DO BETTER MULTI-TASKERS MAKE BETTER PILOTS?

V. Alan Spiker
Anacapa Sciences, Inc.
Santa Barbara, CA
vaspiker@anacapasciences.com

M. Ron Karp
Arizona State University
Mesa, AZ 85212
ron.karp@asu.edu

Tricia Mautone, Susan Fischer
Anacapa Sciences
Santa Barbara, CA
tmautone@anacapasciences.com
sfischer@anacapasciences.com

ABSTRACT

Aircraft operation involves many facets of multi-tasking (MT), where breakdowns in task management have serious implications for performance and safety. There is a need to develop valid, predictive tests of MT ability and provide practical criterion measures of that performance. Thirty advanced students (15 pilot-copilot pairs) in a university flight training program were tested in a medium-fidelity King Air simulator. The 30-min. scenario was designed to task-load both pilot and copilot during an event-filled instrument approach to an unpublished holding point, challenging procedure turn, and steep descent with loss of glideslope and worsening weather, culminating in a missed approach. Two observers independently scored subjects' performance on 65 MT-relevant behavioral events and rated them on six process dimensions of MT. Subjects also took a battery of predictive tests, including two specifically-designed MT tests, and tests of fluid intelligence, processing speed, and aviation mathematics. The MT tests assessed ability to simultaneously monitor multiple visual fields and hold multi-dimensional concepts in memory. The scenario challenged MT for all crews, with 40% of the events showing evidence of disruption. Inter-observer reliability for MT ratings and sub-event performance was high (r=.85-.91) and MT test reliability coefficients exceeded .80. MT criterion measures showed a complex, but fascinating, relationship to the predictive tests. One test of MT was significantly related to performance for copilots but not pilots. The other test of MT predicted performance on two key archival measures of student proficiency (hours to instrument rating, number of extra flights). The paper closes with study implications for developing other criterion measures of MT, adapting MT tests for student placement, and developing MT training programs for "at-risk" student-pilots.

ABOUT THE AUTHORS

Alan Spiker has been a principal scientist at Anacapa Sciences since 1982. He is responsible for human factors research associated with advanced technology and has conducted field studies on aviation training and performance for industry and all branches of service. Dr. Spiker received a Ph.D. in Experimental Psychology from the University of New Mexico in 1978. Dr. Spiker received the Human Factors Society's Alexander C. Williams Award (1991) for outstanding human factors contributions to the design of a major operational system. Since 2000, Dr. Spiker has been conducting collaborative research with ASU-Mesa to determine methods for improving the training proficiency of student pilots in higher-order cognitive competencies that underlie safe, effective airline operations.

Merrill R. (Ron) Karp is Professor of Aviation and Associate Department Chair, Department of Aeronautical Management Technology, Arizona State University, Polytechnic Campus. Since 1994, he has developed and managed the ASU/Mesa Airlines Airline Bridge Training Program using airline high-fidelity motion-based and fixed-based simulators and scenario-based training. Dr. Karp received his Ph.D. in Administration & Management from Walden University in 1996. He was a graduate from the U.S. National War College in 1985 and the Armed Forces Staff College in 1981. He served as a Fighter Wing Commander of F-4G Wild Weasels in Operation Desert Storm, flew the F-16 as a Wing Vice Commander and DO, and served three combat tours in the F-4 in Vietnam.

Tricia Mautone has been a senior scientist at Anacapa Sciences, Inc. since 2003. Her work primarily focuses on the development and evaluation of effective multimedia and game-based instruction and training, as well as the development of computer-based multi-tasking assessments. Dr. Mautone received her Ph.D. in Cognitive Psychology from the University of California, Santa Barbara in 2005.

Susan Fischer is a principal scientist with Anacapa Sciences and has been with the company since 1989. She received her PhD in Cognitive Psychology from the University of California, Santa Barbara in 1990. Dr. Fischer specializes in measurement, test development, and training cognitive processes in individuals. She recently was principal investigator of a project to develop web-based training in critical thinking for Army officers, and is currently directing the ONR-funded project that produced the *MTAT* test discussed in this paper.

DO BETTER MULTI-TASKERS MAKE BETTER PILOTS?

V. Alan Spiker
Anacapa Sciences, Inc.
Santa Barbara, CA
vaspiker@anacapasciences.com

M. Ron Karp
Arizona State University
Mesa, AZ 85212
ron.karp@asu.edu

Tricia Mautone, Susan Fischer
Anacapa Sciences
Santa Barbara, CA
tmautone@anacapasciences.com
sfischer@anacapasciences.com

INTRODUCTION

Operating a commercial airliner involves many facets of multi-tasking (MT), where breakdowns in task management have serous implications for performance and safety (Iani & Wickens, 2007). Historically, flight simulators have proven an excellent platform to test theoretical concepts in dual task performance, as pilots perform multiple tasks (e.g., flight control, instrument scan, checklists, radio operation) both simultaneously and in rapid succession, under tight time constraints, with frequent interruptions and constantly changing parameters (Wickens, 1999). Given the strong role that MT plays in normal aircraft operations, the present study was conducted to determine if a psychometrically robust test of MT, developed in another context, would predict pilot performance in a simulated flight scenario. To the extent that it does, we will have an affirmative answer to the question posed in the title of this paper.

Multi-Tasking Demands on Airline Pilots

Observations of pilot performance under high MT demands have yielded data on how tasks are juggled and shared between crew positions, the types of MT errors made, and the behavioral correlates of breakdowns in MT. The MT-related tasks that pilots encounter during flight often include a mix of both anticipated procedures, such as handling heading changes, as well as unexpected events, such as responding to a caution light or dealing with equipment failure (Colvin, 1999). There is also variability in the timing or urgency of tasks. These factors play a role in the effective management of multi-tasking demands. For example, some tasks, like running checklists, can be performed at any time, so their occurrence can be scheduled to minimize task demands during high workload periods of the flight, such as during an approach.

Failure to effectively manage MT demands can lead to serious errors. An analysis of the frequency of different types of MT-related aircrew errors and aircraft incidents (Funk, Chou, & Madhavan, 1999) noted that typical problems include descending too late, reconfiguring the aircraft (e.g., setting flaps, lowering landing gear) too late, failing to tune navigation and communication radios, overshooting altitude, and being overly distracted by local aircraft traffic. Since pilots may be juggling the performance of anywhere from three to six tasks at any one time, MT ability would seem to play a key role in successful flight performance.

What factors contribute to breakdowns in multi-tasking? Dismukes (2003) examined pilot errors from the standpoint of managing interruptions, distractions, and concurrent task demands. He found that crews are particularly vulnerable to forgetting or omitting procedural steps when their normal tasks are interrupted by concurrent tasks. This poses special problems when the external demands arrive at unpredictable times, where these conditions may force task elements to be performed out of their normal sequence. The list of possible errors that can occur under these conditions is quite long. Examples include being preoccupied with a new departure clearance, resulting in taking off with the flaps not set; examining an annunciator light while failing to note a significant wind shift; and forgetting to complete checklists while receiving multiple messages from ATC.

Observing pilots in their natural flight environment should yield a wealth of information concerning MT ability and MT breakdowns. In this paper, we describe how we tailored an MT flight scenario that we used with students in the collegiate flight training program at Arizona State University to gather behavioral measures of MT ability, and how we then compared these data to various measures of multi-tasking and related abilities. The goal of the study was to identify and measure aspects of MT behavior in a flight environment. Ultimately, the knowledge gained from this study can be used not only to refine assessment of MT, but to develop training programs that can perhaps target specific deficiencies in MT ability.

Characteristics of Multi-Tasking Environments

As a psychological phenomenon, MT is viewed as the ability to concurrently perform or interleave multiple tasks. In a study of various demanding professions, including nurses, emergency medical technicians, and pilots, Fischer et al. (2003, 2005) identified ten defining characteristics of an MT environment. These include: (1) multiple discrete tasks are performed;

(2) not all tasks can be performed simultaneously; (3) important or urgent tasks cannot be shed or significantly postponed; (4) the environment does not always signal or cue task initiation; (5) the environment is dynamic and includes interruptions; (6) tasks differ in terms of priority, difficulty, and length of time; (7) feedback is not provided for some tasks; (8) most tasks are performed in the order of seconds to minutes; (9) the environment is time pressured; and (10) the tasks vary in the amount of cognitive resources they demand.

These characteristics should be present in any scenario designed to assess MT performance. Because MT demands impact success in the work place, it is both practically useful and theoretically necessary to have a predictive assessment instrument to gauge how individuals differ on this construct before they are placed into the job environment. Aircraft operation clearly encompasses all these defining MT characteristics, so application of a predictive test of MT ability would seem to be of great value here.

Cognitive Demands of Multi-Tasking

In analyzing the MT demands of various job domains, Fischer et al. (2003) concluded that MT is principally a cognitive activity. Consequently, development of a predictive assessment instrument requires identifying the underlying cognitive functions needed to successfully perform in a MT environment. Following originating work by Burgess (2000), 12 cognitive functions were identified as essential for MT behaviors. These functions and examples of corresponding pilot tasks are listed in Table 1.

These functions can be grouped into four classes: storing information in memory, controlling attention, logic and reasoning ability (i.e., fluid intelligence), and processing speed. MT ability can be defined as the combined and integrated use of memory, attentional control, reasoning, and processing speed. From a behavioral standpoint, a pilot's ability to MT involves juggling more than one activity at a time, keeping multiple things in mind so they aren't forgotten, and managing tasks through shedding, resource allocation, and task delaying.

Table 1. Cognitive Functions underlying Key Pilot Tasks

Cognitive Functions	**Illustrative pilot tasks requiring those functions**
Short Term Memory (STM)	Remember radio frequency of next VOR/DME or ATIS station
Long Term Memory (LTM)	Draw on knowledge of aerodynamics, aircraft systems, and aviator geometry
Prospective memory	Remember to return to interrupted checklist task and radio communications
Monitoring output	Check correctness of switch settings, radio frequency, ATIS channels
Working memory updating	Continuously update situation awareness (e.g., information about heading, speed, next waypoint, location)
Mental set switching	Switch among very different types of tasks such as entering data into the flight computer or calculating descent angle or vertical velocity
Classification	Classify new information (wind shift, ATC-directed routes) as to whether it requires immediate attention or is of only passing interest
Rehearsal for memory storage	Rehearse STM stores such as new ATIS frequency, new wind data, or new runway ID long enough to put into LTM or until able to write them down
Selective attention	Must attend *only* to: flight control task when taking off or descending, terrain obstacles when flying low, instruments and switch settings when reading checklists
Divided attention	Fly aircraft and enter flight leg data at same time, scan horizon and talk to other pilot at same time, and operate radio and provide backup to other pilot at same time
Prioritizing	Prioritize maneuvering of the aircraft (aviate) ahead of determining current/desired position (navigate) ahead of talking on radio (communicate)
Deductive logic	Logically determine: descent angle based on current airspeed, altitude to lose, and distance to travel; vertical velocity based on descent angle and total distance traveled

Multi-Tasking Flight Scenario

Based on this analysis of the cognitive functions underlying MT ability and how it applies to piloting, we constructed a flight scenario that was designed to tap into these specific processes. The goal was to have an initial period where several "stressor" events imposed unusually high (but still within normal bounds) MT demands on both the pilot flying (PF) and the copilot or pilot-not-flying (PNF). This was followed by a stabilization period of reduced workload, giving pilots time to collect their thoughts and plan for the events in the ensuing MT-laden final period. We also developed specific measures that assessed both *process* (e.g., general MT dimensions, such as communication, task prioritization, etc.) as well as *performance* (e.g. how well or whether the pilot executed MT-related tasks).

The 30-minute MT flight scenario began with the aircraft airborne and enroute to Williams Gateway (IAW). The aircraft was at 12,000 ft and pilots were instructed to fly an instrument approach and land at Runway 30. They were given 5 min. to study the Instrument Approach Chart associated with an ILS landing at that runway. During their approach to the terminal area, participants were told by Phoenix Approach to descend to 6000' and enter an unpublished holding point (SNOWL) that corresponded to a VOR station near the runway. While flying to SNOWL, pilots were told that Approach's radar was out, so they would need to provide standard position reports (of time, altitude) upon entering the holding area. These two unanticipated events, flying to an unpublished holding point and providing position reports, constituted the primary MT stressors of Period 1. The key scenario events and expected pilot actions, broken out by altitude, are displayed in Figure 1.

In the stabilization period, participants were to orbit around SNOWL. They were to maintain 6000', execute standard left hand turns, and be ready to receive the request to descend after several minutes. While orbiting, they were to stay to the west of SNOWL, to avoid leaving protected airspace and encountering surrounding air traffic.

Once stable in the holding pattern, pilots were told to descend, marking the start of Period 2. They were to execute the published Procedure Turn, which entails a staged descent to 4600' and then to 3000', where a glideslope was to be intercepted. The procedure was complicated because the Initial Approach Fix (IAF) and the Final Approach Fix (FAF) were at the same physical point, SNOWL, but at different altitudes (4600' and 2800'). The descent requires a circling approach while losing altitude, and in turn, forces the pilots to make demanding mental calculations of descent angle and vertical velocity while operating the radios and monitoring present position. Precise airspeed and altitude control (+/- 10 kts and 100 ft, respectively) are essential, given traffic in the IAW terminal area.

To increase MT demands, as the aircraft reached 3000' the pilots were told that IAW's glideslope and DME were "off the air." This forced the pilots to rely on timing to determine their missed approach point (by procedure, they should have hacked the clock while reaching the FAF). To compound MT demands, the weather worsened with a lowered ceiling and reduced visibility, requiring the aircraft go "missed approach" at the last minute.

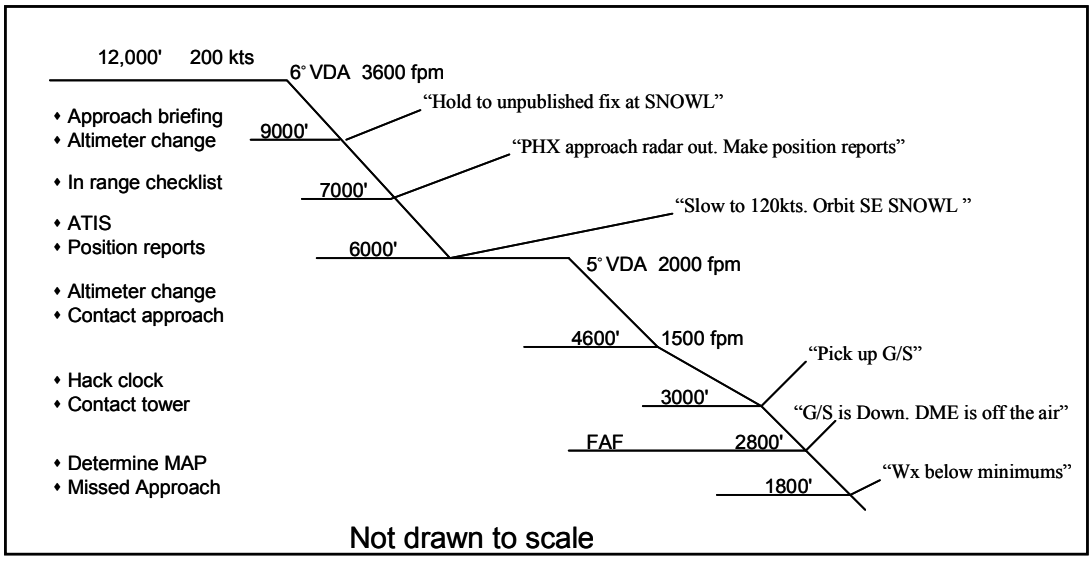

Figure 1. Key MT scenario events by altitude.

Flight Scenario MT Process Scales

Six MT process dimensions were rated: communication, task prioritization, task management, crew resource management (CRM), decision-making, and situation awareness (SA). All have been shown to be important for successful performance (Nullmeyer, Spiker, Golas, Logan, & Clemons 2006). Each dimension was rated on a five-point scale: 5=exceptional, 4=above average, 3=average, 2=below average, and 1=poor. The dimensions were further decomposed into a set of observable behaviors that could be checked present/absent. Examples of task prioritization behaviors include: Failed to do most important task first, failed to hand over less important tasks, did unnecessary tasks at expense of more important ones, failed to do less critical tasks in low workload periods, failed to prioritize IAW aviate-navigate-communicate, failed to anticipate upcoming task needs, and failed to drop or transfer less important tasks when overloaded.

The checklist let observers record both positive and negative instances of these behaviors; these can be summed to produce potentially useful indices for correlating with predictive tests of MT ability. Each observer completed one set of MT process ratings during the scenario for each crew; separate ratings were made for PNF and PF.

Flight Scenario Behavioral Events

A second set of criterion measures was the behavioral event (BE) scores collected throughout the scenario by the observers. These BEs ranged from short, discrete events (e.g., changing the altimeter, changing radio frequency) to more extended/continuous streams of behavior (control vertical velocity during descent). A five-page instrument was used by each observer to record these events, a segment of which is shown in Figure 2.

The instrument is divided into two parts. The left portion is the "script" of events that unfold during the scenario. The verbatim statements (quotes in bolded font) were read by the observer while role-playing Phoenix Approach Control and other participating agents. The quotes in italics are statements the PF or PNF were *supposed* to make during the scenario. Similarly, normal (non-bolded, non-italicized) text indicates actions and thought processes that would be expected from a well-performing crew. This information was used by the observers to keep track of where participants were in the session. The section on the right provides space for scoring the features of each, and are positioned adjacent to the corresponding scripted event. The scoring sheet was used to record 64 BEs performed by one or both pilots during the scenario. Examples of behavioral events include responding to communications from Approach or Tower; running checklists; changing airspeed, altitude, and heading; maintaining position awareness; and changing and comparing altimeter settings.

♦ Phoenix Approach Control: **"Air Shuttle 222, turn right to heading of 065° to intercept Stanfield 025° radial outbound, then direct to SNOWL IAF via the Stanfield 025° Radial. Expect ILS Rwy 30C at Williams Gateway"**	Turn heading (065) and intercept 025 radial Start: _____ End: _____ Acc: _____ Helped? ____
♦ Co-Pilot Responds: *"AS 222; roger, cleared to SNOWL via Stanfield 025°radial"*	Acknowledge PHX Approach Start: _____ End: _____ Re-xmit ____ Acc: _____ Helped? ____
♦ Pilot calls for "*In range checklist*" prior to descending from cruise altitude. Co-pilot reads the "In range checklist" to the "line." "In range checklist, below the line" is continued 15 miles to destination.	In-range checklist Start: _____ End: _____ Acc: _____ Helped? ____
♦ After intercepting 025°R, PHX App: ***"AS 222, descend at pilot's discretion to be at 6000' by SNOWL. Altimeter 29.99. Squawk 6452."***	Begin descent to 6000 Start: _____ End: _____ Acc: _____ Helped? ____
♦ Co-Pilot Responds: *AS 222 cleared to 6000, pilot's discretion, altimeter 29.99*	Change to local altimeter (29.99) Start: _____ End: _____ Acc: _____ Helped? ____
♦ ATC clearance given at approx 17 DME from TFD. Pilot computes descent angle & resulting VSI: Alt to lose in FLs / Distance to travel= 120-60 / 27-17= 6° (must descend at 6° to be at 6000' by SNOWL). At 240 KTAS (4nm/min), 6° X 6 X 100 = 3600 fpm	Compute DA (6 deg) and VSI (3600 fpm) Start: _____ End: _____ Acc: _____ Helped? ____
	PF & PNF compare alt. after changing setting Start: _____ End: _____ Acc: _____ Helped? ____

Figure 2. Segment from the Behavioral Event Scoring Instrument.

METHOD

Participants

Thirty subjects (28 male, 2 female) participated in the study. All were students currently enrolled in the ASU Aeronautical Management Technology flight training program. Their average age was 23.7 years, with an average of 384.5 flight hours. All participants had been in the ASU program at least two years, earning a private pilot's license, certified flight instructor (CFI) rating, instrument rating, multi-engine rating, commercial rating, and certified flight instructor instrument (CFII) rating, the latter either awarded or in process. Participants were tested in pairs, with the more experienced student serving as the pilot flying (PF) and the less experienced student as pilot not flying (PNF). Subjects were assigned to one of 15 "crews" at random, subject to the restriction that PF and PNF slots would be filled by more experienced (CFII awarded) and less experienced (CFII in process) students, respectively.

Criterion Measures of Pilot Performance

Pilot performance was measured in the context of the scenario described above, which was carried out in a moderate fidelity simulator. The technological limitations of the ASU King Air simulator precluded the ability to automatically record aircraft performance data. Consequently, two observers, working independently, were charged with collecting all of the process and performance data. The scenario was kept fairly short, with a nominal duration of 30 minutes.

For *process measures* of MT, observers rated (using a 5-point scale) six dimensions of MT (task prioritization, task management, time management, etc.). Each process rating was accompanied by a behavioral checklist. For *performance measures*, a set of 64 "behavioral events" (BEs) was collected throughout the scenario. These BEs ran the gamut from noting discrete events (check altimeter, acknowledge message) to tasks more extended in time (in-range checklist) as well as more continuous tasks (discrepancies in altitude, airspeed, heading).

In addition to the scenario-based process and performance data, we collected three *archival measures of proficiency* that assessed how well students progressed through the program. Training folders for each subject were reviewed to extract: number of hours required to obtain an instrument rating, total number of "extra" flights required to achieve the various ratings offered by the program, and cumulative grade point average (GPA) within the program.

Predictive Tests of MT Ability

Given the cognitive demands inherent in a MT environment, we amassed a battery of five predictive ability tests to correlate with performance in a criterion flight simulation scenario: *MTAT*, *SYNWIN*, Pattern Comparison Test, Raven Advanced Progressive Matrices, and Aviation Mathematical Reasoning. The first two are tests explicitly designed to measure MT ability. The other tests were employed to assess subjects' ability in the related areas of fluid intelligence, processing speed, and mathematical aptitude. Each test is briefly described below.

The *Multi-Tasking Aptitude Test* or *MTAT* is a web-based assessment developed by Anacapa Sciences (Fischer et. al, 2005) in which participants type in commands to query the attributes of a series of unseen objects and then, based on the outcome of the queries, assign the objects to one of four bins. Each bin has specific rules about what types of objects it accepts. For example, bin #1 may accept only small, red, triangles; bin #2 may accept medium squares of any color, and so on. Placing an object in a correct bin earns points, while attempting to place an object in a non-matching bin results in negative penalty points. The bins are worth different points depending on how restrictive their membership requirements are. For example, a bin that accepts only small, red, circles is more restrictive and thus worth more points than a bin that accepts red circles of any size, which in turn is worth more points than one that accepts any red object. The objective of the task is to assign objects to bins as quickly as possible, while attempting to optimize points earned. To do well, participants must (1) pay attention to when an object is presented (and therefore available for querying and assigning), (2) keep track of objects' identification numbers, (3) initiate appropriate queries, (4) remember outcomes of the queries, and (5) assign each object to a bin that matches the object's attributes, all the while dealing with interruptions. The MTAT requires no prior domain knowledge and takes approximately 30 minutes to administer, including instructions, a practice session, and three 5-minute test sessions.

SYNWIN requires participants to simultaneously monitor and respond to a set of four tasks that are simultaneously presented on a computer screen (Elsmore, 1994). The upper left quadrant of the screen displays a letter recall task in which participants click a button to indicate whether a probe letter was in a previously displayed set of letters. The upper right quadrant presents an arithmetic task, where participants solve simple, randomly-generated three-digit addition problems. A visual monitoring task is in the lower left, where participants click on a gauge to reset a slowly

moving pointer before it reaches the zero mark. The lower right quadrant has an auditory monitoring task where participants listen to a series of high and low frequency tones, and click a button when they hear a high frequency tone. In the current study, one practice session and three 5-minute test sessions were administered.

The *Pattern Comparison* test (Salthouse & Babcock, 1991) is a 2-page paper-and-pencil *processing speed* test. Each page displays 30 pairs of line-segment patterns, with a horizontal line between each pair. Participants are to write an "S" on the horizontal line if the line pattern pairs are the same or a "D" if they are different. Half of the pairs are the same, half are different, with the "different" pairs nearly identical except for the orientation of one of the line segments. Participants are given 30 seconds per page to complete as many items as possible.

The *Raven Advanced Progressive Matrices* (Raven, 2000) is a paper and pencil test of fluid intelligence and inductive reasoning. Each item has eight black and white geometric figures occupying all but the lower right cell in a 3 x 3 matrix. The figures change systematically in one or more dimensions across the columns and down the rows. Participants indicate which of eight possible options best completes the pattern and fits into the missing bottom right cell. The practice consisted of six problems from Set I; the test, which in the current study took 30 minutes, consisted of the 36 Set II items.

The *Aviation Mathematical Reasoning Test* assesses speed and accuracy in solving aviation-related mathematical problems. A typical problem is "At 1200 KTAS, 1 degree of pitch change alters your vertical speed by 200 fpm. If you were at level flight and indicated a 3 degree negative pitch change, what would be your VSI?" Each problem requires mathematical reasoning, but simple computations. Participants were given 3 minutes to complete as many items as possible.

Procedure

The testing took place at ASU facilities in three separate sessions; participants were paid $25 for each session.

MT Predictive Tests

Participants were given the computer-based MT tests in small groups several weeks before completing the flight scenario. The tests were administered according to the test instructions and included a brief explanation of how the test worked, a practice session, then the actual test session. An experimenter was present to answer questions during the explanation and practice periods and to ensure fidelity of the testing environment. The paper and pencil assessment tests were administered in a separate session, several weeks after the flight scenario session. For each of the three tests, the experimenter provided a brief explanation, reviewed the practice problem, and then timed the participants as they completed the actual tests.

Multi-tasking Flight Scenario

Participants were tested in pairs, with one subject serving as pilot (PF) and the other as co-pilot (PNF). All advanced students (PF) had experience using the King Air simulator (Figure 3) as part of class

Figure 3. King Air Simulator.

assignments and practical exercises for the Air Navigation and Airline Instrument Procedure classes. The less experienced students (PNF) had been given practical instruction in operating the simulator in preparation for this study. Per airline procedure, the pilot was responsible for aircraft control and flight decisions, with the co-pilot operating the radios and performing actions in the cockpit (e.g., lowering landing gear) under pilot direction. The pilot had the option of flying the aircraft under autopilot, but at times needed to disengage the autopilot and fly manually.

Using hard copies of the Process rating instrument and the Behavioral Event Sheets, the two observers independently collected data for each participant during the course of the scenario. The first observer, an experienced flight instructor, also controlled the simulator from the instructor operator's station and role-played the duties of Phoenix Approach and Gateway Tower. The second observer recorded total elapsed time to complete the scenario, as well the split times needed to complete each scoring period. Weather information was provided automatically by an Air Traffic Information Service (ATIS) generator software package.

When the session was over, participants were debriefed on their performance as a crew and individually by the first observer. The two observers then convened privately to discuss the behavioral event recordings of the session just completed, compare their respective

ratings/notes, and correct any disagreements. But, the original assignment of behavioral event scores and MT ratings for each rater were used in the calculations of inter-rater reliability described below.

RESULTS

Inter-Rater and Test Reliability

The inter-rater reliability of the flight scenario measures was examined in two ways. The ratings assigned to the 30 participants were correlated for each of the six MT process dimensions (communication, task prioritization, task management, CRM, decision-making, and SA). All correlations were high, ranging from .89 (decision-making) to .97 (task management). In no case did the two raters disagree by more than 1 rating scale point. Inter-rater agreement was then computed for the 64 behavioral events (BEs) that were scored across the three measurement periods. Events were scored as either agree or disagree. Overall, the two raters agreed 91% of the time. Across crews, the rates of agreement ranged from a low of 84% to a 100%.

Reliability was also estimated for the *MTAT* and *SYNWIN*. Two assessments were made for *MTAT*, one for the response time measure and one for accuracy. In each case, correlations were computed between pairs of test sessions, where 3 sessions were administered for each test. All reliability estimates were corrected for attenuation in test length with the Spearman-Brown formula. For the *MTAT* accuracy measure, the inter-session correlations were .86, .65, and .60 for sessions 1-2, 1-3, and 2-3, respectively. The corresponding correlations for *MTAT* response time were .93, .82, and .80. As in previous studies, *MTAT* response time is a more stable measure than the accuracy measure, though all were statistically significant ($p < .01$ or .001). Average accuracy and response time measures were also highly correlated ($r = -.86$, $p < .001$). Inter-session correlations for the *SYNWIN* scores were stable as well, with $r = .82$, .86, and .83 for sessions 1-2, 1-3, and 2-3, respectively.

Typical Breakdowns of Multi-Tasking Behavior in the Flight Scenario

It was clear from our observations that the MT scenario was quite challenging and taxed subjects' multi-tasking ability. This was evident both from the MT process ratings and the BE data. Indeed, most crews had trouble with some aspect of the profile, whether it was finding SNOWL, maintaining altitude and airspeed, capturing the glideslope, or performing the necessary procedural actions at the proper time. Coupled with the added demands of worsening weather, loss of DME and glideslope, and making position reports, these factors accumulated to create a highly tasked cockpit environment. Throughout the profile, both pilots in every crew made mistakes, such as forgetting to change altimeter settings, not acknowledging a radio call, or losing positional awareness.

Ratings on all six MT process dimensions correlated highly with performance, the highest being task prioritization at $r = .95$. The others exhibited substantial correlations as well, including task management (.86), CRM (.81), situation awareness (.78), decision making (.73), and communication (.68). A particularly informative analysis entails tallying the proportion of subjects (out of 30) for whom a given observable MT behavior was checked as deficient. The nine most frequent behavioral errors are listed below, with the proportion of subjects presented on the right.

- Failed to anticipate upcoming task needs .67
- Failed to use standard terminology during radio transmissions or callouts .57
- Had trouble staying "ahead of the aircraft" .57
- Did not anticipate upcoming procedural events .57
- Failed to juggle radio operation and other duties .53
- Let things drop out of his/her scan while doing other duties .53
- Not able to get into a "rhythm" to perform tasks .50
- Had trouble keeping track of present position .50
- Poor judgment in making rapid decisions .50

Creating Composite Indices for Statistical Analysis

To set the stage for statistical analysis, the raw performance and process data were transformed into composite indices. The Flight Scenario MT Process measures were averaged to provide three composite measures for each participant. The first, *Ratings Average*, was calculated by averaging the anchored rating scale scores on each of the six process dimensions. The *Positive Behaviors Average* was calculated by counting the number of positive behaviors checked on the rating checklists for each of the MT process dimensions, and then averaging across the six dimensions. The *Negative Behaviors Average* was calculated in a similar fashion.

In computing a composite BE score for the three scenario periods, we deleted BEs from the tally if they were (1) primarily reflecting overall basic knowledge and skill level (vs MT), or (2) strongly skill-based (e.g., executing a procedure turn IAW with published approach). Fourteen of the original BEs were thus dropped, leaving 50. The remaining BEs were then scored in a binary fashion: behaviors within established performance standards were scored as 0 and those below standard were a 1. The data were then combined to create an overall *BE Error* score for each subject in each of the three scenario periods, with higher values indicating lower performance.

During this analysis, our original sample size, 30, was reduced for the following reasons. One participant was unable to complete testing and was dropped. Another participant was unable to complete the three paper-and-pencil tests, but his data for the scenario and MT tests were retained. Archival data for six participants were missing as they had entered the ASU program some time back and their records were no longer available. Finally, the Mathematical Reasoning scores for two participants were not obtained because of time constraints.

Criterion Measure Data

Table 2 shows the means and standard deviations for the archival measures, MT process ratings, and BE composite scores. Note that the average process rating, 2.5, is below the scale mid-point, and the BE composite scores are around .50, indicating that half of the BEs overall were scored below average. This is consistent with the observations of scenario difficulty described above. Correlations were then computed among the nine criterion measures in Table 2. Because the study sample comprises a substantial percentage of the student pilot population at ASU, the estimated variance of the sample mean was reduced (by 7%) using a finite-population correction coefficient (Winkler & Hays, 1975).

The three archival data measures were significantly correlated to one another, with values ranging from .36-.77. However, only GPA was significantly correlated with the process or BE data. Likewise, the six process and BE measures were significantly correlated with each other, with r's between .40-.75. This suggests that two archival measures, hours to instrument rating and number of extra flights, which inherently include a time component, are qualitatively different from the flight scenario data. These two measures may thus be tapping somewhat different skills or abilities than the flight scenario data.

Table 2. Means and Standard Deviations of Criterion Measures for both Crew Positions

Measure	Mean	STD DEV	N
Archival Data			
Hours to Instrument Rating	46.8	1.67	23
No. Extra Flights	11.8	1.33	23
GPA	3.67	0.05	23
Scenario Process Measures			
Ratings Average	2.5	0.15	29
Pos. Behaviors Average	0.6	0.13	29
Neg. Behaviors Average	2.1	0.28	29
Scenario Performance Measures			
Period 1: BE Score	0.41	0..4	29
Stabilization: BE Score	0.53	0.05	29
Period 2: BE Score	0.48	0.04	29

Predictive Validity

The predictive validity of our five tests was measured through correlations with the criterion variables listed above. These correlations, presented in Table 3, indicate that *MTAT* accuracy significantly predicted two of the archival measures of pilot performance, number of

Table 3. Correlations between Criterion Measures and Predictor Variables

	Predictor Variable					
Criterion Measure	MTAT Avg Acc	MTAT Avg RT	SYNWIN Avg	Raven	Processing Speed	Math Reasoning
Archival Data						
Hours to Instr. Rating	-.38*	.20	-.46*	-.62**	.15	.00
Number Extra Flights	-.45*	.30	-.36*	-.60**	.35	-.09
GPA	-.02	.16	-.01	.20	-.37	.21
Flight Scenario Process Data						
Ratings Average	-.15	-.04	.03	-.10	-.39*	.13
Positive Behaviors Ave	-.06	.01	.25	-.12	-.26	-.01
Negative Behaviors Ave	.19	.02	.00	.07	.48**	-.07
Flight Scenario BE Data						
Period 1 Error Score	.11	.05	-.13	.02	.16	-.14
Stabilization Error	.02	.12	.09	-.12	.03	-.06
Period 2 Error Score	-.15	.10	-.36*	.01	.00	-.40

*p < .05, ** p<.01

extra flights and hours taken to get their instrument rating. Students who scored higher on the *MTAT* required fewer hours and fewer extra flights to complete their instrument rating. *MTAT* response time marginally predicted number of extra flights, with faster times associated with fewer flights. *SYNWIN* significantly predicted hours to instrument rating and number of extra flights, as did the Raven test of fluid intelligence. However, the only MT measure that correlated with any of the scenario measures was *SYNWIN*, which predicted BE errors in Period 2. Math ability and fluid intelligence also failed to predict the flight scenario data. Yet, processing speed was positively correlated with the number of negative behaviors exhibited by participants and inversely correlated with the evaluator ratings of their performance. Thus, it appears that the *MTAT* does have some predictive power with respect to pilot training performance gathered over a long period of time, but it did not predict observer MT ratings nor participants' positive and negative MT behaviors.

To determine whether participants serving in the PF and PNF roles show differential sensitivity to MT demands, separate analyses were conducted on the BE data for PF and PNF participants. The two positions do entail different demands, as many of the PF tasks are non-cued (initiating checklists and monitoring status) while those of the PNF are cued (responding to a communication). Thus, the BE data, broken out by crew position, might serve as a more sensitive indicator of MT demands. Table 4 shows the correlations between the predictor variables and the BE criterion variables for PF and PNF.

There is a significant correlation between *SYNWIN* and the Period 2 BE score for those assigned the PNF role, but not for PF. For those in the PNF role, the lower the score on *SYNWIN*, the higher Period 2 BE Error score.

There is also a significant correlation between the Raven and Period 1 BE Error Scores for both PF and PNF, where the former correlation is in the opposite predicted direction.

CONCLUSIONS

The scenario employed in this study was purposely designed to impose varied, intense demands on both the PF and PNF. The intent was to challenge both crew positions with unexpected events in the hope that opportunities to engage in intensive MT would occur and be captured by the two observers. To aid in that assessment, a specialized profile of behavioral events was constructed, so observers could target "slices" of behavior where those demands would be most apparent. In this regard, all aspects of the data analysis showed that the scenario was successful in imposing multiple task demands, confusions, errors, and difficulties of virtually every stripe.

Given the difficulty of the scenario and its clear requirements to multi-task, the lack of a consistent statistical relationship between our predictive tests and scenario performance was admittedly disappointing. The *MTAT* did not exhibit a predictive pattern with any of the composite indices of flight scenario performance for either crew position. *SYNWIN* was a significant predictor of student performance in the second (and most difficult) period of the scenario for the PNF position, though not for PFs. We believe that part of the problem was due to the inherently small sample size, where the presence of even one outlier – in our case, the student who had the highest MTAT score did poorly in the flight scenario due to frustration and other non-MT factors – can obscure a true effect. On the other hand, the accuracy component of *MTAT* was significantly related to two of the archival criterion measures, number of extra flights and number of hours to instrument rating. *MTAT's* response time metric also

Table 4. Correlations between Behavioral Event Scores and Predictor Variables by Crew Position

	Predictor Variable					
Criterion Measure	MTAT Avg Acc	MTAT Avg RT	SYNWIN Avg	Raven	Processing Speed	Math Reasoning
PF Position (N=14)						
Period 1 BE Score	.22	-.07	-.07	.59*	.39	.00
Stabilization BE Score	.02	.12	-.15	-.11	.14	-.44**
Period 2 BE Score	.16	-.07	-.18	.21	.28	-.46*
PNF Position (N=15)						
Period 1 BE Score	.18	.01	-.18	-.44*	.05	-.09
Stabilization BE Score	.02	.12	.30	-.16	-.09	.39
Period 2 BE Score	-.27	.07	-.60*	-.31	-.21	-.23

*p < .05, ** p<.01

predicted number of extra flights at a marginally significant level. Both archival criterion measures were also significantly related to the Raven's measure of fluid intelligence. The patterns exhibited in the correlation matrix suggest that the variance in the MT tests is not fully shared with the Raven test.

Three major conclusions can be drawn from the study. First, the technical inability to precisely record the temporal flow of behavioral events impeded our ability to obtain true *efficiency* measures aggregated over a period of time, which might have been more sensitive indicators of MT performance. For example, the number of BEs performed correctly per unit time, or the average time it takes to perform a task in this context, would more accurately represent efficiency, which we now believe is a better metric for MT behavior. From a cognitive standpoint, MT ability may be tantamount to the efficiency with which information can be processed and tasks performed in a time-compressed manner. For example, *MTAT* and *SYNWIN* predicted two archival measures of pilot performance (number of hours and extra flights to achieve instrument rating) that, owing to an inherent time component, are measures of efficiency. Yet, we also see those same MT measures failed to predict GPA, which is *not* an efficiency metric. This view is also consistent with the *MTAT's* response time metric being more predictive than points earned (accuracy).

Second, MT assessment should be tailored to a given job position. We found the MT demands imposed on PF and PNF to be quite different, and not surprisingly, there was evidence in our data that our different MT tests were differentially sensitive to performance in the two positions. From a broader perspective, this suggests that different versions of the *MTAT* could be developed, with each variant emphasizing different cognitive attributes of a job.

Finally, it is apparent that the most important aspect of MT assessment is having a detailed understanding of the types of behaviors that must be multi-tasked and the frequency with which MT behaviors are successfully performed or are deficient. In this study, we quantified the frequency with which MT behaviors failed to meet acceptable flight standards during the scenario. Across both crew positions, participants experienced one or more breakdowns in such MT behaviors as: anticipating upcoming task needs, staying "ahead of the aircraft," anticipating upcoming procedural events, juggling radio operations and other duties, not letting things drop of visual scan while doing other duties, prioritizing tasks in accordance with aviate-communicate-navigation, and doing the most important tasks first. Viewing this list, it is apparent that many of the student pilots experienced problems associated with anticipating or thinking about what was going to happen next. Indeed, when faced with the competing demands of an intensive MT environment, participants had difficulty mentally projecting what subsequent activities would be optimal given the current task load. We believe these MT skills to be trainable, however, and we will be working with ASU to develop such a training curriculum in the future.

REFERENCES

Burgess, P. (2000). Strategy application disorder: The role of the frontal lobes in human multi-tasking. *Psychological research*, *63*, 279-288.

Colvin, K. (1999, December 23). *Initial identification of factors that affect task prioritization on the flight deck*. Informal report, 1-26. San Luis Obispo, CA: Dept. of Industrial & Manufacturing Engineering.

Dismukes, K. (2003, October). *Managing interruptions, distractions and concurrent task demands*. ATA AQP Annual Conference.

Elsmore, T. F. (1994). SYNWORK1: A PC-based tool for assessment of performance in a simulated work environment. *Behavior Research Methods, Instruments, & Computers,* 26, 421-426.

Fischer, S. C., Morrin, K. A., & Joslyn, S. (2003). *Measuring Multi-tasking Ability*. (DTIC # ADA417039)

Fischer, S. C., Mautone, P. D., Morrin, K. A., & Joslyn, S. (2005). *Development of an abstract test of multi-tasking ability*. Technical Report to ONR.

Funk, K., Chou, C., & Madhavan, D. (1999, December 23). *Preliminary cockpit task management research at Oregon State University*. Informal report, 1-11. Corvallis, OR: Industrial and Manufacturing Engineering.

Iani, C. & Wickens, C.D. (2007). Factors affecting task management in aviation. *Human Factors*, *49*, 16-24.

Nullmeyer, R.T., Spiker, V.A., Golas, K.C., Logan, R.C., & Clemons, L. (2006, November). The effectiveness of a PC-based C-130 crew resource management aircrew training device. 27th I/ITSEC.

Raven, J. C. (2000). Advanced Progressive Matrices, Set I and Set II. Oxford: Oxford Psychologists Press.

Salthouse, T. A. & Babcock, R. L. (1991) Decomposing Adult Age Differences in Working Memory. *Developmental Psychology,* 27, 763-776

Wickens, C.D. (1999). Cognitive factors in aviation. In F.T. Durso (Ed.) *Handbook of Applied Cognition*, 247-282. Chichester, England: John Wiley & Sons.

Winkler, R.L. & Hays, W.L. (1975). *Statistics: Probability, inference, and decision*. New York: Holt.