

UTILIZING CWS TO TRACK THE LONGITUDINAL DEVELOPMENT OF EXPERTISE

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

The CWS expert performance index was applied to participants' skill development in CTEAM (Controller Teamwork Evaluation and Assessment Methodology), an air traffic control microworld environment. CWS integrates discrimination and consistency such that larger CWS scores indicate better performance. Over eight weeks, participants gained mastery over the task, which involved routing aircraft to assigned destinations. Aircraft Density and Restricted Airspace affected performance such that more complex scenarios led to lower CWS scores. Furthermore, CWS scores increased with practice. Moderate correlations between CWS scores and objective measures of performance were obtained, validating the index. Potential applications to training are discussed.

The Cochran-Weiss-Shanteau (CWS) index of expert performance (Shanteau, Weiss, Thomas, & Pounds, in press) was tested to determine whether it could be used for assessing skill development in dynamic stimulus environments. 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), i.e., 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 consistent 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 CWS acronym.

CWS has been successfully applied as a performance measure 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). The stimuli in these studies were unchanging, in that participants' behaviors did not influence what was presented to them. In contrast, many stimulus environments that experts

work in are dynamic, in that the stimuli change in response to the expert's actions. One such domain is air traffic control (ATC). To this point, it is unclear whether CWS could be successfully applied to dynamic stimulus environments. Thus, the purpose of this study was to determine whether CWS could index performance and skill development in a simulated 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. The study also included two manipulations in scenario complexity. The first involved three levels of Aircraft Density (Low, Medium, High), whereas the second involved the presence or absence of Restricted Airspace (RA). Three dependent measures were collected: (1) the number of Separation Errors (aircraft within five scale miles of the sector boundary or another aircraft at the same altitude) made by participants, (2) the number of Crashes, and (3) Time Through Sector (the amount of time between the aircraft's first appearance on the screen to when it reached its destination). The last measure was converted into CWS scores.

We predicted that CWS scores would vary as a function of scenario complexity. That is, lower CWS scores should be obtained in the High Aircraft Density scenarios than in Low and Medium Aircraft Density scenarios. CWS scores should also be lower in RA scenarios, as participants should have a more

difficult time routing aircraft to their destinations with this obstacle present. We also predicted that CWS scores should increase with practice. Finally, if CWS scores reflect CTEAM performance, they should be negatively correlated with errors.

METHOD

Participants

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

Apparatus, Design, and Procedure

In a single-sector version of CTEAM, participants were presented with six scenario types created by crossing three levels of Aircraft Density (Low = 12, Medium = 24, High = 36 aircraft) with two levels of RA (Present or Absent). 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 monitors connected to NCR-3230 486 computers. The participants' task was to direct aircraft to exit gates or airports using Kensington Expert Mouse trackballs.

The study lasted eight weeks. Participants completed three repetitions (hereafter called replicates) of the same scenario in each session. Participant completed three sessions per week, yielding 24 sessions per participant. Two 2-week blocks of scenarios were created, each presented twice, as indicated in Table 1. For all participants, the order of scenario presentation was the same.

Table 1. Order of scenario presentation.

| Block | Week | Day 1 | Day 2 | Day 3 |
|-------|------|---------|-------|---------|
| 1 | 1 | Low | Med. | Low-RA |
| | 2 | Med.-RA | Low | Med. |
| 1 | 3 | Low | Med. | Low-RA |
| | 4 | Med.-RA | Low | Med. |
| 2 | 5 | Med. | High | Med.-RA |
| | 6 | High-RA | Med. | High |
| 2 | 7 | Med. | High | Med.-RA |
| | 8 | High-RA | Med. | High |

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

RESULTS AND DISCUSSION

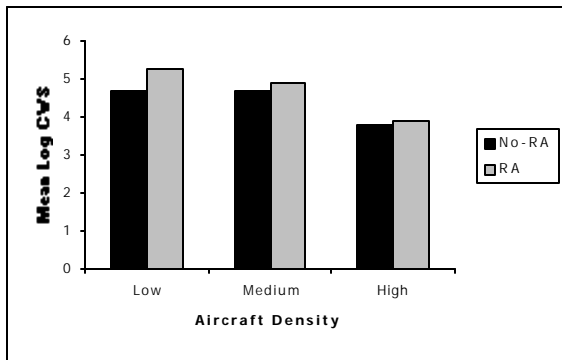
Two participants dropped out of the study, one after 13 sessions, the other after 23 sessions. The data that they provided were included in all subsequent analyses. CWS scores were calculated using Time Through Sector. For each participant and each session, Discrimination was calculated using the variance between different aircraft, whereas Inconsistency was assessed using the variance within each aircraft over the three replicates. The Discrimination-to-Inconsistency ratio formed the CWS score. Only those 12 aircraft 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 they no longer violated the homoscedasticity assumption of Analysis of Variance (ANOVA). Objective measures of performance (i.e., Crashes and Separation Errors) were also collected, so 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 (RA) repeated measures ANOVA. The main effects of Aircraft Density, $F(2, 20) = 19.35, p < .05, \eta^2 = .66$, and RA, $F(1, 10) = 5.08, p < .05, \eta^2 = .34$, 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 with RA yielded higher CWS scores. It was predicted that such scenarios would lead to lower CWS scores, because participants had an obstacle to route aircraft around. However, there are two possible explanations for this result. One is that RA scenarios were always presented after corresponding scenarios without RA. Thus, experience with a particular level of aircraft density may account for the higher CWS scores in the RA-Present conditions. As shown below, this is consistent with the improvements shown by participants with increased practice. Another possible explanation regards the nature of the CWS

score, namely the calculation of Inconsistency. The presence of RA leaves participants with fewer response options to direct aircraft. Thus, Inconsistency 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 with 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 RA was not significant, $F(2, 20) < 1$, $H^2 = .08$.

Figure 1. Mean log CWS scores as 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. Due to space limitations, only three of these results are presented (Figures 2-4). 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-RA scenarios, $F(1, 10) = 5.17, p < .05$, but not for Low-RA, $F(1, 11) < 1$, High, $F(1, 8) = 1.28, p > .20$, or High-RA scenarios, $F(1, 9) < 1$.

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

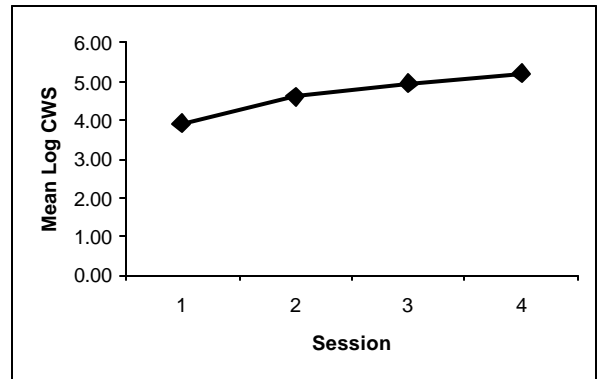


Figure 3. Mean log CWS scores as a function of repeated Medium Aircraft Density sessions.

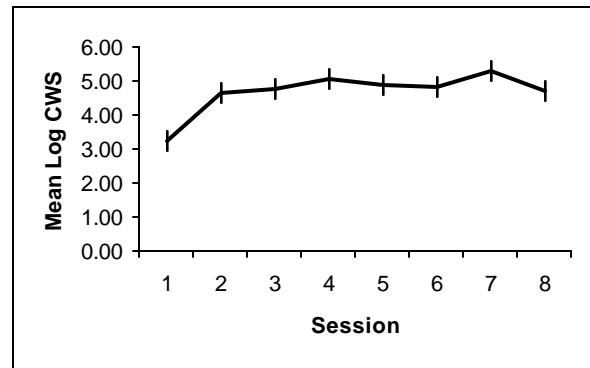
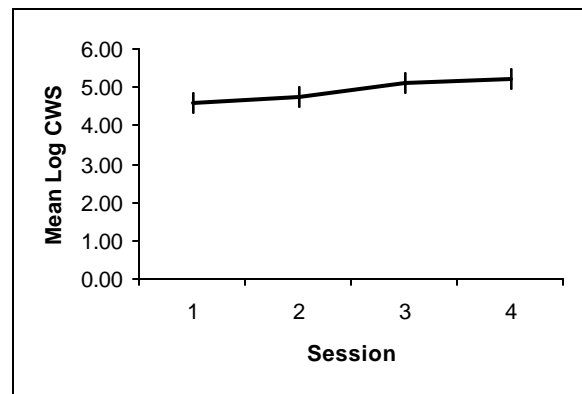


Figure 4. Mean log CWS scores as a function of repeated Medium Aircraft Density sessions with Restricted Airspace.



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-RA and High Aircraft Density-RA

scenarios were presented to participants, not allowing enough time for participants to show performance improvements. Given the relative difficulty of the High Aircraft Density session (participants made an average of 4.10 errors per session, second most to High Aircraft Density-RA), it is plausible that four sessions were insufficient to allow for performance improvements in this scenario.

Validation of CWS

To determine whether CWS truly captured performance in CTEAM, scores were correlated with the number of Crashes, Separation Errors, and Total Errors (Crashes plus Separation Errors), as presented in Table 2.

Table 2. Spearman rank-order correlations between CWS and objective measures of performance.

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

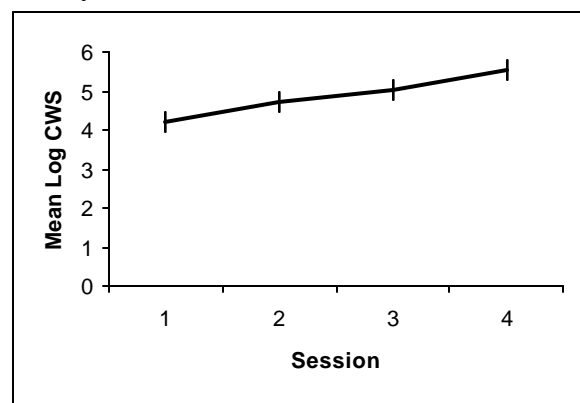
In the preceding analyses, CWS scores were computed such that aircraft involved in crashes were omitted from the analyses. These CWS scores showed moderate negative correlations with errors, indicating that as participants made fewer errors, CWS scores increased. When aircraft involved in crashes were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

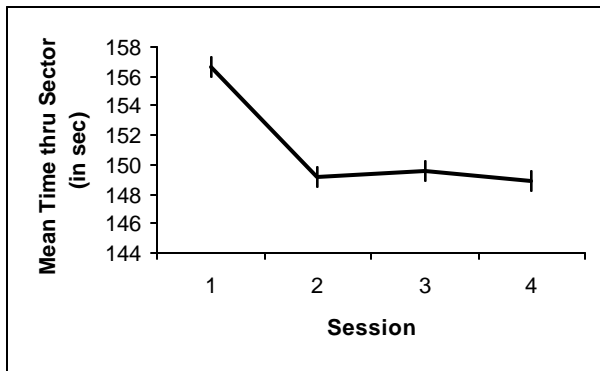
the present study, CWS scores increased in scenarios where no errors were made. An example is presented in Figure 5 for Low Aircraft Density Scenarios. This scenario condition was chosen due to the relatively low number of errors made (0.71 errors per session per participant) and to the ease of the scenario allowing for improvement within a relatively small number of sessions. The increasing scores suggest that CWS captures aspects of performance that error data are insensitive to.

Figure 5. 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 insufficient for assessing later performance improvements. An example of raw Time Through Sector measures is presented in Figure 6. As compared with Figure 2, Time Through Sector asymptotes well before log CWS reaches a maximum.

Figure 6. 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 3.

Table 3. Spearman rank-order correlations among log CWS, raw Time Through Sector, and errors.

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

The results presented above demonstrate that the CWS index of expert performance can be applied to assessing skill development in dynamic environments. Participants' CWS scores increased with practice, suggesting that the index is sensitive to performance improvements. CWS scores were also sensitive to the scenario manipulations in CTEAM, revealing that participants had more difficulty with more complex scenarios. Finally, the finding that CWS scores were negatively correlated with objective measures (i.e., crashes and separation errors) demonstrates that CWS does capture elements of performance.

The above results also suggest that CWS may be useful for training situations. The finding that CWS scores were sensitive to performance improvements in the present study suggests that the index could be used to assess skill development in trainees. In the present study, CWS scores also revealed aspects of development that error and raw Time Through Sector data were unable to capture, suggesting that it could be used to supplement objective measures in assessing training effectiveness. CWS may also be useful for assessing

skill development in comparisons of various training methods. For example, rates of trainees' CWS score improvements could be used to determine whether a "sink-or-swim" approach is preferable to instruction during training.

Future research with CWS will involve development of a real-time, continuous version and its application to comparisons of training methods.

ACKNOWLEDGEMENTS

This research was supported in part by the Federal Aviation Administration, Department of Transportation (Grant 90-G-026). 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*.