

Running Head: IDENTIFYING PERFORMANCE

Identifying Performance in Complex Dynamic Environments

Rickey P. Thomas
Kansas State University
Department of Psychology
1100 Mid-Campus Drive
Manhattan, KS 66506
Phone: (785) 532-6851
e-mail: rickey@ksu.edu

Julia Pounds, Ph.D.
FAA Civil Aeromedical Institute
FAA/CAMI/AAM-510
P.O. Box 25082
Oklahoma City, OK 73125
Phone: (405) 954-1199
Fax: (405) 954-4852
e-mail: julia.pounds@faa.gov

Abstract

The study tested the Cochran-Weiss-Shanteau (CWS) index, a performance-based index of expertise which integrates discrimination and consistency of judgments. Performance of teams and individual operators in an air traffic control (ATC) microworld environment were assessed to determine if the results of CWS analyses would be in the expected directions. Results showed that higher CWS scores were associated with superior individual and team performance. The CWS indices were also sensitive to changes in task complexity, further testing the index. This research represents our first extension of CWS analysis to evaluation of performance in complex dynamic environments.

Introduction

Scientists have long been challenged to develop valid measures of expert performance. To that end, performance measures for air traffic control have been developed within the FAA for over thirty years (Manning et al., in press; Sollenberger, Stein, & Gromelski, 1997). In that time, various performance measures have been proposed and applied to ATC tasks. However, although human performance is clearly critical in joint systems (Hollnagel, Cacciabue, & Hoc, 1995), an objective index has remained elusive¹, particularly for performance in dynamic environments.²

To address these problems, a project was undertaken to develop and test the ability of the Cochran-Weiss-Shanteau (CWS) statistic to meaningfully analyze data from a complex, heavily cognitive task environment. The data was collected using a low fidelity simulation of an air traffic control task. This paper describes the results of these analyses.

Background

Bailey (1989) defined human performance as “the *result* of a pattern of *actions* carried out to satisfy an *objective* according to some *standard*.” He noted that “[t]hings change when people perform” (pg. 6). Although actions can be either observable behaviors or non-observable mental processes, identification of standards representing what is meant by adequate performance have

¹ Ward Edwards has rightly asked us why an index of performance should be free of subjectivity. He pointed out that all sources of subjectivity are not the same and questioned whether one source of data is more acceptable than another for the purposes of CWS computations and evaluation of performance.

² Dynamic environments (e.g., Freed, 1998) are characterized as real-time, realistically complex, uncertain task environments in which multiple, potentially interacting tasks need to be managed to achieve one or multiple goals. The environment typically includes a time-pressure to prioritize, shifting goal states, and interrupted planning. Often, events in dynamic environments continue to develop whether or not the decision maker intervenes. Conversely, a static environment can often be characterized as a task environment where the situation is not evolving over time independently from the person. A static environment may also not include a time

been difficult, often depending on the task objectives and the situation in which the performance takes place. Typical standards of human performance include user accuracy, user speed, skill development time, user satisfaction, etc. (Bailey, 1989). Bailey (1989) pointed out that performance is influenced by human strengths and weaknesses as well as the specific work activities to be performed and the context in which the performance is required.

One difficulty is that performance can be thought of in several ways. In the most general sense, performance is behavior. Particular behaviors in specific contexts are often compared with specific performance criteria. This is useful when performance involves motor activities (Bailey, 1989; Fleischmann & Parker, 1962; Grandjean, 1971). However, performance can also be thought of as either overt, observable behavior or covert, hypothesized mental processes, or as a result of both.

In many domains, such as air traffic control (ATC), performance depends largely on covert cognitive processes, such as attending, deciding, and predicting that culminate in observable actions such as issuing an altitude change to an aircraft. Because of the difficulty associated with measuring covert processes in dynamic environments such as air traffic control, understanding and measuring the relationship between an air traffic controller's mental processes and performance has been difficult.

According to Manning et al. (in press), three difficulties arise in the development of measures of Air Traffic Control Specialists' (ATCS) performance. First, the complex dynamic environment of ATC does not lend itself well to standard measurement techniques. Thus, performance in dynamic domains such as air traffic control makes the use of traditional judgment and rating tasks difficult. The problem with finding an independent index of performance for the ATC domain is that methods used to determine performance in static judgment tasks may not be

pressure element thus allowing more long-term planning.

appropriate for judging how performance changes in dynamic environments (Manning et al., 1999). Second, because of the difficulties in developing direct behavioral measures of operator performance, researchers have relied instead on the observations of subject matter experts (SMEs) to evaluate performance (Manning et al., in press). However, as is the case with experts in other domains (Shanteau, in press), ATC SMEs often disagree in their evaluations and actions--controllers in the same or similar situations may use different strategies to solve the identical problem. Thus, an index of performance should be as free of as many different sources of subjectivity as possible¹ and it should be amenable to quantitative comparisons. Third, the underlying cognitive processes that mediate controller actions are difficult to specify.

To address these problems, a project was undertaken to identify a performance index for complex, heavily cognitive tasks such as air traffic control. Following Gilbert (1978), an index for such a complex subject should be judged by its usefulness, its simplicity, and its coherence--that is, by its utility, parsimony, and elegance.

Cochran-Weiss-Shanteau (CWS) Index

CWS was first developed and successfully tested against existing data sets from expert judgments of static stimuli. Analyses were conducted to determine if CWS could be used to analyze expert performance in dynamic scenarios with characteristics resembling air traffic control. Following a brief review of the CWS methodology, this paper presents results from this test.

CWS is based on the premise that evaluative skill underlies all expertise and, further, that expert evaluative skill must satisfy the two necessary criteria of discrimination and consistency (Shanteau et al., in press). This performance index parallels Cochran's (1943) suggestion that a

discrimination/inconsistency ratio can be used to measure the effectiveness of a response instrument. Cochran argued that an effective response instrument is one that allows the subject to express perceived differences among the stimuli in a consistent way. Shanteau et al. (in press) propose that similar reasoning be applied to people; an expert judge should discriminate consistently. That is, expert performance must satisfy two basic criteria: (1) experts must be able to discriminate among the relevant stimuli in their domain; and (2) those discriminations must be expressed consistently.

Using CWS, the researcher can determine a putative expert's evaluative skill by repeatedly presenting a set of stimulus objects. The candidate's responses are analyzed to generate measures of discrimination and inconsistency. Examining variation in the candidate's responses to different stimuli gauges discrimination. Inconsistency is assessed by variation in the candidate's responses to the same stimuli. The CWS index is the ratio of discrimination to inconsistency; the larger the value of the index, the greater the exhibited degree of expertise.

It is crucial to note that the CWS index functions only in a comparative way; the index is tied to the particular set of stimuli employed in the study. CWS should not be used to compare experts across studies using different stimulus sets or across domains.

The CWS index of expert performance has been successfully applied to many domains of expertise, including auditing, livestock judging, and personnel selection (for further details see Shanteau et al., in press). The index distinguished the performance of acknowledged experts from their less experienced counterparts. For example, CWS analysis distinguished between the performance of expert and novice auditors. It also distinguished between experts in different specialties (i.e., swine vs. cattle) within the livestock domain for judgments of the breeding quality of swine.

Although these analyses have been fruitful for the conceptual development of CWS, they need to be interpreted carefully. The data used for these CWS analyses reflected participants' evaluations about static stimuli. For example, the livestock judges were shown line drawings of hogs and asked to give each a rating of breeding quality. Auditors were presented financial cases; each described by a list of 16 cues and gave an assessment of "going concern" for each business.

These analyses were conducted for the purposes of testing CWS against data from studies using known groups of experts and novices. The goal was to determine if CWS would give results reflecting the original analyses. Any added value to the original authors' analyses is regarded as serendipitous. Although CWS was validated against competent performances in these cases, the levels of expertise were determined by the original designs, not by CWS analysis. The hazard here is to infer that one's level of experience can be used to infer some similar level of expertise.

However, these findings *do* suggested that the utility of the CWS procedure for static judgment tasks might be extended to make the problem of describing human performance in dynamic situations more tractable. However, it is not appropriate to study inherently dynamic domains, such as ATC, relying on static stimuli such as snapshot pictures (Manning et al., in press). Two types of problems had to be solved before CWS could be conducted on data collected from dynamic situations: (1) development of methods to repeat stimuli within individuals and (2) identification of dependent variables appropriately reflecting performance.

Stimulus Requirements

There are two requirements that must be satisfied to apply the CWS methodology to a set of stimulus objects. The first requirement, that individual objects vary on some dimension, is easy to satisfy in dynamic environments. Stimuli within dynamic environments are constantly changing, and differences in behavior across changes in stimuli can be used to assess the discriminating power of an operator.

The second requirement, that stimuli be repeated within individuals, is difficult to satisfy in dynamic environments. In dynamic environments the stimuli change with or without intervention by the participant, although the participant's actions can contribute to the changes, making it difficult to repeat the identical stimulus from beginning to end within (or between) participants. Moreover, in dynamic environments operators must maintain timely control in situations that will continue to change with or without their inputs, making it virtually impossible to assess the consistency of the operators.

Appropriate Dependent Variables

Selecting appropriate dependent variables to represent performance is important. For a dependent variable to be an appropriate candidate for CWS analyses, it must have high task validity. That is, the variable should reflect the performance required by the operator in response to the environment in which he or she typically works. One strength of the CWS approach is that the discrimination and the inconsistency of the operator's behavior can be assessed without also having to identify the specific underlying cognitive processes that precede the behavior.

Extending CWS to Dynamic Environments

To objectively describe the performance of experts who work on complex, changeable tasks, it was necessary to first extend the CWS methodology to dynamic environments. The extension of CWS to dynamic environments relied on data collected using a computer-based *microworld* methodology that bridges the laboratory and real-time environments. A microworld environment allows researchers to simulate real-world problems while still maintaining a certain degree of control. The use of microworlds to study problems in dynamic environments was popularized by the European research community, and is becoming more popular because of growing dissatisfaction with the inability of existing psychological theories to address applied issues (Pounds, 1997). See Buchner (1995) for a discussion of this problem. One reason for this could be that most psychological theories are developed in the laboratory, with little thought given to extending their application to real-world problems.

To date, microworld environments have been developed to study problems in domains ranging from resource allocation (Buchner, 1995) to fire fighting (Omodei & Wearing, 1995a). Microworlds allow the researcher to bring complex systems into the lab, simplify complex systems enough to allow analysis, and assess performance based on relevant elements. Advantages of using microworlds are that they allow researchers to control stimulus manipulations and collect measurements while not sacrificing complex system characteristics. Disadvantages include the possibility that the researchers might lose or ignore important task characteristics.

Using a microworld for research also requires the researcher to consider a variety of trade-offs, such as the naturalness and fidelity of the task to be used, how clearly the goals are defined by the researcher and understood by participants, the transparency of the task to

participants, the autonomy of the stimulus, and the comparability of the microworld task to the real-world task.

New computer technologies facilitate the use of microworld environments in controlled empirical studies of dynamic complex environments. Further, computerized microworlds allow researchers to study both the process of accomplishing an end state performance level and the nature of performance outcomes themselves. The use of a microworld to test CWS permitted the stimulus requirements for CWS analysis to be met while testing extensions of the requirements in a dynamic domain.

Controller Teamwork Evaluation and Assessment Methodology (CTEAM)

The Controller Teamwork Evaluation and Assessment Methodology (CTEAM) microworld was developed by the Human Resources Research Division of the Civil Aeromedical Institute (CAMI) as a multi-sector research platform that affords a low-fidelity simulation of radar-based air traffic control (ATC) tasks (Bailey, Broach, Thompson, & Enos, 1999). It is engaging to participants, much like playing a video game. No air traffic control domain expertise is necessary. However, participants can become adept at controlling CTEAM aircraft within the microworld procedures. For this reason, in these studies that test CWS in dynamic environments, participants needed not be air traffic controllers for the researchers to examine basic cognitive processes related to how aircraft are maneuvered through the microworld's airspace. This also permits a more efficient use of personnel resources than if professional air traffic controllers were required.

Recall that CWS analysis requires stimuli that can be varied on some dimension relevant to the evaluations and that the stimuli be replicable. CTEAM permits the experimenter to

configure the stimuli on dimensions deemed relevant to the study. The software also enables replication of the scenario. Following guidelines suggested by Buchner (1995) and by Frensch and Funk (1995), CTEAM was determined to be an appropriate microworld to test the utility of CWS in dynamic tasks (Appendix A). The software records several individual and team performance scores.

Figure 1 illustrates a CTEAM “controller’s” or “operator’s” view of the airspace (a single sector view). The sector information panel (left side) displays performance feedback. The sector operator issues heading, altitude, speed, and handoff clearances using the command panel (right side). The inter-sector message panel (bottom) allows the operator to communicate with other operators who control the adjacent sectors.

Each CTEAM sector operator views only his or her sector. CTEAM operators must issue control actions in order to maintain separation of aircraft in their sector and to direct the aircraft through their prescribed route. Furthermore, the team members are responsible for coordinating handoffs (relinquishing control or gaining control of aircraft) from operators responsible for adjacent sectors. For aircraft to reach their destinations, the controllers must work as a team. Further description of the system is given in Appendix B.

Stimulus Requirements: Freeze-Frame vs. Free-Frame Approaches

Two possible solutions to the problem of repeatability of stimuli in dynamic domains have been conceptualized: the “freeze-frame” and “free-frame” approaches. These approaches lie on a continuum. At one end of the continuum is the traditional experimental psychology laboratory study. In such studies, the researcher uses snapshot or static stimuli. At the other end

is the dynamic real-world, environmentally intact field study. In such domains, researchers can observe behavior, but have no control over relevant independent variables.

Freeze-Frame presentation of stimuli. In this approach, the task runs to a specific point and then stops, at which point the participant is queried for a response, i.e., a control action. This resembles the Situation Awareness Global Assessment Technique(SAGAT) procedure (Endsley, 1988b) used to measure situation awareness (SA). However, it differs in that SAGAT requires the participant to turn away from the display to answer a series of questions based on their memory. In the freeze-frame approach, in contrast, controllers are asked for their next control actions, evaluation, communication, etc. The freeze-frame approach allows the same stimuli to be repeated both between and within participants. Essentially, the freeze-frame approach takes a continuous dynamic environment and divides it into discrete elements.

Free-Frame presentation of stimuli. Using this approach, the participants' responses are tracked continuously throughout the task, without stopping the scenarios. Applying the free-frame approach, the experimenter sets the parameters of the task (often according to an experimental design). However, the experimenter does not attempt to control for the dependence among the responses. In the free-frame approach, both the response as well as the overall pattern of responding are available. If the participant is discriminating accurately, then his/her behavior should change as the environment changes. If the participant is consistent, then the behavior should be similar across similar environments (i.e., replications). Therefore, the submeasures needed by CWS (discrimination and inconsistency) are available in a dynamic setting.

Research Question

The goal of the present study was to determine whether the CWS method of analysis could be adapted to analyze data from a dynamic task environment with characteristics and constraints similar to air traffic control. If so, this would provide evidence that CWS analysis could be extended to data collected during dynamic tasks and thus could possibly be further developed for high-fidelity dynamic scenarios.

Method

An archival data set was used to test the CWS analysis methodology. The original data were collected by researchers at the FAA Civil Aeronautical Medical Institute (CAMI) in Oklahoma City, Oklahoma. The original study examined teamwork, learning, and performance using a low-fidelity air traffic control microworld. Details are available in Bailey (1998) and Pounds and Bailey (1999). The methods from the original study that provided the data for the analysis are first reviewed, followed by the tests of the CWS index.

Archival Data

The archival data were collected using the Controller Teamwork Evaluation and Assessment Methodology (CTEAM).

Participants. Participants recruited from the local area participated in the original study. After screening for basic computer skills and aviation experience (no training as a pilot or an ATCS), two-hundred-thirty-six people were randomly assigned to one of 59 four-person teams. Participants were then screened to eliminate from further participation those candidates who could not grasp a basic understanding of the procedures required to carry out the air traffic

control task (e.g., make appropriate changes in aircraft direction, speed, and altitude).

Participants completed three, ten-minute screening scenarios. Those who met specified performance criteria (80% of aircraft reached the assigned destination during at least one of three, ten minute scenarios) were asked to return for the practice and experimental sessions.

Stimulus scenarios. Three predefined scenarios were used. They represented three levels of aircraft density (Low, Medium, and High). Aircraft density was defined as the number of aircraft per unit time presented to the participant in an inactive state throughout the duration of the simulation. In a previous study aircraft density was found to be the main determinant of CTEAM sector complexity (Broach et. al., 1998).

Task. Returning participants received practice on three ten-minute scenarios, one at each level of density. Individuals were randomly assigned to teams, and each team was then randomly assigned to one of the three levels of density. Each team was tested on two replications of the experimental scenario.

Dependent measures. CTEAM records data that allows analysis at both the individual and team levels. Individual performance measures include (1) the number of aircraft-to-aircraft separation errors and (2) the number of aircraft-to-boundary separation errors. A separation error is recorded when an aircraft travels within five miles of another aircraft (at the same altitude) or sector boundary (as calculated using CTEAM sector coordinates).

CTEAM also records the aircraft that reach their assigned destination airport. For example, an aircraft started in sector A might have a designated destination airport in sector D. CTEAM operators must work as a team for an aircraft to reach its final destination because the aircraft must travel through all four sectors successfully. Reaching the destination requires skillful controlling of aircraft in the sectors and coordinated handoffs between sectors. In the

CTEAM environment, successful performance of a team can be characterized by the percent of aircraft that successfully reach their final destinations.

Evaluation of the CWS Index³

Dependent measures used to calculate CWS can be derived from a variety of types of measures recorded by the system. For each CWS calculation, discrimination and inconsistency elements are calculated relative to the same dependent measure. In the following cases, CWS was calculated separately for variables 'Time at start' and for 'Number of control actions.'

Dependent measure: Time at start

CWS analyses were conducted to examine both individual and team performance. CWS was calculated using the time an inactive aircraft was started (activated). The dependent measure was chosen because of its task validity and suitability. The participants in the Pounds and Bailey experiment were told to start aircraft in the order in which they were presented and to start them as quickly as possible. They were also told not to compromise safety (i.e., avoid separation errors or crashes). A CWS analysis of the time an inactive aircraft is started captures important aspects of ATCS performance. Moving the aircraft from an inactive state to an active state in the appropriate order maximizes the discrimination score. Starting the aircraft at the same time in repeated simulations minimizes inconsistency.

One can view CWS for start time as a measure of cognitive workload or vigilance. The task calls for starting an aircraft as quickly as possible. If this does not happen, it is reasonable to

³ For calculation algorithms see Shanteau, Weiss, Thomas, and Pounds (2000). For the theoretical and methodological rationale see Shanteau, Weiss, Thomas, and Pounds (in press).

infer that the controller was engaged in some other task (such as maintaining separation) that diverted attention. In addition, the participant must recall the order in which the inactive aircraft appeared. If these cues are not recalled, the aircraft will be started out of order. Our rationale for choosing this particular approach is similar to that of Seven (1989), who proposed that ATC performance on secondary tasks provides an unobtrusive method to measure cognitive workload.

This approach assumes that performance is related to an activation strategy adopted by the participant to cope with the demands of the scenario. For example, to be meaningfully indexed to the dependent variable 'time to activate an inactive aircraft' high performance depends on a consistent activation strategy. That is, inconsistency will increase if the controller adopts a different strategy from time 1 to time 2, either within the scenario for different aircraft or between scenarios. Such inconsistencies include starting the aircraft in different orders or starting the aircraft at vastly different times.

Dependent measure: Number of control actions

CWS was calculated using the average number of control actions issued to aircraft. Control actions are the instructions issued to aircraft, such as changes in altitude, vector, and speed. These are routinely recorded by the CTEAM software and can also be retrieved from recordings in high fidelity environments. Aircraft have different routes that require different numbers of control actions. The extent to which aircraft with different routes are issued different numbers of control actions captures discrimination. Inconsistency is demonstrated when a CTEAM controller issues different numbers of control actions to the same aircraft across repeated runs of the scenario. As aircraft density between scenarios increases, the increasing

workload demand put on the controllers would be expected to increase the inconsistency with which control actions were applied, thus resulting in a lower CWS index.

Method

A *free-frame* method was used to present the stimulus scenarios. Within each scenario, the aircraft were presented to each controller in the same relative position in the sector at identical times in simulation. However, more aircraft were presented in higher aircraft density scenarios, so individual controllers, and consequently teams, were presented different numbers of aircraft.

The aircraft analyzed for the CWS analyses 1) originated in the controller's own sector, and 2) were presented to every controller in the experiment regardless of the scenario. For instance, at 660 seconds into the simulation every controller in the experiment was presented aircraft 111 at coordinates (x, y) on their respective sector display panel. All aspects of the aircraft, including its route, were identical for all controllers in the study. This method of analysis satisfies the constraint of similar environments, thus making behavior across these aircraft comparable.

A participant's application of same or different actions while controlling different aircraft in the sector reflects the discrimination component of CWS. On the other hand, inconsistency captures variation in responding to the aircraft, e.g., whether the CTEAM operator starts the aircraft close to the same time in repeated scenarios or issues a similar number of control actions in repeated scenarios.

Results (Time at Start)

Individual Performance

At the individual level, the CWS index is interpreted as a participant's ability to discriminate and perform consistently. To test this with the archival data, CWS was calculated for each participant using start time as the dependent variable.

The resulting pattern of CWS scores suggested that CWS was sensitive to the manipulation of aircraft density. Figure 2 shows that CWS decreases as the aircraft density of the scenarios increase, $F(2, 233) = 46.10$, $\text{Eta}^2 = .283$, $p < .001$. As the complexity of the task increased the activation strategy of the operators became disrupted.

We tested the extent that CWS scores captured performance relative to the CTEAM individual performance measures: aircraft-to-aircraft separation errors and aircraft-to-boundary separation errors. As shown in Table 1, higher CWS indices did correspond to better individual performance. Fewer separation errors (Table 1) were reflected in higher CWS indices. Thus, CWS did capture individual operator performance.

Team Performance

CWS analysis were also conducted for teams, using start time as the dependent measure. As shown in Figure 3, CWS was sensitive to the changes in aircraft density, $F(2,56) = 26.41$, $\text{Eta}^2 = .485$, $p < .001$. As the complexity of the task increased, team activation strategies became disrupted.

We tested the extent that CWS captured performance relative to the standard CTEAM team performance measure: percent of aircraft reaching destination. As shown in Table 2, higher CWS indices were associated with a higher percentage of aircraft successfully reaching their

destinations. Thus, CWS did capture superior team performance although it is based on a different aspect of the behavior.

Results (Control Actions)

Individual Performance

CWS analyses based on the number of control actions suggested corresponding sensitivity to the level of aircraft density (Figure 4); although this effect was not statistically significant, $F(2, 232) = 1.60, p = .20$. The aircraft density of the scenario may not be an appropriate measure of the level of aircraft density experienced by the individual operator. It is true that a team of CTEAM operators in a higher density scenario do start and control more aircraft than teams in lower density scenarios. However, individual operators within a team have control over their own workload; they may decide not to start all of their aircraft or reject control of aircraft from neighboring sectors. The aircraft density of the scenario is more appropriate for describing the level of aircraft density experienced by the team than describing the level of aircraft density experienced by the individual operator working within a team.

We tested the extent that CWS captured performance relative to the CTEAM individual performance measures: aircraft-to-aircraft separation errors and aircraft-to-boundary separation errors. As shown in Table 3, higher CWS indices did correspond to better individual performance. Fewer separation errors (Table 3) were reflected in higher CWS indices. Thus, CWS did capture individual operator performance. Again, the magnitude of the relation between CWS and the CTEAM individual performance measures was probably underestimated because the aircraft density of the scenarios may not be an appropriate measure of the level of aircraft density experienced by individual operators within a team.

Team Performance

CWS analysis were conducted for teams using control actions as the dependent measure. As shown in Figure 5, CWS was sensitive to the changes in aircraft density, $F(2,56) = 12.87$, $\text{Eta}^2 = .315$, $p < .001$. As the complexity of the task increased the teams' control strategies became disrupted.

We also tested the extent that CWS captured performance relative to the CTEAM team performance measure (percent of aircraft reaching destination). As shown in Table 4, higher CWS indices did correspond to a higher percentage of aircraft successfully reaching their destinations. Thus, CWS did capture superior team performance.

In sum, these results demonstrated that the test of the CWS index in a dynamic environment was successful. Results showed that CWS could be calculated from data collected in a dynamic, free-frame presentation and that the CWS index reflected better and/or worse performance in meaningful directions, e.g., better performance according to CTEAM measures corresponded to higher CWS indices. Moreover, CWS captured differences in performance resulting from manipulations of the stimulus.

Discussion

The results indicate that meaningful CWS indices can be computed from data collected in a dynamic environment. These results extend those reported by Shanteau, Weiss, Thomas, and Pounds (in press) that demonstrated the ability of CWS to measure expert performance in static judgment tasks.

Results also demonstrated that CWS could be calculated from variables extracted from free-frame stimulus replications presented in the CTEAM microworld, eliminating the

requirement of static stimuli for CWS calculation. These analyses demonstrated that dynamic scenarios could satisfy the CWS requirement for analysis—that stimuli be replicated—thereby providing the inconsistency component of CWS.

Using the free-frame approach, the scenario variable of aircraft density could be experimentally manipulated in the CTEAM microworld environment. From performance variables derived thusly, CWS analysis demonstrated that the index was sensitive to the stimulus manipulation—a requirement that provides the discrimination component of CWS.

Moreover, the CWS indices described the performance of both individuals and teams in the appropriate directions. The CWS score for an individual decreased as the number of separation errors he or she committed increased. Similarly, the computed CWS scores for the teams that managed to get a greater percentage of their aircraft to the assigned airports resulted in larger CWS scores compared to teams who did not.

Once it was demonstrated that CWS could be calculated from data gathered in a dynamic environment, the number of control actions was examined for its utility as a performance measure. This variable is one that is also available in higher-fidelity simulations such as the FAA Academy's Simulation, Integration of Air and Ground Links (SIGNAL) and dynamic simulation (DYSIM) trainers.

CWS analysis using number of control actions as the performance measure performed in the expected direction. This demonstration is further evidence that CWS is an analytic tool that can also be used in high-fidelity environments. Studies to this end will be conducted.

An application for CWS would be to determine whether the index can be used to describe team strategies for coordination. For instance, when calculated for team performance, a CWS index based on (for example) the total number of control actions issued to all aircraft, the

measure is assumed to indicate how well the individual operators' control strategies complement one another. One hypothesis would be that high performance teams develop stable strategies, which are congruent with other team members and are applied consistently to manage workload. If the four CTEAM operators use comparable strategies, then they would probably move aircraft through their sectors similarly. This similarity in policies could then influence the distribution of workload (aircraft density) more evenly throughout the four sectors. An alternate hypothesis is that teams performing well have developed compensatory strategies to aid their team members.

We want to explore whether CWS can be used to track changes in performance during learning. For example, as operators improve in performance CWS indices should increase correspondingly. We have conducted a pilot study that suggests that inconsistency decreases with training. It may be that discrimination is less labile, and that limitations in discrimination ability can be identified and used in selection. Any discrepancies can be evaluated to determine if the lack of progress is due to either a lack of discrimination ability or a lack of consistency. For example, when calculated for individual performance, research might be able to determine whether it is the individual's skill in discriminating between stimuli, the consistency of judgment, or both that contribute to high performance. This knowledge can then be used to inform training.

References

- Bailey, L. L. (March, 1998). *A look inside group dynamics: The role of group feedback*. Presentation at the Oklahoma-Kansas Judgment and Decision Making Conference. Stillwater, Oklahoma.
- Bailey, L. L., Broach, D. M., Thompson, R. C., & Enos, R. J. (1999). *Controller teamwork evaluation and assessment methodology: (CTEAM): A scenario calibration study*. (DOT/FAA/AAM-99/24). Washington, DC: Federal Aviation Administration Office of Aviation Medicine. Available from: National Technical Information Service, Springfield, VA 22161.
- Bailey, R. W. (1989). *Human performance engineering*. Englewood Cliffs, NJ: Prentice-Hall.
- Buchner, A. (1995). Basic topics and approaches to the study of complex problem solving. In P. A. Frensch, & J. Funke, (Eds.), *Complex problem solving* (pp. 27-63). Hillsdale, NJ: Lawrence Erlbaum.
- Cochran, W. G. (1943). The comparison of different scales of measurement for experimental results. *Annals of Mathematical Statistics*, 14, 205-216.
- Endsley (1998b). Situation awareness global assessment technique (SAGAT). In *Proceedings of the IEEE 1998 National Aerospace and Electronics Conference-NAECON 1988*, 3, 789-795. New York: Institute of Electrical and Electronics Engineers.
- Fleischmann, E. A., & Parker, J. F. (1962). Factors in the retention and relearning of perceptual motor skill. *Journal of Experimental Psychology*, 64, 215-226.
- Freed, M. (1998). *Managing Multiple Tasks in Complex, Dynamic Environments*. Unpublished. Available at: <http://olias.arc.nasa.gov/cognition/papers/freed/aaai98.html>.

Frensch, P.,A., & Funke, J. (1995). *Complex problem solving*. Hillsdale, NJ: Lawrence Erlbaum.

Gilbert, T. F. (1978). *Human competence. Engineering worthy performance*. New York: McGraw-Hill.

Hollnagel, E., Cacciabue, P. C., & Hoc, J. (1995). Work with technology: Some fundamental issues. In J. Hoc, P. C. Cacciabue, & E. Hollnagel (Eds.), *Expertise and technology* (pp. 1-15). Hillsdale, NJ: Lawrence Erlbaum.

Manning, C. A., Mills, S., Mogilka, H., Hedge, J., Bruskiwicz, Pfliederer, E., (in preparation) Prediction of subjective ratings of air traffic controller performance by computer-derived measures and behavioral observations.

Omodei, M. M., & Wearing, A. J., (1995a). The Fire Chief microworld generating program: An illustration of computer-simulated microworlds as an experimental paradigm for studying complex decision-making behavior. *Behavior Research Methods, Instruments and Computers*, 27, 303-316.

Pounds, J. & Bailey, L. L. (1999). *Cognitive style and learning: Performance of Adaptors and Innovators in a novel dynamic task*. (DOT/FAA/AAM-99/12). Washington, DC: Federal Aviation Administration Office of Aviation Medicine. Available from: National Technical Information Service, Springfield, VA 22161

Shanteau, J. (in press). What does it mean when experts disagree? In Klein, et al., (Eds.) *To appear in: Naturalistic decision making*.

Shanteau, J., Weiss, D. J., Thomas, R., & Pounds, J. (in press). Performance-based Assessment of Expertise: How can you tell if someone is an expert? *European Journal of Operations Research*.

Shanteau, J., Weiss, D. J., Thomas, R. P., Pounds, J. (August, 2000). *Analysis of expertise: Application to medical decision making*. Presented at the Annual Meeting of the Mathematical Psychology Society. Kingston, Ontario.

Sollenberger, R. L., Stein, E. S., & Gromelski, S. (1997). *The development and evaluation of a behaviorally based rating form for assessing air traffic controller performance*. (FAA Technical Note). Atlantic City: William J. Hughes Technical Center. DOT/FAA/CT-TN96/16.

Appendix A

The system satisfies several of the concerns mentioned above (e.g., goal definition, performance measures) and the system conceptually “sits” between the laboratory and real-world extremes described by Frensch and Funk (1995), as illustrated by the following.

1. A scenario generation tool enables experimenters to control and manipulate variables within each scenario.
2. *Experimental control* of airspace configuration and air traffic is possible (e.g., number of airports, airport locations, positions, and directions, when and where new aircraft enter a given sector, the route of flight, the length of a given scenario, no-fly zones).
3. *Autonomy* of the system: the system presents preprogrammed stimuli (aircraft) to the participant, whom then controls the activation and trajectories of the stimuli. Once activated, the aircraft cannot be paused in flight by the participant.
4. *System effectiveness measures* are recorded (loss of separation, crashes, aircraft delay time, and proportion of aircraft reaching their destination) and can be used *as performance outcome measures*.
5. *Task-specific behaviors* are identified (issuance of aircraft heading, altitude, and speed changes).
6. *Clear rules* can be specified (e.g., land at airports only at speed “slow”).
7. Clearly specifiable *goals* exist (e.g., minimize delay, minimize separation errors and crashes, and maximize the number of aircraft reaching their assigned destination).

8. Both *individual and team-oriented behaviors* can be measured (e.g., communication exchanges and the timing of those exchanges between controllers as they negotiate the transfer of aircraft from one sector to the next).
9. Aircraft commands and controller communications are initiated by a mouse-activated communication panel.
10. A computer replay file is generated for replay later.
11. Data extraction is possible for individual and team performance.

Appendix B

The CTEAM research platform consists of five (four clients and a server) 80486/DX2 66 or faster personal computers operating under Windows NT 3.51. The software also runs under the Windows 95 & 98 operating platforms.

There are four airspace quadrants, referred to as sectors. Each sector has four gates. The gates that connect adjacent sectors are called *inter-sector gates* or *handoff gates*. Gates that do not connect to adjacent sectors are called *exit gates*. The airports are represented as circles with the rectangular cut in the airport representing the runway. The aircraft are represented as directional arrows where the direction of the arrow indicates the current heading of the aircraft. The data block above the aircraft displays information about the aircraft including the call sign, speed, altitude, and routing information.

Table 1

Correlation between CWS and CTEAM individual performance measures.

| Individual performance measures | CWS (Start Time) |
|------------------------------------------------------|------------------|
| <i>Average Aircraft-to-Aircraft Separation Error</i> | -.25** |
| <i>Average Aircraft-to-Boundary Separation Error</i> | -.41** |

Note. Correlations were calculated using Spearman's procedure, ** $p < .001$.

Table 2

Correlation between CWS and CTEAM team performance measure.

| Team performance measure | CWS (Start Time) |
|-----------------------------------------------------------|------------------|
| <i>Percent of Aircraft Reaching Final Destination</i> | .78** |

Note. Correlation was calculated using Spearman's procedure, ** $p < .001$.

Table 3

Correlation between CWS and CTEAM individual performance measures.

| Individual performance measures | CWS (Control Actions) |
|------------------------------------------------------|-----------------------|
| <i>Average Aircraft-to-Aircraft Separation Error</i> | -.15** |
| <i>Average Aircraft-to-Boundary Separation Error</i> | -.18** |

Note. Correlations were calculated using Spearman's procedure, ** $p < .001$.

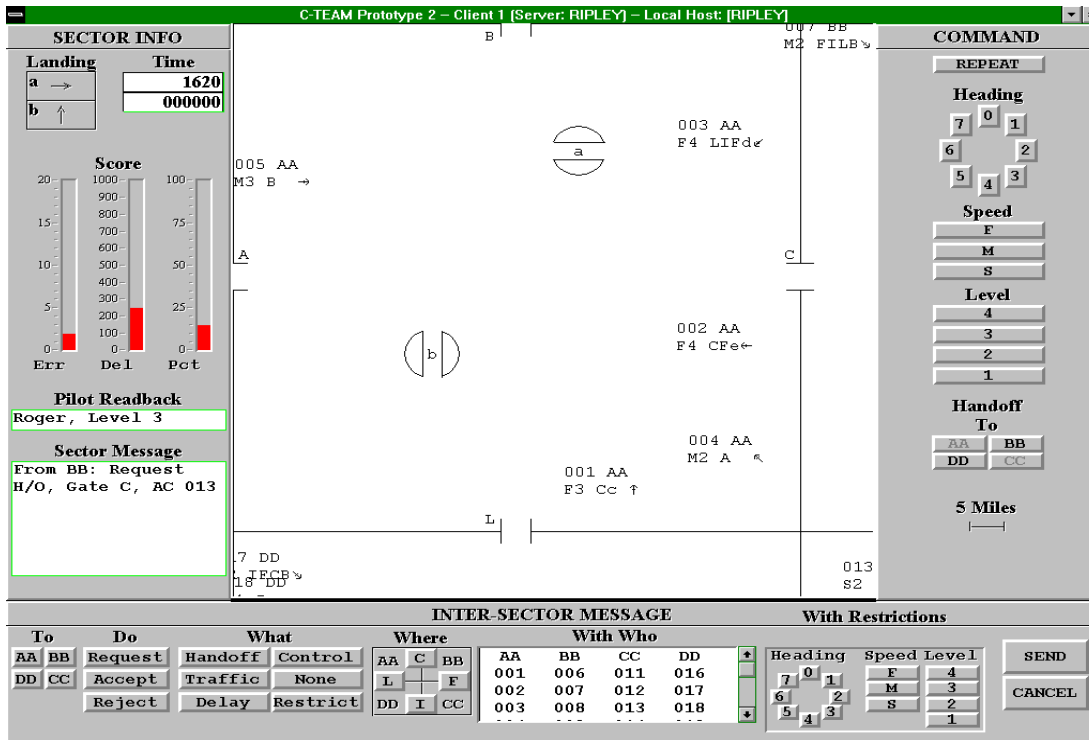
Table 4

Correlation between CWS and CTEAM team performance measure.

| Team performance measure | CWS (Control Actions) |
|-----------------------------------------------------------|-----------------------|
| <i>Percent of Aircraft Reaching Final Destination</i> | .55** |

Note. Correlation was calculated using Spearman's procedure, ** $p < .001$.

Figure 1



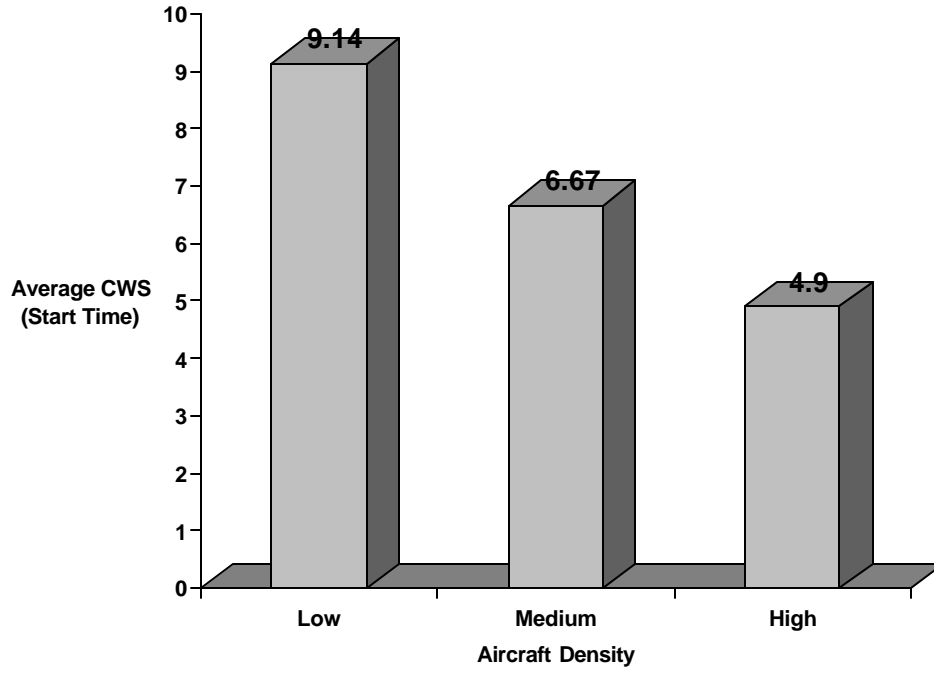


Figure 2

Individual performance (Start Time) decreases as scenario aircraft density increased.

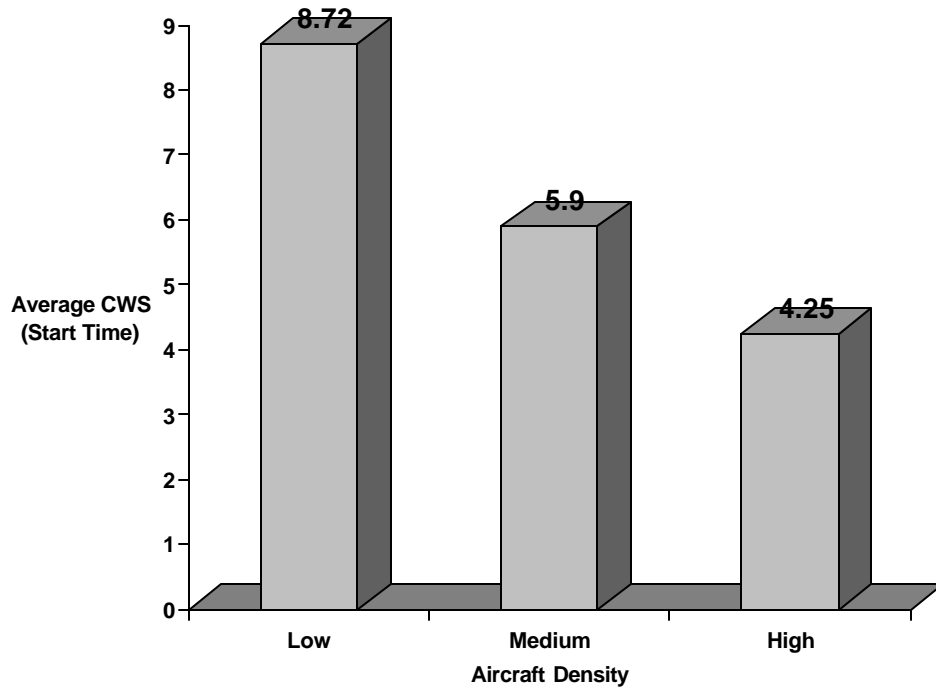


Figure 3

Team performance (Start Time) decreased as scenario aircraft density increased.

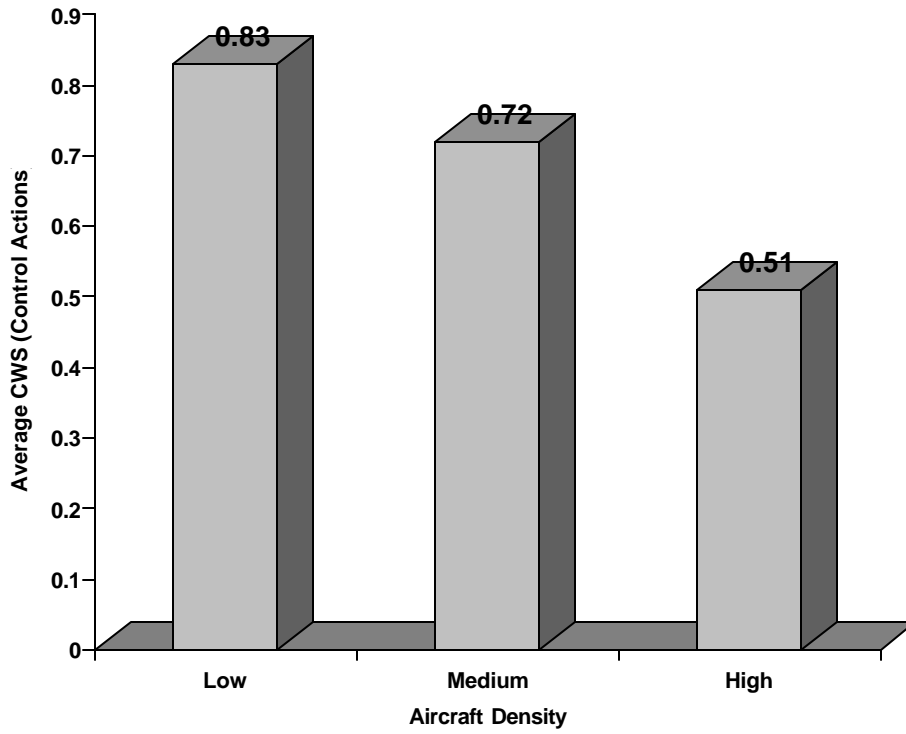


Figure 4

Individual performance (Control Actions) decreased as scenario aircraft density increased.

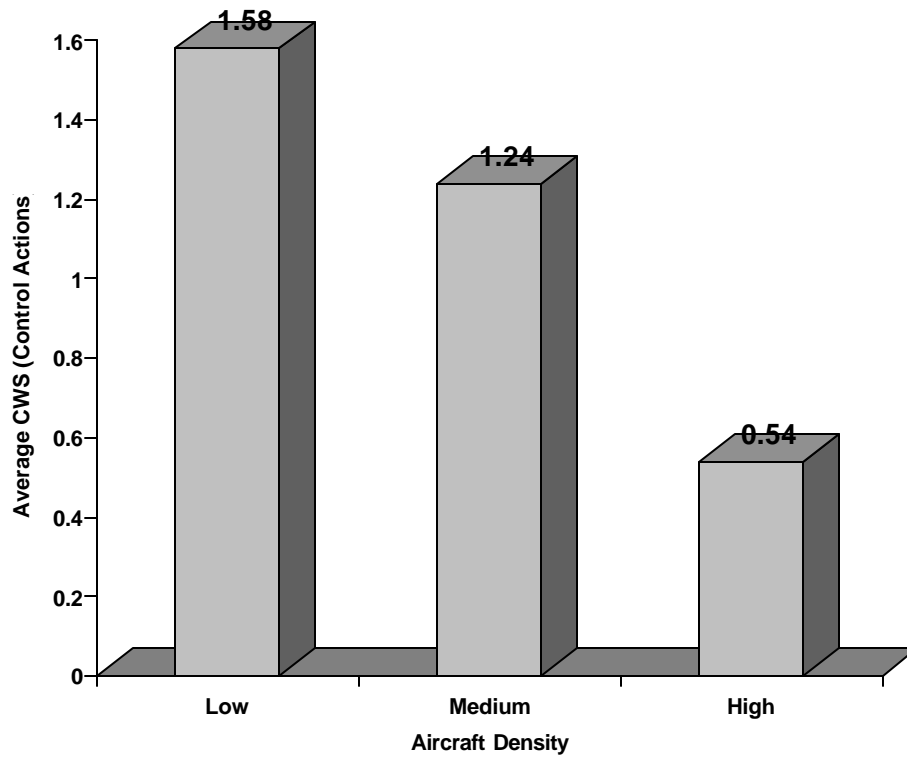


Figure 5

Team performance (Control Actions) decreased as scenario aircraft density increased.
