Erschienen in: Human Factors ; 50 (2008), 4. - S. 643-651

Feedback Design for the Control of a Dynamic Multitasking System: Dissociating Outcome Feedback From Control Feedback

Hansjörg Neth, Sangeet S. Khemlani, and **Wayne D. Gray,** Rensselaer Polytechnic Institute, Troy, New York

Objective: We distinguish outcome feedback from control feedback to show that suboptimal performance in a dynamic multitasking system may be caused by limits inherent to the information provided rather than human resource limits. **Background:** Tardast is a paradigm for investigating human multitasking behavior, complex system management, and supervisory control. Prior research attributed the suboptimal performance of Tardast operators to poor strategic task management. **Methods:** We varied the nature of performance feedback in the Tardast paradigm to compare continuous, cumulative feedback (global feedback) on performance outcome with feedback limited to the most recent system state (local feedback). **Results:** Participants in both conditions improved with practice, but those with local feedback performed better than those with global feedback. An eye gaze analysis showed increased visual attention directed toward the feedback display in the local feedback condition. **Conclusion:** Predicting performance in the control of a dynamic multitasking system requires understanding the interactions between embodied cognition, the task being performed, and characteristics of performance feedback. In the current case, at least part of what had been diagnosed as a deficit caused by limited cognitive resources has been shown to be data limited. **Application:** Perfect outcome feedback can provide inadequate control feedback. Instances of suboptimal performance can be alleviated by better feedback design that takes into account the temporal dynamics of the human-system interaction.

We routinely manage complex systems in the pursuit of specific outcomes: achieve a good GPA by graduation, maximize the monthly amount of energy produced by a power plant, or race a sailboat as far as possible within a day. Ironically, whereas accurate and reliable feedback on how well these goals are being achieved may help us to assess degrees of success or failure, such feedback by itself may be inadequate to guide progress toward achieving our goals.

In this article, we propose a functional distinction between outcome feedback and control feedback, where *outcome feedback* enables us to assess how well we have done in achieving our long-term goals, and *control feedback* allows us to set short-term goals for immediate action. These two types of feedback correspond to different levels of aggregation, with the former providing a global summary of performance and the latter representing a local (smaller and more recent) performance interval.

We dissociate outcome feedback from control feedback in the context of the Tardast task environment (Shakeri, 2003; Shakeri & Funk, 2007). Tardast is a dynamic multitasking system that requires human operators to maximize the overall output of six concurrent subtasks over a specific period of performance. The original Tardast system continuously provided outcome feedback by directly reporting the operator's overall success as measured by the current value of the outcome variable of interest. Studies with this system yielded the conclusion that human performance was suboptimal compared with the near-optimal solution of a machine-learning algorithm.

The finding of stable suboptimal human performance (Fu & Gray, 2004) is the beginning of a research program, not its end. We will show that

Address correspondence to Hansjörg Neth, Cognitive Science Department, Rensselaer Polytechnic Institute, Carnegie 108, 110 8th Street, Troy, NY 12180; nethh@rpi.edu. ***HUMAN FACTORS,*** Vol. 50, No. 4, August 2008, pp. 643–651. DOI 10.1518/001872008X288583.

control feedback, by virtue of being more responsive to local system changes, yields better performance by allocating actions more adaptively. Consequently, we conclude that performance on Tardast was limited not only by the cognitive capacity of its human operators but also by data limits inherent to its original feedback score.

In the following section, we introduce the Tardast task environment and our notion of performance feedback before elaborating on the feedback types and hypotheses of our study. We then present an experiment that compares two human groups with different feedback scores and contrast their performance with two mathematical models that provide theoretical benchmarks. Our results demonstrate that local control feedback yields better performance than global outcome feedback and is used more extensively, as shown by an eye gaze analysis. In the concluding section, we sketch a general framework for the functional analysis of performance feedback and explore the implications of our results for the design of dynamic feedback mechanisms.

BACKGROUND

The Tardast Task Environment

Tardast (Shakeri, 2003; Shakeri & Funk, 2007) is a novel paradigm that allows the study of human multitasking by mapping a variety of behavioral scenarios to an abstract framework. Named after the Persian term for "juggler," the system is based on the analogy that a juggler's feat of spinning plates on vertical poles can represent the concurrent management of multiple tasks that coexist without predefined completion criteria. As multitasking and the related tasks of monitoring, supervisory control, and complex system management (e.g., Berry & Broadbent, 1988; Moray, 1986) are poorly understood, a synthetic task environment that captures their essential elements while abstracting away from domain-specific details is an important research tool (Gray, 2002).

At the core of any multitasking situation lies a resource allocation problem: Because cognitive, perceptual, and action resources are limited, humans have to negotiate trade-offs when deciding which task to do when. In Tardast, n vertical bars abstractly represent competing tasks (see Figure 1). The height of the ith bar indicates the corresponding task's current *satisfaction level* (SL_i) and decreases at a constant *deviation rate* (DR_i) whenever not acted upon. A task's status is improved by pressing a button underneath the bar, which increases the task's SL_i at a constant *correction rate* (CR_i). A serial bottleneck exists in that only a single task can be acted upon at any time. As tasks can differ in *weight* (W_i) and rate parameters (DR_i/CR_i) and the system state is subject to constant changes, maximizing overall performance is nontrivial.

Shakeri and Funk (2007) contrasted human performance in several Tardast scenarios with the near-optimal performance of a machine-learning algorithm (see Glover, 1990, for details on Tabu search) and found human operators to be lacking in comparison. This shortcoming was attributed to poor strategic task management. As the complexity of the system exceeded human resource limitations, operators failed to prioritize important tasks.

Neth, Khemlani, Oppermann, and Gray (2006) replicated the basic phenomenon of stable suboptimal performance with additional experimental controls. By assessing operators' improvement over time, we verified that performance indeed asymptotes at a suboptimal level. Although we argued that the observed performance differences between scenarios have mostly been a function of the task environment, we essentially confirmed Shakeri and Funk's (2007) conclusions.

Our present emphasis is different. Struck by the gap between human and optimal performance, we wonder why such a relatively simple system seems to exceed human processing capacities: The number of concurrent tasks is small, all relevant parameters are accessible, the scoring scheme is explained in detail, the temporal demands are moderate, and there are no hidden dependencies or delayed consequences of actions. We now believe that the feedback provided to operators is partly responsible for their poor performance.

Cognitive processes can generally be constrained by limited processing resources and limits to the data that are processed (Norman & Bobrow, 1975). If a change in the feedback mechanism results in improved system management, the original performance was limited not only by human resources but also by the system's design. Simply put, our diagnosis is that the original feedback score made it difficult to judge which states of high quality can be achieved and sustained. To substantiate this claim, we need to consider the general effects of feedback on performance and

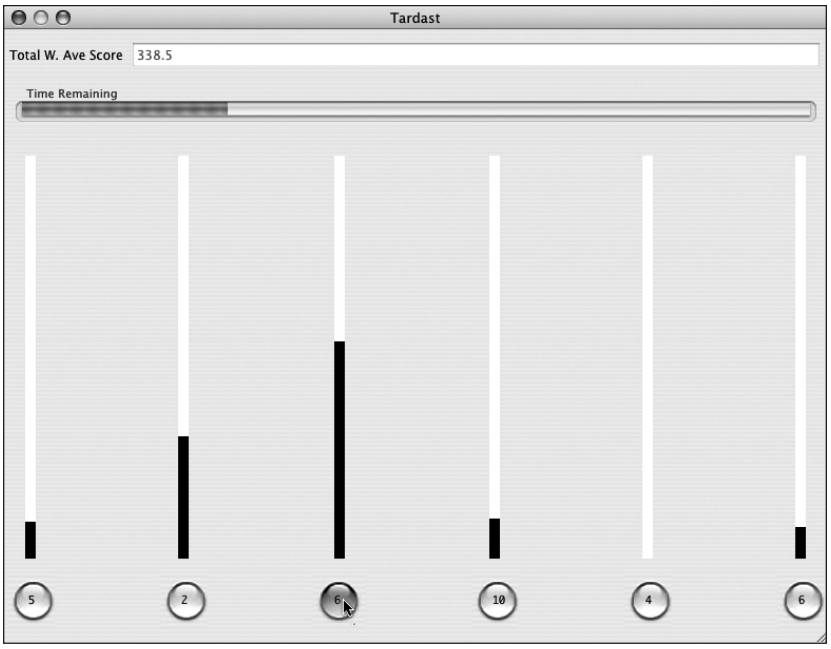

Figure 1. Our Tardast interface showing six concurrent tasks. Each vertical bar represents a task, and the height of the black portion indicates its current satisfaction level (*SL*). Pressing one of the buttons underneath a bar increases the *SL* of the corresponding task. The numbers (5, 2, 6, etc.) on the buttons provide each task's weight *W*. At each system cycle, the *W*s and current *SL*s of all tasks are integrated to update a numerical feedback score at the top (see Equations 1–3). The horizontal progress bar below the score indicates the time remaining for the scenario.

specify our functional analysis of feedback for the Tardast system.

Dimensions of Feedback

The use of feedback to regulate the behavior of dynamic systems – by passing back some aspect of the system output as input – is one of the pillars of cybernetics and control theory (Wiener, 1948). In the terminology of a closed-loop control system, a feedback score serves as a sensor that provides a reference signal to an operator, who monitors discrepancies between actual and desired states, initiates corrective actions, and evaluates the effects of actions on the system to meet certain criteria.

When invoked as an explanatory concept in technological, biological, or social systems, the definition of *feedback* may become so vague and flexible as to be "artificial and of little use" (Ashby, 1956, p. 54). Within psychology, the notion includes factors as diverse as reward and punishment, encouragement and criticism, and other verbal or nonverbal information about different aspects of performance. In the context of this study, we view feedback as an interface between system states and an operator's perceptions and actions. To avoid confusion, we restrict the term *feedback* to denote a numerical index of performance. In addition, we focus on feedback that integrates multiple aspects of performance into a single summary value. Such composite measures contrast with concurrent displays of multiple indices that highlight different aspects of the system.

Although feedback needs to accurately reflect performance, the design of any particular feedback signal is selective. For dynamic systems, three important dimensions are the scope, frequency, and update lag of any feedback signal:

- *Scope* is the amount of performance considered when one computes the numeric feedback value. A key distinction is between local and global feedback. Whereas *local* feedback is limited in scope (e.g., a single assessment of system state), *global* feedback integrates over multiple measurements (e.g., the sum or average of all states so far).
- *Frequency* refers to the interval at which the feedback signal is updated. A common case is that feedback frequency corresponds to the system cycle time, but both shorter and longer intervals are possible.
- *Update lag* is the delay between a change in system state and the update of the feedback signal.

Providing Feedback in Tardast

Tardast is updated every 100 ms, and an aggregate feedback score is recomputed and displayed at the top of the screen (see Figure 1) on each system update. Characterized by our three dimensions, this score is global in scope and high in frequency, with minimal update lag. As these latter two properties provide a continuous and instantaneous feedback signal implemented in many technical systems (e.g., odometers and speedometers), we focus on variations of feedback scope.

A Tardast operator's goal is to manage multiple concurrent tasks so as to maximize the total weighted average score (TWAS), which is computed as the weighted and normalized average of all n tasks' satisfaction levels $SL_i(t)$ averaged over all T time steps elapsed so far:

$$\mathrm{TWAS}(T) = \frac{1}{T}\sum_{t=1}^{T}\left[\frac{\sum_{i=1}^{n} w_i SL_i(t)}{\sum_{i=1}^{n} w_i}\right] \quad (1)$$

(For the sake of clarity we do not explicitly represent the fact that tasks with an *SL* of zero incur a 20% penalty.) To comprehend the implications of this score, one should interpret the term in brackets as an index of system *quality* $Q(t)$ at a particular time *t:*

$$Q(t) = \frac{\sum_{i=1}^{n} w_i SL_i(t)}{\sum_{i=1}^{n} w_i} \quad (2)$$

Substituting $Q(t)$ into Equation 1 yields

$$\mathrm{TWAS}(T) = \frac{1}{T}\sum_{t=1}^{T} Q(t) \quad (3)$$

Thus, TWAS represents global system quality by averaging over all local quality values Q. Because TWAS provides the criterion by which performance in Tardast is ultimately assessed, it also provides perfect outcome feedback.

Analysis and Hypotheses

Under a functional framework, the utility of any feedback signal depends on its uses. Hence, perfect outcome feedback may provide suboptimal control feedback. Our case against TWAS as a control feedback score rests on two intuitions. First, by being a cumulative average, TWAS becomes increasingly insensitive to quality fluctuations. The same change in Q will have smaller effects on TWAS if it occurs later in a scenario. In addition to its increasing inertia, TWAS poorly represents the direction of changes in Q. If Q increases from step t to step $t + 1$, TWAS will still decrease if $Q(t + 1) < \mathrm{TWAS}(t)$. Similarly, it is possible for TWAS to increase while Q decreases or to change while Q remains constant (and vice versa). Thus, neither the magnitude nor the direction of changes in TWAS necessarily coincides with corresponding changes in Q.

But why would monitoring Q benefit the operator when Tardast performance is evaluated by TWAS? Our second intuition relies on Equation 3: to maximize TWAS, an operator's objective is to achieve a state of high system quality for as often and long as possible. As an analogy, consider the task of setting a new 24-h speed-sailing record. The dynamic system of ship and sea depends on many variables (e.g., the weather, the vessel's weight and shape), is subject to sudden changes (of winds and currents), and allows for a variety of actions (steering, setting sails, adjusting the keel, etc.). Crucially, the effects of each action are not obvious and depend on the interactions between multiple factors. To ratify a record, one must obtain a precise outcome measure: What total distance did the ship traverse? But such an aggregate measure may not be the most useful for the sailor, who is aiming to maximize the ship's speed at any moment of the journey. Tardast operators are in the same boat as our sailor: By preserving the history of all previous states in its aggregate value, TWAS provides global outcome feedback at the expense of responsiveness to local changes in system quality.

We hypothesize that control feedback provided by Q will yield better performance and be more relied on than is TWAS. Although Q is memoryless, it is potentially more action relevant as it accurately reflects the magnitude and direction of momentary changes in system quality. A counterintuitive consequence of this prediction is that Q will increase performance outcome by not providing outcome feedback.

EXPERIMENT

Our study contrasts two extremes: local feedback that provides only snapshots of current system

quality versus global feedback that provides an aggregate measure of overall performance. We implemented this manipulation of feedback scope between groups by making the score at the top of the Tardast interface (see Figure 1) either $Q(t)$ or TWAS(T).

In addition, we contrast human performance with that of two artificial agents that were programmed in Lisp and interact with the same software environment as human participants. A *random agent* does not have any task-specific knowledge and establishes a performance baseline by randomly selecting a task every 10 s and completing each scenario 100 times. By contrast, our *heuristic agent* executes a simple deterministic strategy that could, in principle, be adopted by humans. At the beginning of a scenario, it establishes a preference order of tasks by ranking tasks based on the perceptually salient parameters *W* and *DR,* using *DR* only when weights are equal. It then raises its most favored task to a threshold (90% *SL*) before engaging less favored ones in the order of their rank. Every 3 s, it inspects the *SL* of its current and any more preferred task and selects the most preferred task below threshold. This effectively leads to a behavior that keeps two to three tasks at high levels (of approximately 85% *SL*). (We elaborate later on the implications of the 3-s cycle for human performance.) Both artificial agents explore the properties of the task environment and provide reference values for anchoring human performance.

METHOD

Participants

Twenty-four undergraduate students of Rensselaer Polytechnic Institute (mean age of 19.0) volunteered to participate for course credit.

Apparatus and Materials

The experiment was run on an Apple G4 computer (running Mac-OS 10.4) at a 1024-by-768 screen resolution. Participants' eye data were collected using an LC Technologies tracker at a 16-Hz sampling rate. A chinrest was used to stabilize head movements and ensure a fixed viewing distance of 60 cm.

Our version of Tardast was implemented in LispWorks 4.4 and matched all functional characteristics of the original software. In addition, we used the five scenarios that were used in previous Tardast studies. All tasks of one scenario have identical parameter settings, whereas three scenarios vary along one parameter dimension (*DR, CR,* and *W,* respectively), and the tasks of a fifth scenario vary along all three dimensions (see Shakeri & Funk, 2007, for values and underlying rationale). We also adopted the original conventions for the initialization (all tasks at 50% *SL*) and duration (5 min) of scenarios but randomized the six task positions on the screen for each scenario to avoid confounds of parameter values with task positions.

During performance, participants received feedback through a numerical display at the top left of the task interface (Figure 1). The local feedback condition presented $Q(t)$, whereas the global feedback condition presented TWAS(T). Both scores were continuously updated (every 100 ms) and were labeled *current quality score* or *total weighted average score,* respectively.

Design

All participants performed two blocks of five 5-min scenarios. The order of scenarios was randomized for each participant on the first block and repeated for the second block. As we do not have any hypotheses about scenario-specific effects, and because Shakeri and Funk (2007) advise against comparisons between scenarios, we will average across scenarios (i.e., treat them as samples from the population of possible scenarios). Thus, our study includes the between-subjects factor of feedback scope (local vs. global) and the within-subjects factor of block (1 vs. 2).

Procedure

Participants were tested individually. Instructions were modeled on those of Shakeri (2003). All participants first read a one-page instruction sheet that described the task and contained an explicit verbal description of their respective feedback score. They then watched a short movie that demonstrated the interaction with Tardast without conveying a particular strategy. Participants were instructed to maximize their score at all times (on each scenario and over all 10 scenarios) but were not informed about the repetition of scenarios between blocks. Instructions were followed by a 30-s eye-tracking calibration sequence.

After each 5-min scenario, all participants received feedback on the basis of TWAS and their average TWAS score over all scenarios completed

so far. Participants took a short (2- to 5-min) break between blocks. The experiment was completed within approximately 80 min, including instructions.

RESULTS

We present an analysis of three sets of data: first, empirical data collected from human participants; second, predictions yielded by our two software agents and their comparison with human data; and third, eye gaze data collected from human participants.

Human Performance by Feedback Condition

The mean performance score (TWAS) of human participants in the global feedback condition was 232.2 on the first block and 337.5 on the second block. Participants in the local feedback condition achieved mean scores of 339.9 and 413.1 (Figure 2).

A mixed analysis of variance (ANOVA) of mean performance scores by feedback scope and block yielded a significant main effect of block, $F(1, 22) = 47.03$, $MSE = 2032.33$, $p < .001$; the hypothesized significant main effect of feedback scope, $F(1, 22) = 11.32$, $MSE = 8896.84$, $p = .003$; and no significant interaction, $F(1, 22) = 1.52$, $MSE = 2032.33$, *ns*. Thus, participants' mean performance increased over blocks, and participants in the local feedback condition reliably outperformed those in the global feedback condition.

Human Versus Agent Performance

Figure 2 also contains two theoretical benchmarks derived from our software agents. A lower performance baseline with a mean TWAS score of 231 was established by running the random agent 100 times on each of the five scenarios. The heuristic agent performed at an average TWAS of 429. As our agents have no built-in learning mechanisms, their performance is identical on both blocks.

Compared with these benchmarks, participants in the global feedback condition performed at baseline level on the first block. Despite improving significantly on the second block, they failed to reach the performance of the simple heuristic strategy. By contrast, participants in the local feedback condition exceeded baseline performance on the first block and reached the level of the heuristic agent on the second block.

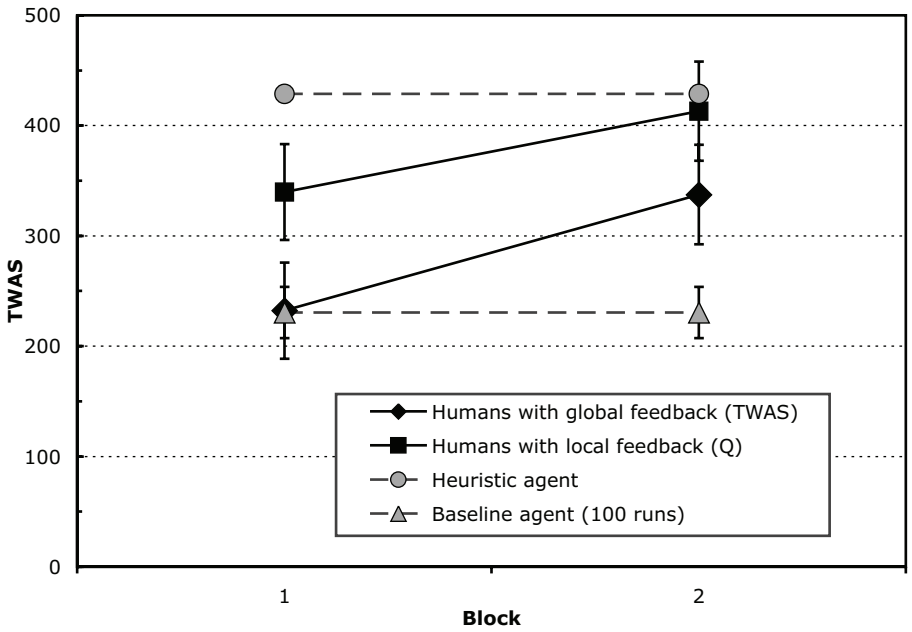

Figure 2. Performance comparison between human participants and artificial agents by block. Mean performance scores of humans in the global (TWAS) and local (*Q*) feedback conditions are compared with each other and with the performance of baseline and heuristic agents. All assessments of performance employ the TWAS metric. (Error bars denote 95% confidence intervals. As the heuristic agent is fully deterministic, it has no variance.)

Eye Gaze Measures

Eye gaze data provide additional evidence that local feedback was more instrumental in guiding behavior than global feedback. To assess the amount of visual attention that participants directed toward feedback, we counted the proportion of eye gaze samples that fell in the rectangular (7-by-3 cm) region surrounding the numerical score. (Two participants, one from each condition, were excluded from this analysis because of eye tracker calibration problems.)

On average, 9.1% of participants' eye gaze samples in the global feedback condition fell in the score region. The corresponding percentage in the local feedback condition was 15.6%. A mixed ANOVA of eye gaze samples in the score region by feedback scope and block yielded a significant main effect of feedback scope, $F(1, 20) = 5.6$, $MSE = 84.60$, $p < .028$, but no significant main effect of block, $F(1, 20) < 1$, $MSE = 17.20$, *ns,* or interaction, $F(1, 20) < 1$, $MSE = 17.20$, *ns.*

In summary, participants generally improved with experience, but those in the local feedback condition performed better and looked more at the numerical feedback score than those in the global feedback condition. Only participants with local feedback outperformed the baseline agent in Block 1 and reached the level of the heuristic agent in Block 2.

DISCUSSION

All participants interacted with the same system and had the same set of actions at their disposal. Hence, the superior performance in the local feedback condition supports our hypothesis that local feedback allows better control of a dynamic multitasking system than global outcome feedback. Our attribution of the performance gap to our feedback scope manipulation is corroborated by eye gaze data: Participants in the local feedback condition paid more attention to the feedback display than those in the global feedback condition.

Despite these encouraging results, we must not overinterpret the benefits of local instead of global feedback. Although performance with *Q* in the second block was in the range of our heuristic agent, a mean score of 413.1 is still suboptimal when compared with the near-optimal solution of Tabu search, which achieved an average score of 539.3 (Shakeri, 2003). However, as Tabu represents a machine-learning approach to system control, it may be unreasonable to expect boundedly rational humans to perform at the level of a normative machine-learning solution (Geisler, 2004; Simon, 1992). Although human performance is still suboptimal, it improved significantly by a small change in the numerical feedback signal. This finding confirms our conjecture that performance on the original Tardast system was limited not only by system complexity but also by feedback design.

GENERAL DISCUSSION

The design of any feedback signal entails choices and requires trade-offs. The same information can be accumulated to show how well we have done in achieving our goals or used to provide snapshots that assess the results of recent actions. Even when performance is ultimately evaluated by a global criterion, an outcome measure that continuously updates global performance is not necessarily a good reference signal for controlling the system. In the remaining sections, we will explore the implications of this gap between outcome and control feedback and flesh out our functional perspective on performance feedback.

A Functional Framework for Performance Feedback

We originally defined the difference between local and global feedback in terms of aggregation. However, an important consequence of this difference in aggregation is the differential temporal dynamics of the two measures. As human control and Tardast each represents complex, dynamic systems, an important component of the success of *Q* lies in the match between its temporal dynamics and those of its human users. To analyze this interaction, we consider three elements: (a) how the dynamics of system changes are reflected in the feedback measure, (b) the dynamics of the control of interactive behavior, and (c) the interaction of feedback with human control.

First, the parameters of a Tardast scenario limit the rate at which *Q* can change. Over most of its range, *Q* changes gradually and can maximally vary around 3% per second. This moderate rate of change implies that the lower digits of its three-digit numerical representation change rapidly, whereas the higher digits remain relatively stable. Interestingly, such temporal considerations have no predictive validity for human performance

unless they are related to the time scale of human activity (see Newell, 1990, chap. 3, for an extended discussion).

Second, the basic decision cycle for routine interactive behavior, unlike deliberative problem solving, is defined by the *unit task.* The unit task "partitions the behavior stream" (Card, Moran, & Newell, 1983, p. 385) beneath the level at which the task hierarchy is defined by the task itself and at the level at which task structure is defined by the control problems faced by the user. Unit tasks range in duration from 3 to 30 s with an internal structure composed of steps that range in duration from 1/3 to 3 s.

We assume that with more experience, as performance becomes routine, the basic human decision cycle in Tardast tends toward the lower end of the unit task range. A decision cycle of around 3 s seems like a good match to the rate of change in the *Q* score. Both TWAS and *Q* are updated instantaneously and continuously available upon demand. Just as with our analogy of the sailor, controlling Tardast requires its operator to frequently assess the system's current state, observe trends, and compare the quality of different states over time. *Q* directly enables each of these tasks, whereas TWAS does not.

Third, as a consequence, control feedback provided by *Q* allows the unit task to bridge Norman's (1988) gulfs between intentions and their realizations. The *gulf of evaluation* is the degree to which the system provides representations that can be perceived and interpreted in terms of the operator's expectations and intentions. While Tardast operators control the system, their intention is to achieve and maintain states of high system quality, not evaluate overall success. If the operator's perceptions and actions are out of sync with the dynamics of the system, its control becomes difficult or impossible.

The *gulf of execution* is the difference between the operator's intentions and the actions supported by the system. Tardast operators constantly face the decision as to whether a different action would be better than the current one. One way to determine this is to act and then evaluate the effects of one's actions. Thus, in tasks that require repeated decision-act cycles, bridging the gulf of evaluation is a prerequisite to bridging the gulf of execution. Control feedback allows the operator to evaluate the current system state and to judge whether recent actions achieved that goal and hence should be continued. In other words, *Q* provides both feedback and feed-forward information.

In summary, our local feedback score *Q* accurately reflects current system quality and changes slowly enough to not overwhelm human operators and fast enough to allow the perception of trends and the effects of interventions. Human performance is facilitated if all perceptual, cognitive, and motor elements required to decide whether to continue the current action or initiate a new action occur at the unit task level. As *Q* expresses the basic outcome measure on a temporal scale that human operators can meaningfully interact with, it is action oriented and has the prerequisites for providing excellent control feedback.

Limitations and Applications

Alas, by basing the usefulness of feedback on its uses and users, our functional framework offers no simple recipes for the design of optimal control feedback. Clearly, designing feedback requires more than just finding the right grain size of information aggregation. Although we presented our study as a dichotomy between outcome and control feedback and showed that a local score provides better control feedback for Tardast than a global score, we do not suggest that *Q* is the only or even an optimal feedback score. Rather, *Q* and TWAS are extremes on a continuum that allows for many alternative designs (e.g., intermediate summary scores that average over time while prioritizing recent states or combinations and integrations of multiple scores). Which functions are best served by each alternative is a question that goes beyond our present study.

Nonetheless, we trust that our functional view of performance feedback will enable better theories and better designs. From a theoretical viewpoint, a more systematic exploration of the ways in which the factors of feedback scope, frequency, and update lag serve various functions will allow researchers to develop better theories about how different feedback designs mediate human-technology interactions. For practical purposes, our functional view promises better engineered solutions. Although we cannot provide simple recipes, we will have succeeded if we convey that feedback needs to be designed with the same care as other aspects of complex interactive systems. Any design has to be evaluated with respect to its intended task and domain, but whenever humans

are in the control loop, matching feedback signals to human operator characteristics is essential.

Conclusion

Tardast is a flexible tool for investigating human multitasking, complex system management, and supervisory control. We went beyond the system inventors' report (Shakeri & Funk, 2007) to explore the influence of local versus global feedback on human performance. However, the implications of our analyses extend beyond the microcosm of Tardast. Instances of stable suboptimal performance (Fu & Gray, 2004) do not automatically reflect human capacity limits but may also result from data limits (Norman & Bobrow, 1975) that can be alleviated by better design. Our demonstration that perfect outcome feedback can provide inadequate control feedback highlights the importance of feedback design when one tries to push human performance toward the limits of bounded cognition (Simon, 1992).

The cognitive engineering goal of predicting human performance in controlling dynamic multitasking systems sometimes seems to recede as the complexity of our systems increases. However, the alternative is more theory, not more empirical trial and error. Understanding how system characteristics and feedback characteristics combine to influence human performance cannot be achieved in isolation; rather, the process of understanding how these factors influence human performance must be mediated by an understanding of the control of the human operator's integrated cognitive system.

ACKNOWLEDGMENTS

Supplemental materials to this article are available at www.cogsci.rpi.edu/cogworks/tardast.

We thank Shakib Shakeri and Ken Funk for their comments and help on our reimplementation of Tardast and are grateful to Dario Salvucci, an anonymous reviewer, Michael J. Schoelles, Chris R. Sims, and Christopher W. Myers for many constructive suggestions. Special thanks to Brittney Oppermann for her efforts in organizing and running the study.

The work reported was supported by grants from the Air Force Office of Scientific Research (AFOSR #FA9550-06-1-0074; Dr. Jun Zhang, program officer) and the Office of Naval Research (ONR #N000140710033; Dr. Ray Perez, program officer).

REFERENCES

Ashby, W. R. (1956). *Introduction to cybernetics.* London: Methuen.

Berry, D. C., & Broadbent, D. E. (1988). Interactive tasks and the implicit-explicit distinction. *British Journal of Psychology, 79,* 251–272.

Card, S. K., Moran, T. P., & Newell, A. (1983). *The psychology of human-computer interaction.* Hillsdale, NJ: Erlbaum.

Fu, W.-T., & Gray, W. D. (2004). Resolving the paradox of the active user: Stable suboptimal performance in interactive tasks. *Cognitive Science, 28,* 901–935.

Geisler, W. S. (2004). Ideal observer analysis. In L. M. Chalupa & J. S. Werner (Eds.), *The visual neurosciences* (pp. 825–837). Cambridge, MA: MIT Press.

Glover, F. (1990). Tabu search: A tutorial. *Interfaces, 20,* 74–94.

Gray, W. D. (2002). Simulated task environments: The role of high-fidelity simulations, scaled worlds, synthetic environments, and microworlds in basic and applied cognitive research. *Cognitive Science Quarterly, 2,* 205–227.

Moray, N. (1986). *Monitoring behavior and supervisory control* (Vol. II). New York: Wiley.

Neth, H., Khemlani, S. S., Oppermann, B., & Gray, W. D. (2006). Juggling multiple tasks: A rational analysis of multitasking in a synthetic task environment. In *Proceedings of the Human Factors and Ergonomics Society 50th Annual Meeting – 2006* (pp. 1142–1146). Santa Monica, CA: Human Factors and Ergonomics Society.

Newell, A. (1990). *Unified theories of cognition.* Cambridge, MA: Harvard University Press.

Norman, D. A. (1988). *The psychology of everyday things.* New York: Basic Books.

Norman, D. A., & Bobrow, D. G. (1975). On data-limited and resource-limited processes. *Cognitive Psychology, 7,* 44–64.

Shakeri, S. (2003). *A mathematical modeling framework for scheduling and managing multiple concurrent tasks.* Unpublished doctoral dissertation, Oregon State University.

Shakeri, S., & Funk, K. (2007). A comparison of human and near-optimal task management behavior. *Human Factors, 49,* 400–416.

Simon, H. A. (1992). What is an "explanation" of behavior? *Psychological Science, 3,* 150–161.

Wiener, N. (1948). *Cybernetics: Or control and communication in the animal and the machine.* Cambridge, MA: MIT Press.

Hansjörg Neth is a research assistant professor in Rensselaer Polytechnic Institute's Cognitive Science Department, Troy, New York. He received is Ph.D. in psychology from Cardiff University, Cardiff, UK, in 2004.

Sangeet S. Khemlani is a Ph.D. candidate in the Psychology Department at Princeton University. He received his B.S. in psychology and electronic arts from Rensselaer Polytechnic Institute, Troy, New York, in 2006.

Wayne D. Gray is a professor of cognitive science at Rensselaer Polytechnic Institute, Troy, New York. He received his Ph.D. in psychology from the University of California at Berkeley in 1979.

Date received: February 14, 2007
Date accepted: March 2, 2008