System Latency Guidelines Then and Now – Is Zero Latency Really Considered Necessary?

Christiane Attig[1(✉)], Nadine Rauh[1], Thomas Franke[2], and Josef F. Krems[1]

[1] Department of Psychology, Cognitive and Engineering Psychology, Chemnitz University of Technology, Chemnitz, Germany
{christiane.attig,nadine.rauh, josef.krems}@psychologie.tu-chemnitz.de
[2] Institute for Multimedia and Interactive Systems, Engineering Psychology and Cognitive Ergonomics, Universität zu Lübeck, Lübeck, Germany
franke@imis.uni-luebeck.de

Abstract. Latency or system response time (i.e., the delay between user input and system response) is a fundamental factor affecting human-computer interaction (HCI). If latency exceeds a critical threshold, user performance and experience get impaired. Therefore, several design guidelines giving recommendations on maximum latencies for an optimal user experience have been developed within the last five centuries. Concentrating on the lower boundary latencies, these guidelines are critically reviewed and contrasted with recent empirical findings. Results of the review reveal that latencies below 100 ms were seldom considered in guidelines so far even though smaller latencies have been shown to be perceivable to the user and impact user performance negatively. Thus, empirical evidence suggests a need for updated guidelines for designing latency in HCI.

Keywords: System response time · Latency · User experience · Design guidelines · Human-computer interaction

1 Introduction

Even though many technological advances aiming at fulfilling the quest for zero latency have emerged in recent years (e.g., regarding hardware and software speed, communication bandwidth), system latency still remains an inevitable aspect of human-computer interaction (HCI). If latency or system response time (SRT; i.e., the time interval between user input and system response), also known as lag or delay, exceeds a certain threshold, users are able to perceive and become aware of latency (e.g., [18]). If it increases even further, user experience (e.g., [35]) and satisfaction (e.g., [12]) can be impaired. Finally, also users' performance can be negatively affected by latency (e.g., [5]), even by latencies below the perceptional threshold [22].

For enabling system engineers and interface designers to create systems with the best user experience possible, several design guidelines for various applications have been established in the last 45 years (for overviews see e.g. [3, 9]). All these guidelines try to

D. Harris (Ed.): EPCE 2017, Part II, LNAI 10276, pp. 3–14, 2017.
DOI: 10.1007/978-3-319-58475-1_1

answer the core question: Where are the latency thresholds? However, different guidelines for different aspects of HCI have to be distinguished. While some guidelines deal with human perception (e.g., what is the upper level of latency that users will just not notice?), others deal with user experience (e.g., what is the minimum latency where users start to get annoyed?). In this review, classic (i.e., before 1999) and more recent (i.e., since 2000) latency guidelines for designing interactive systems are examined. In the light of technical advances striving for zero-latency systems, our central question is: Are latencies close to zero considered necessary in these guidelines? Therefore, we concentrate on the lower latency limits that are specified in the reviewed latency guidelines (see Table 1).

Table 1. Latency guidelines and their lower limit latency recommendations.

Guideline	Smallest latency threshold	Characterization
Miller [23]	100–200 ms	• Latency guidelines for 17 different types of HCI • 100–200 ms is the longest acceptable latency for control activations • Based on the author's expert estimation
Shneiderman and Plaisant [31]	50–150 ms	• Latency guidelines for different task complexity levels • 50–150 ms is the longest acceptable latency for basic, repetitive tasks • Based on empirical data
Card et al. [7]	100 ms	• Latency guidelines representing human perceptual limits • 100 ms is the maximum latency for creating the illusion that a system runs instantaneously • Generalized from classic psychophysical experiments
Seow [29]	100–200 ms	• Latency guidelines for different user expectations • 100–200 ms is the longest acceptable latency for system responses that the user expects to be instantaneous • Foundations not clearly stated
Tolia et al. [32]	150 ms	• Latency guidelines for interactions with thin clients • Below 150 ms user performance will not be negatively influenced and the user will not notice the latency • Based on previous guidelines and empirical data
Kaaresoja et al. [19]	visual: 30–85 ms audio: 20–70 ms tactile: 5–50 ms	• Latency guidelines for different feedback modalities after touchscreen button presses • Perceived button quality will decrease with latencies above the thresholds • Based on empirical data
Kaaresoja [20]	visual-audio visual: 90 ms audio: 70 ms visual-tactile visual: 100 ms tactile: 55 ms tactile-audio tactile: 25 ms audio: 100 ms	• Latency guidelines for bimodal feedback after touchscreen button presses • Perceived button quality will decrease with latencies above the thresholds • Based on empirical data
Doherty and Sorenson [11]	300 ms	• Latency guidelines for different user expectations and attentional states • Below 300 ms the users will feel as if they are in direct control • Based on previous guidelines and empirical data

2 Classic Latency Guidelines

The first author to determine latency thresholds was Miller in 1968 [23]. His design recommendations for various types of HCI were based on "the best calculated guesses by the author" ([23], p. 271), that is, they were not based on systematic empirical investigations (see also [5]). These early guidelines, which were focused on user acceptance (i.e., acceptable latencies), were theoretically grounded on two pillars: (1) common expectancies in interpersonal communication (i.e., typical patterns of interpersonal communication) and (2) memory research. Regarding the first aspect, according to Miller, in a conversation between two people an answer is expected within a few seconds. If the response delay exceeds four seconds, the thread of communication breaks [23]. Miller applied this pattern to HCI, which he viewed as a conversational act similar to a dialogue between two people, and defined maximum SRTs for 17 different kinds of conversational acts between the user and the system. Regarding the second aspect, due to the limited capacity of short-term memory, human thought and problem solving processes are interrupted if the SRT exceeds a certain threshold. The longer a chunk has to be kept active in short-term memory, the more likely are the chances of errors or forgetting (e.g., the chances of forgetting an e-mail address rise with increasing delay in loading the e-mail software). According to Miller, the longest acceptable latency for the system response in the most basic interactions (control activations, i.e., feedback that signals physical activation, e.g., an audible mouse click) is 100–200 ms. SRTs below 100 ms are not mentioned by Miller. Being aware that his recommendations can only be a starting point, Miller urged the need for empirical validation of his guidelines. Nevertheless, they constituted a first valuable guidance for practitioners and were used as reference in research on SRT and its effects on user experience.

In his review on SRT and human performance, Shneiderman [30] summarized experimental research on SRT and underlined the importance of users' expectations for the acceptance of latencies. Expectancies are influenced by three factors [30, 31]: (1) prior experience, (2) an individual's tolerance for and adaptability to delays, and (3) task complexity. First, prior experience with a certain kind of task shapes a user's expectations regarding the same or similar tasks in the future (e.g., if a user learns that the delay between a search query in Google and the display of results is 300 ms, s/he will expect future search processes to take the same amount of time). Second, several person variables (e.g., age, professional experience, mood) determine a user's willingness to wait. Moreover, people can adapt to long SRTs (e.g., by fulfilling other tasks while waiting). Third, with increasing task complexity, users are willing to accept longer SRTs. An experiment investigating simple, repetitive control tasks [15], which Shneiderman [30] referred to, showed SRTs below 1 s (i.e., 160 ms, 720 ms) to be superior for user performance (in contrast to 1149 ms). Regarding more complex problem solving tasks, the picture is less clear: While users had a more favorable attitude towards a low-latency system (330 ms), they made fewer errors with a longer latency (1250 ms; [33]). Furthermore, the higher the complexity, the higher users' adaptation to the latency [30]. In sum, for simple and repetitive tasks, users have a higher satisfaction and better performance if SRTs are short. In contrast, users can adapt to longer SRTs in complex tasks, but their satisfaction decreases with increasing SRT [30, 31]. Based on these empirical results,

Shneiderman and Plaisant [31] defined task-centered latency guidelines regarding user acceptance for tasks with different complexity levels. According to the authors, the most basic, repetitive tasks (e.g., single keystrokes and mouse clicks) require SRTs from 50-150 ms to keep the user satisfied. However, the theoretical basis for the lower boundary of 50 ms remains unclear. Moreover, it is not explicitly stated for which kind of tasks latencies below 100 ms are required, thus, it can only be assumed that users with high prior task experience and a low tolerance for delays prefer very small latencies in simple tasks (i.e., below 100 ms). Yet, as Dabrowski and Munson [9] point out, a definition of task complexity is missing in Shneiderman's classification, thus, it remains unclear what exactly makes a task complex.

Choosing a different approach, Card, Robertson, and Mackinlay [7] referred to psychophysical experiments investigating human perception thresholds (e.g., regarding apparent motion; [6]) and applied those results to HCI. According to the authors, for creating the illusion that a system runs instantaneously, a maximum SRT of 100 ms has to be applied, otherwise the user will notice the delay (e.g., distinct lights on a graphical user interface instead of a single light in motion; [6])[1]. This 100 ms threshold of perceptual processing was later made popular by Nielsen ([24]; see also [29]). Together with the early work by Miller [23], the work of Card et al. [6, 7] made the 100 ms threshold a frequently cited design rule implying that longer SRTs are not acceptable to the user [27].

However, in the 100 ms rule of thumb empirical data regarding perceptual thresholds [6] and subjective estimates regarding user acceptance [23] are somehow entangled. In guidelines based on empirical data regarding user latency acceptance also latencies below 100 ms are mentioned, at least for the most basic computer tasks [31]. Nevertheless, as we will see in the next section, 100 ms remained the lower bottom SRT guideline even in modern design guidelines, implying that SRTs below this threshold should not affect users markedly.

3 Recent Latency Guidelines

In his book on time perception in HCI, Seow [29] emphasized the importance of user expectations for establishing latency guidelines. He stated, similar to Shneiderman [30], that latency acceptance is relative to users' expectations and the nature of the task (i.e., longer latencies are acceptable for tasks with higher complexity as these are expected to require more computing capacity, and therefore, more time). In contrast to Shneiderman [31], he did not derive guidelines for different levels of task complexity but for different user expectations (i.e., instead of task-centered, his guidelines are user-centered with a stronger focus on the interaction). According to Seow [29], users have certain expectations regarding the responsiveness of the system if a certain task is conducted. For instance, tasks that mimic events in the physical world with instantaneous responses (e.g., pressing a virtual button which mimics

[1] It has to be emphasized that Card et al. referred to classic experiments investigating apparent motions. In these, influences of different framerates – and not input latency – on human perception were investigated.

pressing a physical button) should also show instantaneous responses (e.g., an audible click). For this very basic kind of task, the user expects the system to respond instantaneous, which means that a maximum SRT of 100 ms is required for very simple feedback (e.g., audible click after a virtual button press), respectively 200 ms for slightly more complex feedback (e.g., visual drop down menu). The next category, labelled "immediate", concerns situations in which the user expects the system to respond by performing an action initiated by the user (e.g., the display of a letter after a keystroke) and requires a maximum SRT of 500–1000 ms [29]. It remains unclear on what data these latency thresholds are grounded on as no empirical data are presented.

Different from these universal guidelines, some guidelines for single use cases have been developed. Tolia, Andersen, and Satyanarayanan [32] defined latency guidelines for thin clients (i.e., lightweight computers using remote access to a server to run applications). In this case, besides the latency within the application, the end-to-end communication from user to server and back produces additional latency. This is a particular challenge for system engineers, because users are nowadays used to systems without perceivable delay [32]. Based on prior empirical work and latency guidelines [23, 31], the authors concluded that user performance is not negatively influenced by SRTs below 150 ms. Therefore, in order to perceive the thin client's system output as immediate, the SRT (here: end-to-end latency meaning the time it takes from user input to server and back until the display of system output) must not exceed 150 ms, otherwise, the delay will get noticeable (>150 ms) and, finally, the interaction becomes annoying (>1000 ms). Thus, this guideline contains recommendations both for latency perception and user experience.

In contrast, Kaaresoja, Brewster, and Lantz [19] made a clear distinction between perception and user experience by empirically investigating both variables independently and deriving latency guidelines for another specific use case: touchscreen button presses. By experimentally manipulating the latency between the first finger touch and system feedback as well as feedback modality (visual, audio, tactile), the authors calculated the point of subjective simultaneity (PSS) for each feedback modality and, in addition, assessed users' perceived quality of the touchscreen button. Combinations of the three different feedback modalities and nine different latency conditions (ranging from 0 to 300 ms, in addition to the baseline system latency) were presented. Users had to state if the feedback appeared simultaneously with their touch and, in a later but similar phase, how s/he would rate the quality of the button (from 1 = low quality to 7 = high quality). It was reported that the PSS for visual feedback was 32 ms, for audio feedback 19 ms and for tactile feedback 5 ms. Thus, the participants were able to perceive very small latencies, especially for tactile feedback. Significant drops in the perceived quality scores were found at 100–150 ms for visual, and 70–100 ms for audio as well as for tactile feedback. Moreover, buttons with any feedback with a 300 ms latency were rated significantly lower than the buttons with any feedback with latencies ranging from 0 to 150 ms. According to the guidelines by Kaaresoja et al. [19], latencies for visual feedback should lie between 30–85 ms, for audio feedback between 20–70 ms and for tactile feedback between 5–50 ms. Hence, their guidelines were the first to explicitly incorporate latencies smaller than 50–100 ms, if only for a very specific use case.

Using a similar experimental approach, Kaaresoja [20] expanded his guidelines for bimodal feedback (i.e., visual-audio, visual-tactile and tactile-audio). It was found that for different feedback pairs different levels of symmetry between the two feedback modality latencies emerge, as follows. For the combination of visual and audio feedback, the visual feedback latency should not be greater than 90 ms while the audio feedback should not exceed 70 ms. For the combination of visual and tactile feedback, the visual feedback latency should not be greater than 100 ms while the tactile feedback should not exceed 55 ms. And lastly, for the combination of tactile and audio feedback, the tactile feedback latency should not be greater than 25 ms while the audio feedback should not exceed 100 ms.[2] The empirical results [19, 20] suggest a high sensitivity for delay of tactile feedback in tactile HCI. This finding is in line with the suggestion that interactions which mimic events in the physical world (e.g., tactile feedback after virtual button touch) require very small latencies to be perceived as instantaneous [28].

In their review, Doherty and Sorenson [11] updated and expanded the existing general latency guidelines [29, 30] with a special focus on the flow experience [8]. The authors argue that in the usage of today's frequently used interactive systems (e.g., smartphones, tablets) short interactions (e.g., menu navigation, scrolling) are predominant. As it has been pointed out before, small latencies will get noticed or even annoy the user especially in very short and basic interactions (e.g., [19, 23, 29, 31, 32]). One negative influence of perceived latency is that users' interaction with the system can be interrupted, thus, users' flow gets broken ([11]; see also [29]). Incorporating empirical results on user expectations, perceived task complexity and perceptual limits, Doherty and Sorenson's guidelines [11] represent the most elaborate latency guidelines for an optimal user experience so far. However, the authors raised the lower boundary latency threshold for instantaneous responses to 300 ms. This figure was incorporated because of Kaaresoja's [20] finding that the perceived quality of the touchscreen button was significantly lower with 300 ms latency in contrast to 0–150 ms. Thus, "[…] depending on the input modality (mouse, keyboard, touchscreen, air, gesture, speech, etc.), the perception of what a user would consider instantaneous will vary." ([11], p. 4390). While the lower limit of 300 ms gives the guideline a higher generalizability, it also decreases its accuracy for very short interactions.

It becomes apparent that latencies below 100 ms do not play a role in most design guidelines. The only general guideline that explicitly mentioned a latency threshold smaller than 100 ms was the one by Shneiderman and Plaisant [31], but it was not explicitly stated under which conditions (e.g., task demands, user status) a latency has to be as small as 50 ms to be acceptable. The only other guideline recommending maximum latencies below 100 ms is the one by Kaaresoja et al. [20], suggesting that in very basic interactions (i.e., control tasks; [9]) – the ones that Miller [23] called "control activations" and Seow [29] expected to be "instantaneous" – user experience gets significantly impaired by latencies below 100 ms. Still, following the majority of guidelines, zero-latency systems do not seem necessary for optimal user experience. But is this really the case?

[2] Note that the dependent variable was a PSS judgment. Thus, based on these results, users will notice the delay if one of the feedback modalities exceeds the latency thresholds. If and when the participants perceived an asynchronicity between the two feedbacks was not assessed.

4 Empirical Evidence for the Perception of Latencies Below 100 ms and their Impact on HCI

Within recent years, there has been a considerable growth of studies examining latency effects in HCI even below the 100 ms threshold, possibly also because of increasing technical potentialities (e.g., high-speed cameras). In several studies, system latency was experimentally varied and perceptual limits were tested by applying classic psychophysical methods (i.e., estimating the just noticeable difference regarding perceived latency between two identical tasks with different latencies). These studies, which are presented in the following, indicate that users are indeed able to perceive latencies well below 100 ms. In addition, other studies show that even such small latencies can have negative effects on user performance – even when the latencies are below the perceptional threshold. Moreover, influencing factors on the perception of latencies are investigated, implying that latency perception is dependent on user and task variables.

During a digital inking task using a stylus [2], users were able to perceive latencies between input (i.e., the touch of the stylus on the screen) and visual feedback (i.e., the appearance of the digital ink) down to 50 ms with slightly higher perception thresholds for tasks that require more attentional resources (i.e., cause a greater workload). In a direct dragging task on a touchscreen, users were even able to notice latencies down to 11 ms [10], 6 ms [27] and even down to 2 ms under specific circumstances [25]. And even in a direct tapping task on a touchscreen (i.e., button press) where relatively few data are available to make latencies salient, a perceptional threshold of 64 ms was found [18].

So far, these results all refer to zero-order tasks. Zero order is one type of control order, that is, the way that the system responds to a change of the position of the control [34]. In zero-order tasks, a change in the position of the control (e.g., the mouse on the mousepad) leads to a change in the position of the displayed system output (e.g., the cursor on the screen; [17, 34]). In contrast, first-order control tasks require velocity control [34]. Here, a change in the control position leads to a constant change of velocity (e.g., a button press on a DVD remote control to raise up the playback speed to 2x). Finally, second-order control tasks deal with a change of acceleration (i.e., changes in the rate of velocity) and require more cognitive resources than zero- and first-order control tasks. One example in the field of vehicle control is the relationship between steering wheel position and the vehicle's lateral position in the lane. Here, a constant change in the steering wheel position leads to an increasing rate of change in the lateral position [34]. In second-order tasks, when the input is set to zero, the output continues to change and is not instantly set to zero as it is the case in zero- and first-order tasks [17]. Such a more demanding, second-order task was applied in an own study [22]. Using a virtual balance task, it has been shown that performance was already impaired by an added latency of 49 ms (technical base latency: 10.8 ms). However, participants perceived only the added latency from 97 ms on. Hence, even though users were not able to perceive the latency, it had an effect on their performance.

The effect of latency on user performance was also examined more closely in recent years. For instance, Brady et al. [4] applied an indirect mouse movement task and found that an added latency of 33 ms significantly impaired user performance.

In a pointing task, latency began to affect performance at 16 ms [14]. In a 3D game environment, a latency of 41 ms impaired user performance in an aiming task [16].

5 Factors Affecting Latency Perception

The studies presented so far were concerned with identifying latency thresholds for perception and performance. Other studies examined effects of influencing factors on latency perception, suggesting that latency thresholds are not cast in stone, yet, are system-, task- and person-dependent. Hence, a key task from an engineering psychology perspective is to structure relevant variables affecting latency perception. In the following, empirical results as well as assumptions regarding (1) system characteristics, (2), task characteristics, and (3) person characteristics will be discussed.

First, concerning system characteristics, different *input modalities* can be distinguished. When comparing direct (e.g., via touchscreen) and indirect input (e.g., with conventional input devices such as a mouse), sensitivity to latencies is higher in direct interaction [10, 26]. This can likely be ascribed to a higher salience of the latency because the visual attention is located within the same place as the system input. Another factor is the *output modality*: The latency perception thresholds differ with respect to the modality of the feedback after a virtual button press. Users are extremely sensitive to a latency in tactile feedback (when compared to audio and visual feedback) when the input is also tactile [19, 20]. According to Seow [29], a tactile feedback after a virtual button press is very similar to the press of a real physical button, therefore the user expects an instantaneous response and might be more sensitive to interaction delays. Moreover, the *number of feedbacks* seems to play a role in latency perception. When two feedbacks are provided in contrast to just one, latency sensitivity is lower [18, 20, 28]. One explanation for this effect might be an additional information-processing step which is needed to integrate the two feedbacks [25], however, this remains speculative at the present time. In visual dragging tasks, the *size ratio between physical reference and visual feedback* affects latency perception. If the size of the physical reference (e.g., a stylus nib) and the visual feedback are more similar, latency perception is improved [25]. Possibly this can also be attributed to the higher similarity to an interaction in the physical world [29].

Second, regarding task characteristics, an important factor that has already been incorporated in guidelines is *task complexity*. By experimentally varying task complexity, two studies found that users perceive smaller latencies in simple tasks (i.e., dragging tasks) compared to slightly more complex, thus, demanding tasks (i.e., scribbling tasks; [2, 25]. Moreover, *interaction speed* affects latency perception in dragging tasks. The faster the user's hand motion in a dragging task, the better the latency perception [26]. This finding is attributed to the visual effect of a fast hand motion in a dragging task which creates the illusion that the displayed square is "attached to a rubber band to the user's finger" ([26], p. 453). This effect makes latency visible and salient to the user. The latency perception model [1], which describes the process of latency perception, postulates that the user utilizes a referent to make latency judgments. More specifically, a referent is a stimulus within the interaction (e.g., a stylus nib, the user's finger) that

the user compares to the system response to evaluate the latency magnitude [1]. One example is the user's hand in a dragging task as described before [26]. According to [1], the *presence of a referent* affects latency perception. If the hand is made invisible and can therefore not be used as a referent, latency sensitivity is diminished in a scribbling task [2]. Further, the *modality of the referent* is discussed as a factor influencing latency perception [1].

Finally, regarding person characteristics, domain specific *experience* seems to be an important factor for the perception of latencies. The experience with highly dynamic computer games (i.e., action games, racing games, first person shooter games) was found to correlate positively with latency perception in a dragging task [13]. Experience with a specific musical instrument might also affect the perception of audio latencies when playing it [21]. Moreover, *age* has been suggested as a factor affecting latency perception, with younger users perceiving smaller latencies than older users [21]. Closely connected to task complexity is *cognitive load*. The higher the task demands (e.g., because of higher task complexity, secondary tasks or environmental variables), the higher the user's cognitive load. This factor has been discussed with respect to latency perception in several studies [2, 19, 25].

6 Conclusion and Implications

To conclude, while several design guidelines recommend a maximum latency of 100 ms for an optimal user experience in basic interactions, empirical results suggest that latency thresholds for different tasks lay substantially lower. Users are indeed able to perceive latencies down to single milliseconds in specific tasks. Moreover, performance in zero-order and more demanding second-order tasks already gets impaired by latencies between 16–60 ms. Therefore, the lower boundary of 100 ms as mentioned in several design guidelines appears outdated. Especially interactions that are very similar to physical interactions require substantially smaller maximum acceptable latencies. Furthermore, several factors affect latency perception and consequently user performance and tolerance. Hence, a need for updated, evidence-based latency guidelines incorporating system-, task-, and person characteristics emerges.

The literature review revealed further implications. First, the majority of tasks that were utilized in empirical investigations on latency perception were zero-order tasks. However, latency can also impair user performance and experience in first- and second-order tasks. Especially in the emerging field of human-robot-interactions, virtual environments and remote-controlled systems, influences of latency should be further investigated in more complex tasks. Second, the study of factors affecting latency perception, user performance, and user experience needs to be intensified. Besides replicating previous studies and examining several variables more deeply (e.g., domain-specific experience, learning effects, attentional focus, motivational aspects), this also involves assessing age-diverse samples with varying usage experience of the utilized devices and highly dynamic computer games. Moreover, with technical progress aiming at increasingly reducing latencies, users likely get accustomed to hardly perceivable delays. This could lead to a higher sensitivity for very short latencies in users with much experience

with such modern systems and is probably one factor why guidelines from the 20th century are not applicable anymore.

Updated latency guidelines that give specific recommendations for different user groups and use cases will constitute a fruitful information source for interaction designers and system engineers and will enable a more precise and differentiated evaluation of the question: Is zero-latency really necessary?

References

1. Annett, M.: The fundamental issues of pen-based interaction with tablet devices. Dissertation, University of Alberta (2014)
2. Annett, M., Ng, A., Dietz, P., Bischof, W., Gupta, A.: How low should we go? Understanding the perception of latency while inking. In: Graphics Interface Conference 2014, 7–9 May, Montreal, Canada, pp. 167–174. Canadian Human-Computer Communications Society (2014)
3. Boucsein, W.: Forty years of research on system response times – what did we learn from it? In: Schlick, C.M. (ed.) Industrial Engineering and Ergonomics, pp. 575–593. Springer, Heidelberg (2009). doi:10.1007/978-3-642-01293-8_42
4. Brady, K., Wu, B., Sim, S.H., Enquobahrie, A., Ortiz, R., Arikatla, S.: Modeling reduced user experience caused by visual latency. In: Soares, M., Falcão, C., Ahram, T.Z. (eds.) Advances in Ergonomics Modeling, Usability & Special Populations. Advances in Intelligent Systems and Computing, pp. 267–277. Springer, Cham (2017). doi:10.1007/978-3-319-41685-4_24
5. Butler, T.W.: Computer response time and user performance. In: Janda, A. (ed.) CHI 1983, Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 58–62. ACM, New York (1983). doi:10.1145/800045.801581
6. Card, S.K., Moran, T.P., Newell, A.: The Psychology of Human-Computer Interaction. Lawrence Erlbaum Associates, Hillsdale (1983)
7. Card, S.K., Robertson, G.G., Mackinlay, J.D.: The information visualizer, an information workspace. In: CHI 1991, Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 181–186. ACM, New York (1991). doi:10.1145/108844.108874
8. Csikszentmihalyi, M.: Flow and the Foundations of Positive Psychology. Springer, Dordrecht (2014). doi:10.1007/978-94-017-9088-8
9. Dabrowski, J., Munson, E.V.: 40 Years of searching for the best computer system response time. Interact. Comput. **23**, 555–564 (2011). doi:10.1016/j.intcom.2011.05.008
10. Deber, J., Jota, R., Forlines, C., Wigdor, D.: How much faster is fast enough? User perception of latency & latency improvements in direct and indirect touch. In: CHI 2015, April 18–23, 2015, Seoul, Republic of Korea, pp. 1827–1836. ACM, New York (2015). doi:10.1145/2702123.2702300
11. Doherty, R.A., Sorenson, P.: Keeping users in the flow: mapping system responsiveness with user experience. Proc Man **3**, 4384–4391 (2015). doi:10.1016/j.promfg.2015.07.436
12. Fischer, A.R.H., Blommaert, F.J.J., Midden, C.J.H.: Monitoring and evaluation of time delay. Int. J. Hum.-Comput. Int. **19**, 163–180 (2005). doi:10.1207/s15327590ijhc1902_1
13. Forch, V., Franke, T., Rauh, N., Krems, J.F.: Are 100 milliseconds fast enough? Characterizing latency perception thresholds in mouse-based interaction. Paper presented at the 19th International Conference on Human-Computer Interaction, Vancouver, Canada, 9–14 July 2017 (2017)
14. Friston, S., Karlström, P., Steed, A.: The effects of low latency on pointing and steering tasks. IEEE Trans. Vis. Comput. Graph. **22**, 1605–1615 (2015). doi:10.1109/TVCG.2015.2446467

15. Goodman, T.J., Spence, R.: The effect of computer system response time on interactive computer aided problem solving. ACM SIGGRAPH Comput. Graph. **12**, 100–104 (1978). doi:10.1145/965139.807378
16. Ivkovic, Z., Stavness, I., Gutwin, C., Sutcliffe, S.: Quantifying and mitigating the negative effects of local latencies on aiming in 3D shooter games. In: CHI 2015, April 18–23, 2015, Seoul, Republic of Korea, pp. 135–144. ACM, New York (2015). doi: 10.1145/2702123.2702432
17. Jagacinski, R.J., Flach, J.M.: Control Theory for Humans. Lawrence Erlbaum, Mahwah (2003)
18. 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 CHI 2013 Conference on Human Factors in Computing, April 27–May 2, 2013, Paris, France, pp. 2291–2300. ACM, New York (2013). doi:10.1145/2470654.2481317
19. Kaaresoja, T., Brewster, S., Lantz, V.: Towards the temporally perfect virtual button: touch-feedback simultaneity and perceived quality in mobile touchscreen press interactions. ACM Trans. Appl. Percept. **11**, 9:1–9:25 (2014). doi:10.1145/2611387
20. Kaaresoja, T.: Latency guidelines for touchscreen virtual button feedback. Dissertation, University of Glasgow (2016)
21. Mäki-Patola, T., Hämäläinen, P.: Latency tolerance for gesture controlled continuous sound instrument without tactile feedback. In: International Computer Music Conference Proceedings, 2004 (2004)
22. Martens, J., Franke, T., Rauh, N., Krems, J.F.: Effects of low-range latency on performance and perception in a virtual, unstable second-order control task (2016). Manuscript submitted for publication
23. Miller, R.B.: Response time in man-computer conversational transactions. In: Proceedings of the AFIPS 1968, December 9–11, 1968, pp. 267–277. ACM, New York (1968). doi: 10.1145/1476589.1476628
24. Nielsen, J.: Usability Engineering. Academic Press, San Diego (2003)
25. Ng, A., Annett, M., Dietz, P., Gupta, A., Bischof, W.F.: In the blink of an eye: investigating latency perception during stylus interaction. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 1103–1112. ACM, New York (2014). doi: 10.1145/2556288.2557037
26. Ng, A., Dietz, P.H.: The effects of latency and motion blur on touch screen user experience. J. SID **22**, 449–456 (2015). doi:10.1002/jsid.243
27. Ng, A., Lepinski, J., Wigdor, D., Sanders, S., Dietz, P.: Designing for low-latency direct-touch input. In: UIST 2012, October 7–10, 2012, Cambridge, Massachusetts, USA, pp. 453–464 (2012). doi:10.1145/2380116.2380174
28. Nordahl, R.: Self-induced footsteps sounds in virtual reality: latency, recognition, quality and presence. In: The 8th Annual International Workshop on Presence, PRESENCE 2005, Conference Proceedings, 21–23 September 2005, London, United Kingdom, pp. 353–355 (2005)
29. Seow, S.C.: Designing and Engineering Time: The Psychology of Time Perception in Software. Addison-Wesley Professional, Indianapolis (2008)
30. Shneiderman, B.: Response time and display rate in human performance with computers. Comput. Surv. **16**, 265–285 (1984). doi:10.1145/2514.2517
31. Shneiderman, B., Plaisant, C.: Designing the User Interface: Strategies for Effective Human-Computer Interaction. Addison-Wesley Publ. Co., Reading (1987)
32. Tolia, N., Andersen, D.G., Satyanarayanan, M.: Quantifying interactive user experience on thin clients. Computer **39**(3), 46–52 (2006). doi:10.1109/MC.2006.101

33. Weinberg, S.: Learning effectiveness: the impact of response time. ACM SIGSOC Bull. **13**, 140 (1981). doi:10.1145/1015579.810983
34. Wickens, C.D., Hollands, J.G., Banbury, S., Parasuraman, R.: Engineering Psychology and Human Performance. Routledge, Oxford (2013)
35. Zhou, R., Shao, S., Li, W., Zhou, L.: How to define the user's tolerance of response time in using mobile applications. In: 2016 IEEE International Conference on Industrial Engineering and Engineering Management, pp. 281–285. (2016). doi:10.1109/IEEM.2016.7797881