A Comparison of Human and Near-Optimal Task Management Behavior

Shakib Shakeri and **Ken Funk,** Oregon State University, Corvallis, Oregon

Objective: The primary contribution of this work is the development of an abstract framework to which a variety of multitasking scenarios can be mapped. The metaphor of a juggler spinning plates was introduced to represent an operator performing multiple concurrent tasks. **Background:** This allowed seeking a quantitative model for management of multiple continuous tasks instead of a model for completing multiple discrete tasks, which was considered in previous studies. **Methods:** The multitasking performance of 10 participants in five scenarios was measured in a low-fidelity simulator (named Tardast), which was developed based on the concept of the juggler metaphor. This performance was then compared with a normative model, which was a near-optimal solution to a mathematical programming problem found by tabu search heuristics. **Results:** Tabu outperformed the participants overall, although the best individual performance nearly equaled that of tabu. It was also observed that participants initially tended to manage numerous tasks poorly but that they gradually learned to handle fewer tasks and excel in them. **Conclusion:** This suggests that they initially overreacted to the penalization associated with poor performance in the software. Participants' strategic task management (e.g., what tasks to handle) was more significant in obtaining a good score than their tactical task management (e.g., how often to switch between two tasks). **Application:** Potential applications include better design of equipment, procedures, and training of operators of complex systems.

INTRODUCTION

Imagine a juggler who has several plates in a row, each on a vertical pole. The juggler's goal is to have all the plates spinning as smoothly and as long as possible. Each plate requires the juggler's frequent attendance and monitoring. The more or less a plate is attended to, the more smoothly or wobbly it turns. The task of attending to a plate is never completed solely because it was attended to once earlier. Therefore, the number of completed tasks would not represent the juggler's competence. A better measure would be how smoothly the plates have been spinning on average over time.

In this paradigm, the juggler represents an operator and spinning a plate represents performing a task. A *satisfaction level* (SL) is assigned to each task for its current state. For plates, a minimum (zero) SL equals not spinning at all, or the poorest task state, and a maximum (100%) SL equals spinning perfectly, or the desired task state. The rate that a task deviates from its desired state if not attended to is the *deviation rate* (DR). Conversely, the rate that it approaches the desired state while attended is the *correction rate* (CR). Tasks may also have different values or *weights* (Ws) – for example, a fine china plate is more valuable than a plastic plate. Therefore, the objective of the operator would translate into maximizing the weighted average SL across tasks over time. In this study, the term *attendance* is used for working on a task in the sense of increasing its SL. Moreover, the term *monitoring* is used to refer to the careful observation of a task to evaluate its SL and parameters (DR, CR, and W).

A wide variety of multitasking situations can be described by this metaphor. A pilot who manually maintains the altitude, speed, and heading of a plane at desired values can be viewed as a

Address correspondence to Shakib Shakeri, Nuclear Safety Solutions Limited, 4th Floor, 700 University Ave., Toronto, ON M5G 1X6, Canada; shakib.shakeri@lifetime.oregonstate.edu. ***HUMAN FACTORS,*** Vol. 49, No. 3, June 2007, pp. 400–416. DOI 10.1518/001872007X197026.

juggler with three plates. For example, when the altitude is adjusted exactly to the air traffic controller's clearance, it holds a 100% SL. The less attention this task receives, the more it deviates until it reaches the zero (minimum) SL. In this example, DR represents the rate at which a parameter moves away from its desired value when the pilot does not attend to the task; CR represents how responsive the parameter is to action of the pilot or the amount of time or effort required for returning the current task state back up to the desired state. Relative value or weight (W) is the relative importance assigned to a task, with the W for an altitude control task being higher than the W for, say, a task to adjust a radio frequency.

The common factor among a pilot, a nuclear power plant operator, a commander in charge of several units, and a corporate manager is that a single person has to monitor multiple concurrent tasks and make decisions about them. As the complexity of systems increase, so does the importance of good task management. In their study of cockpit accidents and incidents, Chou, Madhavan, and Funk (1996) demonstrated that cockpit task management error occurred as a frequent and significant type of human error in the aviation domain. This error manifests itself as attending to a less important (and/or less urgent) task at the expense of a more important (and/or more urgent) task. To avoid such errors, researchers have tried to analyze how well, in general, humans manage multiple concurrent tasks. To find a standard of comparison (a normative model) for human multitasking, several researchers have made use of mathematical models.

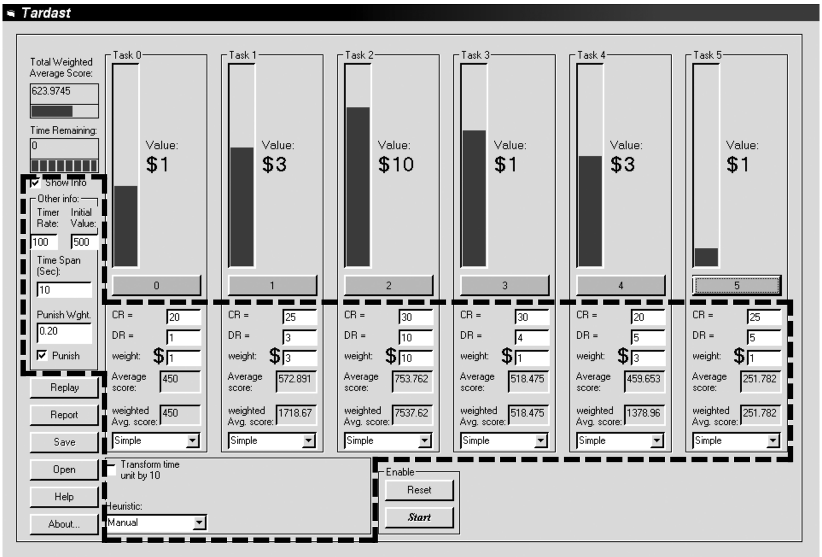
In the next section (Task Description), we discuss a software program that was developed based on the concept of the juggler metaphor for conducting task management experiments. In the subsequent section (Mathematical Programming: A Normative Standard of Comparison), we review other researchers' mathematical approaches and then explain our normative model for human multitasking, which focuses on quality of performance rather than task accomplishment. In the Method section, we discuss how the software program was used in the experiments to measure the performance of participants. In the Results section, we statistically compare the participants' performance data with those of the normative model and then qualitatively analyze the participants' performance graphs. Finally, in the last section, we give a summary of the research, along with our conclusions, general observations, and recommendations. We believe the framework presented in this paper is the primary contribution of this research, as it can be used to model many instances of real-life multitasking environments.

TASK DESCRIPTION

A computer program named Tardast (Persian for juggler) was developed to allow participants to manage tasks in a low-fidelity task management environment. The software had the capability of recording the performance of participants for further analysis. The participants observed six tasks represented by bars in the software interface (Figure 1). The dotted area in Figure 1 was used only by the experimenter, and it was hidden from the participants during an experiment.

A task was attended by simply depressing the button underneath the bar using the mouse. As long as the left mouse button was held down, the SL of that bar rose at a CR up to a maximum of 100% SL. On the contrary, not attending to a task allowed the bar to drop at a DR down to a minimum of zero SL. If a task reached or stayed at zero SL, however, the software penalized it, and the penalty increased proportional to the length of stay. The details of the scoring mechanism and penalization will be discussed further.

The requirement to hold down the button (e.g., as opposed to making a single click) was to enforce a closer relationship between a task that is attended and a task that is merely monitored, although it was still possible to work on one task and monitor the others. Also, as long as the mouse button was held down, the participant could move the mouse pointer to another task without interrupting the SL improvement in the first task. This allowed the switching cost, in terms of the time it takes between ending attendance to one task and starting attendance to another task, to be as small as zero. To display the rise and fall in a task's SL continuously and smoothly, the software updated the system state every 0.1 s. As a result, the participants had to spend at least 0.1 s continuously on a task attended, or in other words, they could theoretically switch between the tasks as quickly as 10 tasks/s. In terms of the decision-making, this allowed the participants to decide every 0.1 s whether



Total Weighted Average Score: 623.9745

Time Remaining: 0

☑ Show Info

Other info:

Timer Rate: 100

Initial Value: 500

Time Span (Sec): 10

Punish Wght. 0.20

☑ Punish

	Task 0	Task 1	Task 2	Task 3	Task 4	Task 5
Value:	$1	$3	$10	$1	$3	$1
CR =	20	25	30	30	20	25
DR =	1	3	10	4	5	5
weight: $	1	3	10	1	3	1
Average score:	450	572.891	753.762	518.475	459.653	251.782
weighted Avg. score:	450	1718.67	7537.62	518.475	1378.96	251.782
	Simple	Simple	Simple	Simple	Simple	Simple

□ Transform time unit by 10

Heuristic: Manual

Figure 1. Tardast interface; the dotted area was hidden from participants. Adapted from *2003 IIE Annual Conference Proceedings,* S. Shakeri & K. Funk, A comparison between humans and mathematical search-based solutions in managing multiple concurrent tasks, [CD-ROM] (2003), with permission from the Institute of Industrial Engineers.

to switch tasks or continue attending to the same task.

A task's W was visible to the participants as "Value" with a dollar amount. However, the DR and CR for every task had to be estimated by operating the software (game). The participant's score was a function of the average SL of tasks, how long tasks were penalized, and the balance of attention between higher and lower W tasks. This score and the time remaining to the end of the game were continuously recalculated and displayed on the top left side of the screen.

The scoring mechanism can be mathematically stated as

$$z = \frac{1}{(3001)(\sum_{i=0}^{5} W_i)} \times \sum_{t=0}^{3000} \sum_{i=0}^{5} (W_i)[SL_{it} - 0.20(1000)Int(1 - \frac{SL_{it}}{1000})], \qquad (1)$$

in which z is the participant's final score; W_i is the W of task i, $i = 0, 1, 2, 3, 4, 5$; and SL_{it} = SL of task i at time t, $i = 0, 1, 2, 3, 4, 5$, $t = 0, 1, 2, \ldots, 3000$.

In other words, the scoring mechanism consisted of a virtual snapshot of the screen at every time unit, considering the SL of each task as the score of that task for that time, which could vary between 0 and 1000 (100%). For each task, the summation of these scores across time divided by the time elapsed, or the number of snapshots taken up to that moment, was called the average score across time for that task (shown as "Average score" in Figure 1). This number, when multiplied by the task's W, was called the weighted average score across time for that task (shown as "weighted Avg. score" in Figure 1).

The summation of weighted average scores across time across tasks divided by the total W of all tasks (which can be calculated as 19 in Figure 1) was displayed as "Total Weighted Average Score" on the top left side of the interface. This number, again with a maximum of 1000, was the participant's score, which was to be maximized. This method of scoring captured the essence of concurrent task management performance by linearly rewarding satisfactory task performance proportional to the task's W. All tasks started at 50% SL at time zero (i.e., $SL_{i0} = 500$).

A task at zero SL not only added no value but was penalized by being assigned a negative score (instead of zero) for every time unit that it remained in that state. The penalty factor had to be determined carefully, as it could change the optimal behavior. An excessively low penalty would suggest ignoring low-W tasks in favor of excelling in higher W ones. On the other hand, too high a penalty might encourage barely avoiding zero SL in as many tasks as possible. This approach would result in excelling in no task or performing poorly in all. A moderate penalty factor of 20% of the task's worth was chosen for this study to allow a balance between the two extreme conditions. For example, a task with W = 4 would have a negative weighted score of –(20%)(4)(1000) = –800 for one time unit the instant it hit zero SL and for every time unit it continued to stay at zero SL. However, the same task would have a positive weighted score of +(40%)(4)(1000) = +1600 for one time unit if it were at 40% SL.

MATHEMATICAL PROGRAMMING: A NORMATIVE STANDARD OF COMPARISON

Literature Review

Pattipati and Kleinman (1991) and the National Research Council (1998, pp. 112–128) reviewed engineering models and theories of multitasking. The focus of engineering theories, unlike psychological theories, is on the outcome of a human decision rather than its underlying psychological or cognitive causes. Note that in the multitasking models we will review as well as the model for the juggler metaphor introduced in this paper, the focus is not on *why* an operator attends to certain tasks and the effect of factors such as memory, psychomotor skills, or stress on the operator's decision. Rather, the interest is in *how* the human operator attends to multiple tasks – that is, which tasks and how many are handled, in what order they are handled, how well they are handled, and how the operator performs overall.

The reviews of quantitative multitasking models begin with queuing theory, which is the mathematical study of systems described in terms of a few servers and many customers who arrive randomly and wait in lines (queues) to be serviced. Two typical application areas are the design of traffic control and phone answering systems, in which intersections and operators are the servers, and cars and phone calls, respectively, are the customers. The designer of such systems typically wants to minimize the mean customer waiting time for being serviced while employing as few servers as possible, for economic reasons. Carbonell, Ward, and Senders (1968) used queuing theory to model a human operator's visual attention as a single server and different displays to be monitored as customers in the queue. Other researchers continued along this line, modeling more generalized server service times and/or task (or customer) arrival times (Chu & Rouse, 1979; Greenstein & Rouse, 1982; Walden & Rouse, 1978).

Tulga and Sheridan (1980), on the other hand, used dynamic programming with the branch-and-bound approach as a form of deterministic combinatorial optimization to model a multitasking environment. Dynamic programming is used to solve an optimization problem over discrete stages in time in which, at each stage, a decision is made out of a few alternatives; branch and bound is a method for finding optimal solutions in such discrete optimization problems. In their model, tasks arrived in groups with random rewards in different queues. The solution to the model was the task attendance that maximized the operator's aggregate reward earned for completing tasks.

A different form of deterministic quantitative theory, known as scheduling theory, has been advocated by Dessouky, Moray, and Kijowski (1995) for the analysis of human strategic decision making. Scheduling theory addresses the problem of optimally sequencing jobs to be processed by one or more machines to achieve a certain goal, such as minimizing the completion time of all jobs. Moray, Dessouky, Kijowski, and Adapathya (1991) used this theory to find the optimal attendance sequence for a human operator to attend to tasks. They considered, in scheduling terminology, the operator as a single machine and the tasks to be processed as jobs with specific due dates. The objective was to maximize the number of tasks completed by their due dates. Partially completed tasks were not counted, which also meant ignoring tasks with no chance of completion on time.

A number of researchers have used models based on estimation theory and optimal control theory (Kleinman, Baron, & Levision, 1970; McRuer, 1980; Pattipati & Kleinman, 1991; Rouse, 1980). These theories estimate the future state of the system and find the optimal method of reaching a desired state so as to maximize or minimize certain system criteria. A typical example is to optimally bring a vehicle to a straight line by determining the turning angles of the steering wheel to the left and right, in a series of actions, based on the feedback received from the vehicle's position.

Unlike those theories, however, the assumption in this research is that as long as a task is attended, it will progress toward the desired state with a constant rate, but not necessarily optimally. It is which task to attend at any time that is of concern here, not how a particular task should be attended to improve its state. This aspect of decision making is relevant in the realm of human supervisory control, as opposed to the manual control domain.

The quantitative theory used for the purpose of this research is scheduling theory. It is superficially similar to that of Moray et al. (1991) in considering the operator as a single machine that can preemptively attend to tasks, one at a time. However, it is significantly different from the studies previously noted in that it assumes the juggler paradigm for the multitasking environment – that is, it seeks a model for management of multiple continuous tasks instead of a model for completing multiple discrete tasks.

Normative Model: Standard of Comparison

Normative models based on quantitative methods have increasingly been of interest to behavioral researchers in relatively recent years. The National Research Council (1998) reported the findings of a multidisciplinary team of qualitative and quantitative researchers on human behavior representation in military simulation applications. Also, federal agencies such as the Air Force Office of Scientific Research (2007, p. 27) have been allocating funds for the research to quantitatively model different aspects of human performance, in order to provide optimal rules for performing tasks such as human decision making and resource scheduling while subject to constraints of attention and memory.

In this research, to determine how well humans manage multiple tasks, we required a standard of comparison, a normative model of, ideally, how operators should manage multiple concurrent tasks (Wickens, Lee, Liu, & Gordon-Becker, 2004, pp. 157–158). We sought a mathematical model yielding optimal or (if the optimum was mathematically intractable) near-optimal allocation of attention among tasks over time so as to maximize the total weighted average score over time. The comparison of human performance with optimal or near-optimal performance could not only tell us how well our participants performed but also give insights into good task management strategies. In order to provide a meaningful comparison, the normative model should be subject to common human limitations – for example, a normative model that switches tasks every 0.1 s for duration of 3000 s performs beyond reasonable human performance.

A mathematical model for the multitasking environment discussed earlier (the juggler paradigm) was developed, formally a mixed (binary) integer linear programming (MILP) model (Shakeri, 2003; Shakeri & Logendran, 2007). Although the details are beyond the scope of this paper, MILP can be briefly described as a subcategory of linear programming, which in turn is a subcategory of mathematical programming (MP). MP is a technique in operations research to solve a mathematically described problem (mathematical model) with the purpose of finding an optimal value. The goal of MP, represented as an objective function, is commonly expressed in terms of maximizing benefit (or minimizing cost), subject to several feasibility constraints. The constraints usually reflect limited capacities (Hillier & Lieberman, 2005).

The mathematical model's objective function in this study was to maximize the weighted average SL across tasks over time, hereafter called the score. The model's constraints limited the maximum and minimum SL of tasks to 100% and zero, respectively. The constraints also enforced task behaviors (i.e., to improve by a CR when attended or to deteriorate by a DR when not attended). Additionally, the constraints deducted a penalty factor from the score for every time unit that a task stayed at zero SL. This was to lower the value of the model's solutions that suggested ignoring a task for too long.

This mathematical model is believed to belong to the class of NP-hard (nondeterministic polynomial time – hard) problems, for which the computational complexity of the solution algorithm increases as a nonpolynomial (e.g., exponential) function of the problem size. This characteristic causes an exceedingly long wait for computing the optimal solution by the computer as the problem size grows. Such problems are generally solved by specific heuristics nearly (vs. truly) optimally, but in a reasonable time. One such heuristic is the *tabu search* method, which is a complex iterative procedure to solve hard optimization problems nearly optimally. The tabu search heuristic has several parameters, which can be manipulated by trial and error to effectively search and evaluate the solution space using the search history without falling into the trap of local optima. The details of the tabu search meta-heuristic are described by Glover (1990).

Previous studies in the manufacturing domain have shown excellent performance of tabu search heuristics in other mathematically formulated complex scheduling problems. In the current research, an algorithm based on the tabu search concept was developed and implemented in Microsoft Visual Basic 6.0 to solve the problem at hand, which was similar to scheduling problems. The near-optimal solutions found by this algorithm were proven to be very close to the true optimal solutions, where they were available (Shakeri, 2003; Shakeri & Logendran, 2007). These near-optimal solutions will hereafter be called *tabu solutions.*

The tabu solutions became a standard of comparison for the operator's competence in management of multiple concurrent tasks. A solution was a time series of ordered pairs in which an ordered pair (*i*, *t*) meant attendance to task *i* at time *t* for one time unit; only one task could be attended at a time. For instance, attendance to three tasks (0, 1, and 2) for 10 time units might be represented in a solution as {(2, 0), (2, 1), (0, 2), (1, 3), (1, 4), (—, 5), (0, 6), (1, 7), (1, 8), (1, 9)}. The ordered pair (0, 2) means attendance to Task 0 at Time 2 for one time unit, and (—, 5) means no task was attended at Time 5 for one time unit. The actual problems solved in this study had six tasks and 3000 time units. As the time unit in the game was designed to be 0.1 s, this resulted in a task management standard for a time span of 5 min. Details of the problems solved in different scenarios, and how they were compared with human performance, will be explained in the next section.

METHOD

Apparatus

The Tardast software, as described in the Task Description section, was coded in Microsoft Visual Basic 6.0. The computer used for the experiment had a Pentium II/300 MHz processor, 64 MB RAM, and the Windows NT 4.0 operating system. The software recorded the attendance of the participants to tasks and states of the tasks every 0.1 s. These data were later used to replay and review the participant's performance on the computer in real-time, faster than real-time, and slower than real-time rates. The fast mode had a zoom-out effect and helped to determine the participant's overall strategy. In contrast, the slow mode had a zoom-in effect and helped to identify the parameters behind a certain action – for example, why the user switched from one task to another at a specific time. Further, the data were used to generate graphs of attendance using a Visual Basic Macro in a Microsoft Excel 2000 spreadsheet (see Figures 2–4). Statistical analysis of the data was performed using the STATGRAPHICS Plus 5.0 statistical package.

Participants

Ten students at Oregon State University, two women and eight men not selected based on gender, participated in this research voluntarily. They signed a consent letter that explained matters such as the intent of the experiment, risks and benefits, and confidentiality of identity. The participants were observed and their data collected while they played the game. At the end of the experiment, they were given a short questionnaire on their task attendance strategy in the experiment. For each participant, the total length of the experiment was less than 2 hr, with repeated breaks every 5 min. The total net time of the data collection for playing the game on the computer was 75 min.

Procedure

The participants were given written instructions on the mechanics, scoring, score penalization, and intent of the experiment. Verbal comments were also given before the start of the game on how to play it, but following a script to maintain consistency. The participants were told to try to obtain the highest possible score using any strategy they desired. The experiment consisted of five different scenarios.

For one scenario, each of the three parameters (DR, CR, and W) varied across the six tasks, making it the most complex scenario (see Table 1). For three of the scenarios, two parameters were fixed but one parameter (DR, CR, or W) was varied. This allowed investigation of the effect of varying one parameter. Finally, for one scenario, all three parameters were fixed across six identical tasks, making the effect of task selection minimal,

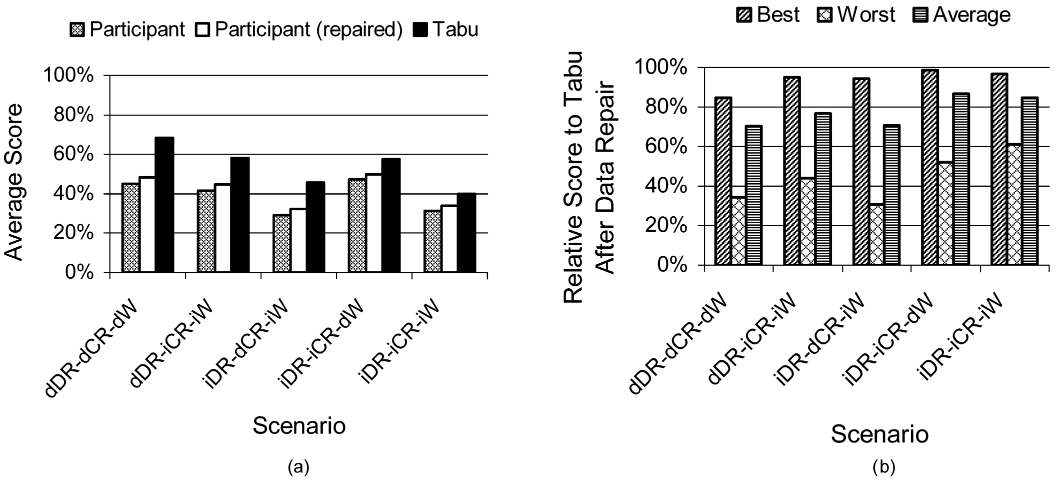


Figure 2. Summary of participants' scores in each scenario (*dDR-dCR-dW*: different deviation rates, correction rates, and weights across tasks; *dDR*-iCR-iW: different deviation rates across tasks; iDR-*dCR*-iW: different correction rates across tasks; iDR-iCR-*dW*: different weights across tasks; and iDR-iCR-iW: identical tasks). (a) Average original and repaired participants' scores and corresponding tabu scores. (b) Participants' best, worst, and average scores relative to tabu scores after the data repair.

and this was used as a baseline scenario. The goal of these scenarios was to study the reaction of the participants to variations in each of the parameters across six tasks. It was also of interest to see how close the participants could get to the near-optimal tabu behavior as the complexity of the scenarios changed.

All of the scenarios were assigned to each participant in random order to avoid a learning effect in the data across the scenarios. In the transition from one scenario to the next, the participants were informed only that the scenario had changed. They were not provided any further information on what had changed because the characteristics of each scenario were meant to be learned by playing the game. Outlined in Table 1 are the design of scenarios, the numeric structure of the parameters for six tasks (0, 1, 2, 3, 4, 5) in each scenario, and the most successful task attendance strategy in each scenario. For example, in *dDR*-iCR-iW (short for Scenario *dDR*-iCR-iW), different (*d*) *DRs* for Tasks 0, 1, 2, 3, 4, 5 were, in order, 2, 1, 3, 3, 4, 2, whereas their identical (i) CRs were 9 and their identical (i) Ws were 5. The DR and CR numbers in Table 1 are rates per 0.1 s, and W is the relative task importance.

It can be seen with careful examination of the numerical structure in Table 1 that the seemingly arbitrary numbers for each scenario have the same average of parameters, approximately DR = 3, CR = 9, and W = 5. It will be shown in the results section that this structure ensured that roughly four out of six tasks in each scenario could be kept at a high SL if the other two tasks were ignored (shed). Therefore, there was no great variation among the scenarios as to the number of tasks that could be handled. It also ensured that in each scenario, a successful strategy, which had to be discovered by the participant, necessitated ignoring the two tasks with the lowest contribution to a high score (i.e., tasks with low Ws, high DRs, and low CRs).

Scenarios were each 5 min long, but prior to every scenario the participants had up to 10 practice trials, each 1 min long. The trials were designed to let the participants become familiar with dynamics of each scenario and to provide them a chance to find the most successful strategy of attendance for that scenario. The details of attendance in practice trials were not recorded, but the final score in each trial was recorded to monitor the progress of the participant. If the two highest scores in practice trials were within 2% of each other, the practice trials were halted if the participant desired; only 6% of the possible practice trials were not taken.

After the practice trials, data were collected for one 5-min session of playing the game in the same scenario. These detailed data included attendance of the participant (i.e., which task at what time was attended by depressing its button) and the state (SL) of the tasks for every 0.1 s. Finally,

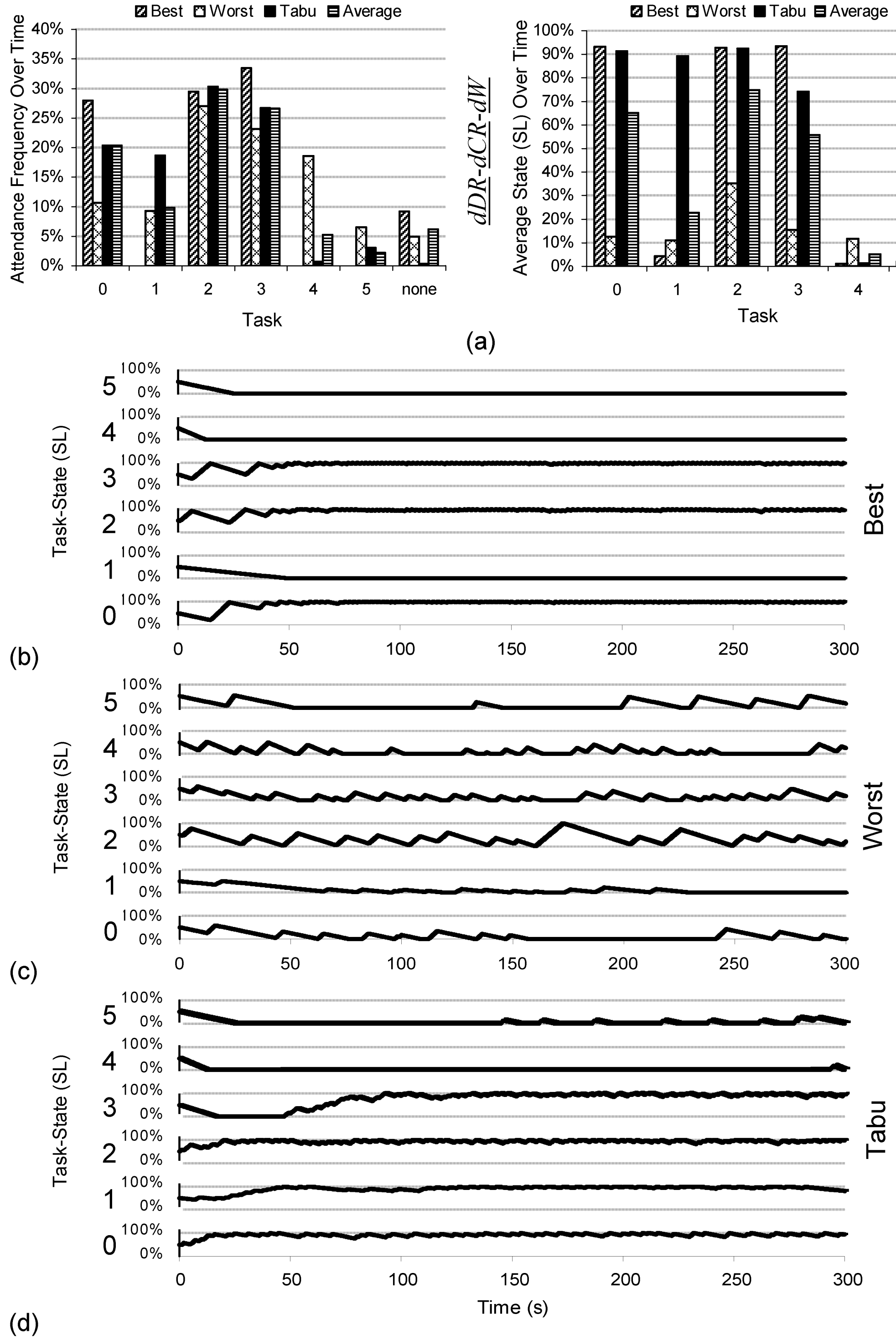


Figure 3. (a) Attendance frequency (%) and average state (SL) over time for each task for the best, worst, tabu, and average participant performance. Moment-to-moment (b) best participant performance (score: 570) and (c) worst participant performance (score: 156). (d) Tabu performance (score: 685) over time (seconds) in Scenario *dDR-dCR-dW* (different deviation rates, correction rates, and weights across tasks). Adapted from *2003 IIE Annual Conference Proceedings,* S. Shakeri & K. Funk, A comparison between humans and mathematical search-based solutions in managing multiple concurrent tasks, [CD-ROM] (2003), with permission from the Institute of Industrial Engineers.

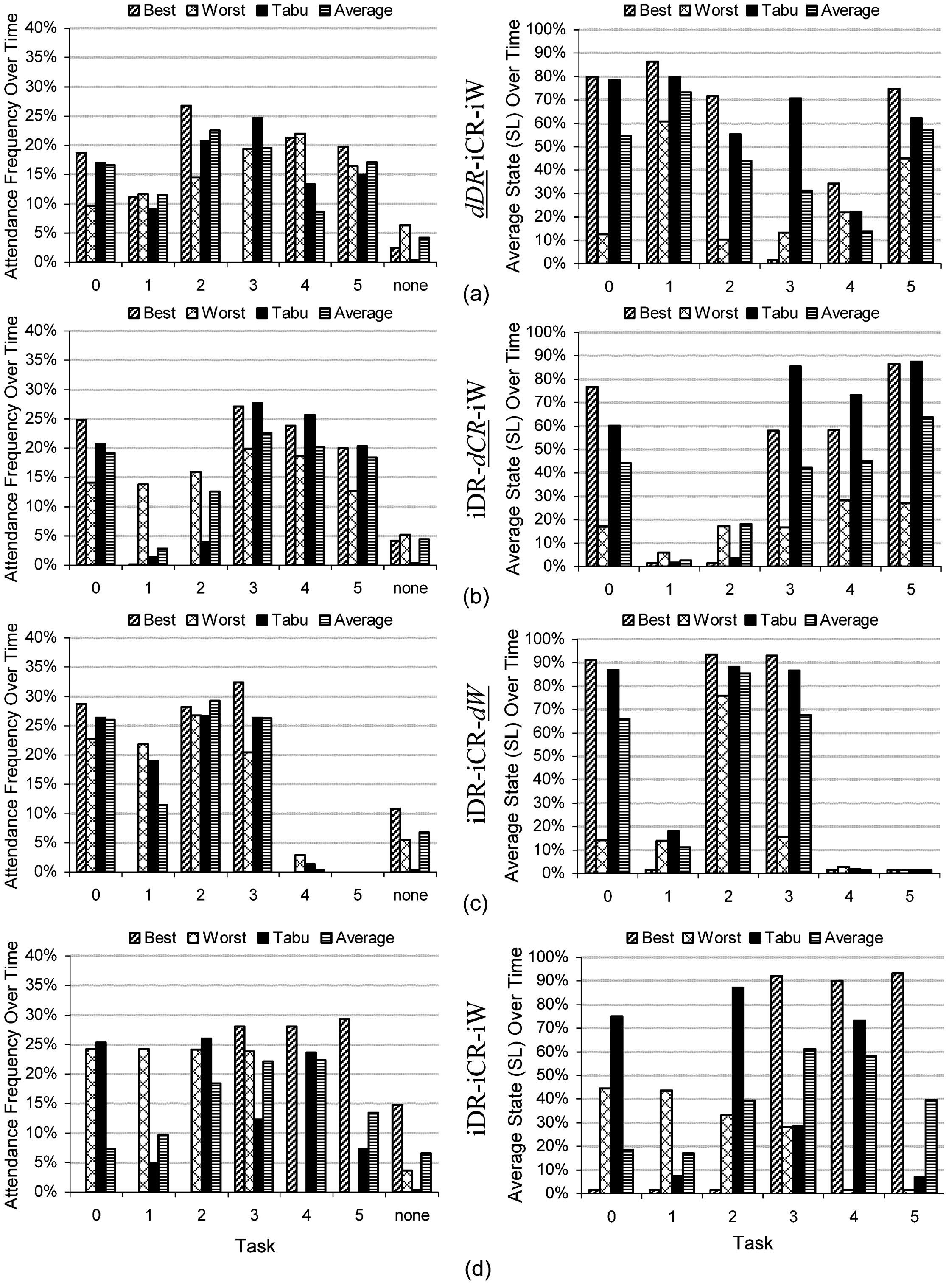


Figure 4. Attendance frequency (%) and average state (SL) over time for each task for the best, worst, tabu, and average participant performance in Scenario (a) *dDR*-iCR-iW (different deviation rates across tasks), (b) iDR-*dCR*-iW (different correction rates across tasks), (c) iDR-iCR-*dW* (different weights across tasks), and (d) iDR-iCR-iW (identical tasks).

TABLE 1: Design of Scenarios With Regard to Six Tasks' Deviation Rates (DRs), Correction Rates (CRs), and Weights (Ws) and the Most Successful Strategy Recommended for Each Scenario

	Parameter Across Tasks			
Scenario	DRs	CRs	Ws	Most Successful Strategy
dDR-dCR-dW	Different [2, 1, 3, 3, 4, 2]	Different [9, 5, 8, 8, 9, 13]	Different [6, 5, 10, 6, 4, 2]	Attend to the most important (highest W) tasks if there is no significant difference between their DRs and CRs.
dDR-iCR-iW	Different [2, 1, 3, 3, 4, 2]	Identical 9	Identical 5	Discover those tasks that deteriorate least quickly (lowest DRs) and attend to them; attend to others only as time permits.
iDR-*dCR*-iW	Identical 3	Different [9, 5, 8, 8, 9, 13]	Identical 5	Discover the tasks that are most responsive to your efforts (highest CRs) and attend to them; attend to others only as time permits.
iDR-iCR-*dW*	Identical 3	Identical 9	Different [6, 5, 10, 6, 4, 2]	Attend to the most important (highest W) tasks; attend to others only as time permits.
iDR-iCR-iW	Identical 3	Identical 9	Identical 5	Pick a few tasks and keep them satisfactory.

Adapted from *2003 IIE Annual Conference Proceedings,* S. Shakeri & K. Funk, A comparison between humans and mathematical search-based solutions in managing multiple concurrent tasks, [CD-ROM] (2003), with permission from the Institute of Industrial Engineers.

after the participants completed all of the scenarios in the experiment, they filled out a questionnaire on their age and hours per week of computer use and driving (a type of multitasking). At the end of the questionnaire, they explained the strategy they used (if any) when attending to tasks in the experiment.

RESULTS AND DISCUSSION

Data Adjustments and Considerations

The tabu search algorithm was run two to four times in each scenario with different search parameters. This allowed the algorithm either to intensify the search in spaces where historically good solutions were found or to diversify the search to spaces where solutions were historically less explored. The best score obtained was selected as the tabu score for that scenario, and therefore, unlike the participants' scores, the tabu score had a zero variance within each scenario.

The participant's score was the numerical score calculated by the computer after 5 min of playing the game. Two adjustments were made to enable drawing reasonable inferences from the comparison of participants (human) with tabu search (computer). These adjustments focused the comparison on the decisions made regarding which tasks to handle and how they were handled instead of on human delays in movement or lapses. The first adjustment was to handicap the program by limiting tabu's task switching rate to a maximum rate of 1/s, which was 10 times slower than the rate allowed for the participants. In other words, tabu made a decision as to whether or not to switch tasks every 1 s, versus every 0.1 s allowed for humans. This adjustment compensated for relatively slow human movement and decision making, so that a very skilled human would have a chance of achieving the tabu performance.

Because tabu, unlike the participants, had lost no time while attending to the tasks in the software, the second adjustment (extra bonus) was to *repair* the participants' data for those times (1.5%–14.7% of the time, $M = 5.6\%$) that the participants did not attend to any task. Although technically the software allowed the switching cost (or moving time between two tasks) to be as small as zero, nonattendance resulted from the participants' uncertainty of which task to attend to next or just simply failing to hold down the mouse button. This gap in the data was repaired with the number of the task that the participants actually attended to next, which effectively reduced switching time to zero.

For example, the gap in the time series 0, 0, [?], 1, 3, … indicated that no task was attended for one time unit while switching from Task 0 to Task 1; this gap was filled with "1" for Task 1. In turn, a new numerical score was recalculated based on the repaired data and was called the *repaired score*. This score was on average 10.1% higher than the original score, and it was used as the participant's score in the statistical comparisons henceforward.

One consideration in reflecting on the experiments is that the order in which scenarios (see Table 1) were assigned to participants was random. However, for each scenario, the physical left-to-right order of tasks (i.e., bars in Figure 1) was fixed; in retrospect, it should have been random. The fixed order of tasks may have contributed to the participants gaining better scores (relative to those of the other participants) as they advanced through the scenarios, $F(1, 48) = 23.13, p < 0.01$, $r_s = .57$, for a test of the null hypothesis that the slope of the fitted line is zero. Nonetheless, this research does not focus on how much or why a participant's score improved by going through the scenarios. The focus is on the best strategies the participants developed (as to which tasks to handle or shed) and the effectiveness of their strategies when compared with the optimal strategies (as judged by tabu) for each scenario. Therefore, the primary statistical and general behavioral conclusions in this study should be robust and reliable for the population of participants.

Statistical Comparison of Tabu and Participants

Because the parameters (DR, CR, and W) of the scenarios were different, tabu and participants also obtained different scores between the scenarios. Therefore, the scores could be compared only by a pairwise comparison between that of the participant and tabu in each scenario. The following two null hypotheses were statistically tested using a two-sample pairwise comparison at a 95% confidence level: (a) The mean for the pairwise difference between the participant's score and the tabu score is zero when combining all of the scenarios ($n = 50$, $M = 121.83$, $SD = 105.12$); and (b) the mean for the pairwise difference between the scores of participants and that of tabu is zero when evaluating each scenario ($n = 10$) independently. In top-to-bottom order of scenarios in Table 1, means (and *SD*s) were 203.66 (127.46), 134.78 (101.41), 133.65 (92.65), 76.55 (90.68), and 60.51 (50.86), respectively.

The first null hypothesis was rejected with 95% confidence, $t(49) = 8.20, p < .01$ (two sided); that is, there is substantial evidence that tabu outperformed the participants. The second null hypothesis was also rejected for every scenario with 95% confidence. The t statistic (and two-sided p) for every scenario in Table 1, in order, was $t(9) = 5.05$ ($p < .01$), 4.20 ($< .01$), 4.56 ($< .01$), 2.67 ($= .03$), and 3.76 ($< .01$); that is, there is strong evidence that tabu outperformed the participants in every scenario independently. It should be noted that these results were obtained despite repairing the participants' data and effectively "allowing" them to switch between the tasks up to 10 times faster than tabu.

Figure 2a illustrates the differences between the participants' original, repaired, and corresponding tabu scores in each scenario. These scores should be compared within a scenario, but not between the scenarios, as each scenario had its own parameters. Figure 2b shows that the best participant's score in each scenario came within 1% to 15% of the tabu score. More discussion of the data in Figure 2b will follow, in the General Discussion section.

Performance Analysis in Scenarios

Besides the statistical comparison of scores, there was yet much to be learned from the comparison of the pattern of attendance between the participant and tabu search. This comparison was made using several different methods. The attendance of all 10 participants was observed while they played the game, and it was replayed and reviewed on the computer after the game. Further, the tabu performance was simulated in the game environment. That is, the tabu solution was used to drive the program in real time so that it could be visualized and qualitatively compared with human performance. Additionally, the participants' verbal comments while playing the game and their written comments on the questionnaire after the game were reviewed with respect to the task attendance strategy that they had in mind while attending to tasks.

Finally, several different types of graphs were generated to summarize different aspects of attendance (see Figures 3 and 4). When generating the graphs for performance analysis, we intentionally used the participants' original data (not repaired) to preserve the uniqueness of the participants' behavior. Performance is depicted in two forms: attendance frequency and average SL. The former is the percentage of time (out of overall 300 s) that a task was attended altogether; the latter is the average SL gained for a task over time resulting from the attendance. It should be noted that two identical attendance frequencies may not result in the same average SLs for two identical tasks. The reason is that one task might be attended only when its SL is low, thus resulting in low average, whereas the other task gains a high average because it is usually attended when its SL is already high. To address this aspect of performance, we generated moment-to-moment performance graphs for all scenarios, but here we present only those for *dDR-dCR-dW,* which was the most complex scenario (see Figures 3b, 3c, and 3d).

The best, worst, and average performance analysis of the participants and that of tabu are presented in the following passages for each scenario. Further, looking at the performance of all 10 participants and that of tabu in each scenario, we provide a strategy that is thought to give the best results, called the "most successful strategy" for that scenario.

Scenario dDR-dCR-dW: *Different DRs, CRs, and Ws.* This was the scenario in which all of the parameters were different (DR = [2, 1, 3, 3, 4, 2], CR = [9, 5, 8, 8, 9, 13], W = [6, 5, 10, 6, 4, 2]). The DR, CR, and W of the tasks were identical to *dDR*-iCR-iW, iDR-*dCR*-iW, and iDR-iCR-*dW,* respectively. The performance of the participants and that of tabu are illustrated in Figure 3. Figure 3a shows the attendance frequency (percentage) and average state (SL) of tasks over time in two graphs. Figures 3b, 3c, and 3d illustrate, in order, the moment-to-moment task attendance for the best performer, worst performer, and tabu. Each of those three graphs consists of six parallel, smaller, horizontal graphs, in which the zigzag line between 0% and 100% shows the variation of SL for each task over time. The start and end of an upward line in the graph represent the start and end of continuous attendance to a task. Because no more than one task could be attended at any point in time, with the SL line for one task being upward, the rest of the lines should be downward or steady at zero SL.

The best performer attended only to the three highest W tasks (Task 0, W = 6; Task 2, W = 10; Task 3, W = 6) without being influenced by their differences in DR and CR. It can be seen in Figure 3b that although these three tasks were attended equally from the beginning, it took about 50 s to be able to keep them at a high SL. Each of these tasks gained a high average SL, whereas the rest of the tasks were ignored.

In contrast, the worst performer attended to too many tasks in order to avoid penalization and never took the time to raise (and maintain) a few of the tasks to (and at) a high SL, which resulted in a poor overall score. This behavior can be clearly seen in Figure 3c, in that all tasks were largely attended only when they hit the zero SL line (or penalization line). Task 2 differed from the other tasks only in that more time was spent to take it to a higher SL whenever attended. Therefore, Task 2 (the highest W task, W = 10) had a modest average SL, whereas the rest of the tasks had a poor average SL (see Figure 3a).

The tabu strategy was to initially bring the two highest W tasks (Task 0, W = 6; Task 2, W = 10) up to near 100% SL while maintaining around 50% SL for the lowest DR task (Task 1, DR = 1, W = 5). When the two high-W tasks (Task 0, W = 6; Task 2, W = 10) were stable at high SLs, the third task (Task 1, W = 5) also was improved to a near 100% SL. The selection of Task 1 (DR = 1, CR = 5, W = 5) in favor of Task 3 (DR = 3, CR = 8, W = 6), which had a higher W and CR, may be justified by its very low DR. After all the three tasks (Task 0, W = 6; Task 1, W = 5; Task 2, W = 10) were stable at a high SL, the fourth task (Task 3, W = 6), which had the second highest W, was also attended to around the 50th s. By the 100th s, all four of these tasks (Task 0, W = 6; Task 1, W = 5; Task 2, W = 10; Task 3, W = 6) reached a steady state near 100% SL.

Figure 3d indicates the success of this strategy in managing the four highest W tasks, which was one task more than the best human performance. Although Task 5 (DR = 2, CR = 13, W = 2) had the lowest W, it was sporadically attended probably because of its very high CR and low DR. The average performance of the participants had a relatively higher SL in the three highest W tasks (0, 2, 3) in comparison with the rest of the tasks. Most successful strategy: For heterogeneous tasks, attend to the most important (highest W) tasks if there is no significant difference between their DRs and CRs.

Scenarios dDR-*iCR-iW, iDR*-dCR-*iW, and* *iDR-iCR*-dW: *In order, different DRs, CRs, and Ws*. In each of these scenarios, tasks had two identical parameters while differing in only one parameter. In order, the parameters were DR = [2, 1, 3, 3, 4, 2], CR = 9, W = 5; DR = 3, CR = [9, 5, 8, 8, 9, 13], W = 5; and DR = 3, CR = 9, W = [6, 5, 10, 6, 4, 2]. Figures 4a, 4b, and 4c show the attendance frequency and average state (SL) of tasks in these scenarios. The tasks with the highest contribution to the score (*effective tasks,* in short) in each scenario were the tasks with the lowest DRs, highest CRs, and highest Ws. The first few effective tasks in each scenario were, respectively, Task 0 (DR = 2), Task 1 (DR = 1), Task 5 (DR = 2); Task 0 (CR = 9), Task 3 (CR = 8), Task 4 (CR = 9), Task 5 (CR = 13); and Task 0 (W = 6), Task 2 (W = 10), Task 3 (W = 6).

The best performers in these scenarios discovered the effective tasks and focused on them. Initially in each scenario, two of these tasks were attended, and then the third and sometimes the fourth one. These participants maintained good to excellent average SLs in the two most effective tasks; fair to good average SLs in the next one or two effective tasks; and poor average SLs in the remaining tasks. The attendance frequency to three or four of the tasks focused on was roughly within a 10% range, and in the rest of the tasks, one up to three tasks were completely ignored. Note that tasks with high DRs (or low CRs) had to be attended more often to maintain the same average SL in tasks with low DRs (or high CRs).

The worst performers in these scenarios were overly concerned with penalization and tried to avoid it by attending to all tasks and not ignoring any task. This hypothesis was supported by the participants' verbal comments, the fact that tasks were maintained barely above zero SL, and that tasks were largely attended for a short time immediately after they hit the zero SL line. Overall, these participants maintained a poor average SL in all tasks by not ignoring (or not focusing on) a few of them – for example, attending to all tasks uniformly in a cyclic, sequential order gave poor results. Although a relatively better average SL was gained in more effective tasks in *dDR*-iCR-iW and iDR-*dCR*-iW, that is believed to be largely attributable to the parameters of these tasks. For example, attending to a task with low DR (or high CR) gains a better average SL as compared with identically attending to a task with high DR (or low CR).

The tabu performance in these scenarios was fairly similar to that of the best performers in that it focused on two or three most effective tasks and ignored the rest. It was slightly different, however, in that after the first few tasks were stable at high SLs, the model gradually moved on to the next effective task or tasks while maintaining the first few tasks at high SLs. Considering that a task's SL could not go beyond 100%, some of the best performers put too much effort in perfecting an already great task state instead of moving on to the next task or tasks being penalized. The average performance of the participants, in these scenarios, had a relatively higher average SL in the effective tasks in comparison with the rest of the tasks. Most successful strategy: Attend to the most effective tasks; attend to others only as time permits.

Scenario iDR-iCR-iW: Identical tasks. In this scenario, tasks were identical (DR = 3, CR = 9, W = 5) and task management had the least effect (see Figure 4d). The best performer primarily attended to three tasks while keeping a high average SL, and practically ignored the rest of the tasks. The worst performer kept a modest average SL in the four tasks equally attended and ignored the rest of the tasks. No task was kept at a high SL long enough to improve the score much. The tabu performance was similar to the best participant performance in that primarily three of the tasks were attended. Other tasks were occasionally attended for just a short time and mostly when they were at zero SL, which can be explained in terms of the desire to avoid penalization. The average performance of the participants had a relatively higher attendance frequency to each of the tasks in the middle of the screen (3 and 4) and a lower attendance frequency to tasks on the left side of the screen (0 and 1). Most successful strategy: For identical tasks, pick a few tasks and keep them satisfactory.

General Discussion

Initial overreaction to penalization. The penalty factor in this study was set to 20%. If it had been much higher, it would probably have been wise for participants to avoid penalization at all costs. The participants initially attempted to handle too many (more than four) tasks, usually attending to them just when they hit zero SL (penalization line) to avoid penalization. The drawback of this approach was that none of the tasks could draw enough attention to approach a high SL and, in turn, contribute to raising the score.

The results of fitting a linear model indicate a significant negative relationship between the number of tasks handled (i.e., tasks attended over 5% of the time) and the scenario order to which participants were exposed, $F(1, 48) = 10.47$, $p < .01$, $r_s = -.42$, for a test of the null hypothesis that the slope of the fitted line is zero. That is, the majority of participants learned to handle fewer tasks (or ignore more tasks) as they advanced through the scenarios. This resulted in better scores.

Therefore, the pattern observed in this research suggests that humans initially overestimate the magnitude and effect of a penalty but gradually learn its true effect via practice. In a real-life multitasking environment, the penalty of performing poorly in a task might be further exaggerated by light, noise, vibration, and other salient stimuli. In such an environment, it would be extremely hard for the operator to ignore the salient task in favor of other tasks, even if that is the right thing to do. This suggests that design should explicitly consider the salience of likely stimuli or that training should be provided to help operators understand the real consequences of unsatisfactory task performance. At the same time, it is important for operators to dynamically evaluate which tasks to shed, the value of continuing a task, and the consequences of shedding a task to achieve the best results.

Importance of strategic versus tactical task management. The results of fitting a linear model indicate a convincing relationship between the score of participants relative to that of tabu and the scenario order to which the participants were exposed, $F(1, 48) = 30.04$, $p < .01$, $r_s = .62$, for a test of the null hypothesis that the slope of the fitted line is zero. That is, the score of participants relative to that of tabu improved as the participants advanced through the scenarios. Note that earlier we also provided convincing statistical evidence that the participants learned to handle fewer tasks as they advanced through the scenarios.

These results are indirectly in agreement with the observation of the performances within each scenario that the participants who worked on the same tasks, with regard to type and number, and showed the same level of reaction to penalization, had very similar scores: good or bad. This relationship was observed despite the fact that the participants' moment-to-moment attendance to tasks was quite different.

The choices of which tasks to handle, which tasks not to handle, and the trade-off between avoiding penalization for all tasks versus excelling in a few tasks (at the cost of penalization for some other tasks) were considered as the participants' *strategic* task management. The participants' moment-to-moment behavior was considered as their *tactical* task management, which can be attributed to task-switching frequency, order of attendance to tasks, length of continuous time spent on a task, or the number of times attended to a task.

Participants appeared to consciously determine their strategy (many times expressed verbally) just prior to starting a scenario based on their experiences in practice trials and previous scenarios; a chosen strategy was rarely changed during a scenario. However, given a selected strategy, the participants seemed to change tactics intuitively throughout that scenario.

Attendance to effective tasks. As mentioned earlier, effective tasks are high-W tasks that have low DRs and high CRs and thus do not need much attention to attain high weighted average SLs. The general rule of priority (or the biggest bang for the buck) was to first attend to and maintain a high average SL in such tasks because they provide the maximum benefit (score) for the least effort (attendance).

The results of fitting a linear model indicates a convincing relationship between the task average SL and the tasks in the increasing order of Ws, DRs (negative relationship), and CRs within iDR-iCR-*dW*, *dDR*-iCR-iW, and iDR-*dCR*-iW, respectively. In order, the corresponding statistics for a test of the null hypothesis that the slope of the fitted line is zero are $F(1, 58) = 106.10$, $p < .01$, $r_s = .88$; $F(1, 58) = 40.12$, $p < .01$, $r_s = -.64$; and $F(1, 58) = 37.34$, $p < .01$, $r_s = .63$. That is, the participants performed better in those tasks that had a higher effectiveness in each of these scenarios.

This result is also in agreement with the participant performance in *dDR-dCR-dW* (see Figure 3), but it is not applicable to iDR-iCR-iW, as all tasks had the same effectiveness. It is speculated that if a real-life operator is not attending to effective tasks, it might be because of his or her misperception of the relative value of the W, DR, or CR of the tasks – for example, attending to a low-W task just because it is visually more salient.

Importance of parameter conspicuity in attending to tasks. The decreasing order of conspicuity for the parameters was in the following order: W, DR, and CR. That is, the participants had perfect knowledge of the Ws as they were displayed on the screen next to each task (as dollar values). They also had good knowledge of the DRs, as adjacent tasks could be compared while deteriorating simultaneously. However, they had only a fair knowledge of the CRs because simultaneous comparison of tasks was not possible, so CRs could be estimated only one at a time, while a task was attended.

Interestingly, it also appeared that the majority of participants, when deciding on which task to attend, considered the system parameters in the order of W, DR, and CR. That is, if the Ws of the tasks differed significantly, the higher W tasks were attended to with little concern regarding their DR or CR (see the performance analyses for iDR-iCR-*dW* and *dDR-dCR-dW*). When there was no significant difference among the Ws, tasks with lower DRs were attended to with little or no concern for CR (see the performance analysis for *dDR*-iCR-iW). If Ws and DRs were equal or nearly equal, tasks with high CRs were attended to (see the performance analysis for iDR-*dCR*-iW).

Although no conclusive evidence can be provided for this observation, it is in general agreement with that of other researchers, that participants choose the path of least cognitive effort (Gray & Fu, 2004; Payne, Bettman, & Johnson, 1993). Although in practice, it might be harder to perceive a task's W than in this study, operators usually have a good understanding of a task's W or which task is more important. In contrast, it is more difficult for operators to understand the subtle differences in task DRs and, even more so, the differences in task CRs.

Performance in scenarios with distinguishable tasks. The participants considered that distinguishing tasks was easiest in *dDR-dCR-dW* and most difficult in iDR-iCR-iW. Ironically, distinguishable tasks did not help the participants' scores relative to that of tabu in *dDR-dCR-dW* (see Figure 2b). On the other hand, the participant scores relative to those of tabu in iDR-*dCR*-iW, *dDR*-iCR-iW, and iDR-iCR-*dW* improved as the varying task parameter became more conspicuous and tasks became more distinguishable, $F(1, 28) = 4.03, p = .05, r_s = .36$, for a test of the null hypothesis that the slope of the fitted line is zero.

There is no statistically significant evidence of a difference among the mean scores of the participants relative to those of tabu in the five scenarios, $F(4, 45) = 1.99, p = .11$. However, the Fisher's LSD method indicates that the mean score of participants relative to that of tabu in iDR-iCR-*dW* was significantly higher than those gained in iDR-*dCR*-iW and *dDR-dCR-dW:* respectively, 15.8 and 16.3, ±15.46, 95% LSD.

It was noticed in the experiments that as the varying task parameter became more conspicuous, tasks became more distinguishable. However, it was also noticed that as the number of varying parameters increased, tasks became more distinguishable. Therefore the conspicuity of the varying parameter (or paramenters) and the number of varying parameters were confounding factors in this experiment.

That is, the scenario with the least distinguishable tasks (iDR-iCR-iW) had no varying parameter and, obviously, no conspicuous varying parameter. In contrast, the scenario with the most distinguishable tasks (*dDR-dCR-dW*) had three varying parameters with three levels of conspicuity. This might be the reason that in the three scenarios with one varying parameter, the score of participants relative to that of tabu improved as the tasks became more distinguishable. However, this relationship was no longer statistically significant when all five scenarios were considered together.

SUMMARY, CONCLUSIONS, AND RECOMMENDATIONS

The metaphor of a juggler spinning plates was introduced to model an operator responsible for multiple concurrent tasks. This paradigm is particularly applicable to those environments in which tasks are ongoing and do not have a clear completion point, and consequently the number of completed tasks is not a good measure of performance. The mathematical model developed for this paradigm was solved by tabu search heuristic methods, and the tabu solution found was considered as a normative model or a near-optimal method of managing multiple concurrent tasks. The tabu performance was then compared with human performance.

A software environment was developed to measure and analyze human performance while managing six concurrent tasks. Five different scenarios, each 5 min long, were developed based on the variations of DR, CR, and W of the tasks. The performance of the 10 study participants was recorded, replayed, reviewed, and analyzed statistically and graphically.

In the statistical comparison, there is substantial evidence that tabu outperformed the participants in all scenarios together and for each of the five scenarios considered individually. Although the participants could not beat the tabu score, the best performers and the participants on average gained scores that were, respectively, as close as 82% to 99% and 71% to 87% of the tabu score in each scenario. These numbers show that there is chance for a skilled participant to beat the tabu score with additional practice and experience.

In general, it appeared that the participants' strategy with respect to the level of reaction to penalization and the selection of type and number of tasks attended was more important in attaining a high score than the participants' intuitive, moment-to-moment tactics on how to attend to tasks. There is substantial statistical evidence that as the participants advanced through the scenarios, they handled fewer tasks and their scores improved. However, a rather focused experiment is needed to convincingly support the importance of strategic versus tactical task management.

The generic approach of tabu in attending to tasks seemed to be a stepwise strategy. Three of the tasks were first raised concurrently to a high SL; then the fourth and sometimes a fifth task were attended to while maintaining the high state of the previous tasks. The majority of the participants had poor performance in this regard. They either attempted to handle too many tasks all at once (as overreaction to penalization in early scenarios) or handled fewer tasks than they were capable of handling (in late scenarios).

As expected, the participants identified and attended to effective tasks (i.e., tasks with high Ws, low DRs, and high CRs); good performers were more successful in doing so. The conspicuity of the parameters in this experiment decreased in the order of W, DR, and CR. The participants seemed to consider the more conspicuous parameter as the primary attendance decision factor, although no statistical evidence is provided. However, there is strong statistical evidence that the participants' scores relative to those of tabu increased as the conspicuity of the varying parameter increased in the three scenarios with only one varying parameter.

In the design of the scenarios in this experiment, the number of varying parameters and the conspicuity of the varying parameters were confounding factors, and both affected the distinctiveness of a scenario in the eyes of the participants. Therefore, a more thorough study is recommended that includes rather isolated variations of parameter conspicuity and/or the number of varying parameters.

This study could be extended in numerous ways for further research. One research extension could be defined by varying the software design. Examples include changing the task parameters (DR, CR, W, and penalty factor) and their salience, introducing discrete and/or stochastic changes in a task's SL instead of continuous changes, and comparing easy scenarios with a manageable number of tasks and reasonable CR/DR rates versus difficult scenarios. However, in this category the most critical experiment would be to assess how closely this software environment represents real-life task management by validation against real-world or high-fidelity simulated environments. Another research extension could be related to the participants in the experiment, such as the investigation of the effects of gender, age, fatigue, instruction, practice, or expertise in different multitasking fields. The last suggested extension to this research is to investigate different scoring mechanisms and penalty factors, which could very well lead to different optimal behaviors.

The main contribution of the present work is the development of an abstract but flexible framework to which many multitasking environments can be mapped. The software environment developed based upon this framework facilitates the measurement, analysis, and comparison of human multitasking performance with that of others or with optimal (or near-optimal) performance. It should be noted that the good, bad, or near-optimal task attendance strategies we have discussed are highly dependent on the structure of the experiment and the parameters chosen. Therefore, any generalization of the conclusions in this study to other environments with different parameters has to be done cautiously.

General Observations

Humans can use information about system dynamics (DR, CR) and task weight (W) in managing multiple, concurrent tasks, and those who invest the time to explore these parameters can perform surprisingly well (almost optimally when human performance data were repaired and compared against a normative model that was handicapped to account for human limits). Operators who ignore this information (DR, CR, and W), however, allow themselves to be distracted by tasks that do not contribute much value when attended or that do not result in extremely bad consequences when ignored.

General Recommendations

Design equipment, procedures, and training so as to keep operators comfortably aware of the rate at which the state of tasks deteriorates when not attended (DR), the responsiveness of tasks to operators' efforts (CR), the value or weight (W) of performing tasks at satisfactory levels, and the consequence of shedding tasks (penalty). To avoid the possible detrimental consequences of information overload, it might be best to provide this information in type layers (i.e., the capability to view or hide the task DRs, CRs, and Ws in three separate layers) and/or provide them in ranges (e.g., high, medium, or low DR). This information is complex, highly related to other system and environment parameters, and very dynamic. Thus, providing this awareness will be very challenging, but even modest success in achieving it could have significant benefits.

ACKNOWLEDGMENTS

The development of the Tardast software in this research was partially funded by NASA Grant NAG-1287, "Training Pilots to Prioritize Tasks." The authors would like to thank Dr. R. Key Dismukes for his encouragement and substantive comments and the anonymous reviewers for their patience, interest, and constructive criticism.

REFERENCES

Air Force Office of Scientific Research. (2007). *AFOSR Research Interest Brochure and Broad Agency Announcement 2007-1.* Retrieved January 15, 2007, from http://www.afosr.af.mil/pdfs/afosr_baa_2007_1.pdf

Carbonell, J. R., Ward, J. L., & Senders, J. W. (1968). A queuing model of visual sampling: Experimental validation. *IEEE Transactions on Man-Machine Systems, 9*(3), 82–87.

Chou, C., Madhavan, D., & Funk, K. (1996). Studies of cockpit task management errors. *International Journal of Aviation Psychology, 6,* 307–320.

Chu, Y.-Y., & Rouse, W. B. (1979). Adaptive allocation of decision making responsibility between human and computer in multitask situations. *IEEE Transactions on Systems, Man, and Cybernetics, 9,* 769–777.

Dessouky, M. I., Moray, N., & Kijowski, B. (1995). Taxonomy of scheduling systems as a basis for the study of strategic behavior. *Human Factors, 37,* 443–472.

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

Gray, W. D., & Fu, W.-T. (2004). Soft constraints in interactive behavior: The case of ignoring perfect knowledge in-the-world for imperfect knowledge in-the-head. *Cognitive Science, 28,* 359–382.

Greenstein, J. S., & Rouse, W. B. (1982). A model of human decision making in multiple process monitoring situations. *IEEE Transactions on Systems, Man, and Cybernetics, 12,* 182–193.

Hillier, S. H., & Lieberman, G. J. (2005). *Introduction to operations research* (8th ed.). New York: McGraw-Hill.

Kleinman, D. L., Baron, S., & Levision, W. H. (1970). Optimal control model of human response: Part I. Theory and validation. *Automatica, 6,* 357–369.

McRuer, D. (1980). Human dynamics in man-machine systems. *Automatica, 16,* 237–253.

Moray, N., Dessouky, M. I., Kijowski, B. A., & Adapathya, R. (1991). Strategic behavior, workload, and performance in task scheduling. *Human Factors, 33,* 607–629.

National Research Council. (1998). Attention and Multitasking. In R. W. Pew & A. S. Mavor (Eds.), *Modeling human and organizational behavior: Applications to military simulations* (pp. 112–128). Washington, DC: National Academy Press.

Pattipati, K. R., & Kleinman, D. L. (1991). A review of the engineering models of information-processing and decision-making in multi-task supervisory control. In D. L. Damos (Ed.), *Multiple task performance* (pp. 35–68). London: Taylor & Francis.

Payne, J. W., Bettman, J. R., & Johnson, E. J. (1993). *The adaptive decision maker.* New York: Cambridge University Press.

Rouse, W. B. (1980). *Systems engineering models of human-machine interaction.* New York: Elsevier/North Holland.

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

Shakeri, S., & Funk, K. (2003). A comparison between humans and mathematical search-based solutions in managing multiple concurrent tasks. In *2003 IIE Annual Conference Proceedings* [CD-ROM]. Norcross, GA: Institute of Industrial Engineers.

Shakeri, S., & Logendran, R. (2007). A mathematical programming-based scheduling framework for multitasking environments. *European Journal of Operational Research, 176,* 193–209.

Tulga, M. K., & Sheridan, T. B. (1980). Dynamic decisions and workload in multitask supervisory control. *IEEE Transaction on Systems, Man, and Cybernetics, 10,* 217–232.

Walden, R. S., & Rouse, W. B. (1978). A queuing model of pilot decision making in a multitask flight management situation. *IEEE Transactions on Systems, Man, and Cybernetics, 8,* 867–874.

Wickens, C. D., Lee, J., Liu, Y. D., & Gordon-Becker, S. E. (2004). *An introduction to human factors engineering* (2nd ed.). Upper Saddle River, NJ: Pearson/Prentice Hall.

Shakib Shakeri is an associate analyst at Nuclear Safety Solutions, Ltd., in Toronto, Ontario, Canada. He earned his Ph.D. in industrial engineering from Oregon State University in 2003.

Ken Funk is an associate professor in the Department of Industrial and Manufacturing Engineering at Oregon State University. He received his Ph.D. in industrial and systems engineering from The Ohio State University in 1980.

Date received: January 23, 2004
Date accepted: February 19, 2006