Improving the Human–Computer Dialogue With Increased Temporal Predictability

Florian Weber and **Carola Haering**, University of Wuerzburg, Wuerzburg, Germany, and **Roland Thomaschke**, Universität Regensburg, Regensburg, Germany

Objective: An experiment was conducted to investigate the impacts of length and variability of system response time (SRT) on user behavior and user experience (UX) in sequential computing tasks.

Background: Length is widely considered to be the most important aspect of SRTs in human–computer interaction. Research on temporal attention shows that humans adjust to temporal structures and that performance substantially improves with temporal predictability.

Method: Participants performed a sequential task with simulated office software. Duration and variability, that is, the number of different SRTs, was manipulated. Lower variability came at the expense of on average higher durations. User response times, task execution times, and failure rates were measured to assess user performance. UX was measured with a questionnaire.

Results: A reduction in variability improved user performance significantly. Whereas task load and failure rates remained constant, responses were significantly faster. Although a reduction in variability came along with, on average, increased SRTs, no difference in UX was found.

Conclusion: Considering SRT variability when designing software can yield considerable performance benefits for the users. Although reduced variability comes at the expense of overall longer SRTs, the interface is not subjectively evaluated to be less satisfactory or demanding. Time design should aim not only at reducing average SRT length but also at finding the optimum balance of length and variability.

Application: Our findings can easily be applied in any user interface for sequential tasks. User performance can be improved without loss of satisfaction by selectively prolonging particular SRTs to reduce variability.

Keywords: system response times, temporal variability, temporal attention, subjective task load, waiting times, human–computer interaction, user behavior, user experience

Address correspondence to Carola Haering, University of Wuerzburg, Department of Psychology, Roentgenring 11, 97070 Wuerzburg, Germany; e-mail: haering@psychologie.uni-wuerzburg.de.

HUMAN FACTORS
Vol. 55, No. 5, October 2013, pp. 881-892
DOI:10.1177/0018720813475812

INTRODUCTION

In a seminal study, Kubovy and Pomerantz (1981) stated that one dimension of human perception has largely been neglected, even though it is absolutely essential to human perception and action: time.

In computing systems, people encounter time in the form of system response times (SRTs), which are defined as the time elapsed from entering a command until its completion (Miller, 1968). For example, when one clicks on the "Save" button or on a hyperlink, the system needs some time to process the task before the prompt window asking where to save the file appears or before the linked file or webpage is loaded. SRTs are determined by system characteristics, such as processing capacity and network bandwidth, as well as situational factors, such as the complexity of the computational processes at a given time and processor or network load. SRTs can affect user response time (URT), the time the user needs to perceive and process the computer output and enter a further command after the system has responded, by two determinants: the average length of the SRTs and the variation of SRTs.

SRT Duration

Early studies on SRTs focused on the impact of SRT length on user performance. From conditioning experiments of the 1950s, Miller (1968) derived a critical upper SRT boundary of 2 s for human performance within one task, which has been widely diffused in application (Nielsen, 1999; Shneiderman & Plaisant, 2009). Participants had to search for a blank target space between two letters, mark it, correct it, and wait for the next trial, which had two different average SRT durations (2 or 8 s) and two different variability modes (1 fixed SRT vs. 7 SRTs). Although task execution time (TET) did not depend on SRT duration, failure rate decreased with increasing SRTs, but at the same time, physiological and subjective stress levels

increased (see also Schaefer, 1990; Schleifer & Amick, 1989).

However, there is also evidence that does not support a reduction of SRT length. No performance differences were found between short and long SRTs (5 vs. 10 s) when programmers, debugging computer code, waited for the next line of code to be editable (Dannenbring, 1983). Furthermore, users have been shown to adopt different strategies on the speed–accuracy continuum depending on the pace of the interface (Teal & Rudnicky, 1992). Fast interfaces were found to increase error and stress in simple routine tasks (Kohlisch & Kuhmann, 1997) and slow response speed (Thum, Boucsein, Kuhmann, & Ray, 1995; Zijlstra, Roe, Leonora, & Krediet, 1999). In some studies (Kuhmann, Schaefer, & Boucsein, 1989; Zijlstra et al., 1999), performance advantages were even found with short disruptions versus continuous interaction without disruptions.

The aforementioned evidence can be accounted for by Seow's (2008) classification of SRTs. Seow suggests that users form expectations of the speed of response of a computing system depending on certain types of tasks. In simple tasks, such as key presses, delays of more than 100 to 200 ms will feel interruptive. An immediate response—0.5 to 1 s—is expected in tasks such as a mouse click to view the next page on the screen. More complex interactions will be perceived as continuous in time ranges from 2 to 5 s. At SRTs of 7 to 10 s, the user will begin to give up on the task if no feedback occurs. Likewise, more and more evidence suggests that the time a user is willing to wait for a task largely depends on a multitude of factors, such as the complexity of the task (Caldwell & Wang, 2009; Dabrowski & Munson, 2011); environmental factors, such as time pressure (Caldwell & Wang, 2009); and the expertise of the user (Caldwell, 2008).

In sum, extended SRTs are, in general, assumed to negatively affect speed as well as the accuracy of task execution and user satisfaction because the man–machine interaction process is interrupted. However, as computer systems and networks increase in speed, applications and the number of processes become more resource demanding, and network resources are not evenly available to different users, the question remains of how to cope with inevitable delays (Dabrowski & Munson, 2011). Changing the variability of SRTs may be a way to do so.

Variability of SRTs

Although the SRT duration has been extensively investigated, the variability of SRTs for the same task has not gained much attention. Accordingly, common models, such as Seow's (2008) and the cost-benefit model of information and communications technology interaction (Caldwell, 2008), do not take SRT variability into account. SRT variability stems from many situational influences on SRTs, such as network congestion, or the number and nature of concurrently running system operations in the background, such as automatic background saving, system updates, or virus scanners, which are prevalent in today's multitasking enabled systems (Cota-Robles & Held, 1999; Flautner, Uhlig, Reinhardt, & Mudge, 2000; Yates, Kurose, Towsley, & Hluchyj, 1993).

Classic research on delays before a target stimulus generally showed faster user responses when the target always appears after a given delay than when it appears after two varying delays (see Niemi & Näätänen, 1981, for review). Generally, humans have been shown to be able to use temporal regularities (Correa, Lupiáñez, Milliken, & Tudela, 2004; Coull & Nobre, 1998; Haering & Kiesel, 2012; Kingstone, 1992; Thomaschke & Dreisbach, 2013; Thomaschke, Kiesel, & Hoffmann, 2011; Thomaschke, Wagener, Kiesel, & Hoffmann, 2011a, 2011b). That is, when one repeatedly experiences the same temporal structures, one adapts to these structures so that one orients attention to specific points in time and can thus respond more quickly when events occur at the expected time.

In human–computer interaction (HCI) research, some authors point out that temporal variability may impair user performance and increase users' stress level. In this context, variability means unpredictability, and it leads to temporal uncertainty and stress (Hui & Tse, 1996; Osuna, 1985). Accordingly, Kuhmann, Boucsein, Schaefer, and Alexander (1987) hypothesized that variable SRTs should

influence physiological stress measures and user performance. However, temporal variability neither increased physiological stress markers nor performance measures. Another study used a continuous computer game–like task consisting of several steps that were either contiguous or separated by variable SRTs. When SRTs were variable, URTs were increased and those blocks were enjoyed less (Szameitat, Rummel, Szameitat, & Sterr, 2009).

In sum, even if actual SRTs are commonly subjected to tremendous variability, far less research has been devoted to the impact of SRT variability on users than to the impact of length. Findings on the benefits of decreased variability cast doubt on the main focus on shortening SRTs to improve HCI. Users have been found to accept longer SRTs when they are predictable and accepted as reasonable (Caldwell & Wang, 2009). Focusing on variability as a means to heighten predictability may be a fruitful approach.

There is, nevertheless, a technical trade-off between variability and mean SRT duration. Minimal SRT durations can be achieved when a system responds as fast as possible. However, computers are not always maximally responsive, which is why minimally possible average SRTs come at the expense of variability. As shortening SRTs is beyond the scope of an interaction designer, reducing variability as a means to heighten predictability can be achieved only by artificially prolonging short SRTs on a given computer system so that they are the same length as longer SRTs. Doing so, however, naturally results in a trade-off between lower variability on one hand and longer average SRTs on the other.

Aim of This Study

To the best of our knowledge, no research has been done to assess how the trade-off between variability and duration of SRTs influences user behavior and user experience. As humans build temporal expectations and, by doing so, exploit temporal structures, users should also be able to generate expectancies concerning the speed of a computer program and the time they usually have to wait after certain queries. In this article, we examine whether designing for low variability of SRTs at the cost of slower average SRTs is worth considering.

In the present study, we manipulated the temporal regularity of an e-mail client's SRTs. Performance, subjective task load, and likeability of interaction were compared for two SRT distributions. The first is an approximately continuous Poisson distribution, which is the distribution predicted by queuing theory (Kleinrock, 1975) and actually measured SRTs on modern computers best (Cota-Robles & Held, 1999; Flautner et al., 2000; Yates et al., 1993). For the second distribution, all SRTs from the first distribution were accumulated to two SRTs, a medium one and the longest one. We expected that a reduction of variability would decrease URTs and TETs, since participants should be able to implicitly adapt to the given temporal structures. With regard to likeability of the e-mail system, two outcomes are possible. A reduction of variability could either improve likeability or decrease likeability, as the low variability comes with a greater average SRT duration.

METHOD

Participants

A total of 22 participants (1st-year psychology students or recruited from a participant database) took part for €12 or course credit. Exclusion criteria were either a failure rate above 15% or an average URT 2.5 standard deviations above the average of all other participants. Because of these criteria, 2 of 22 participants had to be excluded from data analysis. The average age of the remaining 20 participants (13 female, 19 right-handed) was 22.12 years (SD = 2.59).

Task

The participants' task was to assume the position of a management assistant, responsible for their manager's e-mails. The task was accomplished with the right hand and a mouse. At the beginning of the trial, the e-mail client's in-box displayed two e-mails (see Figure 1). Participants had to check whether the top e-mail was relevant or spam (veridically labeled in the rightmost column as *spam* or *not spam*) and to forward

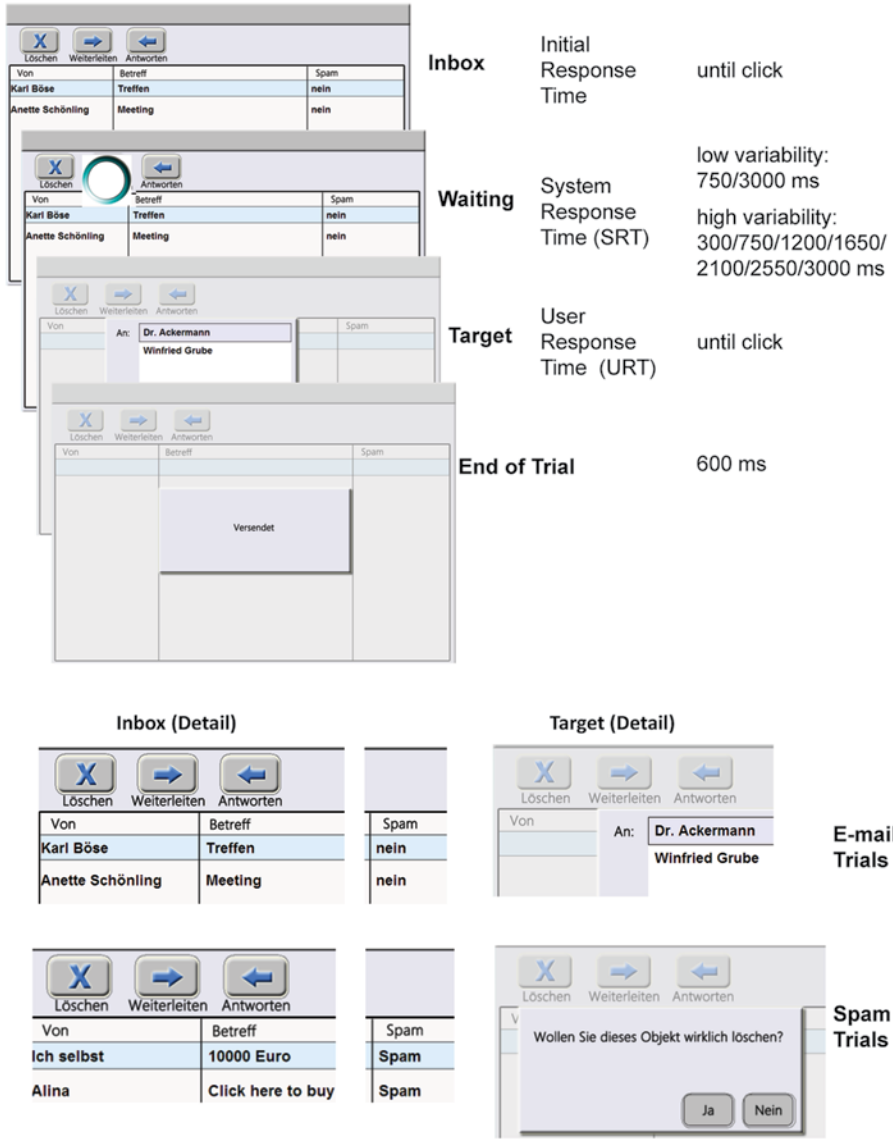

Figure 1. Schematic sketch of a trial. In the in-box, participants have to forward (click "*Weiterleiten*") the top e-mail if it is not spam or delete (click "*Löschen*") the e-mail if it is spam. After the system response time, the target appears. In e-mail trials, participants have to click their boss's name (Dr. Ackermann), and in spam trials, they have to confirm deleting (click "*Ja*"). Spam trials are depicted only in detail pictures here (lower half).

relevant e-mails to their manager and delete spam. To initiate the chosen action, participants had to click the "*Löschen*" (delete) or "*Weiterleiten*" (forward) button. After clicking, participants had to wait until they could select the correct recipient or confirm deleting. An animated activity indicator indicated the SRT between the click and the target screen. After the SRT, a pop-up window for selecting the recipient (always the top name) appeared for the *forward* response. For the *delete* response, a pop-up asked the participants to confirm their selection.

After the user response, a pop-up confirmed the sending or deletion for 600 ms before the next trial started.

Spam e-mails and relevant ones were distributed 20% and 80%. Spam trials were included to ensure that participants had to discriminate the e-mails in the in-box and would not just blindly click as soon as any stimulus appeared. To prevent participants from constantly clicking until target onset, more than two clicks before target onset triggered the error message "*Input Fehler—zu viele Mausklicks. Bitte warten Sie, bis das System neu startet*" ("Input error—too many mouse clicks. Please wait until the system reboots") for 10 s.

Apparatus and Stimuli

The experimental setting was created with Adobe Photoshop (Adobe Systems, Mountain View, CA) for the visual design and E-Prime2 (Schneider, Eschman, & Zuccolotto, 2002) for adding interactivity and collecting data. Data were collected on a Windows PC with 17-in. CRT display (screen resolution 1,024 × 768 pixels). Participants' responses were recorded with the use of a standard optical mouse.

SRTs were chosen with the assumption that SRTs follow a Poisson distribution (Kleinrock, 1975). SRT range from 300 to 3,000 ms was chosen on the basis of the nature of this task. According to Seow (2008), this study's task would be categorized as "immediate." Therefore, time intervals of 0.5 to 1 s would be most appropriate, whereas SRTs of 2 to 5 s would still be perceived by the users as continuous. Additionally, user satisfaction has been found to be highest with SRTs between 2 and 4 s (Galletta, Henry, McCoy, & Polak, 2004; Nah, 2004). In the high-variability condition (seven SRTs), which represented modern computing systems, SRTs followed approximately a continuous Poisson distribution with $\lambda = 1/1{,}500$ ms. (The expression *variability* in this article refers to the number of different SRTs that are presented to the participant. It should not be confounded with the statistical term *variance*.) In this distribution, more than 75% of all trials still remain in the "tolerable" category of below 2 s (Nah, 2004) and nearly 50% were even below 1 s.

TABLE 1: Number of Trials With Corresponding System Response Time (SRT) in Low-Variability and High-Variability Conditions

	Number of Trials	
SRT (ms)	Low Variability	High Variability
300		10
750	23	13
1,200		8
1,650		7
2,100		5
2,550		4
3,000	27	3

For reasons of experimental practicability, the continuous distribution was divided into seven categories with 450-ms distance (Table 1). The number of trials for each SRT was chosen on the basis of the probability of the given SRT in the Poisson distribution. The SRTs in the low-variability (two SRTs) condition were chosen so that all other SRTs could be extended to one of them and the occurrence of both SRTs was evenly probable. In this condition, approximately 50% of all trials were outside of this tolerable range. We thus compare user performance in an on-average faster condition with more variability (seven SRTs) with performance in a condition with only two on-average longer SRTs.

Questionnaires

A German translation of the NASA–Task Load Index (NASA-TLX) questionnaire (see Table 2; Hart & Staveland, 1988; Pfendler, 1990) to measure subjective task load and the AttrakDiff questionnaire to measure likeability (see Hassenzahl, Burmester, & Koller, 2003; Table 3) were administered after each session. No other questionnaires were administered.

Procedure

Participants attended two sessions on different days within 1 week. Each session differed in the variability of SRTs. The order of the variability conditions was counterbalanced across participants.

TABLE 2: NASA–Task Load Index Questionnaire Scale Definitions

Scale	End Points	Descriptions
Mental Demand	Low–high	How much mental and perceptual activity was required (e.g., thinking, deciding, calculating, remembering, looking, searching, etc.)? Was the task easy or demanding, simple or complex, exacting or forgiving?
Physical Demand	Low–high	How much physical activity was required (e.g., pushing, pulling, turning, controlling, activating, etc.)? Was the task easy or demanding, slow or brisk, slack or strenuous, restful or laborious?
Temporal Demand	Low–high	How much time pressure did you feel due to the rate or pace at which the tasks or task elements occurred? Was the pace slow and leisurely or rapid and frantic?
Performance	Good–poor	How successful do you think you were in accomplishing the goals of the task set by the experimenter (or yourself)? How satisfied were you with your performance in accomplishing these goals?
Effort	Low–high	How hard did you have to work (mentally and physically) to accomplish your level of performance?
Frustration Level	Low–high	How insecure, discouraged, irritated, stressed, and annoyed versus secure, gratified, content, relaxed, and complacent did you feel during the task?

Note. Adapted from Hart & Staveland (1988, p. 168).

In each session, participants completed a practice block of 20 trials, with a constant SRT of 500 ms, independent of the variability condition. The practice block was followed by 13 experimental blocks of 50 trials each. Trial sequence was randomized within blocks. After each block, a feedback screen informed participants of their average speed and accuracy, encouraging them to work faster and more precisely. At the end of each session, participants completed the two questionnaires on their subjective evaluation of the system.

RESULTS

Preliminary Analysis

In learning experiments, experience with different procedures in the first session might affect performance differently in the second session (Greenwald, 1976). Such carryover effects would confound a within-subject analysis, because they cannot be accounted for by counterbalancing. Thus we performed a carryover check prior to analysis.

In particular, we compared the average target URTs after SRTs appearing in both conditions (750 ms and 3,000 ms) in the learning blocks and the last three blocks. The order of sessions interacted with the variability condition, $F(1, 18) = 53.56$, $p < .001$, $\eta_p^2 = .748$. Both groups became faster, that is, improved in the task, throughout Session 1 ($ps \leq .001$). However, no group showed a significant effect of learning in terms of speeding up throughout Session 2, neither the group starting with high variability, $t(9) = -0.82$, $p = .432$ (15 ms faster on average), nor the group starting with low variability, $t(9) = 0.81$, $p = .438$, who got even numerically slower (10 ms on average). This order effect indicates that both procedures differed in their effects in the consecutive session.

In that case, a reliable within-subject analysis is not possible and it is recommended that the second session be dropped and only the first half be analyzed as between-subjects design (Cook & Campbell, 1979). To check whether a between-subjects analysis of the first session

TABLE 3: AttrakDiff Questionnaire Scales

Scale	Description
Pragmatic Quality	Perceived ability of a product to reach goals by providing useful and usable functions
Hedonic Quality–Stimulation	Ability of a product to satisfy the need for improvement of one's skills and knowledge
Hedonic Quality–Identity	Ability of a product to communicate self-worth-improving messages to relevant others
Attractiveness	Global positive–negative rating of a product

only would be appropriate or whether it would be confounded by a priori group differences, we compared URTs after SRTs appearing in both conditions and the average time users spent on the whole task in all SRT conditions (TET) for the first experimental block. Neither a mixed-measures ANOVA for URT (Variability Group [low and high] × SRT [750 ms and 3,000 ms]), $F(1, 18) = 0.75$, $p = .398$, $\eta_p^2 = .040$, nor a between-subjects ANOVA for TET (Variability Group [low, high]), $F(1, 18) = 1.343$, $p = .262$, $\eta_p^2 = .069$, indicated initial differences between both groups. We have no reason to assume that participants differed prior to the experiment and thus conjecture that carryover effects occurred between experimental conditions.

So, due to the confounding order effect, requirements for a within-subject analysis were not met, but requirements for a between-subjects analysis were. Thus, we report only data (URTs, TETs error, and questionnaire data) of the between-subjects analysis of Session 1. For all further analyses, the practice block and the first three experimental blocks were considered learning blocks and, therefore, not included into the analyses.

Target URTs

For the analysis of URTs, we analyzed only URTs in the two SRTs that appeared in both the high- and the low-variability conditions (750 ms and 3,000 ms). This criterion was necessary because the duration before a target's appearance itself influences URT (Niemi & Näätänen, 1981), and therefore, URT differences across different SRTs could be caused by experimental manipulation as well as by the length of the SRT itself. Only e-mail trials, a priori defined as relevant

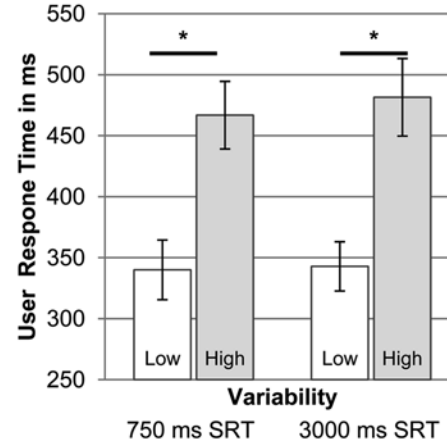

Figure 2. Target user response time with variability analyzed as between-subjects factor for system response times of 750 ms and 3,000 ms.

trials, were analyzed because of an insufficient number of spam trials, requiring another movement for statistical analysis. We excluded statistical outliers greater than 3 standard deviations above or below each participant's individual average in each condition (1.5%) from this analysis to reduce beta errors (Bush, Hess, & Wolford, 1993; Ratcliff, 1993). We also removed failure trials (3.6%). In total, 5.1% of all trials were removed.

A two-way, mixed-measures ANOVA, with the between-subjects factor variability (low, high) and the within-subject factor SRT (750 ms, 3,000 ms), showed that participants experiencing low variability responded significantly faster than did participants experiencing high variability, $F(1, 18) = 13.32$, $p = .002$, $\eta_p^2 = .43$ (see Figure 2 and Table 4). No other main effects or interactions were significant, all ps > .3.

TABLE 4: Means for All Measures

Measure	Low Variability	High Variability
Mean URT with SRT = 750 (ms)	340 (77)	343 (64)
Mean URT with SRT = 3,000 (ms)	466 (88)	482 (101)
Total time on task (ms)	3,039 (139)	2,660 (146)
Task execution time (ms)	1,306 (144)	1,537 (150)
Failure rate with SRT = 750 ms (%)	5.40 (2.84)	5.53 (3.94)
Failure rate with SRT = 3,000 ms (%)	5.55 (3.00)	6.00 (11.12)
NASA–Task Load Index total score	9.27 (3.60)	8.99 (2.28)
AttrakDiff Pragmatic Quality	4.24 (0.50)	4.20 (0.53)
AttrakDiff Hedonic Quality–Stimulation	3.87 (0.69)	4.26 (0.52)
AttrakDiff Hedonic Quality–Identity	4.13 (0.33)	3.97 (0.33)
AttrakDiff Attractiveness	4.10 (0.27)	4.14 (0.37)

Note. Standard deviations shown in parentheses. URT = user response time; SRT = system response time.

Total Time on Task and TET

In this article, TET is defined according to Dannenbring (1983) as the measure for the human component in completing the whole task. Total time on task refers to the entire time that the man–computer system needs to process a task.

Although we analyzed URTs in a statistically "clean" way, we are aware that for a "real" application, stakeholders would be more interested in overall performance gains. Therefore, we included all trials in our analysis of TETs and total time on task to have a realistic measure of how long our participants spent on an average trial after the learning phase. To do so, we calculated the time spent on the task as the average time between onset and offset of each trial, including the initial response and the target response (see Figure 1) for the low-variability and the high-variability condition. Both spam and relevant e-mail trials were included in this analysis as well as all different SRTs. Trials with total time on task higher or lower than 3 standard deviations above or below each participant's individual average in each condition were eliminated from analysis (1.3%). Failure trials were included, as failure rate was intended to be part of the performance measure. Thus, two average times on task were obtained for each participant, corresponding to the two variability conditions.

Because total time on task includes, on average, higher SRTs in the low-variability condition, we also calculated TET (Dannenbring, 1983) as a measure of participants' average performance. That is, we subtracted SRTs from the time on task to gain TET as the "human part" of interaction time.

A one-way between-subjects ANOVA on total time on task with the factor variability (low, high) revealed that the overall duration of a trial was longer in the low-variability condition than in the high-variability condition, $F(1, 18) = 35.37$, $p < .001$, $\eta_p^2 = .66$ (see Figure 3 and Table 4). An ANOVA with the between-subjects factor variability (low, high) for TET revealed a significantly shortened TET for low variability, $F(1, 18) = 12.40$, $p = .002$, $\eta_p^2 = .41$.

Failure Rate

Spam and relevant trials were included in this measure. Wrong classification of relevant e-mails as spam and classification of spam e-mails as relevant were considered errors. Furthermore, trials with more than one click on the target screen were considered target click errors, because those trials indicated either a too-early response or a lack of visual attention to the screen. Classification and target click errors were averaged for each participant for 750-ms and 3,000-ms SRTs. A two-way mixed-measures ANOVA (Variability [low, high] ×

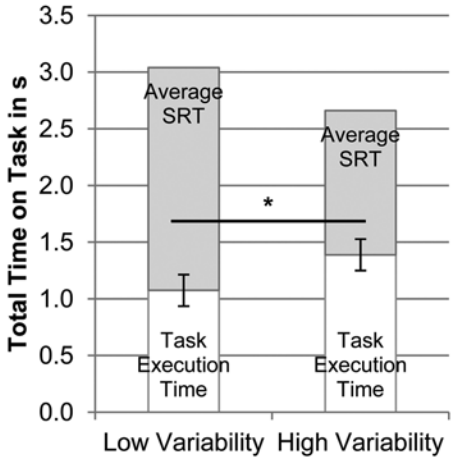

Figure 3. Task execution time with variability analyzed as between-subjects factor. Task execution time averaged across system response times, spam and relevant e-mail trials. Error bars indicate standard deviation.

SRT [750 ms, 3,000 ms]) showed no difference in percentage of error, $F(1, 18)$ for SRT < 2, all other Fs < 0.1 (see also Table 4).

NASA-TLX

NASA-TLX questionnaire data were analyzed according to the guidelines (Hart & Staveland, 1988). The quantitative ratings on each individual scale were multiplied with the weight, obtained from the pairwise comparisons. We calculated a full score by summing up the weighted scale scores and dividing the final score by the sum of its weights. A t test on overall task load on the NASA-TLX questionnaire showed no significant difference between the low- and the high-variability condition, $t(18) = 0.20$, $p = .842$, $d = 0.091$ (see Table 4).

AttrakDiff

For each participant, we calculated scores on all four scales by averaging the score of the items belonging to each individual scale (see Table 4). The t tests on all individual scales of AttrakDiff revealed no significant difference on any of the four scales, all ps > .4. The mean overall for the scales was $M = 4.16$ ($SD = 0.18$).

DISCUSSION

In accordance with our hypotheses, participants performed faster in a sequential task interrupted by SRTs when the variability of the SRTs was low, compared with when the variability was high. Although reduced variability of SRTs came at the expense of, on average, elongated waiting times, the overall reduction in TET reveals that the human response component of the task is accelerated (as already predicted, but not found, by Kuhmann et al., 1987). Although responses were faster in the low-variability condition, error rates do not differ between the conditions (and numerically point to the same direction as URTs). Therefore, increased response speed cannot be attributed to decreased accuracy.

User frustration or decreased user satisfaction cannot account for URT differences, as neither subjective task load nor likeability differed between the conditions, although SRTs were on average 693 ms longer (with 54% SRTs > 2 s) in the low-variability condition than in the high-variability condition (24% SRTs > 2 s). This finding challenges Fischer, Blommaert, and Midden's (2005) finding of a linear negative relationship between total time on task and user satisfaction and contradicts Miller's (1968; Nielsen, 1999; Shneiderman & Plaisant, 2009) widely spread SRT boundary of 2 s to maximize user satisfaction and user performance.

The general suggestion to minimize SRTs to obtain maximal user performance and satisfaction has already been criticized because of the risk of more errors after very short SRTs (Shneiderman & Plaisant, 2009). We conclude that the relationship between time to task completion and performance, as well as user satisfaction, cannot always be a linear one. Instead, the results of the present study suggest that both factors might be moderated by variability and, thus, temporal predictability. In light of this possibility, focusing on variability in interface design might be a promising approach.

However, the null effects have to be interpreted carefully, as we did not have sufficient statistical power to rule out potential effects. Authors of further studies should attempt to use

more sensitive measures of user satisfaction and include physiological measures.

We assume temporal expectancy to be the cognitive mechanism underlying the effect. Humans build up expectations as to when the next event will happen (Awramoff, 1903; Wundt, 1874). Previous experience shapes these expectancies. When the intervals are more or less constant, one can anticipate the time of target occurrence (Cardoso-Leite, Mamassian, & Gorea, 2009; Los & Schut, 2008) and, hence, respond relatively quickly and accurately. However, when intervals are variable, exact anticipation of the next event is much less precise, and response quality suffers (Elithorn & Lawrence, 1955). Thus, we successfully demonstrated that implicit adaptations to temporal structures (Olson & Chun, 2001; Wagener & Hoffmann, 2010) can be used to decrease URTs and TETs in HCI.

We had unexpectedly strong carryover effects. Participants from both groups did not learn in the second session. We speculate that participants overlearned either the predictability or unpredictability (depending on the starting condition) of the system in the first session, suggesting that temporal expectations are formed quickly and then held even if they no longer apply. Alternatively, a lack of motivation "to do the same boring task again," uttered by some participants at the beginning of the second session, might indicate that it was a lack of participants' compliance. Although groups did not differ in the first block of the experiment, we cannot rule out that different user strategies existed prior to the experiment that contributed to this effect. Further research should address the importance of constancy of temporal patterns.

To our knowledge, this study revealed behavioral benefits for temporally predictable user interfaces for the first time. Further research should replicate and generalize these findings. Several factors could influence how temporal expectancies determine user performance and user satisfaction. First, with different characteristics of the task, such as greater complexity or a lower average SRT duration, participants might be able to fully compensate the time loss caused by elongated SRTs in the low-variability condition. Second, we assume that different interaction devices and interaction types, for instance, workplace applications versus network-dependent Internet applications, might benefit differently from SRT regularities. For touch screen devices, temporal predictability might help to lower the particularly high error rates (Brewster, Chohan, & Brown, 2007; Hoggan, Brewster, & Johnston, 2008). For mobile devices, temporal expectancies could foster interaction as users shift their attention away from the screen, and therefore the task, with SRTs between 4 and 8 s (Roto & Oulasvirta, 2005). Knowing when a process will be finished might, therefore, improve the users' attendance to the task.

Last, we suggest that the role of temporal predictability for different types of users in different situations be investigated (for an overview, see Caldwell, 2008). For example, temporal predictability might help novice users more than experts to understand that the system is actually responding. Therefore, the knowledge of when a response of the system is to be expected could be part of the user's expertise regarding the system and explain why experienced users accept longer SRTs (Caldwell & Paradkar, 1995). Additionally, the detrimental influence of situational factors, such as time pressure (Caldwell & Garett, 2005) or task load, on performance could be alleviated by temporal predictability, as the automatic allocation of attention to a task at its expected time could leave more resources for the actual task.

To conclude, we conjecture that reducing variability provides, because of the ability to adapt to temporal structures, a possibility of time design beyond minimizing SRTs. Computing systems could be deliberately slowed down to enhance temporal predictability. The present effect, however, has to be examined further to deduce general recommendations for interface design. The range of the SRTs' time scale, its parameters, and its impacts on HCI should consequently be topics of further research.

KEY POINTS

- Responsiveness in terms of average system response times is a common measure for quantifying software quality, as it affects user experience.

- This article shows that computer users adapt to temporal structures in computing systems.
- Temporal predictability enhances user performance, even when it comes at the cost of longer absolute waiting times.
- Temporal predictability should be considered when dealing with responsiveness.

REFERENCES

Awramoff, D. (1903). Arbeit und Rhythmus: Der Einfluss des Rhythmus auf die Quantität und Qualität geistiger und körperlicher Arbeit, mit besonderer Berücksichtigung des rhythmischen Schreibens [Work and rhythm: The influence of rhythm on the quantity and quality of mental and physical work, paying particular consideration to rhythmic writing]. *Philosophische Studien*, *18*, 515–562.

Brewster, S., Chohan, F., & Brown, L. (2007). Tactile feedback for mobile interactions. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (pp. 159–162). New York, NY: ACM Press.

Bush, L. K., Hess, U., & Wolford, G. (1993). Transformations for within-subject designs: A Monte-Carlo investigation. *Psychological Bulletin*, *113*, 566–579. doi:10.1037//0033-2909.113.3.566

Caldwell, B. S. (2008). Knowledge sharing and expertise coordination of event response in organizations. *Applied Ergonomics*, *39*, 427–438. doi:10.1016/j.apergo.2008.02.010

Caldwell, B. S., & Garett, S. K. (2005, June). Experience, task cycles and proactive resource allocation in organizational settings. Paper presented at the Eighth Human Factors in Organizational Design and Management (ODAM) Symposium, Maui, HI.

Caldwell, B. S., & Paradkar, P. (1995). Factors affecting user tolerance for voice mail message transmission delays. *International Journal of Human–Computer Interaction*, *7*, 235–248. doi:10.1080/10447319509526123

Caldwell, B. S., & Wang, E. (2009). Delays and user performance in human-computer-network interaction tasks. *Human Factors*, *51*, 813–830. doi:10.1177/0018720809359349

Cardoso-Leite, P., Mamassian, P., & Gorea, A. (2009). Comparison of perceptual and motor latencies via anticipatory and reactive response times. *Attention, Perception, & Psychophysics*, *71*, 82–94. doi:10.3758/app.71.1.82

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis issues for field settings*. Chicago, IL: Rand-McNally.

Correa, Á., Lupiáñez, J., Milliken, B., & Tudela, P. (2004). Endogenous temporal orienting of attention in detection and discrimination tasks. *Perception & Psychophysics*, *66*, 264–278.

Cota-Robles, E., & Held, J. P. (1999, February). A comparison of Windows driver model latency performance on Windows NT and Windows 98. Paper presented at the Third Symposium on Operating Systems Design and Implementation, New Orleans, LA.

Coull, J. T., & Nobre, A. C. (1998). Where and when to pay attention: The neural systems for directing attention to spatial locations and to time intervals as revealed by both PET and fMRI. *Journal of Neuroscience*, *18*, 7426–7435.

Dabrowski, J., & Munson, E. V. (2011). 40 years of searching for the best computer system response time. *Interacting With Computers*, *23*, 555–564. doi:10.1016/j.intcom.2011.05.008

Dannenbring, G. (1983). The effect of computer response time on user performance and satisfaction: A preliminary investigation. *Behavior Research Methods & Instrumentation*, *15*, 213–216. doi:10.3758/bf03203551

Elithorn, A., & Lawrence, C. (1955). Central inhibition: Some refractory observations. *Quarterly Journal of Experimental Psychology*, *7*, 116–127.

Fischer, A. R. H., Blommaert, F. J. J., & Midden, C. J. H. (2005). Monitoring and evaluation of time delay. *International Journal of Human–Computer Interaction*, *19*, 163–180.

Flautner, K., Uhlig, R., Reinhardt, S., & Mudge, T. (2000). Thread-level parallelism and interactive performance of desktop applications. *ACM Sigplan Notices*, *35*, 129–138. doi:10.1145/356989.357001

Galletta, D. F., Henry, R., McCoy, S., & Polak, P. (2004). Web site delays: How tolerant are users? *Journal of the Association for Information Systems*, *5*, 1–28.

Greenwald, A. G. (1976). Within-subjects designs: To use or not to use. *Psychological Bulletin*, *83*, 314–320. doi:10.1037//0033-2909.83.2.314

Haering, C., & Kiesel, A. (2012). Time in action contexts: Learning when an action effect occurs. *Psychological Research*, *76*, 336–344. doi:10.1007/s00426-011-0341-8

Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In P. A. Hancock & N. Meshkati (Eds.), *Human mental workload* (pp. 139–183). Amsterdam, Netherlands: North Holland Press.

Hassenzahl, M., Burmester, M., & Koller, F. (2003). AttrakDiff: Ein Fragebogen zur Messung wahrgenommener hedonischer und pragmatischer Qualität [AttrakDiff: a questionnaire for measuring perceived hedonic and pragmatic quality]. In J. Ziegler & G. Szwillus (Eds.), *Mensch & Computer 2003: Interaktion in Bewegung* (pp. 187–196). Stuttgart, Germany: Teubner.

Hoggan, E., Brewster, S. A., & Johnston, J. (2008). Investigating the effectiveness of tactile feedback for mobile touchscreens. In *Proceedings of the Twenty-Sixth Annual SIGCHI Conference on Human Factors in Computing Systems* (pp. 1573–1582). New York, NY: ACM Press.

Hui, M. K., & Tse, D. K. (1996). What to tell consumers in waits of different lengths: An integrative model of service evaluation. *Journal of Marketing*, *60*, 81–90.

Kingstone, A. (1992). Combining expectancies. *Quarterly Journal of Experimental Psychology*, *44*, 69–104.

Kleinrock, L. (1975). *Queueing systems: Vol. 1: Theory*. New York, NY: Wiley-Interscience.

Kohlisch, O., & Kuhmann, W. (1997). System response time and readiness for task execution: The optimum duration of intertask delays. *Ergonomics*, *40*, 265–280.

Kubovy, M., & Pomerantz, J. R. (1981). *Perceptual organization*. Hillsdale, NJ: Lawrence Erlbaum.

Kuhmann, W., Boucsein, W., Schaefer, F., & Alexander, J. (1987). Experimental investigation of psychophysiological stress-reactions induced by different system response times in human-computer interaction. *Ergonomics*, *30*, 933–943.

Kuhmann, W., Schaefer, F., & Boucsein, W. (1989). *Effekte von Wartezeiten innerhalb einfacher Aufgaben: Eine Analogie zu Wartezeiten in der Mensch-Computer-Interaktion* [Effects of waiting times within simple tasks: An analogy to waiting times in human-computer-interaction]. Wuppertal, Germany: Fach Psychologie, FB 3, Gesamthochschule, Berg. Univ.

Los, S. A., & Schut, M. L. J. (2008). The effective time course of preparation. *Cognitive Psychology*, *57*, 20–55. doi:10.1016/j.cogpsych.2007.11.001

Miller, R. B. (1968, December). Response time in man-computer conversational transactions. Paper presented at the AFIPS '68 Fall Joint Computer Conference, Part 1, San Francisco, CA.

Nah, F. F. H. (2004). A study on tolerable waiting time: How long are web users willing to wait? *Behaviour & Information Technology*, *23*, 153–163. doi:10.1080/01449290410001669914

Nielsen, J. (1999). User interface directions for the web. *Communications of the ACM*, *42*, 65–72. doi:10.1145/291469.291470

Niemi, P., & Näätänen, R. (1981). Foreperiod and simple reaction-time. *Psychological Bulletin*, *89*, 133–162.

Olson, I. R., & Chun, M. M. (2001). Temporal contextual cuing of visual attention. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *27*, 1299–1313.

Osuna, E. E. (1985). The psychological cost of waiting. *Journal of Mathematical Psychology*, *29*, 82–105. doi:10.1016/0022-2496(85)90020-3

Pfendler, C. (1990). Zur Messung der mentalen Beanspruchung mit dem NASA-Task Load Index [Measuring mental load with the NASA-Task Load Index]. *Zeitschrift für Arbeitswissenschaft*, *44*, 158–163.

Ratcliff, R. (1993). Methods for dealing with reaction-time outliers. *Psychological Bulletin*, *114*, 510–532. doi:10.1037/0033-2909.114.3.510

Roto, V., & Oulasvirta, A. (2005, May). Need for non-visual feedback with long response times in mobile HCI. Paper presented at the 2005 International Conference on the World Wide Web, Chiba, Japan.

Schaefer, F. (1990). The effect of system response times on temporal predictability of work flow in human–computer interaction. *Human Performance*, *3*, 173–186.

Schleifer, L. M., & Amick, B. C. (1989). System response time and method of pay: Stress effects in computer based tasks. *International Journal of Human–Computer Interaction*, *1*, 23–39. doi:10.1080/10447318909525955

Schneider, W., Eschman, A., & Zuccolotto, A. (2002). *E-prime user's guide*. Pittsburgh, PA: Psychology Software Tools.

Seow, S. (2008). *Designing and engineering time: The psychology of time perception in software*. Vancouver, Canada: Addison-Wesley Professional.

Shneiderman, B., & Plaisant, C. (2009). *Designing the user interface* (5th ed.). Boston, MA: Addison-Wesley.

Szameitat, A. J., Rummel, J., Szameitat, D. P., & Sterr, A. (2009). Behavioral and emotional consequences of brief delays in human–computer interaction. *International Journal of Human–Computer Studies*, *67*, 561–570. doi:10.1016/j.ijhcs.2009.02.004

Teal, S. L., & Rudnicky, A. I. (1992). A performance model of system delay and user strategy selection. In P. Bauersfeld, J. Bennett, & G. Lynch (Eds.), Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (pp. 295–305). New York, NY: Association for Computing Machinery.

Thomaschke, R., & Dreisbach, G. (2013). Temporal predictability facilitates action, not perception. *Psychological Science*, *24*(7), 1335–1340.

Thomaschke, R., Kiesel, A., & Hoffmann, J. (2011). Response specific temporal expectancy: Evidence from a variable foreperiod paradigm. *Attention, Perception & Psychophysics*, *73*, 2309–2022. doi:10.3758/s13414-011-0179-6

Thomaschke, R., Wagener, A., Kiesel, A., & Hoffmann, J. (2011a). The scope and precision of specific temporal expectancy: Evidence from a variable foreperiod paradigm. *Attention, Perception & Psychophysics*, *73*, 953–964. doi:10.3758/s13414-010-0079-1

Thomaschke, R., Wagener, A., Kiesel, A., & Hoffmann, J. (2011b). The specificity of temporal expectancy: Evidence from a variable foreperiod paradigm. *Quarterly Journal of Experimental Psychology*, *64*, 2289–2300. doi:10.1080/17470218.2011.616212

Thum, M., Boucsein, W., Kuhmann, W., & Ray, W. J. (1995). Standardized task strain and system response times in human–computer interaction. *Ergonomics*, *38*, 1342–1351. doi:10.1080/00140139508925192

Wagener, A., & Hoffmann, J. (2010). Temporal cueing of target-identity and target-location. *Experimental Psychology*, *57*, 436–445. doi:10.1027/1618-3169/a000054

Wundt, W. (1874). *Grundzüge der Physiologischen Psychologie* [Principles of physiological psychology]. Leipzig, Germany: Engelmann.

Yates, D., Kurose, J., Towsley, D., & Hluchyj, M. G. (1993). On per-session end-to-end delay distributions and the call admission problem for real-time applications with QOS requirements. *SIGCOMM Computer Communication Review*, *23*(4), 2–12. doi:10.1145/167954.166238

Zijlstra, F. R. H., Roe, R. A., Leonora, A. B., & Krediet, I. (1999). Temporal factors in mental work: Effects of interrupted activities. *Journal of Occupational and Organizational Psychology*, *72*, 163–185. doi:10.1348/096317999166581

Florian Weber is a usability specialist working in the automotive industry. He received his diploma in psychology in 2011 from the University of Wuerzburg.

Carola Haering is a postdoctoral researcher in the Cognitive Psychology Unit at the University of Wuerzburg, Germany. She received her PhD in psychology in 2010 from the University of Wuerzburg.

Roland Thomaschke is a postdoctoral researcher in the Department of General and Applied Psychology at the University of Regensburg. He received his PhD in psychology in 2009 from Lancaster University.

Date received: December 22, 2011
Date accepted: December 18, 2012