

MEASURING THE PERFORMANCE OF EXPERTS: AN APPLICATION TO AIR TRAFFIC CONTROL

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 Cochran-Weiss-Shanteau (CWS) index, a behavioral-based 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 the performance of controllers operating in high-fidelity simulations of air traffic control (ATC). Large CWS scores were associated with superior performance, e.g., fewer separation errors. The CWS indices were also sensitive to changes in task complexity and controller efficiency, further validating the index. This research extended CWS to real controllers operating in high-fidelity simulations of ATC.

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.

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.

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., 2000). 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. (2000) 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. (2000): auditing (Ettenson, 1984), personnel hiring (Nagy, 1981), and livestock judging (Phelps & Shanteau, 1978). In each of these studies CWS successfully distinguished expert performance. For example, Ettenson (1984) asked two groups of auditors to evaluate a set of financial cases described by a common set of cues. One group of 15 "expert" auditors was recruited from Big Six accounting firms in Omaha, Nebraska. They included audit seniors and partners, with 4 to 25 years of audit experience. For comparison, 15 "novice" accounting students were obtained from two large Midwestern universities.

For each case, participants were asked to make a *Going Concern* assessment. Based on feedback from a senior auditor, the cues were classified as either relevant (e.g., *Net Income*), partially relevant (e.g., *Aging of Receivables*), or irrelevant (e.g., *Prior Audit Results*). Discrimination was estimated from the mean square values for each cue. Consistency was estimated from the average of within-cell variances – low

variance implies high consistency. The ratio of discrimination variance divided by consistency variance was computed to form separate CWS values for relevant, partially relevant, and irrelevant cues.

The results in Table 1 show that CWS values for the expert group drop systematically as the relevance of the cues declines. For the novice group, there is a similar but less pronounced decline. More important, there is a sizable difference between experts and novices for relevant cues. The size of this difference is less for partially relevant cues, and nonexistent for irrelevant cues.

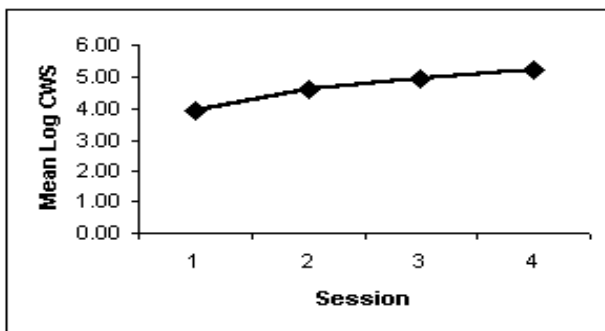
For relevant cues, CWS clearly distinguishes between experts and novices. However, the size of the difference between the groups declines for less relevant cues. These results show that CWS can distinguish between levels of expertise for these two groups.

Table 1: Average CWS Values for Two Groups of Auditors for Three Categories of Cues

	Relevant	Partially Relevant	Irrelevant
Experts	13.10	6.42	3.32
Novices	8.08	5.13	3.03

In the auditor study, however, the stimuli were static. Thomas and Pounds (2001) successfully applied CWS to performance of naïve operators and teams of naïve operators in a dynamic task similar to ATC. Friel, Thomas, Shanteau, and Raacke (2001) conducted a longitudinal study to see whether CWS could track the acquisition of competency as operators gained experience on a low-fidelity simulation of ATC. In the longitudinal study, CWS scores were compared over repeated sessions to determine whether these scores would reflect performance improvements with practice. One such result is presented in Figure 1. Linear trend analyses revealed significant increases in CWS as the operators gained experience on the task $F(1, 10) = 7.37, p < .05$. The results demonstrate that CWS is sensitive to the development of competency in a low-fidelity simulation of ATC.

Figure 1. Mean log CWS scores as a function of repeated Low Aircraft Density sessions.



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 of the experiment is a 2(aircraft density) x 2(conflict type) completely within design. Crossing these factors produces four scenarios. In the low aircraft density scenarios the controllers were presented with an average of seven aircraft per 15 minutes with seven 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 of the scenarios 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 three types of measures: controller performance, controller efficiency, and task complexity. The number of separation errors and the duration of separation errors comprise the simulator performance measures. The efficiency measures include the number of altitude, speed, and heading changes issued by the sector controller. Efficient controllers issue fewer 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.

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

¹ Ben Willem (ACT-530) provided the data set for the CWS analysis.

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 2 will serve to illustrate exactly how CWS was calculated using distance flown.

Table 2. Example of CWS calculation for Distance Flown (nautical miles).

Replicate	ATCS 1		ATCS 2	
	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 2 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 Controller 1's performance was better than that of Controller 2.

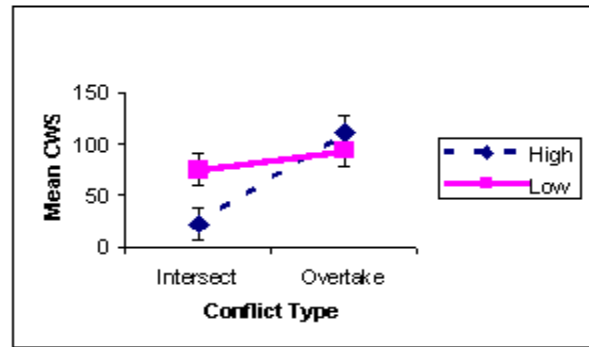
As the scenarios become more difficult we expect the increase in controller workload to increase the inconsistency with which tactics are implemented. Inconsistency in the application of tactics will result in the same aircraft traveling different distances through the sector across the two runs of the scenario. Also, we predict that increased workload will require aircraft with shorter routes to travel longer distances though the sector decreasing discrimination. Thus, more complex scenarios should result in lower consistency, 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 under control as the dependent variable. There was no main effect of aircraft

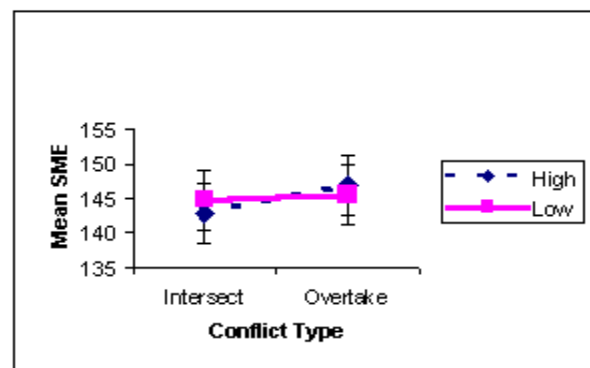
density $F(1, 11) = 1.03, p > .05, \eta^2 = .10, power = .15$. There was a significant main effect of conflict type $F(1, 11) = 9.41, p \leq .01, \eta^2 = .46, power = .80$. However, it is not appropriate to interpret the main effects as there is a significant aircraft density by conflict type interaction, $F(1, 11) = 5.90, p \leq .05, \eta^2 = .35, power = .60$. As illustrated in Figure 2, 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 2. Mean CWS by Conflict Type and A/C Density.



SMEs rated the performance of the controllers as they were performing the tasks. We considered whether the ratings made by the SMEs captured the experimental manipulations. Although the SME ratings (Figure 3) 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 3. Mean SME by Conflict Type and A/C Density



The extent to which discrimination, inconsistency and CWS captured the simulator measures of task complexity, controller efficiency, and controller performance was

evaluated. Table 3 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 controller deteriorated (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 3. 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
<i>Composite performance</i>	-.53*	.51*	-.24*
Efficiency			
Altitude changes	-.35*	.29*	-.37*
Heading changes	-.35*	.23	-.55*
Speed changes	-.47*	-.01	-.56*
<i>Composite efficiency</i>	-.41*	.32*	-.52*
Complexity			
Aircraft handed-off	-.36*	.21	-.77*
Aircraft controlled	-.38*	.23	-.69*
System activity	-.34*	.19	-.78*
<i>Composite complexity</i>	-.39*	.23	-.69*
<i>Spearman's rho, *p < .05</i>			

Note in Table 4 and Table 5 the SME ratings were not related to the simulator's "objective" measures of task complexity, controller performance or controller efficiency. There are many potential explanations why SME judgments were not associated with these measures, e.g., the SMEs may have based their judgments on more global aspects of expertise. Also, the SME ratings are subject to individual differences. The SMEs are acknowledged expert controllers, however, there is no reason to believe they are expert at rating the performance of other controllers. In other words, not all players make good coaches. In one striking example, the SMEs were asked to make a rating concerning 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 operational deviations ($r = .14, p > .05$) committed by the controllers. However, CWS was associated with the number of separation errors ($r = -.47, p < .05$).

Table 4. 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
<i>Composite performance</i>	-.53*	-.25
Efficiency		
Altitude changes	-.35*	-.30*
Heading changes	-.35*	-.29*
Speed changes	-.47*	.18
<i>Composite efficiency</i>	-.41*	-.27*
Complexity		
Aircraft handed-off	-.36*	.08
Aircraft controlled	-.38*	-.14
System activity	-.34*	-.04
<i>Composite complexity</i>	-.39*	-.14
<i>Spearman's rho, *p < .05</i>		

Table 5. 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
<i>Spearman's rho, *p < .05</i>				

BRIEF DISCUSSION

In sum, the results indicate that meaningful and valid CWS indices can be computed from data collected from expert controllers operating in high-fidelity simulations of ATC environments. 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 are drawbacks to relying solely on SME ratings. CWS provides a more 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 including wine judging, medicine, auditing, livestock judging, personnel selection, food tasting, and finally 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.), *Expertise and technology* (pp. 1-15). Hillsdale, NJ: Lawrence Erlbaum.
- Manning, C. A., Mills, S., Mogilka, H., Hedge, J., Bruskiwicz, Pfliegerer, E., (in preparation) Prediction 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*.