When Choice Happens

A Systematic Examination of Mouse Movement Length for Decision Making in Web Search

Lukas Brückner
lukas@lukas-brueckner.de
Independent researcher
Germany

Ioannis Arapakis
ioannis.arapakis@telefonica.com
Telefonica Research
Spain

Luis A. Leiva
name.surname@uni.lu
University of Luxembourg
Luxembourg

ABSTRACT

Searchers often make a choice in a matter of seconds on SERPs. As a result of a dynamic cognitive process, choice is ultimately reflected in motor movement and thus can be modeled by tracking the computer mouse. However, because not all movements have equal value, it is important to understand how do they and, critically, their sequence length impact model performance. We study three different SERP scenarios where searchers (1) noticed an advertisement, (2) abandoned the page, and (3) became frustrated. We model these scenarios with recurrent neural nets and study the effect of mouse sequence padding and truncating to different lengths. We find that it is possible to predict the aforementioned tasks sometimes using just 2 seconds of movement. Ultimately, by efficiently recording the right amount of data, we can save valuable bandwidth and storage, respect the users' privacy, and increase the speed at which machine learning models can be trained and deployed. Considering the web scale, doing so will have a net benefit on our environment.

CCS CONCEPTS

• **Information systems** → *Search interfaces*; • **Computing methodologies** → **Modeling and simulation**.

KEYWORDS

Mouse Cursor Tracking; Decision Making; Deep Learning

1 INTRODUCTION

Searchers are faced with choices and decisions all the time, from the very basic (e.g., whether to click on a link or bookmark the page) to the more informed (e.g., devising a search strategy or judging the relevance of a snippet). Every decision is a commitment to a course of action [42] and it may involve a complex process of considering

and evaluating different pieces of information. Moreover, a decision making process is inherently linked to our sensorimotor system [11] and can be conceptualized as involving two distinct phases [16, 29, 30]: a (pre-movement) *planning* phase, where key parameters of the upcoming action are specified and ready for implementation, and a *control* phase, in which corrective processes fine-tune the movement to ensure successful completion.

In information retrieval (IR) tasks, most of the user interactions happen on the client side only, so it is not possible to infer them on the server side. To find out what content is consumed and how it was consumed, the website can use client-side tracking to record richer interactions [19]. Historically, eye tracking has been used to study user behavior in IR [1, 7, 12, 21]. However, eye tracking requires specialized equipment, ranging from expensive stationary eye trackers to more affordable but noisy webcams [26]. In the recent years, a large body of work has examined the relationship between eye gaze and mouse cursor movements during search [13–15], providing ample evidence on the connection between the two signals and highlighting mouse cursor tracking as a low-cost, scalable alternative to eye tracking. Mouse tracking is easily accessible to researchers and practitioners alike, requiring no dedicated equipment or extensive training to use [20]. It also provides a face-valid, readily interpretable dynamic assessment of choice [34].
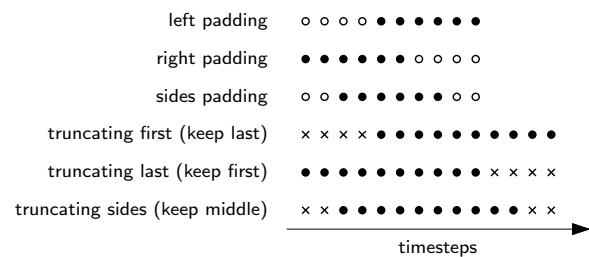
Figure 1: We study the effect of padding and truncating mouse movement sequences using three modes that operate at the beginning of the sequence, end of the sequence, or both parts. Here, all sequences are set to a fixed length of 10 timesteps, so shorter sequences are padded up to 10 timesteps and longer sequences are truncated to that length.

An effective approach to modeling mouse cursor movements are Recurrent Neural Networks (RNNs), which remove the need to manually engineer features [3]. Instead, RNNs will pick up on features that may not be evident yet important for prediction, using raw, unprocessed movement sequences as input. RNNs can handle variable-length sequences, however all sequences in the same batch must have the same length. The usual way to address this is by

truncating longer sequences and padding shorter sequences with dummy values up to a predefined length. Figure 1 illustrates three different ways of implementing both kinds of operations. In this paper we explore all of them, as well as their combinations (e.g. left padding with right truncating). In total, we tested 9 mode combinations (3 padding and 3 truncating modes) for a variable number of sequence lengths; see Figure 2. Our fundamental research question is: How does sequence length impact model performance?

To the best of our knowledge, we are the first to investigate the impact of sequence length on RNN performance in IR tasks. We argue that it is critical for many web actors (e.g. webmasters, developers, researchers, search engines) because not all movements have equal value for the task at hand. By efficiently recording the right amount of mouse movements, it is possible to save bandwidth and storage, which, considering the web scale, eventually it will have a net benefit on our environment. In addition, with shorter mouse sequences it is possible to speed up RNN model training. Finally, because shorter sequence use less information about user interactions, online profiling becomes a more challenging task, hence contributing to preserving the users' privacy.

2 RELATED WORK

Formal approaches to assessing decision making processes are based on the premise that people make decisions and then implement them. Zsambok [44] provides a succinct definition of naturalistic decision making that emphasizes both the context in which decisions are made (e.g. dynamic environments, competing goals, time constraints) [25], as well as the role of prior experience [28]. Tversky and Kahneman [39] describe three heuristics: *availability*, *representativeness*, and *anchoring*, and argue that these explain the patterns of human judgement observed across a wide variety of contexts. Another class of heuristics – fast and frugal [37] – emphasises the need to consider the bounded rational nature of human cognition and the importance of the interaction between the mind and the environment. This crucial mind-environment interaction highlights the role of learning from feedback in well-structured environments, such as Search Engine Results Page (SERPs), as well as the differences that are observed when people make decisions on the basis of described or experienced information.

In the context of IR, much of human decision making involves a complex and lengthy process of considering and evaluating different kinds of relevant information before choosing a course of action. Hence, the user is subjected to various conscious and unconscious processes and stages [41] (e.g., identifying attributes relevant to the decision, deciding how to weight those attributes, obtaining a total utility, or selecting the option with the highest weighted total), that are oftentimes misaligned [24].

One important limitation faced nowadays is the lack of accessible and scalable tools outside research labs. Psychological experiments too often test micro-hypothesis about concrete processing phenomena in tightly controlled laboratory conditions. This limits significantly the amount of actual knowledge on the dynamics of human information processing that is brought into a real-world testing ground [27]. For example, technologies such as electroencephalography or eye tracking require high resource investments and, quite often, sophisticated experimental setups. On the other hand, there are online methods like mouse cursor tracking, which has been shown to be a cost-effective and scalable technique for capturing various user engagement measures, such as search success [13, 14] and satisfaction [22], interest [2, 4], attention [3, 5], or abandonment [8, 40]. Moreover, mouse tracking has been shown to be a sensitive metric to capture the onset, magnitude, and evolution of conflict within decision-making [35, 36].

In this work, we build upon the temporal construal theory [38] and frame our analysis in terms of the effect of mouse movement length as a proxy of how searchers make decisions on SERPs. The temporal construal theory describes the effects of 'psychological distance' on thinking, decision making, and behavior. More specifically, temporal distance has been shown to have important effects on a wide range of decisions, including e.g. consumer's choice [31, 43]. In that respect, it provides a suitable basis for establishing experimentally how much of the decision making process is expressed in the early, middle, or late stages of user interactions on a SERP, as reflected by their mouse cursor movements (represented as multivariate sequential data of x,y coordinates and timestamps). Our work addresses the current knowledge gap in the research literature about the utility of this important signal with respect to its sequence length.

3 EXPERIMENTS

We investigate the effect of mouse movement length in three paradigmatic IR tasks: attention (ads noticeability), page abandonment, and user frustration. These tasks can be tackled with existing public mouse tracking datasets. As can be observed in Table 1, mouse sequences are rather variable in length, which calls for research in this topic.

Table 1: Summary of mouse sequence lengths (number of timesteps) in our analyzed datasets. 'Rate' denotes the average sampling rate (in ms) of the captured mouse movements.

Dataset/task	Min	Max	Mean	Median	SD	Rate
Attention	6	222	20.4	15	18.4	150
Abandonment	2	123	25.0	19	21.6	150
Frustration	6	595	92.6	64	87.7	60

3.1 Predicting Ads Noticeability

For this task we use the attentive cursor dataset [17], which provides 716 mouse sequences on Google SERPs with associated ground-truth labels regarding user's attention to native advertisements (either the user noticed the ad or not) while browsing transactional queries (e.g. "buy rolex watch"). The are 476 cases were users noticed the ad (majority class), according to a binary self-reported rating collected at post-task.

We use the same RNN model and settings provided by Arapakis et al. [5], which outperformed other approaches in this task. The model is a bidirectional long short-term memory (BiLSTM) layer with 0.5 dropout and embedding size of 50 units, followed by a fully-connected layer of 1 neuron with sigmoid activation as output. The BiLSTM layer uses hyperbolic tangent as activation function and sigmoid activation in the recurrent step.
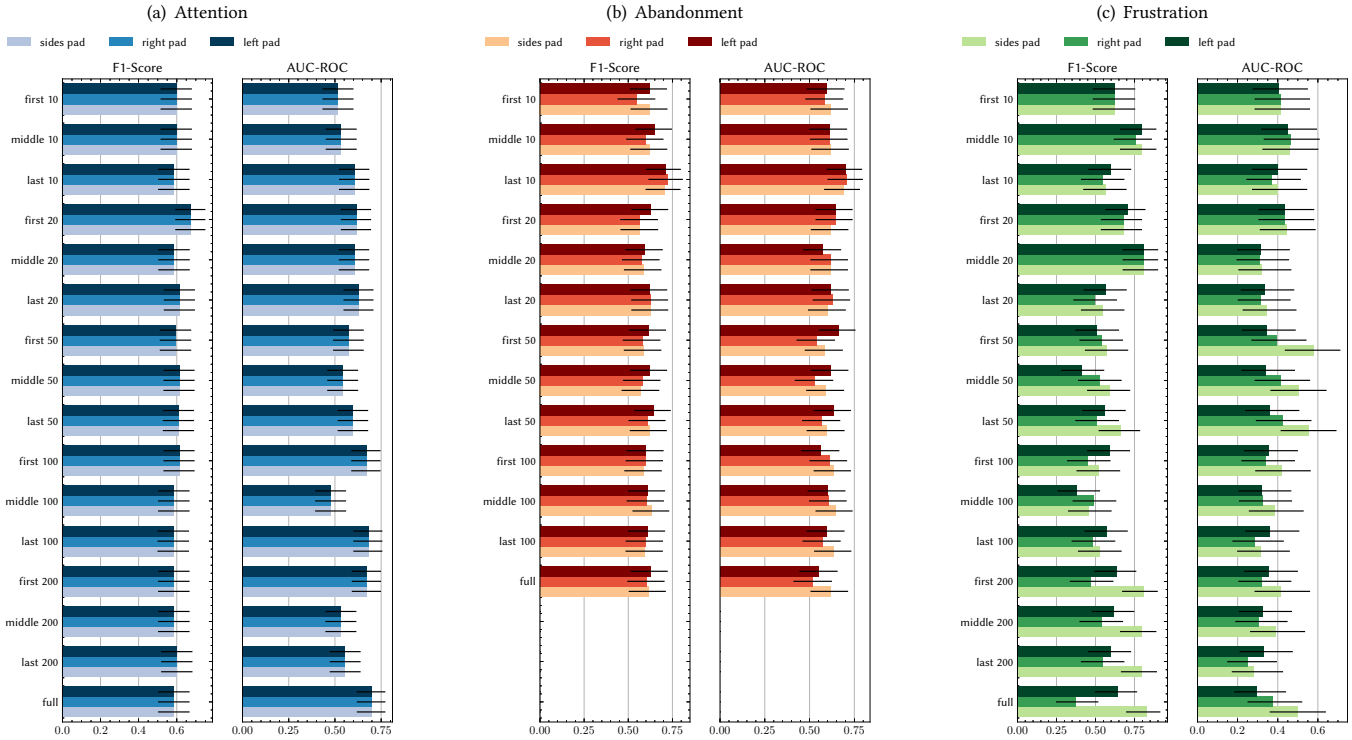
Figure 2: Experimental results. Rows denote a truncation mode (e.g. "first 10" means keeping the first 10 timesteps) and bars denote the padding mode. Error bars denote 95% confidence intervals using the Wilson method for binomial distributions.

3.2 Predicting Page Abandonment

For this task we use the dataset provided by Brückner et al. [8], which comprises 133 mouse sequences on Yahoo SERPs with associated ground-truth labels regarding page abandonment (either good or bad abandonment) while browsing informational queries (e.g. "Brad Pitt's age"). The are 77 good abandonment cases (majority class), where users left the SERP without clicking anywhere.

We use the same RNN model and settings provided by Brückner et al. [8], which outperformed other approaches in this task. The model is a stack of 2 BiLSTM layers with 0.3 dropout, followed by a fully-connected layer of 1 neuron with sigmoid activation as output. The BiLSTM layers use hyperbolic tangent as activation function and sigmoid activation in the recurrent step.

3.3 Predicting Search Frustration

For this task we use the dataset provided by Feild et al. [10], which comprises 259 mouse sequences on several SERPs (Google, Yahoo, Bing, Ask.com) with associated ground-truth labels regarding search frustration on SERPs (either the user was frustrated or not) while browsing both informational and navigational queries. There are 30 cases where users felt frustrated (minority class), according to a self-reported rating ≥ 4 in a scale of 1–5 (1 means no frustration at all). We deliberately ignore ratings of 2 and 3 because the goal of this task is to detected truly frustrated users. Since this task has not been tackled with RNN models before, we use the same

architecture of the abandonment model, because of the similarities between both tasks.

3.4 Procedure

In each scenario, we split the data using stratified sampling, to produce balanced splits that preserve the original class distribution. We use 60% of the data for training, 10% for validation, and 30% for testing. We train the models in batches of 16 mouse sequences. We use the Adam optimizer with learning rate $\eta = 0.0005$ and decay rates $\beta_1 = 0.9$ and $\beta_2 = 0.999$. We set a maximum number of 500 training epochs but use early stopping of 10 epochs by monitoring the validation loss (binary cross-entropy); i.e., if the loss on the validation set does not improve in 10 consecutive epochs, training finishes and the best learned weights are kept. The test set is a held-out partition that simulates unseen data. We run the experiments using 5 different seeds, to avoid potential biases from evaluating on a single run.

Our RNN models are fed with raw mouse sequences as input, i.e. sequences of x, y, t tuples. We use 3 padding modes and 3 truncating modes (Figure 1) for a variable number of sequence timesteps: 10, 20, 50, 100, 200, and full length. In the "full length" case, sequences are padded to the maximum length observed in each dataset (Table 1). In total, we tested 54 length variations for the attention and frustration scenarios, and 48 variations for the abandonment scenario (in this case, since the maximum length is 123 timesteps, we tested up to 5 length cases, as opposed to the 6 cases for the other scenarios).

3.5 Results

The results of our experiments are reported in Figure 2. We report model performance using weighted average F-measure (F1 score) and area under the ROC (AUC). We can see that in some cases using full-length sequences (i.e., forcing all sequences to have the same length of the maximum sequence observed in each dataset) is detrimental for performance. For example, in the ads attention task it is better to use sequences comprising the *first* 100 timesteps instead of padding all sequences to the maximum sequence length. Furthermore, using just the first 20 timesteps is as beneficial as using the first 100 timesteps, as indicated by the higher values of F1 and AUC. Interestingly, for the abandonment scenario it is the *last* 10 timesteps the cases that achieve better performance. For the frustration scenario, however, we observed that in most cases the winning combinations comprise the middle points of the mouse sequences. For example, "middle 10" and "middle 20" with any padding mode are as good as using full sequences with sides padding. In sum, in all scenarios, using shorter mouse sequences may lead to the same or even better performance than using full-length sequences as input to the RNN models.
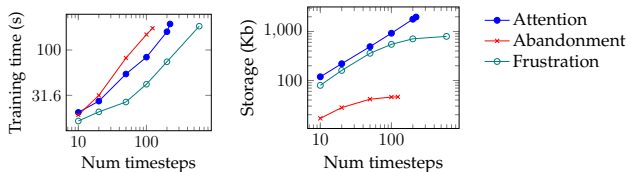

Figure 3: Model training time and dataset storage size as a function of the number of mouse sequence timesteps.

We provide additional results about the savings incurred in terms of training time and storage size in Figure 3, as a function of the number of sequence timesteps. In both plots, the X and Y axes are in logarithmic scale. Training time is measured on a commodity Intel i5 CPU @ 2.4 Hz, whereas storage size is measured by serializing the mouse sequences (multivariate time series) and ground-truth labels as JSON. We can see again that, by using shorter sequences it is possible to speed up training considerably, and this was so without a significant drop in model performance, as noted by our previous experiments. Similarly, the storage savings are clear, sometimes by more than one order of magnitude.

4 DISCUSSION

Built off cognitive models that propose a dynamic interplay between motor movements and underlying cognition [32, 33], mouse tracking provides a dynamic window into how a decision unfolds [23]. Our work is the first to provide empirical evidence for the three paradigmatic SERP scenarios considered: predicting ads noticeability, page abandonment, and search frustration.

4.1 Takeaway Messages

For IR tasks that involve investigating how users pay attention to different parts of the SERP (such as advertisements), it is beneficial to focus on the initial mouse movements. This is justifiable because our visual system is extremely fast and efficient. Hence, upon loading the page, users quickly scan it and locate the different SERP elements (e.g. logo, organic links, cards, etc.) even before moving their mouse. This minimizes the amount of thinking or "working out" that goes into reading and interpreting the page, and simply lets the eyes do their efficient job, then the mouse follows. On the contrary, for predicting SERP abandonment it is beneficial to focus on the final mouse movements. This is also understandable, as by definition the act of abandoning means to give up, which happens toward the end of the search session. Finally, we have observed that users can become frustrated at anytime, even when they have satisfied their search needs on the SERP [10], so it is difficult to predict this outcome by tracking mouse movements alone.

In sum, attention prediction tasks should focus on the initial timesteps of the mouse movement sequence, and abandonment prediction tasks should focus on the final sequence timesteps. It only takes about 20 timesteps (roughly 3 seconds of interaction data) to detect the aforementioned decision making processes with reasonable performance.

4.2 Implications and Future Work

Despite dramatically lower costs for bandwidth and storage, the demand for these resources has continued to exceed the ability of organizations to satisfy it within constrained budgets [9]. It is argued that every gigabyte requires between 3 and 7 kw/h for storage [6]. Therefore, considering the web scale, it is important to record efficiently the right amount of mouse movements. Eventually, doing so will have a net benefit on our environment.

Our work has also implications for user's privacy. It has been shown that mouse movements can disclose sensible demographics information such as gender and age, which could be exploited without the user even noticing it [18]. Therefore, by recording a small subset of mouse movements we argue it would be very difficult to profile the user. This, however, remains to be validated in future work. Further avenues for future work include comparing other model architectures and other mouse tracking datasets, which unfortunately are scarce in the IR literature.

5 CONCLUSION

The psychological study of decision making is replete with important applications to real-world problems and contexts. Our work, which investigated three representative SERP scenarios where searchers made a choice – either consciously (page abandonment), unconsciously (ads noticeability), or both (frustration) – further elucidates how much of the decision making process is expressed in the early, middle, or late stages of user interactions on a SERP, as reflected by their mouse cursor movements. By analyzing the effect of mouse sequence length, we have found that: (1) attentional events happen at the beginning of the search session; (2) SERP abandonment decision happens towards the end of the search session; (3) searcher's frustration can happen at anytime in the search session. We also have found that these choices happen rather quickly, so, taken together, it is not really necessary to record *all* mouse movements on the SERP, thereby saving valuable bandwidth and storage, respecting the users' privacy, and increasing the speed at which machine learning models can be trained and deployed.

REFERENCES

[1] I. Arapakis, M. Lalmas, B. B. Cambazoglu, M.-C. Marcos, and J. M. Jose. 2014. User engagement in online News: Under the scope of sentiment, interest, affect, and gaze. *J. Assoc. Inf. Sci. Technol.* 65, 10 (2014), 1988–2005.

[2] I. Arapakis and L. A. Leiva. 2016. Predicting User Engagement with Direct Displays Using Mouse Cursor Information. In *Proc. SIGIR*. 599–608.

[3] I. Arapakis and L. A. Leiva. 2020. Learning Efficient Representations of Mouse Movements to Predict User Attention. In *Proc. SIGIR*. 1309–1318.

[4] I. Arapakis, L. A. Leiva, and B. B. Cambazoglu. 2015. Know Your Onions: Understanding the User Experience with the Knowledge Module in Web Search. In *Proc. CIKM*. 1695–1698.

[5] I. Arapakis, A. Penta, H. Joho, and L. A. Leiva. 2020. A Price-Per-Attention Auction Scheme Using Mouse Cursor Information. *ACM Trans. Inf. Syst.* 38, 2.

[6] R. Arnold and W. Langelaar. 2019. Getting our heads out of the cloud: Data deletion 101. Victoria University of Wellington, Sustainability Week.

[7] Brightfish, Lumen Research, and Profacts. 2018. From viewable to viewed: using eye tracking to understand the reality of attention to advertising acros smedia. https://effectiveviews.be/files/White_Paper_From_Viewable_to_viewed.pdf

[8] L. Brückner, I. Arapakis, and L. A. Leiva. 2020. Query Abandonment Prediction with Recurrent Neural Models of Mouse Cursor Movements. In *Proc. CIKM*. 1969–1972.

[9] G. Dorworth. 2003. The paradox of cheap storage and bandwidth. Computerworld.

[10] H. A. Feild, J. Allan, and R. Jones. 2010. Predicting Searcher Frustration. In *Proc. SIGIR*. 34–41.

[11] J. P. Gallivan, C. S. Chapman, D. M. Wolpert, and J. R. Flanagan. 2018. Decision-making in sensorimotor control. *Nat. Rev. Neurosci.* 19, 9 (2018), 519–534.

[12] J. H. Goldberg, M. J. Stimson, M. Lewenstein, N. Scott, and A. M. Wichansky. 2002. Eye Tracking in Web Search Tasks: Design Implications. In *Proc. ETRA*. 51–58.

[13] Q. Guo, S. Yuan, and E. Agichtein. 2011. Detecting success in mobile search from interaction. In *Proc. SIGIR*. 1229–1230.

[14] A. Hassan, R. Jones, and K. Klinkner. 2010. Beyond DCG: user behavior as a predictor of a successful search. In *Proc. WSDM*. 221–230.

[15] J. Huang, R. W. White, and S. Dumais. 2011. No Clicks, No Problem: Using Cursor Movements to Understand and Improve Search. In *Proc. CHI*. 1225–1234.

[16] M. Jeannerod. 1988. *Oxford psychology series, No. 15. The neural and behavioural organization of goal-directed movements*. Clarendon Press.

[17] L. A. Leiva and I. Arapakis. 2021. The Attentive Cursor Dataset. *Front. Hum. Neurosci.* 14 (2021).

[18] L. A. Leiva, I. Arapakis, and C. Iordanou. 2021. My Mouse, My Rules: Privacy Issues of Behavioral User Profiling via Mouse Tracking. In *Proc. CHIIR*. 51–61.

[19] L. A. Leiva and J. Huang. 2015. Building a Better Mousetrap: Compressing Mouse Cursor Activity for Web Analytics. *Inf. Process. Manag.* 51, 2 (2015), 114–129.

[20] L. A. Leiva and R. Vivó. 2013. Web Browsing Behavior Analysis and Interactive Hypervideo. *ACM Trans. Web* 7, 4 (2013), 20:1–20:28.

[21] Y. Li, P. Xu, D. Lagun, and V. Navalpakkam. 2017. Towards Measuring and Inferring User Interest from Gaze. In *Proc. WWW '17 Companion*. 525–533.

[22] Y. Liu, Y. Chen, J. Tang, J. Sun, M. Zhang, S. Ma, and X. Zhu. 2015. Different Users, Different Opinions: Predicting Search Satisfaction with Mouse Movement Information. In *Proc. SIGIR*. 493–502.

[23] M. Maldonado, E. Dunbar, and E. Chemla. 2019. Mouse tracking as a window into decision making. *Behav. Res.* 51 (2019), 1085–1101.

[24] S. A. Malkoc, G. Zauberman, and C. Ulu. 2005. Consuming now or later? The interactive effect of timing and attribute alignability. *Psychol. Sci.* 16 (2005), 411–417.

[25] J. Orasanu and T. Connolly. 1993. The reinvention of decision making. In *Decision making in action: Models and methods*, R. Calderwood G.A. Klein, J. Orasanu and C.E. Zsambok (Eds.). 3–20.

[26] A. Papoutsaki, P. Sangkloy, J. Laskey, N. Daskalova, J. Huang, and J. Hays. 2016. WebGazer: Scalable Webcam Eye Tracking Using User Interactions. In *Proc. IJCAI*. 3839–3845.

[27] A. Pereda-Baños and I. Arapakis. 2015. On Human Information Processing in Information Retrieval (Position Paper). In *NeuroIR 2015: SIGIR 2015 Workshop on Neuro-Physiological Methods in IR Research*. 83–88.

[28] J. Pruitt, J. Cannon-Bowers, and E. Salas. 1997. In search of naturalistic decisions. In *Decision making under stress: Emerging themes and application*, M. Strub R. Flin, E. Salas and L. Martin (Eds.). 29–42.

[29] D. A. Rosenbaum. 1991. *Human Motor Control*. Academic Press.

[30] G. S. 2004. Separate visual representations in the planning and control of action. *Behav. Brain Sci.* 27, 1 (2004), 3–24.

[31] D. Soman. 1998. The illusion of delayed incentives: Evaluating future effort-money transactions. *J. Mark. Res.* 35 (1998), 427–438.

[32] J.-H. Song and K. Nakayama. 2006. Role of focal attention on latencies and trajectories of visually guided manual pointing. *J. Vis.* 6 (2006), 982–995.

[33] M. J. Spivey. 2008. *The Continuity of Mind*. Oxford University Press.

[34] P. E. Stillman, I. Krajbich, and M. J. Ferguson. 2020. Using dynamic monitoring of choices to predict and understand risk preferences. *PNAS* 117, 50 (2020).

[35] P. E. Stillman, X. Shen, and M. J. Ferguson. 2018. How mouse-tracking can advance social cognitive theory. *Trends Cogn. Sci.* 22 (2018), 531–543.

[36] N. Sullivan, C. Hutcherson, A. Harris, and A. Rangel. 2015. Dietary self-control is related to the speed with which attributes of healthfulness and tastiness are processed. *Psychol. Sci.* 26 (2015), 122–13.

[37] P. Todd and G. Gigerenzer. 2007. Environments that make us smart. *Curr. Dir. Psychol. Sci.* 16 (2007), 167–171.

[38] Y. Trope and N. Liberman. 2003. Temporal construal. *Psychol. Rev.* 110 (2003), 403–421.

[39] A. Tversky and D. Kahneman. 1974. Judgment under uncertainty: heuristics and biases. *Science* 185 (1974), 1124–1131.

[40] K. Williams and I. Zitouni. 2017. Does That Mean You're Happy? RNN-Based Modeling of User Interaction Sequences to Detect Good Abandonment. In *Proc. CIKM*. 727–736.

[41] G. Wright. 1984. *Behavioral decision theory*. Harmondsworth: Penguin.

[42] J. F. Yates, E. S. Veinott, and A. L. Patalano. 2003. Hard decisions, bad decisions: On decision quality and decision aiding. In *Emerging perspectives on judg- ment and decision research*, S.L. Schneider and J. Shanteau (Eds.). Cambridge University Press, 1–63.

[43] G. Zauberman. 2003. The intertemporal dynamics of consumer lock-in. *J. Consum. Res.* 30 (2003), 405–419.

[44] C. Zsambok. 1997. Naturalistic decision making: where are we now? In *Naturalistic decision making*, C.E. Zsambok and G.A. Klein (Eds.). 3–16.