

CWS APPLIED TO AN AIR TRAFFIC CONTROL SIMULATION TASK (CTEAM)

Brian M. Friel, Rickey P. Thomas, James Shanteau, & John Raacke
Kansas State University
Manhattan, KS

ABSTRACT

Twelve undergraduates participated in an eight-week longitudinal study involving the development of skill in an air traffic control microworld. The Cochran-Weiss-Shanteau (CWS) index of expert performance, which integrates discrimination and consistency, was applied to the data. Results indicated that CWS was sensitive to variations in performance due to changes in scenario complexity, to increased practice, and to individual differences in competence in the task. CWS was also moderately correlated with objective measures of performance, i.e., controlled flights into obstacles and separation errors. This study demonstrates that CWS can be successfully applied to dynamic stimulus environments.

BACKGROUND

The Cochran-Weiss-Shanteau (CWS) index of expert performance was tested to determine whether it could be used for dynamic stimulus environments (Shanteau, Weiss, Thomas, & Pounds, in press). CWS integrates two necessary conditions for expert skill. The first is consistency, as argued by Einhorn (1972, 1974). Experts must make reliable judgments of identical stimuli; unreliable judgments serve as evidence against expertise. The second necessary condition is discrimination ability (Hammond, 1996). That is, experts should be able to differentiate stimuli on the basis of subtle differences that non-experts are typically insensitive to. CWS integrates these two conditions by taking the ratio of discrimination to inconsistency, such that higher CWS scores are more indicative of expert performance. That is, experts should be consistent in their discriminations of stimuli in their domain. This is in line with the suggestion made by Cochran (1943) to use the ratio of between stimulus variance to within stimulus variance for assessing response instrument quality, hence his inclusion in the acronym CWS.

CWS has been successfully applied to several pre-existing datasets, three of which are presented in Shanteau et al. (in press): auditing (Ettenson, 1984), personnel hiring (Nagy, 1981), and livestock judging (Phelps & Shanteau, 1978). However, the stimuli in these studies were unchanging, in the sense that participants' behaviors did not influence what was

presented to them. In contrast expert air traffic controllers deal with dynamic stimulus environments, i.e., those that change when the expert acts on the stimuli. To this point, it is unclear whether CWS could be successfully applied to such environments. Thus, the purpose of this study was to determine whether CWS could be used as a performance index in a simulated air traffic control (ATC) environment, the Controller Teamwork Evaluation and Assessment Methodology (CTEAM) microworld (Bailey, Broach, Thompson, & Enos, 1999).

An eight-week longitudinal study involving naïve participants was conducted to determine whether CWS could serve as an index for performance improvements over time. The study also included two manipulations in scenario complexity. The first involved three levels of Aircraft Density (Low, Medium, and High), whereas the second involved the presence or absence of Restricted Airspace.

For the present study, three dependent measures were collected. These were the number of Separation Errors (i.e., aircraft within five scale miles of another aircraft at the same altitude or within five scale miles of the sector boundary) made by participants, the number of Controlled Flights Into Obstacles (CFIOs; i.e., aircraft collisions with other aircraft, sector boundaries, or restricted airspace), and Time Through Sector (i.e., the amount of time from when the aircraft first appeared on the screen to when it reached its destination). The latter measure was converted into CWS scores. The example in Table 1 will serve to illustrate how this was done.

Table 1. Example of CWS calculation for Time Through Sector.

	Controller 1		Controller 2	
<u>Replicate</u>	<u>Aircraft 1</u>	<u>Aircraft 2</u>	<u>Aircraft 1</u>	<u>Aircraft 2</u>
1	192	132	156	246
2	192	126	210	162
3	198	126	204	156
CWS	6534/12 = 544.500		6/1704 = 0.004	

As Table 1 indicates, Controller 1 consistently discriminated the two aircraft over the three replicates in terms of Time Through Sector. Discrimination was

computed by taking the mean squared deviation between aircraft (6534), where higher figures indicate greater discrimination. Consistency was computed by taking the mean squared deviation between replicates (12), where lower figures indicate greater consistency. Dividing discrimination by consistency yields a CWS score of 544.500. Controller 2 exhibits far less discrimination and consistency, yielding a much lower CWS score of 0.004. Thus, we would say that Controller 1's performance was better than Controller 2's performance for these aircraft over the three replicates.

We predicted that CWS scores would decrease as a function of scenario complexity. That is, lower CWS scores should be found for the High Aircraft Density scenarios than for Low and Medium Aircraft Density scenarios. CWS scores should also be higher for scenarios without Restricted Airspace, as participants were expected to have a more difficult time routing aircraft to their destinations with the presence of this obstacle. We also predicted that CWS scores should increase with repeated sessions of the same scenario (practice). Finally, if CWS scores reflect performance in CTEAM, they should be negatively correlated with errors.

METHOD

Participants

Twelve Kansas State University undergraduates participated in the study in exchange for \$12 per two-hour session.

Apparatus, Design, and Procedure

Participants were presented with six scenario types in a single-sector version of CTEAM. These six scenarios were created by crossing three levels of Aircraft Density (Low = 12, Medium = 24, and High = 36 aircraft) with two levels of Restricted Airspace (Present or Absent). In order to compare CWS scores across Aircraft Density, the 12 aircraft from the Low scenario were embedded in the Medium and High scenarios. Scenarios were presented on 17-inch Sceptre monitor connected to an NCR-3230 486 computer, and the participants' task was to route aircraft from their origins to their ultimate destination points (either an exit gate or an airport) using a Kensington Expert Mouse track ball.

The study lasted 8 weeks, with participants completing three repetitions (herein called replicates) of the same scenario in a given session. Each participant completed three sessions per week, yielding

a total of 24 sessions per participant. The order of scenario presentation was the same for each participant. Two 2-week blocks of scenarios were created, each presented twice, as indicated in Table 2.

Table 2. Order of scenarios presented to participants.

	Day in Week		
	1	2	3
Week 1	Low	Medium	Low-RA
Week 2	Medium-RA	Low	Medium
Week 3	Low	Medium	Low-RA
Week 4	Medium-RA	Low	Medium
Week 5	Medium	High	Medium-RA
Week 6	High-RA	Medium	High
Week 7	Medium	High	Medium-RA
Week 8	High-RA	Medium	High

Note. Low = 12 aircraft, Medium = 24 aircraft, High = 36 aircraft, RA = restricted airspace present.

RESULTS AND DISCUSSION

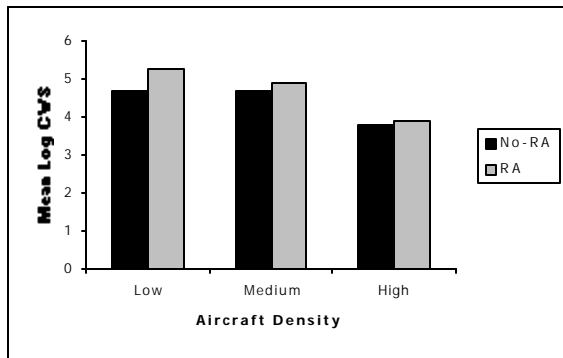
Two participants dropped out of the study, one after 13 sessions, the other after 23 sessions. The data that they provided were included in all subsequent analyses. CWS scores were calculated using the time each aircraft took to reach its destination starting from the time that it first appeared on the screen (Time Through Sector). For each participant and each session, Discrimination was calculated using the mean squared deviation between different aircraft, whereas Consistency was assessed using the mean squared deviation within each aircraft over the three replicates. A CWS score for each participant for each session was then calculated. Only the 12 aircraft that were common to all three aircraft densities were included in the CWS calculations. Because there was a strong positive correlation ($r = +.84$) between the means and variances of CWS scores, the CWS scores were log transformed so that they no longer violated the homoscedasticity assumption of Analysis of Variance (ANOVA). Objective measures of performance, i.e., the number of Controlled Flights Into Obstacles (CFIOs) and Separation Errors, were also collected, so that CWS scores could be validated.

Scenario Complexity Manipulations

Mean log CWS scores for each scenario condition are presented in Figure 1. These data were submitted to a 3 (Aircraft Density) \times 2 (Restricted Airspace) repeated measures ANOVA. The main effects of Aircraft Density, $F(2, 20) = 19.35$, $p < .05$, $\eta^2 = .66$, Observed Power $> .999$, and Restricted Airspace $F(1, 10) = 5.08$, $p < .05$, $\eta^2 = .34$, Observed Power = .53,

were both statistically significant. As Aircraft Density increased, log CWS scores decreased, consistent with the idea that more complex scenarios are more difficult for participants. Oddly, scenarios where restricted airspace was present yielded higher CWS scores. It was predicted that such scenarios would lead to lower CWS scores, because participants had an added obstacle to route aircraft around. However, there are two possible explanations for this result. One is that scenarios with restricted airspace were always presented after corresponding scenarios without restricted airspace. Thus, experience with a particular level of aircraft density may be responsible for the higher CWS scores in the Restricted Airspace Present conditions. As shown below, this is consistent with the improvements shown by participants with increases in experience. Another possible explanation is in regard to the nature of the CWS score, namely the calculation of Consistency. The presence of restricted airspace leaves participants with fewer response options to direct aircraft. Thus, consistency scores may have been lower because aircraft over replicates are less free to vary in terms of Time Through Sector, leading to increased CWS scores. This is, in fact, what we found, as scenarios involving Restricted Airspace (RA) yielded lower values of Inconsistency (i.e., better Consistency) over each level of Aircraft Density (Low-RA: 27.95, Low: 70.75; Medium-RA: 50.03, Medium: 187.85; High-RA: 238.31, High: 260.44). The interaction between Aircraft Density and Restricted Airspace was not significant, $F(2, 20) < 1$, $\eta^2 = .08$, Observed Power = .17.

Figure 1. Mean log CWS scores a functions of Aircraft Density and Restricted Airspace.



Performance Improvements Over Time

Within each level of Aircraft Density, CWS scores were compared over repeated sessions to determine whether these scores would reflect performance improvements with practice. These results are presented in Figures 2-7. Linear trend analyses revealed significant increases for Low, $F(1, 10) = 7.37$,

$p < .05$, Medium, $F(1, 9) = 5.99$, $p < .05$, and Medium with Restricted Airspace scenarios, $F(1, 10) = 5.17$, $p < .05$, but not for Low with Restricted Airspace, $F(1, 11) < 1$, High, $F(1, 8) = 1.28$, $p > .20$, or High with Restricted Airspace scenarios, $F(1, 9) < 1$.

Figure 2. Mean log CWS scores as a function of repeated Low Aircraft Density sessions.

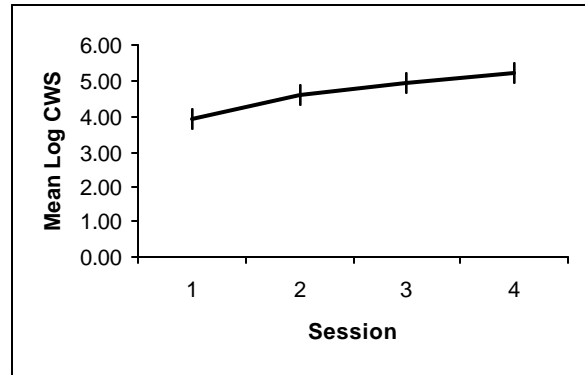


Figure 3. Mean log CWS scores as a function of repeated Low Aircraft Density sessions with Restricted Airspace.

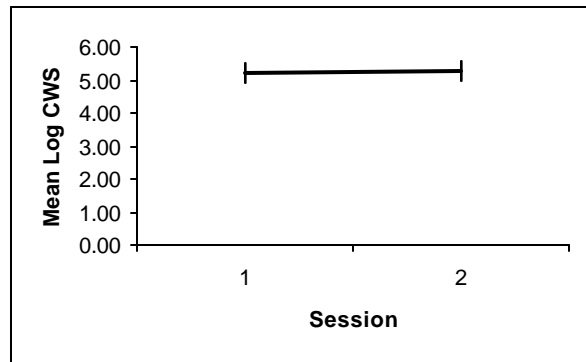


Figure 4. Mean log CWS scores as a function of repeated Medium Aircraft Density sessions.

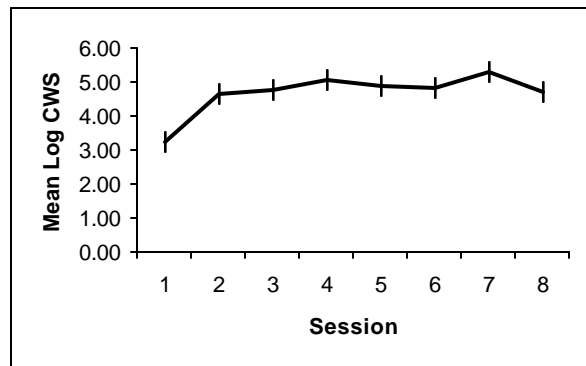


Figure 5. Mean log CWS scores as a function of repeated Medium Aircraft Density sessions with Restricted Airspace.

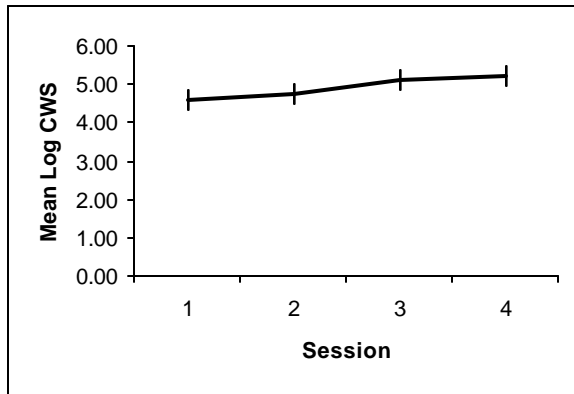


Figure 6. Mean log CWS scores as a function of repeated High Aircraft Density sessions.

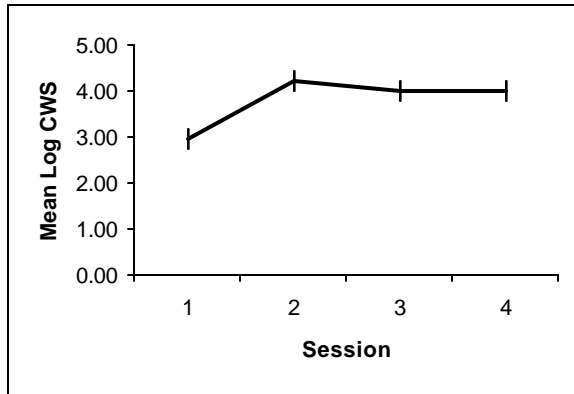
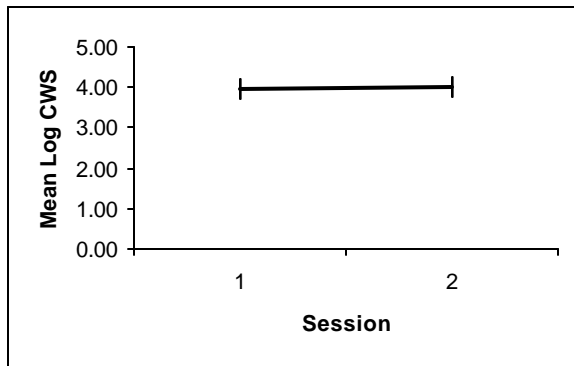


Figure 7. Mean log CWS scores as a function of repeated High Aircraft Density sessions with Restricted Airspace.



As predicted, CWS scores reflected performance improvements for three of the scenarios. However, the other three did not yield significant increases in CWS scores. One possible reason is that only two sessions of the Low Aircraft Density-Restricted Airspace and High Aircraft Density-Restricted Airspace scenarios were

presented to participants, not allowing enough time for participants to show improvements in performance. Given the relative difficulty of the High Aircraft Density session (participants made an average of 4.10 errors, i.e., CFIOs and separation errors, per session, second most to High Aircraft Density with Restricted Airspace), it is plausible that four sessions were not sufficient to allow for performance to improve in this scenario.

Validation of CWS

In order to determine whether CWS was truly capturing performance in CTEAM, scores were correlated with the number of CFIOs, separation errors, and total errors (CFIOs plus separation errors). These results are presented in Table 3.

Table 3. Spearman rank-order correlations between CWS and objective measures of performance.

Error	CWS	CWS _{With CFIOs}
CFIOs	-.24	-.52
Separation Errors	-.34	-.36
Total Errors	-.35	-.49

Note. Controlled Flights into Obstacles (CFIO) indicate that the participant directed an aircraft into a collision with another aircraft, restricted airspace, or the sector boundary.

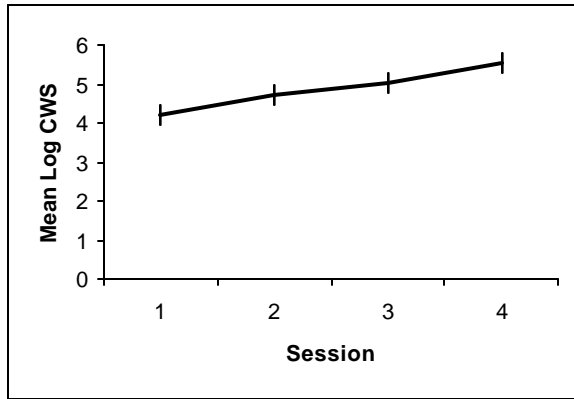
In the preceding analyses, CWS scores were computed such that aircraft involved in CFIOs were omitted from the analyses. These CWS scores showed moderate negative correlations with errors, indicating that as participants made fewer errors, CWS scores increased. When aircraft involving CFIOs were included (by setting the Time Through Sector to the total scenario time—1680 seconds), the correlations increased in magnitude. Thus, CWS appears to capture performance.

CWS vs. Other Measures

Although CWS scores seem to reflect relative performance, the question still remains as to whether they are necessary given that objective measures such as errors and “raw” Time Through Sector are available. We argue that CWS can be used to supplement these measures, providing information that the objective measures do not. For instance, CWS scores are superior to error measurements in situations where few errors are made. With professional air traffic controllers, errors are rarely made, limiting the use of error measures for assessing performance. Secondly, in the present study, CWS scores increased in scenarios where no errors were made. An example is presented for Low Aircraft Density Scenarios below in Figure 8.

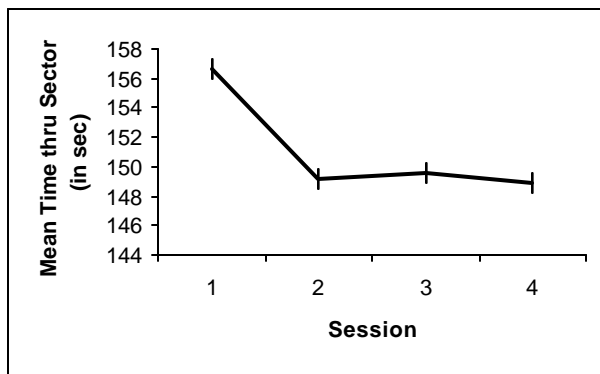
This scenario condition was chosen due to the relatively low number of errors made (0.71 errors per session per participant) and to the ease of the scenario allowing for improvement within a relatively small number of repeated sessions. The increasing scores suggest that CWS captures aspects of performance that error data are not sensitive to.

Figure 8. Mean log CWS scores in Low Aircraft Density sessions in which no errors were made.



CWS scores also provide more information about performance than does Time Through Sector in its raw form (i.e., not submitted to CWS calculations). In assessing performance improvements with practice, Time Through Sector measurements tended to asymptote before CWS scores did, suggesting that raw measures of time are not sufficient for assessing later performance improvements. An example of raw Time Through Sector measures is presented in Figure 9 below. As compared with Figure 2, Time Through Sector asymptotes well before log CWS reaches a maximum.

Figure 9. Raw Time Through Sector as measures of performance improvements.



Furthermore, CWS scores based on Time Through Sector were more strongly correlated with errors than

were raw Time Through Sector Measures, as can be seen in Table 4.

Table 4. Spearman rank-order correlations among log CWS, raw Time Through Sector, and errors.

Error	CWS	CWS _{With CFIOs}	Time-Sector
CFIOs	-.24	-.52	+.21
Separation Errors	-.34	-.36	+.15
Total Errors	-.35	-.49	+.19

CWS and Individual Differences in Competence

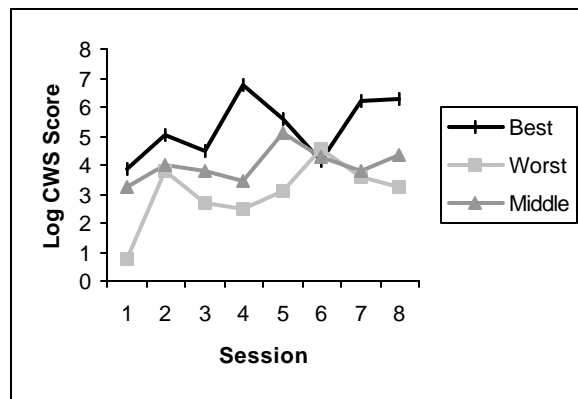
In a preliminary effort to determine whether CWS could be used to differentiate strong and weak performers, three representative participants were selected for a comparison between objective performance measures and CWS scores: our best participant, our worst participant, and one in between in terms of the number of errors made. Their performance is summarized in Table 5.

Table 5. Performance summaries for three example participants.

Participant	CFIOs	Sep. Errors	Total Errors	Mean Log CWS
Best	3	6	9	5.27
Worst	47	88	135	2.96
Middle	6	32	38	4.25

The Mean log CWS scores maintain the same rankings as those defined by errors and maintain the relative positions over nearly all sessions. Those for Medium Aircraft Density scenarios are presented in Figure 10.

Figure 10. Stability of CWS rankings across Medium Aircraft Density scenarios for three example participants.



Thus, at least for these participants, CWS can be used to differentiate more competent performers from less competent performers.

CONCLUSIONS

The above results suggest that the CWS index can be used to assess performance in a dynamic stimulus environment. CWS can also be used to assess the development of skill over time. The CWS scores increased as participants gained experience with the CTEAM microworld. CWS scores showed moderate negative correlations with errors and revealed aspects of development that error data and raw Time Through Sector were unable to capture. Finally, CWS scores were consistent with individual differences in task competence. These results suggest that CWS is an effective measure for assessing performance in dynamic environments.

ACKNOWLEDGEMENTS

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 Brian Friel at Kansas State University, Department of Psychology, 492 Bluemont Hall, 1100 Mid-Campus Drive, Manhattan, KS 66506-5302.

REFERENCES

- 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.
- Cochran, W. G. (1943). The comparison of different scales of measurement for experimental results. *Annals of Mathematical Statistics*, 14, 205-216.
- Einhorn, H. J. (1972). Expert measurement and mechanical combination. *Organizational Behavior and Human Performance*, 7, 86-106.
- Einhorn, H. J. (1974). Expert judgment: Some necessary conditions and an example. *Journal of Applied Psychology*, 59, 562-571.
- 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.

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

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., Weiss, D. J., Thomas, R. P., & Pounds, J. C. (in press). Performance-based assessment of expertise: How to decide if someone is an expert or not. *European Journal of Operations Research*.