Applying Flow Theory to Predict User-Perceived Performance of Tablets

James Scovell(✉) and Rina Doherty

Intel Corporation, Platform Evaluation and Competitive Assessment, Hillsboro, OR, USA
{James.J.Scovell,Rina.A.Doherty}@intel.com

Abstract. A users' perception of interactive device performance is influenced by their feeling of being *in control* and that there is a sense of constant progress. A system will be able to keep users in the flow by meeting expectations and keeping up with their inputs and commands. The concept of flow has been discussed since the 1960's and has been used in the context of computing devices; however, the ability to operationally define and quantitatively measure this construct is limited. This paper describes a study that tested a new framework for measuring flow as it relates to User-Perceived Performance (UPP) of tablets.

Keywords: User-perceived performance · Severity-Duration · Mean Opinion Score (MOS) · User experience (UX) · Flow · Tablet · Computer performance

1 Introduction

Technological advances over the last several years have shifted the way in which humans interact with computing devices. One major shift has been in the use of touchscreens as a primary means of interaction. However, the software and hardware that enables this input modality has created complexities that challenge engineers to maintain the instant response a user has come accustom to like with a mouse and keyboard. New challenges like this have impacted the utility of computer performance measurement techniques and their relevance to user experience. Traditional performance metrics are primarily designated for compute intensive operations as opposed to the shorter, more interactive exchanges. As such, the aspects that largely shape how an end-user perceives the performance of a touchscreen device cannot be measured using these customary methods. Comprehending this distinction has taken time and being able to quantify and rate the perceived performance of a touch interactive device poses many exciting challenges. This paper discusses the use of a new user-centric approach to measuring computer performance developed from a distillation of user research studies, along with a review of published literature. A comparison of the average participant ratings from a tablet study is presented as they compare with predicted average ratings derived from this new approach.

A. Marcus (Ed.): DUXU 2016, Part III, LNCS 9748, pp. 77–87, 2016.
DOI: 10.1007/978-3-319-40406-6_8

2 Background

2.1 Flow

Keeping users in the flow is a key to end-user satisfaction with highly interactive devices like tablet computers. The theory of flow was first introduced in the 1960's and started to be discussed in the literature more in the 1980's when Csikszentmihalyi described the concept as the state of being fully absorbed and motivated toward an activity where a person's attention is so narrowly focused on an activity that time can seem to fade away [1]. Flow is sometimes interchanged with the notion of immersion or being *in the zone*. According to Csikszentmihalyi, flow has four preconditions; (1) goals, (2) clear rules to obtain those goals, (3) clear and immediate feedback to provide certainty, and (4) skill level must be appropriate to achieve a balance of control and challenge [1]. Amongst the different categories that flow has been examined, it has also been studied as it relates to computer performance and user satisfaction; specifically, how poor computer performance impacts flow [2–4]. Especially in the case of highly interactive devices like tablets, poor responsiveness violates the last two preconditions of flow; it creates uncertainty and it diminishes a users' sense of control. Depending on the user request, dissatisfaction can be the result of sub-second latencies or much longer processing delays [4–8]. This unique dimension of time perception increases the challenge associated with determining how to measure user satisfaction. As such, research on how users perceive computer performance has a long history.

2.2 User-Perceived Performance (UPP)

Mangan has been credited with first describing the term, "perceived performance", in a white paper he published in 2003 [10]. He recommended that practitioners shift their focus away from only relying on traditional computational performance measurement and scoring practices to those aspects of system behavior that impact end-users more saliently. Prominent researchers such as Miller [11], Shneiderman [2], Card et al. [12], and Seow [4] have proposed taxonomies of system response requirements centered on memory, task type, user expectations, and task complexity. For a more in-depth understanding, these contributions are described in Anderson et al. [13], Doherty and Sorenson [9], and Dabrowski and Munson [3].

Largely influenced by Mangan and the other researchers noted above, Verheij published a white paper in 2011 describing an approach to quantifying and rating perceived performance of a virtual desktop system application [14]. The goal of his process was to give an indication of how an average user would rate responsiveness. It includes what he calls an ARI (Application Responsiveness Index) and a PPI (Perceived Performance Index). In this approach, the rating of response times is determined by the type of user action. User actions are categorized in three ways and each have a corresponding threshold of time as seen in Table 1. An Apdex [15] calculation is used to quantify the level of perceived performance and maps back to one of these five qualifiers: excellent, good, fair, poor, and unacceptable. The PPI adds an additional weighting function based on the variability of response times; the less variability the

Table 1. Ingmar's response time rating categories and time thresholds

Category name	Threshold
Acknowledgement of command	0.1 s
Simple task	1 s
Complex task	10 s

better the perceived performance rating. He believed this added a good indication of the perceived performance of an application over time.

There have also been other approaches presented in the literature to quantify perceived performance. For example, Tolia et al. [16] described a technique they used to quantify the impact of network latency on what they called "interactive experience" using thin clients (i.e., all application and operating system code is executed on a server). Thin client computing is particularly challenged with providing crisp responses to the basic (but common) interactions like menu navigation and mouse tracking. As such, their quantification and rating categories focused solely on these shorter interactions and system responses that require limited processing. The categories that they placed response times into can be seen in Table 2 below.

Table 2. Tolia et al.'s response time categories

Category name	Time
Crisp	<150 ms
Noticeable to annoying	150 ms to 1 s
Annoying	1 to 2 s
Unacceptable	2 to 5 s
Unusable	>5 s

In 2015, Doherty and Sorenson presented another categorization mapping system response time (SRT) to user satisfaction. Their categorization was an extension of Shneiderman and Seow's work that combined the influence of user expectations and complexity of tasks and added human perceptual limits as a third factor for determining appropriate categories. They also went beyond the attentive (10 s) time frame to include those more compute intensive operations that require SRTs beyond users' attention span to provide a more comprehensive set of SRT ranges to account for any user task flow (see Table 3).

In addition, they proposed that it is necessary to go beyond this simple mapping in order to quantify and rate users' perception of flow. A more comprehensive mapping includes predictive models that quantitatively define how user experience ratings change as a function of SRT changes so that instead of just being able to report if a SRT fell within a given range of user satisfaction, it is possible to calculate a quantifiable rating on a continuous scale. This can provide more robust data for practitioners to make informed decisions around design trade-offs, but also affords the ability to 'add

Table 3. Doherty and Sorenson's SRT framework with category names, time range, and descriptions.

Attention	Category name	SRT range	Category description
Attentive	Instantaneous	<300 ms	User feels like they are in a closed-loop system; as if they are in direct control
	Immediate	300 ms–1 s	Processes perceived by user as easy to perform
	Transient	1 s–5 s	Perceived by user as requiring some simple processing but user feels that they are making continuous progress (appropriate feedback required). It is unlikely a user would disengage from task flow
	Attention span	5 s–10 s	Perceived by users as requiring more processing/wait time but user needs useful and informative feedback to stay closely engaged.
Non-attentive	Non-attentive	10 s–5 min	Perceived by users as requiring more complex processing. Users would be likely to disengage and multi-task during this process. Feedback of progress is necessary
	Walk-away	>5 min	Perceived by users as requiring intensive processing. Users would not stay engaged with this task. Feedback of progress is necessary

up' or aggregate a sequence of interactions and calculate an overall responsiveness experience rating.

A similar example of this can be seen from a study conducted by Anderson et al. [13] where participants were asked to carry out several tasks and then rate the satisfaction of the response times on a five point scale. Participants repeated the tasks and ratings under varying levels of computer performance and as a result the researchers presented trend lines depicting the user ratings as a function of SRT (Fig. 1). These, in fact, represent the early stages of a collection of predictive models that can populate the Table 3 SRT framework.

In past research there has been a strong emphasis on total system response time from the start of a user input to the end (completion) of the system response. However, the type and stages of feedback given can impact a user's perception of being in control and the feeling that constant progress is being made. For example, recognition of a user input and progressive loading can reduce participant anxiety and set expectations as to how long the interaction will take to complete [17].

There are other factors that can degrade user-perceived performance (UPP) while using an interactive device, such as the interface and/or graphics quality, the smoothness of content, or the accuracy of responses to user inputs. Display smoothness and/or poor frame delivery can greatly impact a user's perception of performance, especially for high motion interactions such as gaming and watching videos. Similarly,

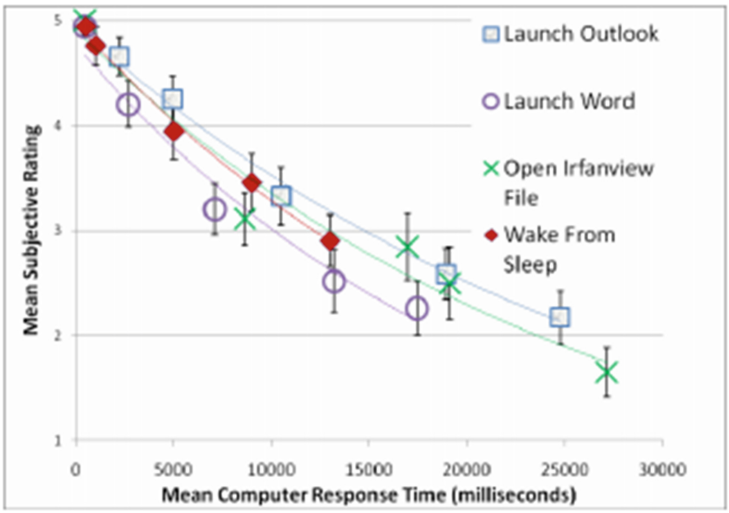

Fig. 1. Mean rating by duration for launching Outlook, Word, an IrfanView file, and wake from sleep. Reprinted from "Diminishing Returns? Revisiting Perception of Computing Performance" by G. Anderson, R. Doherty, E. Baugh, 2011, Proceedings of CHI, p. 2073.

inaccuracies related to touch interactions affect user's performance when she or he is forced to repeat a selection and/or correct an unintended input. Consideration of smoothness and accuracy variables is necessary to predict tablet UPP since they impact users' sense of certainty and control.

The study presented below was designed to collect participant ratings of tablet performance from realistic use cases and interactions. Measures of responsiveness, feedback, smoothness, and accuracy were used to predict overall UPP ratings for each workflow on each device tested. These overall predicted ratings were compared to the average participant ratings to determine if the predicted formula was a good approximation for UPP. By developing representative workflows and measuring multiple stages of feedback it was hypothesized that it would be possible to utilize predictive models to more holistically quantify UPP.

3 Research Methods

3.1 Devices

A total of six tablets of similar screen size were included in this study, two from each of the three common operating systems (iOS, Android, and Windows). Each pair of operation system (OS) devices included one high-end system and one system that had been on the market for two to three years. These devices were selected to understand participant expectations of best in class tablets and ensure there would be variation in the performance of the devices.

3.2 Workflows

Participants completed eight workflows including; email, web browsing, photo editing, video editing, video streaming, two gaming applications (Fruit Ninja and Jetpack Joyride), and mapping. These workflows were chosen to represent common usages of tablets and consisted of a series of interactions that represented capabilities of the software and application being used. Workflows consisted of interactions from basic system level to computationally complex interactions. While some applications, like games, were the same across OS this was not possible for all applications. To represent the most common experience of a given device, default applications such as for email and mapping were used. Using default applications provided the most representative experience. Participants completed all eight workflows on both of the tablets of their personal, primary OS.

3.3 Participants

A total of 51 participants completed the study. Participants were all experienced tablet users who used one or more tablets for more than five hours per week. Participants were screened in an attempt to get an equal number per OS (19 iOS, 16 Windows, 16 Android) and a good distribution across age, gender, and income. Test sessions lasted about one hour and participants were compensated accordingly.

3.4 Procedures

The experimental design and rating procedures followed the MOS (Mean Opinion Score) ITU standards for measurement of subjective assessment [18]. Participants received instructions on how to complete the eight workflows to ensure each participant completed the same interactions. While completing the workflows participants were asked to focus on the device performance and ignore extraneous variables such as comfort of the chair or room environment. A five point scale (5 = excellent, 4 = good, 3 = fair, 2 = poor, 1 = bad) was presented to participants to rate the performance of the device. Participants were asked to give a rating at the end of each workflow and provide an overall rating when they completed all eight workflows on a device. In addition, participants were asked to provide a rating any time they felt the system was not performing at a 5 (excellent). In order to gather more insights into what variables impacted UPP, participants were also asked to comment on what aspects impacted their ratings. These participant comments were transcribed for later analysis.

Participants were presented with one device of their personal, primary OS. Participants were then instructed to complete all eight workflows, one at a time. Upon completion of all eight workflows the participant was given a short break then completed all eight workflows on the other device of the same OS (in the same order as the first device). Participants only interacted with devices that had the OS they were most familiar with to reduce potential learning effect confounds. Device order was counterbalanced between participants to minimize order effects.

All participant sessions were recorded with the device display as the focal point of the camera. These recordings were used to capture the interactions with the device to observe any discrepancies/errors, to capture participant comments, and capture the device latency to participant inputs. The tablet devices were held in place at a 45 degree angle to the participant on a tablet stand. The participants did not hold or pick up the devices during the study.

3.5 Analysis

Quantifying Subjective Comments. SRT was captured during the study objectively using the video capture content. However, video/gaming smoothness and input accuracy issues were collected subjectively. In an effort to quantify these metrics, participant comments were documented and categorized to understand their contribution to negative UPP. Subjective comments were quantified by dividing the number of negative comments in a given category (i.e. input accuracy issues) by the number of participants who completed the given workflow on a given device. This calculation was completed for each device workflow combination so each workflow had a percentage associated with input accuracy and smoothness issues for each device. These percentages were then converted to a five point scale by correlating them with the average participant ratings. This provided an estimation in the absence of objectively measured feasibility at the time of this study.

System Response Time. The video recordings were analyzed to collect the latencies of each participant interaction. Multiple stages of loading were captured for each interaction to capture the aspect of feedback. There were two stages of loading that were measured, start of load (first indication to participant that the system is executing the intended interaction) and end of interaction (load is complete and ready for further input). Average system response times were taken across participants for each phase of load and of each interaction for all eight workflows per device. This provided the average response time for each interaction within each workflow on each device. Doherty and Sorenson's [9] framework was used to categorize each SRT measurement into a perceptual category. Proprietary mathematical models were assigned to each stage of load for each interaction according to human perceptual limits, perceived complexity, and user expectations. Calculating a range of response times across devices provided an indication of what users expect for each interaction. Using these models, a UPP was calculated for each interaction and stage of feedback.

The predicted UPP's were then aggregated according to a concept called Severity-Duration (Fig. 2). Severity-Duration penalizes for the severity of degradation when interactions do not meet user expectations. There is also a penalty for duration (or consecutiveness) of interactions that did not meet user expectations. An example of this can be seen in Table 4. This concept was implemented into the aggregation methodology since using a straight average was not believed to capture the impact of these negative contributors to UPP over the duration of a workflow. Evidence of this has been seen in the literature where negative experiences outweigh positive when reporting overall impressions [19–22].

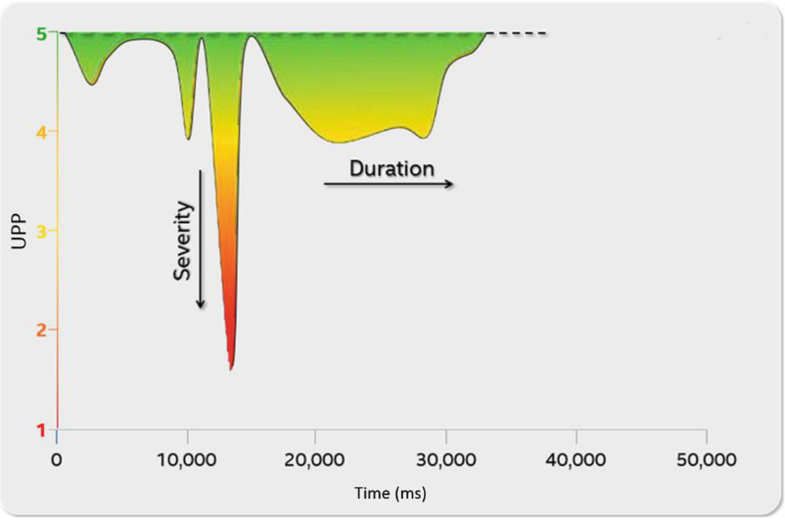

Fig. 2. Severity-Duration concept

Table 4. Severity-Duration example calculation

	SRT	SRT MOS	Severity-Duration MOS
	A	B	C
1	Time (ms)	y = m(A1) + B	=B1
2	Time (ms)	y = m(A2) + B	=(C1 + B2)/2
3	Time (ms)	y = m(A3) + B	=(C2 + B3)/2
4	Time (ms)	y = m(A4) + B	=(C3 + B4)/2
			Average

Metric Aggregation. The overall predicted UPP rating was calculated by averaging the SRT Severity-Duration MOS with relevant subjective variables, depending on the workflow. Gaming and video streaming workflows equally weighted the SRT Severity-Duration MOS, accuracy subjective rating, and smoothness rating. The other five workflows did not include significant video or animation so overall predicted ratings were simply an average of the SRT Severity-Duration MOS and accuracy subjective ratings.

4 Results

Results of a correlational analysis show the overall predicted UPP ratings were highly correlated with the average study participant ratings $r(46) = .783$, $p < .001$ with an average absolute delta of 0.14. Figure 3 shows the results of the regression analysis for each workflow on each device. Although there was an effort to include tablet devices

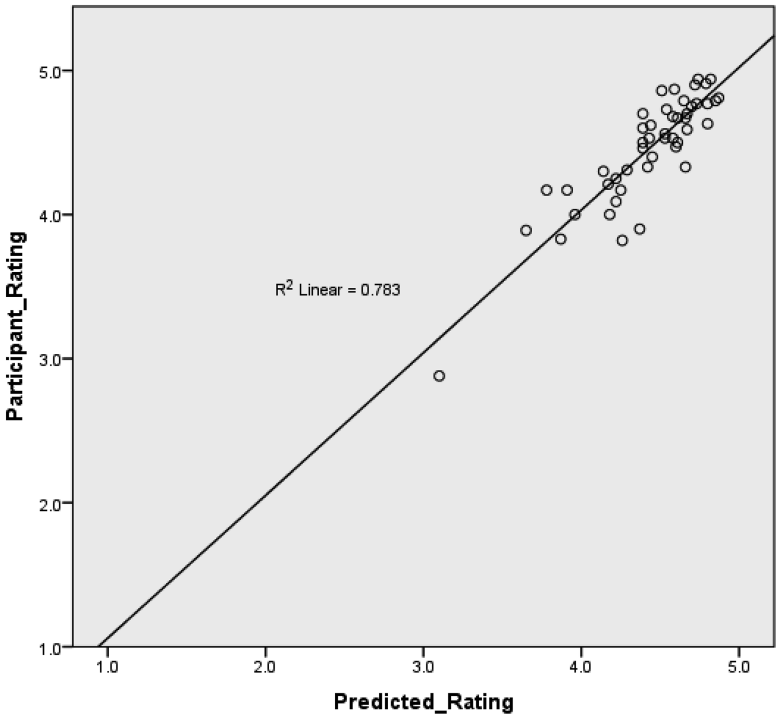

Fig. 3. Correlation of overall predicted ratings with average user ratings

with a range of perceived performance, the results show that the majority of average participant and predicted ratings fell above the 3.5 range.

5 Discussion

The study described in this paper adds to the evolution of flow theory as it relates to computer performance. It provides a new user-centric approach for calculating UPP of tablet devices. This new approach also expands upon traditional UPP metrics by incorporating touch accuracy and video/gaming smoothness metrics. A new aggregation concept was introduced, called Severity-Duration, to provide an alternative approach to simply *averaging* multiple interactions within a workflow to calculate an overall UPP. This methodology was able to accurately predict participant UPP ratings within 0.14, on average (range 0.0–0.47). It should be noted that while an effort was made to get a wide range of UPP tablets, ratings were almost entirely above 3.5. Future research is necessary to expand this range to ensure accuracy across the full ratings scale.

It is critical to develop methodologies that can collect objective metrics for both perceived touch accuracy issues and video/animation smoothness. The transformation of subjective comments to a percentage and then to a five point scale helped confirm

the impact of these variables on UPP; however, more validation from objectively measured metrics along with user rating consistency is required from future research. Subjective data was useful in estimating these metrics, collecting subjective data is not scalable for evaluating new workflows or new devices as they are released. It will also be important that these new objective measurement capabilities can be collected during live user research to ensure the established aggregation methodology holds true to predict participant ratings.

To be truly successful the metrics described would need to be collected in an automated fashion, allowing for minimal researcher interaction and no end user involvement other than the periodic and regular user study to account for user expectations shifting over time. Automated data collection would allow for iterative evaluation during the product development cycle and competitive analysis on existing devices. Upon validating the framework described here for touch devices, this framework can also be tested and applied to other devices with different input modalities. This will allow for a more robust framework where UPP can be objectively measured for all computing devices, bringing a more representative form of performance measurement that is representative of what really matters to end-users.

References

1. Csikszentmihalyi, M.: The flow experience and its significance for human psychology. In: Csikszentmihalyi, M., Csikszentmihalyi, I.S. (eds.) Optimal Experience: Psychological Studies of Slow in Consciousness, pp. 15–35. Cambridge University Press, Cambridge (1988)
2. Shneiderman, B.: Designing the User Interface: Strategies for Effective Human-Computer Interaction, 1st edn. Addison-Wesley, Reading (1987)
3. Dabrowski, J., Munson, E.: 40 Years of searching for the best computer system response time. Interact. with Comput. **23**, 555–564 (2011). Elsevier B.V
4. Seow, S.C.: Designing and Engineering Time: The Psychology of Time Perception in Software. Addison-Wesley, Amsterdam (2008)
5. Zakay, D., Hornik, J.: How much time did you wait in line? A time perception perspective. In: Time and Consumer Behaviour (1991)
6. Angrilu, A., Cherubini, P., Pavase, A., Manfredini, S.: The influence of affective factors on time perception. Percept. Pyschophys. **59**, 972–982 (1997)
7. Goldstein, B.E.: Sensation and Perception, 5th edn. University of Pittsburg, Pittsburg (1999)
8. Jota, R., Ng, A., Dietz, P., Wigdor, D.: How fast is fast enough? A study of the effects of latency in direct touch pointing tasks. In: Proceedings of S1GCHI Conference on Human Factors in Computing Systems (CHI 2013), pp. 2291–2300. ACM, New York, NY, USA (2013)
9. Doherty, R., Sorenson, P.: Keeping users in the flow: mapping system responsiveness with user experience. Procedia Manuf. **3**, 4384–4391 (2015)
10. Mangan, T.: White paper perceived performance. Tuning a system for what really matters. TMurgent Technologies (2003). http://www.tmurgent.com/WhitePapers/PerceivedPerformance.pdf
11. Miller, R.B.: Response time in man-computer conversational transactions. In: International Business Machines (IBM) Corporation, Fall Joint Computer Conference, Poughkeepsie, New York (1968)

12. Card, S.K., Moran, T.P., Newell, A.: The Psychology of Human-Computer Interaction. Lawrence Erlbaum Associates, Hillsdale (1983)
13. Anderson, G., Doherty, R., Baugh, E.: Diminishing returns? Revisiting perception of computing performance. In: Proceedings of CHI, pp. 2703–2706 (2011)
14. Verheij, I.: White paper quantifying and rating perceived performance of a virtual desktop system application (2011). http://www.ingmarverheij.com/wp-content/uploads/downloads/2011/09/Whitepaper-Quantifying-Perceived-Performance-v1.0.pdf
15. Apdex (2007). http://apdex.org/overview.html
16. Tolia, N., Andersen, D.G., Satyanarayanan, M.: Quantifying interactive user experience on thin clients. In: Proceedings of the IEEE. Carnegie Mellon University (2006)
17. Norman, D.: The Design of Everyday Things. Basic Books, New York (2002). ISBN 978-0-465-06710-7
18. ITU-R.: Methodology for the subjective assessment of the quality of television pictures. Recommendation BT.500-13, Geneva (2012)
19. Mittal, V., Ross Jr, W.T., Baldasare, P.M.: The asymmetric impact of negative and positive attribute-level performance on overall satisfaction and repurchase intentions. J. Mark. **62**, 33–47 (1998)
20. Oliva, T.A., Oliver, R.L., Bearden, W.O.: The relationship among consumer satisfaction, involvement, and product performance. Behav. Sci. **40**(April), 104–132 (1995)
21. Oliver, R.L.: Cognitive, affective, and attribute bases of the satisfaction response. J. Consum. Res. **20**(December), 418–430 (1993)
22. Peeters, G., Czapinski, J.: Positive-negative asymmetry in evaluations: the distinction between affective and informational negativity effect. Eur. Rev. Soc. Psychol. **1**, 33–60 (1990)