CWS APPLIED TO CONTROLLERS IN A HIGH FIDELTY SIMULATION OF ATC

Rickey P. Thomas, *Ben Willem, James Shanteau, John Raacke, & Brian Friel

Kansas State University, Manhattan, KS

*William J. Hughes Technical Center FAA, Atlantic City, NJ

ABSTRACT

The study applied the CWS index, a measure of expertise that integrates discrimination and consistency. Larger CWS scores are indicative of better evaluation, i.e., greater discrimination and consistency. CWS was used to assess controller performance operating in high-fidelity simulations of air traffic control (ATC). Large CWS scores were associated with superior performance, e.g., fewer separation errors. CWS indices were also sensitive to changes in task complexity and controller efficiency. The findings indicate that we can calculate valid CWS scores in high-fidelity simulations of ATC that capture the performance of expert controllers.

INTRODUCTION

Scientists have long been challenged to develop valid measures of expert performance. To that end, performance measures for controllers have been developed within the FAA for over thirty years (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 elusive, particularly for performance in dynamic environments.

Performance in ATC 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 challenges associated with measuring covert processes in dynamic environments such as ATC, understanding and measuring the relationship between a controller's mental processes and performance has been difficult.

According to Manning et al. (2000), difficulties arise in the development of measures of controller performance. First, the complex dynamic environment of ATC does not lend itself well to standard measurement techniques. Thus, performance in dynamic domains such as ATC makes the use of traditional judgment and rating tasks problematic. 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 appropriate for studying how performance changes in dynamic, autonomous environments (Manning et al., 2000). Second, because of difficulties in developing direct behavioral measures of operator performance, researchers rely instead on the observations of SMEs to evaluate performance. However, as 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 an identical problem. An index of performance should be free of subjectivity and amenable to quantitative comparisons.

To address these problems, a project was undertaken to adapt CWS, a behavioral-based index of expertise, to complex, cognitive tasks such as ATC. Following Gilbert (1978), an index for such a complex subject should be judged by its usefulness, its simplicity, and its coherence. CWS was first developed and successfully tested against existing data sets from expert judgments of static stimuli (Shanteau, Weiss, Thomas, & Pounds, in press). Thomas, Pounds, & Shanteau (in press) extended the CWS methodology for use in dynamic domains like ATC. The present study concerns the application of CWS to performance of expert controllers operating in high-fidelity simulations of ATC environments.

CWS INDEX

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 stimuli in a consistent way. Shanteau et al. (in press) propose that similar reasoning be applied to expert evaluation. That is, experts must discriminate consistently.

Using CWS, an expert's responses are analyzed to generate measures of discrimination and inconsistency (we typically compute variances). 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 (i.e., larger discrimination and smaller inconsistency) the greater the exhibited degree of expertise.

CWS has been successfully applied to several pre-existing datasets concerning expert evaluation, three of which are presented in Shanteau et al. (in press): auditing (Ettenson, 1984), personnel hiring (Nagy, 1981), and livestock judging (Phelps & Shanteau, 1978). 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 (swine vs. cattle) within the livestock domain for judgments of breeding quality of swine. Thomas, Pounds, & Shanteau (in press) successfully applied CWS to performance of naï ve operators and teams of naïve operators in a dynamic task similar to ATC. Friel, Thomas, Shanteau, and Raacke (2001) conducted a longitudinal study, which demonstrated CWS was sensitive to the development of competency in a low-fidelity simulation of ATC. The primary purpose of this experiment was to apply the CWS methodology to expert controllers working within a high-fidelity simulation of ATC.

METHODOLOGY

An archival data set was used to evaluate the CWS methodology. Researchers at the William J. Hughes Technical Center in Atlantic City, NJ collected the original data.[1]

Participants

The twelve participants in the experiment were active full-performance-level ATCS (Air Traffic Control Specialists) from Level 5 Terminal Radar Approach Control Facilities (TRACONS).

Stimulus scenarios

The design is a completely within-participants 2(aircraft density) x 2(conflict type) experiment. Crossing these factors produces four scenarios. In the low aircraft density scenarios the controllers were presented with an average of 7 aircraft per 15 minutes with 7 aircraft visible on the radar screen at any given time. In the high aircraft density scenarios the controllers were presented with an average of 14 aircraft per 15 minutes with 14 aircraft visible on the radar screen at any given time. There were two types of conflicts: overtaking and intersecting. The scenarios were designed so the built-in conflict (overtaking or intersecting) would occur 6 minutes into the simulation if the controller failed to intervene. Each scenario was replicated. The aircraft identification tags and beacon codes were different between the two replications; however, all other aspects of the scenarios were identical between replications.

Performance measures

The simulator creates high-fidelity ATC environments and provides "objective" measures of controller performance, controller efficiency, and task complexity. The number of separation errors and the duration of separation errors comprise the simulator performance measures. Separation errors of long duration are indicative of poor performance. The efficiency measures include the number of altitude, speed, and heading changes issued by the sector controller. Controllers who issue relatively few control actions are more efficient *ceterus paribus* than controllers who issue many control actions. The task complexity measures include the number of aircraft handled by the ATCS, the duration of time aircraft were under ATCS control, and a measure of system activity. Scenarios with high levels of system activity that require controllers to handle many aircraft for long durations are difficult (complex) scenarios.

SME ratings

The SMEs made 24 ratings of controller performance. The 24 ratings cluster into 6 underlying dimensions: performance, situational awareness, prioritization, efficiency, domain knowledge, and communication.

[1] We are grateful to Ben Willems (ACT-530) for providing the data set for the CWS analysis.

EVALUATION OF CWS

CWS analyses were conducted to examine individual controller performance. CWS was calculated using the distance each aircraft flew (measured in nautical miles) under the controller's command. This dependent measure was chosen because of its task validity and suitability. CWS analysis of distance flown captures important aspects of ATCS performance. Aircraft have different routes that require them to fly different distances within the controller's sector. The discrimination component of CWS measures the extent to which aircraft with different routes fly different distances through the sector. Inconsistency is demonstrated when the same aircraft is controlled to fly different distances through a sector across the two replications of the scenario.

The example provided in Table 1 will serve to illustrate exactly how CWS was calculated using distance flown.

Table 1. Example of CWS calculation for Distance Flown (nautical miles).

	ATCS 1		ATCS 2	
Replicate	A/C 1	A/C 2	A/C 1	A/C 2
1	192	132	156	246
2	192	126	210	162
CWS	3969/9 = 441		441/25 = 0.18	

As Table 1 indicates, Controller 1 consistently discriminated the two aircraft over the two replicates in terms of distance flown (nautical miles under ATCS control). Discrimination was computed by taking the mean squared deviation between aircraft (3969), where greater variation indicates better discrimination. Consistency was computed by taking the mean squared deviation between replicates (9), where lower variance indicates greater consistency (or less inconsistency). Dividing discrimination by inconsistency yields a CWS score of 441 for controller 1. Controller 2 exhibits far less discrimination and consistency, yielding a much lower CWS score of 0.18. Thus, we conclude on the basis of CWS that the performance of Controller 1 was better than that of Controller 2.

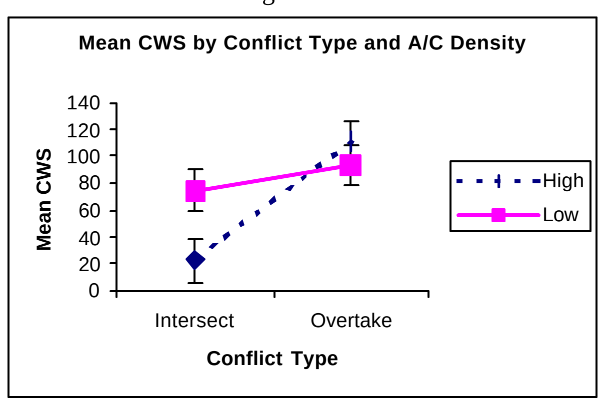
Hypotheses

As the scenarios increase in difficulty we expect increases in controller workload. Increases in controller workload should lead to decreased consistency in the implementation of tactics. Inconsistency in the application of tactics will manifest itself in the data as the same aircraft travel different distances through the sector across the two runs of the scenario-- increasing the inconsistency component of CWS. Also, we predict that increased workload will require aircraft with shorter routes to travel longer distances though the sector—decreasing the discrimination component of CWS. Thus, we hypothesize complex scenarios will result in lower consistency (i.e., higher inconsistency), discrimination, and CWS indices.

RESULTS

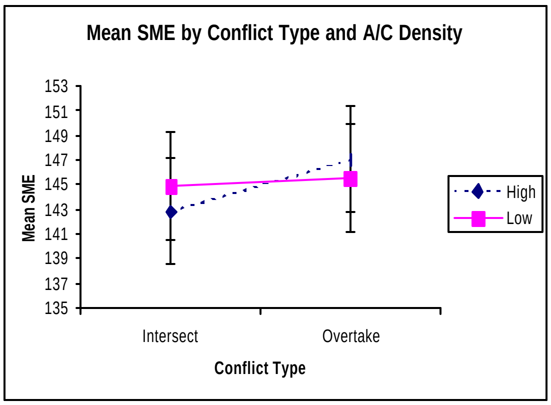
The CWS index is interpreted as the controller's ability to discriminate and perform consistently. To test this with the archival data, CWS was calculated for each controller (for each of the four scenarios) using distance flown as the dependent variable. There was not a significant main effect of aircraft density $F(1, 11) = 1.03$, $p > .05$, $Eta^2 = .10$, $power = .15$. There was a significant main effect of conflict type $F(1, 11) = 9.41$, $p \leq .01$, $Eta^2 = .46$, $power = .80$. However, it is not appropriate to interpret the main effects as there is a significant aircraft density by conflict type interaction, $F(1, 11) = 5.90$, $p \leq .05$, $Eta^2 = .35$, $power = .60$. As illustrated in Figure 1, the controllers had significantly lower CWS scores in the high aircraft density scenario where there was an intersecting conflict. The CWS scores were lowest in this scenario for every controller in the experiment. Thus, the resulting pattern of CWS scores suggests the index was sensitive to the behavioral changes of the controllers as influenced by the manipulations of aircraft density and conflict type.

Figure 1



SMEs rated the performance of the controllers as they were engaged in the tasks. We considered whether the ratings made by the SMEs captured the experimental manipulations. Although the composite SME ratings (Figure 2) show the same general pattern of results as CWS, there were no significant main effects or interactions: conflict type [$F(1,35) < 1.00$, $p \geq .05$, $Eta^2 = .01$], aircraft density [$F(1,35) < 1.00$, $p \geq .05$, $Eta^2 = .00$], conflict type by aircraft density interaction [$F(1,35) = 30.50$, $p \geq .05$, $Eta^2 = .01$]. Thus, the resulting pattern of findings indicates that the SME ratings were less sensitive than CWS to the experimental manipulations of aircraft density and conflict type.

Figure 2



The extent that discrimination, inconsistency and CWS were associated with the simulator measures of task complexity, controller efficiency, and controller performance was evaluated. Table 2 indicates that discrimination, consistency, and CWS are moderately correlated with each of the simulator's "objective" measures in the appropriate direction. As the performance of the controllers decreased (i.e., the number and duration of separation conflicts increased) CWS decreased. In addition, as the controllers became less efficient (i.e., they issued more heading, speed, and altitude changes) their CWS indices tended to decrease. As the complexity of the scenarios increased (i.e., more aircraft were controlled and handed-off) CWS indices decreased. Thus, CWS indices tended to decrease as the complexity of the task increased, the efficiency of the controllers decreased, and the performance of the controllers decreased.

Table 2: Correlations of CWS, Inconsistency (INCON), and Discrimination (DISC) by Simulator Measures of Performance, Efficiency, and (Task) Complexity.

Simulator Measures	CWS	INCON	DISC
Performance			
Separation errors	-.47*	.39*	-.41*
Duration of Separation errors	-.52*	.50*	-.22
Composite performance	-.53*	.51*	-.24*
Efficiency			
Altitude changes	-.35*	.29*	-.37*
Heading changes	-.35*	.23	-.55*
Speed changes	-.47*	-.01	-.56*
Composite efficiency	-.41*	.32*	-.52*
Complexity			
Aircraft handed-off	-.36*	.21	-.77*
Aircraft controlled	-.38*	.23	-.69*
System activity	-.34*	.19	-.78*
Composite complexity	-.39*	.23	-.69*

*Spearman's rho, *p < .05*

Table 3: Correlations of CWS and Composite SME ratings by Simulator Measures of Performance, Efficiency, and (Task) Complexity.

Simulator Measures	CWS	SME
Performance		
Separation errors	-.47*	-.17
Duration of Separation errors	-.52*	-.26
Composite performance	-.53*	-.25
Efficiency		
Altitude changes	-.35*	-.30*

Heading changes	-.35*	-.29*
Speed changes	-.47*	.18
Composite efficiency	-.41*	-.27*
Complexity		
Aircraft handed-off	-.36*	.08
Aircraft controlled	-.38*	-.14
System activity	-.34*	-.04
Composite complexity	-.39*	-.14
*Spearman's rho, *p < .05*		

Table 4: Correlations of SME Rating Dimensions with CWS and Simulator Measures of Performance (Perf.), Efficiency (Eff.), and Complexity (Comp.).

SME RATING DIMENSIONS	*CWS*	*Perf.*	*Eff.*	*Comp.*
Conflict Resolution	.11	-.09	-.10	-.03
Situational Awareness	-.08	-.18	-.13	.04
Prioritization	.11	-.21	-.26	-.14
Efficiency	.12	-.29*	-.23	-.24
Knowledge	-.10	-.15	-.19	-.10
Communication	-.08	-.09	-.11	-.02
*Spearman's rho, *p < .05*				

Note in Table 3 that the composite SME ratings did not correlate with CWS or the simulator measures of performance. We were somewhat baffled by this result. To further evaluate the finding we reduced the composite SME ratings into their six underlying dimensions. Table 4 shows that the SME rating dimensions were not related to CWS or the simulator's "objective" measures of task complexity, controller performance, or controller efficiency. The one exception was the instance in which the SME rating of efficiency was weakly related to the simulator's composite measure of controller performance.

To further investigate the results we looked at the relationship between the 24 SME ratings, CWS, and the simulator measures. There were only a few statistically significant relationships, however, they were weak and of no practical significance. Thus, it seems that the SME ratings did not capture the dimensions they were specifically asked to rate. For example, the SMEs were asked to rate the extent to which the controller maintained separation and resolved potential conflicts. The SME ratings on the separation dimension were not related to the number of separation errors committed by the controllers ($r = .14$, $p > .05$). However, CWS was associated with the number of separation errors ($r = -.47$, $p < .05$).

There are many potential explanations why SME judgments were not associated with CWS or the simulator measures. First, the SMEs may have based their judgments on more global aspects of expertise. Second, the SME ratings are subjective and therefore subject to individual differences. Finally, the SMEs are acknowledged expert controllers, however, there is no reason to believe they are expert at evaluating the performance of other controllers on 24 separate dimensions using an instrument with which they have limited experience. Thus, the SMEs may have engaged in a task (rating controller performance) that lies outside their specific and narrow realm of expertise (controlling traffic). In other words, not all good players make good coaches.

BRIEF DISCUSSION

In sum, the findings indicate that we can calculate valid CWS scores in high-fidelity simulations of ATC that capture the performance of expert controllers. CWS was sensitive to the experimental manipulations and moderately correlated with the simulator's measures of task complexity, controller efficiency, and controller performance. These findings extend those reported by Thomas et al. (in press) that demonstrated the ability of CWS to measure the performance of naïve operators in a dynamic task.

The identification of experts is difficult but vital to any study or application involving expertise. In the absence of any "gold standard" we often rely on acknowledged experts (SMEs) to identify who is (and who is not) performing expertly. However, as the findings of this study indicate there could be drawbacks relying solely on SME ratings. CWS provides an objective mechanism for making determinations concerning performance. Although discrimination and consistency are not sufficient to determine whether one is an expert--they are necessary. Thus, CWS captures important aspects of competent evaluation that all experts must exhibit.

We believe this is the reason why the measure has been successfully applied to several domains of expertise including wine judging, medicine, auditing, livestock judging, personnel selection, food tasting, and air traffic control.

ACKNOWLEDGMENTS

This research was supported, in part, by Grant 90-G-026 from the Federal Aviation Administration, Department of Transportation. Correspondence concerning this research can be addressed to Rick Thomas at Department of Psychology, 492 Bluemont Hall, 1100 Mid-Campus Drive, Kansas State University, Manhattan, KS 66506-5302.

REFERENCES

Cochran, W. G. (1943). The comparison of different scales of measurement for experimental results. *Annals of Mathematical Statistics*, *14*, 205-216.

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

Ettenson, R. (1984). A schematic approach to the examination of the search for and use of information in expert decision making. Unpublished doctoral dissertation, Kansas State University.

Friel, B. M., Thomas, R. P., Shanteau, J., & Raacke, J., (in preparation). CWS applied to an air traffic control simulation task (CTEAM).

Hammond, K. R. (1996). *Human judgment and social policy*. New York: Oxford University Press.

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

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

Nagy, R. H. (1977). How are personnel selection decisions made? An analysis of decision strategies in a simulated personnel selection. Unpublished doctoral dissertation, Kansas State University.

Phelps, R. H., & Shanteau, J. (1978). Livestock judges: How much information can an expert use? *Organizational Behavior and Human Performance*, *21*, 209-219.

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

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.

Thomas, R. P., Pounds, J., & Shanteau, J. (in press). Evaluation of a performance-based measure of expertise in a dynamic complex environment. *Federal Aviation Administration.*