A Mouse (H)Over a Hotspot Survey: An Exploration of Patterns of Hesitation through Cursor Movement Metrics

Lucas Pereira
ITI/LARSyS, M-ITI, and prsma.com
Funchal, Portugal
lucas.pereira@m-iti.org

ABSTRACT

This paper presents the results of an empirical exploration of 10 cursor movement metrics designed to measure respondent hesitation in online surveys. As a use case, this work considers an online survey aimed at exploring how people gauge the electricity consumption of domestic appliances. The cursor metrics were derived computationally from the mouse trajectories when rating the consumption of each appliance and analyzed using Multidimensional Scaling, Jenks Natural Breaks, and the Jaccard Similarity Index techniques. The results show that despite the fact that the metrics measure different aspects of the mouse trajectories, there is an agreement with respect to the appliances that generated higher levels of hesitation. The paper concludes with an outline of future work that should be carried out in order to further understand how cursor trajectories can be used to measure respondent hesitation.

KEYWORDS

mouse tracking; single- and multi-target metrics; hesitation patterns; online survey; user experience

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

CHI'19 Extended Abstracts, May 4–9, 2019, Glasgow, Scotland UK

© 2019 Copyright held by the owner/author(s).

ACM ISBN 978-1-4503-5971-9/19/05.
https://doi.org/10.1145/3290607.3312956

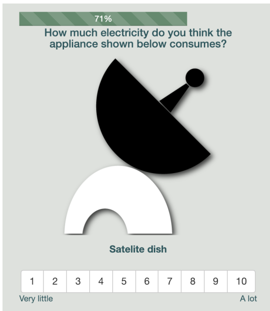

Figure 1: An example of a rating screen.

Table 1: Full list of cursor transitions and hovers from a random respondent when rating a satellite dish.

Rate	Time	Event	Elapsed Time
8	14:48:09.292	leave	
7	14:48:09.290	enter	250 ms
7	14:48:09.542	leave	
7	14:48:09.859	enter	34 ms
7	14:48:09.893	leave	
7	14:48:10.178	enter	384 ms
7	14:48:10.562	leave	
6	14:48.10.562	enter	467 ms
6	14:48.11.029	leave	
7	14:48:11.029	enter	340 ms
7	14:48:11.396	leave	
8	14:48:11.397	enter	283 ms
8	14:48.11.680	leave	
7	14:48:11.681	enter	33 ms
7	14:48:11.714	leave	
6	14:48:11.715	enter	4779 ms
6	14:48:16.494	leave	
7	14:48:16.495	enter	723 ms
7	14:48:17.218	click	

INTRODUCTION

Despite the simplicity and ubiquity of cursor operation, considerable amounts of information can be derived from pointing and clicking. For example, cursor movements and clicks have been studied in the context of the World Wide Web, as a mean to improve the effectiveness of the content presented in the form or a regular web-page [10, 11] or the results of an Internet search [2, 6, 7]. Furthermore, as more tasks are being carried out online, cursor analysis is also being used to infer user behavior. For example, cursor trajectories are used to gauge user uncertainty when filling online forms [8], decision fatigue when completing complex tasks [13], and even to measure self-efficacy and willingness to learn in e-learning and web-based End-user Development systems [5, 9].

A key aspect of cursor analysis research lies in the ability to extract quantifiable patterns from the mouse/trackpad movements and clicks. Among the most commonly extracted patterns are hesitation, reading, and decision process, which are quantified through metrics such as the number of pauses (hesitation), horizontal/vertical movements (reading), and reaction/response time (decision) [1, 3].

This paper focuses on the patterns of hesitation and presents the initial results of an empirical exploration of 10 cursor movement metrics designed to quantify hesitation in online surveys. It is organized as follows: first, the data collection process is described. Second, the data analysis techniques and the proposed methodology are presented. This is followed by the presentation and discussion of the results. The paper ends with the main conclusions and an outline of future work directions.

DATA COLLECTION

This paper uses mouse movement data from a previous work aimed at exploring how people gauge the electricity consumption of 41 domestic appliances [12]. More concretely, 41 screens, each containing a graphical representation of a domestic appliance, were displayed to the participants who were asked to rate their consumption using a relative 1-10 Likert scale (from very low to very high consumption). Figure 1 shows an example of the rating screen with the satellite dish.

Besides the consumption rates, each time an appliance was displayed, the system recorded the mouse movements by keeping track of *mouse enter*, *mouse leave* and *click* events on the screen elements (appliance icon, appliance label and the 10 score buttons of the Likert scale). Mouse transitions and hovers were then derived from the data computationally. Transitions correspond to cursor movements between elements, i.e., the cursor leaving one element and entering another. Hovers refer to the number of times that the cursor is on top of an element. In this paper, a hover is only considered if the hovering time (i.e., the time elapsed between the mouse entering and leaving an element) is at least 100ms as per suggestion of [7].

Table 1 shows the full list of the cursor transitions and hovers from one of the respondents when rating the satellite dish. A graphical representation of the movements is shown in Figure 2. In this

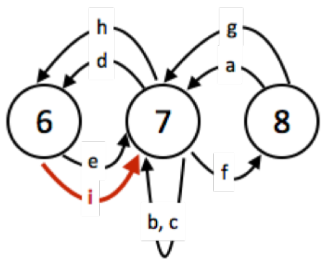

Figure 2: Graphical representation of the cursor movements listed in Table 1.

Table 2: Metric values with respect to the list in Table 1.

Metric	Value
Back and Forth Movements	4
Back and Forth Mov. in Answer	4
Ucliked Hovers	6
Ucliked Hovers in Answer	3
Loops	2
Loops in Answer	2
Pauses	7
Pauses in Answer	4
Total Pause Time	7253 ms
Pause Time Before Click	723 ms

case, the respondent transitioned and hovered a total of three scores before selecting score 7. Note that in the two occasions that the elapsed time was below 100ms hover events were not considered.

Cursor Movement Metrics

According to the literature, cursor metrics to quantify hesitation can be categorized either as: i) metrics that consider at least two targets, or ii) metrics that consider a single target [3]. In this work, five metrics in each category are studied. Table 2 shows the value of each metric with respect to the cursor movements listed in Table 1.

Back-and-forth movements (bnf) and back-and-forth movements in answer (bnf_a). These metrics are multi-target and capture situations where the cursor passes at least twice by one target in the same sequence [3]. The *bnf_a* metric only captures situations where one of the elements is the answer.

Loops (l) and loops in answer (l_a). These metrics capture situations where the cursor leaves and returns to a target without visiting any other targets. The *l_a* metric only captures loops in the answer. *l* is multi-target, and *l_a* is single-target. Note that if an element is looped twice in the same sequence, a *bnf* (or *bnf_a)* is also considered.

Unclicked hovers (u_h) and unclicked hovers in answer (u_h_a). These metrics capture the number of times that a target is hovered but not clicked [7]. These metrics consider only the clickable elements on the screen, i,e, the rating buttons. The *u_h_a* metric only captures situations where the unclicked target is the answer. *u_h* is multi-target, and *u_h_a* is single-target.

Pauses (p) and pauses in answer (p_a). These metrics refer to the number of pauses before providing an answer, where a pause is the lack of movement for longer than 200 milliseconds [5]. In this particular case, a pause is considered whenever a UI element is hovered for at least 200 ms. The *p_a* metric only captures situations where the pause occurs in the answer. *l* is multi-target, and *l_a* is single-target. Note that, by our definition, *p* that occur in rating elements are also *u_h*, and all *p_a* are also *u_h_a*.

Pause time (p_t). This is a multi-target metric and captures the total pause time before an answer is provided [5]. *p_t* is obtained by summing the duration of all the pauses captured by metric *p*.

Pause time before click (p_b_c). This is a single-target metric and captures the elapsed time between the last transition to the target and the click [4], irrespective of the hover duration.

DATA ANALYSIS METHODOLOGY

In order to explore the relationships between the cursor metrics and hesitation levels the data was analyzed using three different techniques: i) Multidimensional Scaling[1] (MDS), ii) Jenks Natural Breaks [2] (JNB), and iii) Jaccard Similarity Index[3] (JSI).

[1] Multidimensional Scaling, https://www.statisticshowto.datasciencecentral.com/multidimensional-scaling/

[2] Jenks Natural Breaks, https://www.ehdp.com/methods/jenks-natural-breaks-1.htm

[3] Jaccard Similarity Index, https://www.statisticshowto.datasciencecentral.com/jaccard-index/

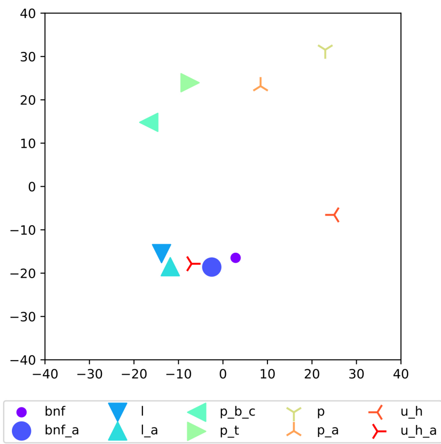

Figure 3: 2D-MDS showing the relative distances between the 10 cursor movement metrics with respect to their normalized values.
Metrics that appear closer to each other are more similar than others (e.g., *bnf_a*, *l_a*, and *u_h_a*). Metrics that have similarities with two or more groups appear closer to each group (e.g., *u_h*). Finally, metrics that are different appear in opposite areas of the graph (e.g., transition- vs pause-based metrics).

MDS was used to provide a visual representation of the distances or dissimilarities between the 10 metrics. Using MDS, metrics that are more similar (i.e., have shorter distances) appear closer together on the graph than those that are less similar (i.e., have longer distances). In this work, 2-dimensional MDS was applied since 2D plots are easier to visualize than their 3D counterparts. Note that before proceeding with the MDS analysis all the metrics were scaled to have unit norm.

The JNB is a clustering technique for 1-Dimensional data and was used to find clusters for each of the metrics. More precisely, given a metric, the JNB algorithm was used to find the k intervals (i.e., clusters) that represented the best split of its values. Then, for each metric, the different appliances were associated with a cluster based on its value in the respective metric. In this work, the number of clusters was set to 5, under the assumption that appliances with higher levels of hesitation would be more prominent in clusters 4 and 5.

The appliances in each cluster were then ranked by the number of occurrences (from larger to smaller), and the top-5 appliances in clusters 4 and 5 were selected as being the ones with higher hesitation levels. This was done for each metric, and the JSI was then used to compare the 10 sets of appliances that were generated. The JSI is a measure of similarity between two sets, with a range from 0 to 1. The higher the index, the more similar the sets are.

RESULTS AND DISCUSSION

These results are based on data from the 252 participants that completed the study in [12] using either a mouse or a trackpad. The remaining 33 participants used touchscreen devices, and thus could not be considered in this work.

The results of the MDS are presented in Figure 3, from which it is possible to see that there is a clear separation between metrics based on pauses (top half) and metrics based on unclicked hovers and transitions (bottom half). In the bottom half of the graph, it is also possible to see that the four transition based metrics appear together. Nevertheless, the most intriguing observation is the fact that the metrics based on unclicked hovers are not close to each other. In fact, *u_h_a* appears between *bnf_a* and *l_a*, which may indicate a strong effect of the fact that all the three metrics have one target in common (i.e., the answer). As for the *u_h* metric, a possible explanation is the fact that pauses in rating elements are also unclicked hovers. As such, *u_h* and *p* have some similarities, which are reflected by their positioning on the right-hand side of the chart.

To further understand the relationships between the metrics, Figure 4 shows the top-5 appliances in clusters 4 and 5 according to each metric. From here, two main observations can be drawn:

(1) **A large number of different appliances selected:** despite the fact that only the top-5 appliances were selected, the results span 34 of the 41 appliances presented in the survey. More concretely, 28 in multi-target metrics, and 30 in the single-rate metrics. The main reason for

	Multiple Elements							Single Element				
	bnf	bnf_a	l	u_h	p	p_t	sr	l_a	u_h_a	p_a	p_b_c	sr
Air Conditioner											5	25%
Aquarium		5	5				33%	1			3	50%
Blender		2				5	33%					
Cable TV Converter			3	5	4	3	67%		4		4	50%
Ceilling Fan		4					17%				3	25%
Cell Phone			5			5	33%				5	25%
Clothes Dryer						4	17%				3	25%
Curling Iron		4		4			33%				4	25%
Desk Fan										5	5	50%
Desktop	2	4	1			5	67%		1	1		50%
Dishwasher											1	25%
DVD Player		4	4			2	50%				4	25%
Electric Blanket				2	1	5	50%		3	2	2	75%
Freezer		5	4				33%			3		25%
Hand Mixer			5				17%	2			5	50%
Heater			2		4		33%					
iPod								2			4	50%
Laptop	1	1	5				50%	2	4			50%
Lightbulb CFL	5		4	1	4		67%		2	4		50%
Lightbulb ICL			5				17%	2				25%
Microwave											5	25%
Playstation			4				17%	2				25%
Popcorn machine	3	5	4	3	3	3	100%	1	4	4		75%
Refrigerator			3				17%	2			5	50%
Router						5	17%					
Satellite Dish	2	3	2	3	2	1	100%	2	4		2	75%
Stero System		4					17%					
Stove		2				5	33%		5		3	50%
Table Mixer			4			4	33%				2	25%
TV CRT	4	4					33%		5			25%
TV Flatscreen		3	5		5		50%			2		25%
Vaccum Cleaner		4	2			5	50%				3	25%
Washing Machine			5				17%		5	3		50%
Water Heater											2	25%
Count	**6**	**15**	**19**	**6**	**7**	**13**	**28**	**9**	**10**	**8**	**20**	**30**

Figure 4: Top-5 appliances selected by each of the 10 cursor movement metrics. Note that the metrics are grouped by category and that for each category *sr* refers to selection rate, i.e., the ratio between the number of times that an appliance was selected and the number of metrics in that category. The higher the *sr*, the more the metrics agree.
Overall, this visualization provides an easy way to inspect how: 1) each metric ranks the appliances in terms of hesitation, 2) the level of agreement between individual metrics, and 3) the level of agreement between the two categories of metrics.

this is the considerably high number of ties for appliances in the fourth and fifth positions in two of the metrics (l and p_b_c). For instance, if only the top-3 appliances were selected, the results would span 26 appliances (11 in multi-target and 23 in single-target).

(2) **Agreement with respect to the appliances with more hesitation:** despite the large number of selected appliances, there is a strong agreement with respect to the appliances with more hesitation. More concretely, the *Satellite Dish* and the *Popcorn Machine* are the top appliances in both categories, selected by 9 out of the 10 metrics. These are followed by four appliances selected by 6 out of the 10 metrics, namely the *Cable TV Converter* (4 multi- + 2 single-target), the Desktop (4 + 2), the Electric Blanket (3 + 4), and the CFL Lightbulb (4 + 2).

Finally, Figure 5 shows the JSI across the top-5 appliance sets selected by each metric. The following two main observations can be drawn:

(1) **The low JSI values overall ($0.04 \leq JSI \leq 0.63$)**: in fact, the JSI is only above 0.5 for two pairs of sets. More concretely, $JSI(p, u\_h) = 0.63$, and $JSI(u\_h\_a, bnf) = 0.6$, which ultimately helps to understand the positioning of the u_h and u_h_a metrics in Figure 3.

(2) **The very low JSI values for the *p_b_c* and *l_a* metrics**: regarding the former, this effect can be justified by the fact that this metric only takes time into consideration, therefore the 0.38 JSI between p_b_c and p_t. As for the latter, this effect can be potentially justified by the low number of loops in the data, since only 6% of the total responses have 1 or more loops, and only 1% have 1 or more loops in the answer. Nevertheless, it should be stressed that despite the low JSI, the l_a metric also ranked the *Satellite Dish* and *Popcorn Machine* as the appliances that generated more hesitation.

CONCLUSION AND FUTURE WORK DIRECTIONS

This work presented an empirical exploration of 10 cursor metrics (six multi- and four single-target), for assessing end-user hesitation when responding to online surveys.

Ultimately, the results show that despite the wide range of selected appliances, there is an agreement with respect to those that generate more hesitation. Nevertheless, since the JSI among the top-5 of each metric is considerably low ($0.04 \leq JSI \leq 0.63$), decisions based on a single metric should be avoided. With respect to the distinction between multi- and single-target based metrics, the lower number of ties with respect to the top-3 appliances suggests that multi-target metrics may have more discriminant power. However, since there is no actual feedback from the respondents with respect to how confident they are with their responses, it is not possible to make such a conclusion.

Against this background, future iterations of this work should seek to include mechanisms to gather self-reported measures of confidence from the participants. Likewise, future work should consider different UI designs, since it is possible that some metrics are affected by this aspect. For instance, it is

	bnf	bnf_a	l	l_a	u_h	u_h_a	p	p_a	p_t	p_b_c
bnf		0,31	0,31	0,25	0,33	0,60	0,30	0,27	0,19	0,04
bnf_a	0,31		0,36	0,20	0,17	0,32	0,16	0,21	0,33	0,25
l	0,25	0,36		0,40	0,19	0,32	0,30	0,29	0,28	0,30
l_a	0,25	0,20	0,40		0,15	0,19	0,14	0,06	0,10	0,21
u_h	0,33	0,17	0,19	0,15		0,45	0,63	0,27	0,27	0,18
u_h_a	0,60	0,32	0,32	0,19	0,45		0,42	0,38	0,35	0,15
p	0,30	0,16	0,30	0,14	0,63	0,42		0,36	0,25	0,13
p_a	0,27	0,21	0,29	0,06	0,27	0,38	0,36		0,17	0,08
p_t	0,19	0,33	0,28	0,10	0,27	0,35	0,25	0,17		0,38
p_b_c	0,04	0,25	0,30	0,21	0,18	0,15	0,13	0,08	0,38	

Figure 5: Pairwise JSI across the top-5 appliance sets selected by each metric (please note that the matrix is symmetric). The higher the JSI, the more the metrics agree.

Overall, this representation provides an easy way to understand how the differences in the metrics, exposed by the MDS, affect the selection of the top-5 appliances in terms of hesitation levels.

possible that the low number of loops in the answer (*l_a*) is a consequence of the sequential layout of the rating buttons, since it may be hard to leave one element without hovering over another. Finally, future work should also consider the effects of regional variation in the hesitation levels, since (and especial in this use case) it is likely that the familiarity with some appliances varies across regions.

ACKNOWLEDGMENT

This research was partially funded by the Portuguese Foundation for Science and Technology (FCT) through the research grant with reference UID/EEA/50009/2019.

REPRODUCIBLE RESEARCH

For research reproducibility purposes, the data and Python scripts used in this work are freely and publicly available to the community. They can be found in https://www.alspereira.info/pubs/chi-2019-a/.

REFERENCES

[1] Catia Cepeda, João Rodrigues, Maria Camila Dias, Diogo Oliveira, Dina Rindlisbacher, Marcus Cheetham, and Hugo Gamboa. 2018. Mouse Tracking Measures and Movement Patterns with Application for Online Surveys. In *Machine Learning and Knowledge Extraction (Lecture Notes in Computer Science)*, Andreas Holzinger, Peter Kieseberg, A Min Tjoa, and Edgar Weippl (Eds.). Springer International Publishing, Hamburg, Germany, 28–42.

[2] Y. Chen, Y. Liu, M. Zhang, and S. Ma. 2017. User Satisfaction Prediction with Mouse Movement Information in Heterogeneous Search Environment. *IEEE Transactions on Knowledge and Data Engineering* 29, 11 (Nov. 2017), 2470–2483. https://doi.org/10.1109/TKDE.2017.2739151

[3] Silvana Churruca. 2011. *Patterns of cursor movement for different devices. Comparative study of cursor movement patterns between a touchpad and a mouse devices.* Master. Universitat Pompeu Fabra, Barcelona, Catalonia, Spain.

[4] Clicktale. 2019. Enterprise Experience Analytics | Conversions Optimization. https://www.clicktale.com/

[5] M. S. R. Dijkstra. 2013. *The diagnosis of self-efficacy using mouse and keyboard input.* Master. Universiteit Utrecht, Utrecht, the Netherlands. http://dspace.library.uu.nl/handle/1874/280262

[6] Qi Guo and Eugene Agichtein. 2008. Exploring Mouse Movements for Inferring Query Intent. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '08).* ACM, New York, NY, USA, 707–708. https://doi.org/10.1145/1390334.1390462

[7] Jeff Huang, Ryen W. White, and Susan Dumais. 2011. No Clicks, No Problem: Using Cursor Movements to Understand and Improve Search. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '11).* ACM, New York, NY, USA, 1225–1234. https://doi.org/10.1145/1978942.1979125

[8] Jeffrey L. Jenkins, Ross Larsen, Robert Bodily, Daniel Sandberg, Parker Williams, Steve Stokes, Spencer Harris, and Joseph S. Valacich. 2015. A multi-experimental examination of analyzing mouse cursor trajectories to gauge subject uncertainty. In *2015 Americas Conference on Information Systems, AMCIS 2015.* Americas Conference on Information Systems, Fajardo, Puerto Rico, 1–14.

[9] Tzafilkou Katerina and Protogeros Nicolaos. 2018. Mouse Behavioral Patterns and Keystroke Dynamics in End-User Development. *Comput. Hum. Behav.* 83, C (June 2018), 288–305. https://doi.org/10.1016/j.chb.2018.02.012 00002.

[10] Florian Mueller and Andrea Lockerd. 2001. Cheese: Tracking Mouse Movement Activity on Websites, a Tool for User Modeling. In *CHI '01 Extended Abstracts on Human Factors in Computing Systems (CHI EA '01).* ACM, New York, NY, USA, 279–280. https://doi.org/10.1145/634067.634233 00198.

[11] Vidhya Navalpakkam and Elizabeth Churchill. 2012. Mouse Tracking: Measuring and Predicting Users' Experience of Web-based Content. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '12).* ACM, New York, NY, USA, 2963–2972. https://doi.org/10.1145/2207676.2208705 00058.

[12] Lucas Pereira and Yoram Chisik. 2017. A Mouse over a Hotspot Survey: An exploration of perceptions of electricity consumption and patterns of indecision. In *Proceedings of the fifth IFIP Conference on Sustainable Internet and ICT for Sustainability.* IEEE / IFIP, Funchal, Portugal, 1–4. https://doi.org/10.23919/SustainIT.2017.8379808

[13] Daniel Reinhardt and Jörn Hurtienne. 2018. Cursor Entropy Reveals Decision Fatigue. In *Proceedings of the 23rd International Conference on Intelligent User Interfaces Companion (IUI '18 Companion).* ACM, New York, NY, USA, 31:1–31:2. https://doi.org/10.1145/3180308.3180340

View publication stats