Erschienen in: Proceedings of the Human Factors and Ergonomics Society 50th annual meeting : October 16-20, 2006, San Francisco. - Santa Monica, Calif. : Human Factors and Ergonomics Soc., 2006. - S. 1142–1146. - ISBN 978-0-945289-29-6

JUGGLING MULTIPLE TASKS: A RATIONAL ANALYSIS OF MULTITASKING IN A SYNTHETIC TASK ENVIRONMENT

Hansjörg Neth (nethh@rpi.edu) **Sangeet S. Khemlani** (khemls@rpi.edu) **Brittney Oppermann** (opperb@rpi.edu) **Wayne D. Gray** (grayw@rpi.edu)

Cognitive Science Department
Rensselaer Polytechnic Institute
Troy, NY 12180 USA

Tardast (Shakeri 2003; Shakeri & Funk, in press) is a new and intriguing paradigm to investigate human multitasking behavior, complex system management, and supervisory control. We present a replication and extension of the original Tardast study that assesses operators' learning curve and explains gains in performance in terms of increased sensitivity to task parameters and a superior ability of better operators to prioritize tasks. We then compare human performance profiles to various artificial software agents that provide benchmarks of optimal and baseline performance and illustrate the outcomes of simple heuristic strategies. Whereas it is not surprising that human operators fail to achieve an ideal criterion of performance, we demonstrate that humans also fall short of a principally achievable standard. Despite significant improvements with practice, Tardast operators exhibit stable sub-optimal performance in their time-to-task allocations.

Human multitasking performance, whether in general terms of multitasking between tasks in different domains or in more specific terms of multitasking between several tasks within a single domain, is poorly understood and addressed by a disparate set of literatures and labels (e.g., *supervisory control*, Moray, 1986; *attention allocation* in dual-task situations, Wickens, 1992; *task switching* and *interruptions*, Rogers & Monsell, 1995; McFarlane & Latorella, 2002). Any multitasking situation essentially poses a resource-allocation problem: limited resources (of time, attention, or action) have to be distributed across multiple tasks in order to meet some criterion of performance.

Human rational behavior is generally constrained by the structure of task environments and the cognitive and perceptual-motor capabilities of human agents (Simon, 1990). To capture the functional relationships of complex tasks while abstracting away from domain specific details, we advocate the use of synthetic task environments (Gray, 2002).

In this paper, we report how the synthetic task environment of Tardast (Shakeri, 2003, Shakeri & Funk, in press) can be used to explore human multitasking behavior and exemplify a methodological framework to relate operator performance to the functional characteristics of complex task environments. Whenever the properties of complex environments and human agents are interdependent and subject to dynamic changes, any principled assessment of the scope and limits of human rationality requires a non-trivial amalgam of theoretical and empirical analyses. Our methodological approach is inspired by Anderson's (1990) notion of *rational analysis*, but promotes an in-depth analysis of *functional task environments* (Gray, Neth, & Schoelles, in press) rather than the evolutionary context to which cognition adapted.

We will first introduce the Tardast system and briefly summarize and critique prior findings. We then present an experiment that replicates and extends the study of Shakeri and Funk (in press). To distinguish environmental limitations from those imposed by human cognitive or perceptual-motor constraints, the performance of human operators will be compared to various software agents. Despite significant learning effects, human performance is shown to be sub-optimal both with respect to a normative ideal and an attainable heuristic strategy.

The Tardast Task Environment

Tardast was introduced by Shakib Shakeri (2003) and captures the essential core of many multitasking situations—but also aspects of supervisory control, monitoring, and complex

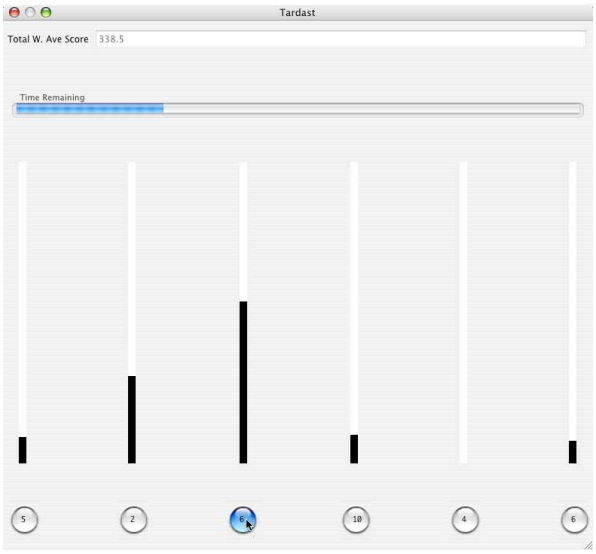

Figure 1: Our version of the Tardast interface. Each bar represents a task and the height of the black bar indicates its satisfaction level (SL). Selecting the buttons at the bottom increases the SL of the corresponding task. The numbers on the buttons indicate the tasks' weights (W).

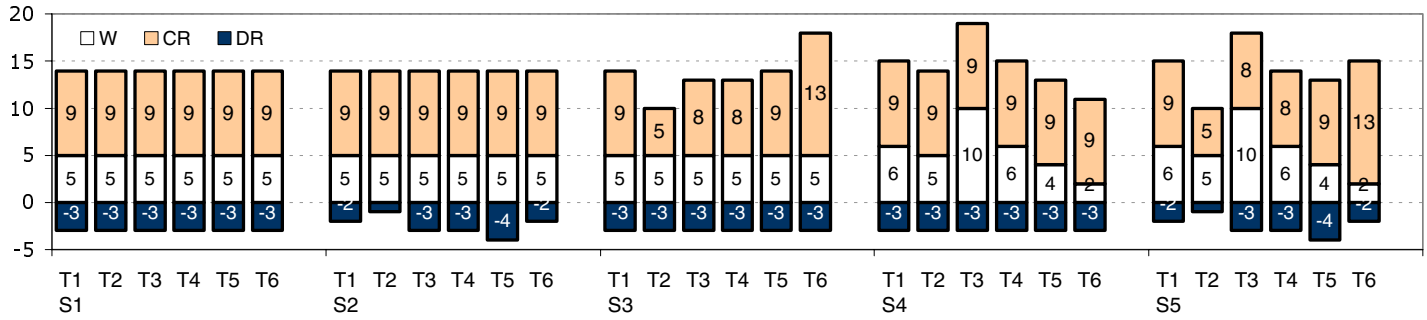

Figure 2: Parameter values for the five scenarios used by Shakeri (2003) and in our experiment. All tasks of Scenario 1 have the same parameter values; Scenarios 2–4 vary parameter values on one dimension; Scenario 5 varies parameter values on all three dimensions.

system management—in a tractable task environment. 'Tardast' is Persian for 'juggler' and was inspired by the metaphor that a juggler's feat of spinning plates on vertical poles can represent the concurrent management of multiple tasks that co-exist without pre-defined completion criteria.

The Tardast *interface* consists of multiple progress bars that represent independent tasks competing for an operator's attention (see Figure 1). The status of any particular task (its *satisfaction level*, SL) is visually represented by a vertical progress bar. At any time, the operator can act on a single task by pressing the button underneath a task's status indicator, which improves the task's SL by increasing the corresponding progress bar.

The behavior of each task is governed by three parameters: its rate of decrease when not acted upon (*deviation rate*, DR), its rate of increase when acted upon (*correction rate*, CR), and its value (*weight*, W). Whereas W values are displayed explicitly for each task, task DRs and CRs have to be inferred by observing the rate at which their SLs change over time. While the DRs of all tasks are visible simultaneously, only the CR of the currently engaged task can be observed.

A Tardast *scenario* is defined by a number of tasks and a parameter triple for each task. Shakeri (2003) constructed five scenarios that each contain six tasks, but vary in terms of their overall complexity (Figure 2). Whereas all tasks of Scenario 1 have identical parameter settings, the next three scenarios each vary along one parameter dimension (DR, CR, and W, for Scenarios 2, 3 and 4, respectively), and the tasks of a Scenario 5 vary along all three parameter dimensions. Parameter values were chosen so that scenarios were comparable with respect to optimal strategies and possible scores.

An operator's goal is to maximize the total weighted average score (TWAS), which is calculated as the average task SLs over time, weighted by their respective W-values (see Shakeri & Funk, in press, for details). As it is impossible to keep all six tasks at a maximum level, achieving a high performance score necessitates trade-offs that are informed by the particular task parameters of a given scenario.

We also adopted the conventions that tasks are initialized at intermediate SLs of 50%, SLs of 0% are penalized by being scored at –20%, and scenarios are updated every deci-second and last for 300 seconds. Both the current score (TWAS) and the time remaining are shown at the top of the task interface.

Critique of Previous Research

Previous research has established that operators performed sub-optimally in comparison to the near-optimal solution of a machine-learning solution (Tabu search, see Shakeri, 2003; Shakeri & Funk, in press, for details). Although our use of the Tardast paradigm reflects our appreciation of the authors' general approach and we agree with most of their conclusions, we feel that their argument is currently lacking some details.

First, although practice is believed to be a key factor in the acquisition of task-specific skills, the authors only report the scores of skilled performers *after* practice. Without practice data, it remains unclear which point of the learning curve operators had reached after limited task exposure. The diagnosis of stable sub-optimality would only be warranted if the gain of performance through practice had already reached asymptote. Yet a main effect of scenario number suggests that operators were still learning during the assessment phase.

A second criticism concerns the appropriate standards of comparison for human performance. Unless there is reason to believe that humans have the capability to perform optimally, a diagnosis of sub-optimality with respect to a normative ideal is no more surprising than that humans cannot calculate as well as computers or outrun cheetahs. Shakeri and Funk address this issue by skewing the comparison in favor of humans. But their efforts to allow a fairer comparison by deliberately impairing the tabu search solution (by limiting it to one task-switch per second) and simultaneously 'repairing' human data (by compensating for unused task switching time) blurs the distinction between actual and optimal performance. In our view, the expressed hope that "there is a chance for a skilled participant to beat the tabu score with additional practice and experience" misconstrues the role of normative benchmarks in the assessment of human operator performance. Whereas it is to be expected that humans with a limited amount of experience fail to solve a NP-hard problem (see Shakeri, 2003, for a proof sketch) it would be highly informative to compare human operators with additional benchmarks like baseline performance levels or the outcomes of specific strategies.

Other criticisms include a lack of experimental control (e.g., due to a mix of training and test trials and a confound of task parameters with spatial positions), a failure to account for scenario effects (by refraining from comparisons between

scenarios), and a tendency to rely on anecdotal evidence (or individual cases) to draw somewhat speculative conclusions.

In summary, human performance at a particular level of expertise has been found to be lacking relative to an optimal criterion of performance. However, without data about practice trials, a more systematic exploration of the task environment, and a comprehensive account of what is learned in the process, it seems premature to conclude that Tardast operators persistently perform sub-optimally. More fundamentally, it remains impossible to judge precisely which aspects of operator behavior changed over time and to what extent performance reflected inherent properties of the specific task environment, as opposed to genuine adaptations to it, lack of practice, or basic human cognitive or perceptual-motor constraints.

We address these concerns on two distinct levels. Empirically, we replicate Shakeri and Funk's study with additional experimental controls, and collect learning data that shows the acquisition of task specific expertise over time.

Theoretically, we extend the role of artificial software agents to further explore properties of the Tardast task environment. In addition to comparing human performance to near-optimal solutions, we use random agents that explore baseline performance levels and heuristic agents that assess the consequences of simple strategies. Beyond bracketing the possible range of human performance, this approach will demonstrate that differences in performance between scenarios largely reflect environmental differences and that humans initially barely perform better than baseline. Contrasting human performance with the pure strategies of simple heuristic agents will reveal an even more dramatic failure to act optimally than the previous comparison with a normative performance benchmark.

EXPERIMENT

The primary purpose of this experiment was to replicate and extend the findings of Shakeri and Funk (in press) with additional experimental controls, and assess the performance profiles of operators over repeated task exposures.

Method

Apparatus. Our interface of Tardast matched all functional characteristics of the original software (see Figure 1). The software was implemented in LispWorks 4, ran on a Macintosh G4 computer, and was displayed on a flat-panel display at a 1024-by-768 resolution. (We also collected eye data, but these results will not be reported in this paper.)

The five scenarios that were used featured the parameter settings illustrated above (Figure 2).

Participants. Twelve undergraduate students of RPI participated for course-credit.

Design. This study employed two within-subjects factors of block (3) and scenario (5). Order of scenarios within blocks and positions of tasks within scenarios were randomized.

Procedure. Participants were instructed as in the original study (Shakeri, 2003), tested individually, and completed three blocks of five scenarios in approximately 100 minutes.

Results

Performance. Operator performance varied both as a function of practice and scenario. Figure 3 illustrates the significant performance increase over blocks, $F(1.2, 40.7) = 19.1$, $MSE = 8475.4$, $p < .001$, Huynh-Feldt correction due to sphericity violation. As error bars denote 95%-confidence intervals, the graph shows that subjects significantly improved their performance from Block 1 to 2, but non-significantly from Block 2 to 3. Also displayed is the mean score reported by Shakeri (2003)'s participants after practice, which does not significantly differ from our scores after Block 1.

In addition to the improvement with practice, mean performance scores also vary as a function of the task scenario, $F(4, 44) = 40.9$, $MSE = 7839.9$, $p < .001$. As illustrated by the confidence interval boundaries of Figure 4, Scenarios 1 and 3 yield lower scores than Scenarios 2, 4 and 5. This basic pattern is consistent throughout all three blocks and the scores for the individual scenarios in Blocks 2 and 3 are within the confidence interval of Shakeri and Funk's results.

These results imply relatively rapid learning and support Shakeri and Funk's conclusions insofar as they suggest that the reported sub-optimality was not merely due to a lack of experience. But the differences between scenarios also raise the question *why* operators were more successful in some scenarios than others. Did operators learn more about those scenarios (i.e., acquire scenario-specific *knowledge*), or do these gradients in scores merely reflect environmental differences (i.e., different *baselines*)?

We address the nature of operators' learning by analyzing the behavioral changes over blocks, as well as by contrasting the behavior of two subgroups (by a median split into 'better' and 'worse' operators).

Process. As neither the total amount of actions nor the total time-on-task differed across blocks or between subgroups (all $p > .1$), two typical measures of overall activity fail to account for the observed differences in performance.

An inspection of performance profiles revealed that subtle differences in time-to-task allocations resulted in large

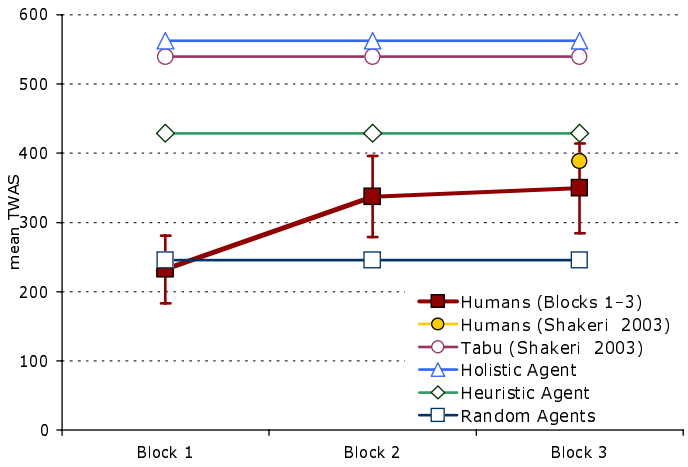

Figure 3: Mean performance of humans and artificial agents by block. (Each block contains 5 scenarios; agents do not learn.)

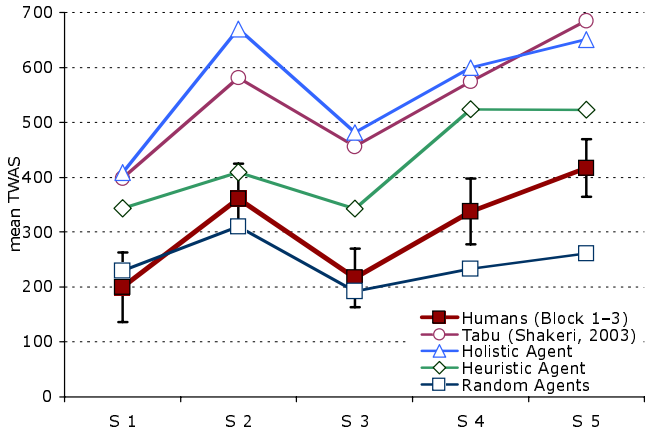

Figure 4: Mean performance of humans and artificial agents by scenario.

differences in scores. Two principle ways in which operators can boost their performance are a) making qualitatively better task choices (by becoming more attuned to important parameter gradients) and b) being quantitatively more selective (by reducing the total number of tasks acted upon).

Both of these (not mutually exclusive) strategies can be shown to be present in our data. To quantify changes in operators' task usage patterns we first computed the relative importance of tasks with varying parameters as their normalized parameter value (to assign an importance of 0 to the 'worst' and 1 to the 'best' task). Weighting these task importance values by the proportion of time allocated to each task and summing the results over all tasks for each scenario yields the proportion of time allocated to important tasks (and can range from 0%, if all time was spent on the least important task, and 100%, if all time was spent on the most important task). To assess whether operators learn to allocate more time to more important tasks and whether this explains the differences between better and worse operators we conducted three ANOVAs with a within-subjects factor of block (3) and a between-subjects factor of overall success (2 subgroups). Significant main effects of block for the scenarios with differences in W and DR show that more experienced operators got better at using gradients in those parameters, but there were no significant effects for scenarios with CR differences. Also, the absence of subgroup effects or interactions indicates that the increased ability to spend more time on more important tasks fails to distinguish between better and worse performers.

As a higher degree of prioritization is identical to exhibiting less entropy, the selectivity in time-to-task allocations over a scenario can be quantified by calculating the *entropy H* of the proportion of time $t(i)$ spent on each task i:

$$H = -\sum_{i=1}^{n} t(i) \cdot \log_2 t(i)$$

A mixed ANOVA to assess the effects of experience and overall success on time-allocation entropy yielded a significant interaction of subgroup × block, in addition to significant main effects of subgroup and block. While all operators became more selective with experience, better operators prioritized even more than worse operators.

Our results so far suggest that operators improved their scores by acquiring task specific knowledge and being increasingly selective, and that the best operators were the most selective ones. However, these analyses do not yet rule out the possibility that differences in task environments may have modulated the results.

Artificial Agents

Any principled assessment of the scope and limits of human behavior must be based on a precise analysis of its environment. Shakeri and Funk (in press) have addressed the issue of normative performance by determining optimal scenario scores through the machine-learning algorithm of tabu search. Although we agree that a normative analysis provides valuable and often indispensable benchmarks (see Gray et al., in press, and Neth et al., 2004, for examples), a critical evaluation of multitasking performance in Tardast requires a more thorough exploration of the task environment.

In addition to determining the optimal performance for specific scenarios in Tardast, we aimed for an assessment of baseline and various intermediate levels of performance by constructing artificial software agents that perform the same tasks as human operators but can be controlled to implement 'pure' strategies. This allows us to further explore the task environment by measuring the impact of strategies unencumbered by the slips and vagaries of human cognition and to systematically explore the effects of variables that may co-determine human performance (like time delays and perceptual-motor constraints).

A Family of Agents. Our artificial agents are characterized on three dimensions: the types of *knowledge* used to evaluate a given situation (e.g., perceptual access to scenario parameters, memory), their *goals* (e.g., maintaining particular SLs), and various *boundary conditions* (like decision cycle times and task-switching parameters) that constrain their performance. We developed three distinct families of agents:

Random agents performed Tardast scenarios by randomly selecting tasks in the absence of any task knowledge. This established performance baselines and their dependency on boundary conditions. Trivially, decision cycle times (or task switching frequency) had an impact on agent performance: choosing a task at every time step would yield a different score than choosing a task only once per trial. As either of these two extremes seems unrealistic, we set the random agent switching frequency to once every ten seconds.

Holistic agents were equipped with super-human abilities and attempted to meet pre-defined performance goals. For instance, the agent in Figures 3 and 4 had perceptual access to all parameters, a rapid decision cycle of .1 sec, and tried to achieve SLs of 99% on prioritized tasks. While not epistemologically plausible, they provide normative (or near-optimal) benchmarks of performance and thus serve the same function as Shakeri and Funk's more sophisticated tabu algorithm. Together, random and holistic agents *bracket* the possible range of human performance.

Heuristic agents assess the effects of specific strategies. Of particular interest are levels of performance of epistemologically plausible agents that implement strategies

that could, in principle, be used by humans. For instance, the heuristic agent shown in Figures 3 and 4 would first 'observe' the rank orders of perceptually salient parameters (DR and W) and then try to raise prioritized tasks to a 90%-SL. Given its moderate decision cycle time of three seconds, human operators could easily have adopted this simple strategy.

Results. Comparisons between humans and random agents reveal that human operators initially perform at baseline (Figure 3). Although their performance across scenarios improves on Blocks 2 and 3, they are not generally better than baseline for Scenarios 1, 2 and 3 (Figure 4).

The parallel pattern of human and random agent data suggests that scenarios are not all created equally. Specifically, high performance scores in Scenario 2 seem to be due to environmental effects rather than operator actions.

Not surprisingly, human operators fail to reach the normative ideal of the holistic agent and tabu scores. More dramatically, humans also fail to reach the level of the simple heuristic agent in all but the second scenario (and this pattern does not change when only considering the human data of Block 3). This failure to match the result of a perfectly achievable heuristic strategy reveals human performance as sub-optimal in an even stronger sense.

CONCLUSIONS

Our explorations of the Tardast task environment support most of Shakeri and Funk's (in press) conclusions, but put them on a firmer empirical and theoretical basis. Operators learn quickly, improve their performance within only two exposures to a scenario, but then tend to asymptote at a sub-optimal level. A contributing factor to this is operators' inability to capitalize on gradients of perceptually non-salient parameters and to focus on only a few prioritized tasks.

Our current conclusions are mixed: on one hand, Tardast has been shown to be an ingenious synthetic task environment that combines many aspects of multitasking behavior, complex system management, and supervisory control in a novel way. As we have only scratched the surface of the paradigm's possibilities, it is a valuable tool for further empirical research.

On the other hand, the complexity of Tardast necessitated a more thorough exploration than previously provided. Only the parallel consideration of human learning data, scenario baselines, and the scores of optimal and heuristic strategies allowed an evaluation of human performance profiles.

Our comparisons with various artificial agents paint human performance results in a more sobering light than the optimistic interpretation provided by Shakeri and Funk (in press). For many scenarios, humans barely performed above baseline and were demonstrably sub-optimal, not only with respect to a normative ideal, but also with respect to a heuristic strategy that operators could easily have implemented.

This leaves us with a puzzle: *Why* did human operators fail to discover this simple heuristic strategy? We recommend that future studies should focus on operators' unwillingness to sacrifice tasks due to fear of penalization, as well as the role and format of the performance feedback provided to operators.

ACKNOWLEDGMENTS

We thank Shakib Shakeri for his help on implementing our version of Tardast, as well as Chris Sims, Mike Schoelles, and Chris Myers for many helpful discussions.

The work reported was supported by grants from the Air Force Office of Scientific Research (AFOSR #F49620-03-1-0143), as well as the Office of Naval Research (ONR #N000140310046).

REFERENCES

Anderson, J. R. (1990). *The adaptive character of thought*. Hillsdale, NJ: Lawrence Erlbaum.

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

Gray, W. D., Neth, H., & Schoelles, M. J. (in press). The functional task environment. In A. Kramer, A. Kirlik & D. Wiegman (Eds.), *Applied attention*. Mahwah, NJ: Lawrence Erlbaum Associates.

Gray, W. D., Sims, C. R., Fu, W.-T. & Schoelles, M. J. (in press). The soft constraints hypothesis: A rational analysis approach to resource allocation for interactive behavior. *Psychological Review.*

McFarlane, D. C., & Latorella, K. A. (2002). The Scope and Importance of Human Interruption in HCI Design. *Human-Computer Interaction, 17 (1),* 1–61.

Moray, N. (1986). Monitoring behavior and supervisory control. In K. R. Boff, L. Kaufman, & J. P. Thomas (Eds.), *Handbook of Perception and Performance*, Vol II (pp 40-1 – 40-51). New York: Wiley & Sons.

Neth, H., Sims, C. R., Veksler, V. & Gray, W. D. (2004). You Can't Play Straight TRACS and Win: Memory Updates in a Dynamic Task Environment. In K. D. Forbus, D. Gentner & T. Regier (Eds.). *Proceedings of the 26th Annual Meeting of the Cognitive Science Society* (pp. 1017–1022). Hillsdale, NJ: Lawrence Erlbaum.

Rogers, R. D., & Monsell, S. (1995). Costs of a predictable switch between simple cognitive tasks. *Journal of Experimental Psychology: General, 124,* 207–231.

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

Shakeri, S., & Funk, K. (in press). A Comparison of Human and Near-Optimal Task Management Behavior. *Human Factors*.

Simon, H. A. (1990). Invariants of human behavior. *Annual Review of Psychology, 41*, 1–19.

Wickens, C. D. (1992). *Engineering Psychology and Human Performance* (2nd ed.). New York, NY: HarperCollins Publishers Inc.