–2570. https://doi.org/10.1145/3025453.3025517

[21] Martez E. Mott and Jacob O. Wobbrock. 2014. Beating the Bubble: Using Kinematic Triggering in the Bubble Lens for Acquiring Small, Dense Targets. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Toronto, Ontario, Canada) *(CHI '14)*. Association for Computing Machinery, New York, NY, USA, 733–742. https://doi.org/10.1145/2556288.2557410

[22] Aanand Nayyar, Utkarsh Dwivedi, Karan Ahuja, Nitendra Rajput, Seema Nagar, and Kuntal Dey. 2017. OptiDwell: Intelligent Adjustment of Dwell Click Time. In *Proceedings of the 22nd International Conference on Intelligent User Interfaces* (Limassol, Cyprus) *(IUI '17)*. Association for Computing Machinery, New York, NY, USA, 193–204. https://doi.org/10.1145/3025171.3025202

[23] Phillip T. Pasqual and Jacob O. Wobbrock. 2014. Mouse Pointing Endpoint Prediction Using Kinematic Template Matching. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Toronto, Ontario, Canada) *(CHI '14)*. Association for Computing Machinery, New York, NY, USA, 743–752. https://doi.org/10.1145/2556288.2557406

[24] Candace E. Peacock, Ben Lafreniere, Ting Zhang, Stephanie Santosa, Hrvoje Benko, and Tanya R. Jonker. 2022. Gaze as an Indicator of Input Recognition Errors. *Proceedings of ACM Human-Computer Interaction* 6, ETRA, Article 142 (may 2022), 18 pages. https://doi.org/10.1145/3530883

[25] Abdul Moiz Penkar, Christof Lutteroth, and Gerald Weber. 2012. Designing for the Eye: Design Parameters for Dwell in Gaze Interaction. In *Proceedings of the 24th Australian Computer-Human Interaction Conference* (Melbourne, Australia) *(OzCHI '12)*. Association for Computing Machinery, New York, NY, USA, 479–488. https://doi.org/10.1145/2414536.2414609

[26] Ken Pfeuffer, Yasmeen Abdrabou, Augusto Esteves, Radiah Rivu, Yomna Abdelrahman, Stefanie Meitner, Amr Saadi, and Florian Alt. 2021. ARtention: A Design Space for Gaze-adaptive User Interfaces in Augmented Reality. *Computers & Graphics* 95 (2021), 1–12. https://doi.org/10.1016/j.cag.2021.01.001

[27] Ken Pfeuffer and Yang Li. 2018. Analysis and Modeling of Grid Performance on Touchscreen Mobile Devices. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3173574.3173862

[28] Jimin. Pi and Bertram. E. Shi. 2017. Probabilistic Adjustment of Dwell Time for Eye Typing. In *10th International Conference on Human System Interactions (HSI)*. IEEE, 251–257.

[29] Kari-Jouko Räihä and Saila Ovaska. 2012. An Exploratory Study of Eye Typing Fundamentals: Dwell Time, Text Entry Rate, Errors, and Workload. In *Proceedings of the 2012 CHI Conference on Human Factors in Computing Systems* (Austin, Texas, USA) *(CHI '12)*. Association for Computing Machinery, New York, NY, USA, 3001–3010. https://doi.org/10.1145/2207676.2208711

[30] Dario D. Salvucci and Joseph H. Goldberg. 2000. Identifying Fixations and Saccades in Eye-Tracking Protocols. In *Proceedings of the 2000 Symposium on Eye Tracking Research & Applications* (Palm Beach Gardens, Florida, USA) *(ETRA '00)*. Association for Computing Machinery, New York, NY, USA, 71–78. https://doi.org/10.1145/355017.355028

[31] Andrew Sears, Julie A. Jacko, Josey Chu, and Francisco Moro. 2001. The Role of Visual Search in the Design of Effective Soft Keyboards. *Behaviour & Information Technology* 20, 3 (2001), 159–166. https://doi.org/10.1080/01449290110049790 arXiv:https://doi.org/10.1080/01449290110049790

[32] Oleg Špakov and Darius Miniotas. 2004. On-Line Adjustment of Dwell Time for Target Selection by Gaze. In *Proceedings of the 2004 Nordic Conference on Human-Computer Interaction (NordiCHI '04)*. Association for Computing Machinery, New York, NY, USA, 203–206. https://doi.org/10.1145/1028014.1028045

[33] Colin Ware and Harutune H. Mikaelian. 1987. An Evaluation of an Eye Tracker As a Device for Computer Input. In *Proceedings of the 1987 CHI/GI Conference on Human Factors in Computing Systems and Graphics Interface* (Toronto, Ontario, Canada) *(CHI '87)*. Association for Computing Machinery, New York, NY, USA, 183–188. https://doi.org/10.1145/29933.275627

[34] Xinyong Zhang, Pianpian Xu, Qing Zhang, and Hongbin Zha. 2011. Speed-Accuracy Trade-off in Dwell-Based Eye Pointing Tasks at Different Cognitive Levels. In *Proceedings of the 1st International Workshop on Pervasive Eye Tracking & Mobile Eye-Based Interaction* (Beijing, China) *(PETMEI '11)*. Association for Computing Machinery, New York, NY, USA, 37–42. https://doi.org/10.1145/2029956.2029967

Received November 2022; revised February 2023; accepted March 2023