DEVELOPMENT OF THE CRITERION TASK SET PERFORMANCE DATA BASE

Robert E. Schlegel Kirby Gilliland* Betina Schlegel

School of Industrial Engineering
and
*Department of Psychology
The University of Oklahoma
Norman, OK 73019

ABSTRACT

The initial phase of a large-scale experimental study was conducted involving the training and testing of fifty human subjects on the Criterion Task Set (Version 1.0). Testing was performed under baseline conditions and the stressors of noise and sleep loss. The resulting data base includes CTS performance data and subjective ratings obtained using the Subjective Workload Assessment Technique (SWAT) for each task, along with information on subject individual differences. This paper presents the methodology used for the data collection and analysis efforts and provides a summary of the performance and subjective assessment information. In general, no performance differences were found under the noise stress condition. Following sleep loss, response times for the central processing tasks deteriorated as did performance on the Unstable Tracking and Interval Production tasks.

INTRODUCTION

The USAF Criterion Task Set (CTS) is part of an effort to develop a standardized workload assessment methodology which will aid in the design and operation of complex systems. The CTS is one of few, if not the only, task battery based on current theoretical models of human information processing. The battery is composed of nine tasks which differentiate between the three primary stages of processing (perceptual input, central processing, and motor output) and further isolate the separate resources associated with the mode of input (visual/auditory), code in which central processing is performed (spatial/symbolic), and the mode of response output (manual/vocal).

The CTS Version 1.0 tasks include Memory Search, Interval Production, Continuous Recall, Linguistic Processing, Probability Monitoring, Grammatical Reasoning, Mathematical Processing, Unstable Tracking, and Spatial Processing. A detailed description of the CTS and each of the component tasks has been provided by Shingledecker (1984). The training characteristics of the entire task battery have been previously investigated (Schlegel, 1986; Schlegel and Shingledecker, 1985).

The CTS has been applied as a test instrument to evaluate the relative sensitivity, reliability and intrusiveness of a variety of available workload measures. It has also been used as a performance assessment battery to evaluate the effects of various stressors on individual components of the human information processing system. In order to support the widespread application of the CTS, a large-scale validation study was undertaken to initiate the development of a comprehensive CTS data base. The objectives of this paper are to present the methodology used for the data collection and analysis efforts and to provide a summary of the performance and subjective assessment information obtained thus far.

CONTENTS OF THE DATA BASE

The data base includes performance, subjective assessment, and individual differences data along with data on performance under noise and sleep loss conditions. The data is stored in a form that allows easy access by the Statistical Analysis System (SAS). Future expansion and analysis will allow evaluation of subject variability in training and overall performance, and provide better definition of the CTS structure through multivariate analysis.

CTS Performance Data

All tasks except Interval Production were run as standard three-minute trials under the subject-paced condition (CTS Menu Option 1) which places a 15-second limit on subject response time for the central processing tasks. The performance measures for the central processing tasks include the mean and standard deviation of response time for correct responses and counts of total stimuli, and correct and incorrect (both errors and missed) responses during each

three-minute trial. These measures are also derived separately for those stimuli with positive ("YES") responses and negative ("NO") responses.

The performance measures for Unstable Tracking include the mean absolute error and total edge violations for the three-minute trial. Measures for Interval Production include the mean and standard deviation of the tapping intervals along with the variability score for the trial.

Subjective Workload Measure

Throughout the study, subjects were asked to provide subjective assessments of the workload presented by the various CTS tasks. The Subjective Workload Assessment Technique (SWAT) was used to assess subjective workload. The SWAT Scale (Reid, 1985; Reid, Eggemeier, and Nygren, 1982) is a psychometric instrument for measuring subjective ratings on three major dimensions of workload: **T**ime pressure, mental **E**ffort, and psychological **S**tress. Given the demands of any specific workload period, subjects rate each dimension of Time, Effort, and Stress on a 1 to 3 Likert-type scale.

The SWAT not only provides a means for obtaining an individual subject's workload ratings, but also provides a method for establishing cross-subject comparability. This is accomplished by having each subject complete a SWAT sort of all 27 combinations of the three rating levels for all three workload dimensions. The sorting order is then subjected to conjoint scaling yielding a standardized rating metric which can be used for comparisons across subjects.

Individual Difference Psychometric Tests

Subjects in the study were also administered a battery of psychometric tests measuring individual difference dimensions that have a known relationship to performance efficiency or are known to be biologically/perceptually based. This battery included measures of generalized arousal (extraversion), sensation seeking, stimulus screening, test anxiety, clinical anxiety, Type A-B behavior, and impulsiveness. Results of this aspect of the study are reported elsewhere in the Proceedings (Gilliland, Schlegel, and Dannels, 1986).

EXPERIMENTAL METHODOLOGY

General Testing Protocol

The testing protocol consisted of two-hour testing sessions conducted once per day for ten days over a two-week testing cycle. Multiple workstations allowed collection of up to four subjects' data per two-hour session. In the majority of cycles, four two-hour sessions were conducted each test day. Thus, approximately 16 subjects were run during a two-week cycle. Numerous two-week testing cycles were conducted to collect the data necessary for the project.

Primary Study

Subjects. Twenty-five men and 25 women served as subjects in this project. With the exception of approximately ten subjects, the volunteers were undergraduate students attending the University of Oklahoma. All subjects were recruited through posted announcements and were paid for their participation. Male subjects ranged in age from 19 to 32 years (mean = 23.6 years), and female subjects ranged from 18 to 43 years (mean = 23.Ø years). All subjects reported 2Ø/2Ø actual or corrected vision, no history of hearing impairment, and no current use of medication.

Test facilities and equipment. The testing location consisted of a three-room laboratory suite at the University of Oklahoma. One room served as a CTS data collection area, another room served as a data management/reduction area, and the third room was a psychophysiological testing area which served several ancillary testing purposes.

Four workstations were installed in the data collection area. The workstations were separated by acoustic panels to reduce noise and subject interaction. Each workstation consisted of a color CRT monitor and the three, standard CTS response controllers. Installed immediately behind the subjects was the experimenter control station which included a Commodore 64 microprocessor with dual floppy disk drives and a color CRT monitor for each subject workstation.

The data management/reduction room contained an additional Commodore 64 microprocessor system for software development, training, and data reduction/transfer functions. Also installed in this room was a terminal to the University IBM 3Ø81 mainframe computer. This terminal provided direct access to larger computing capacity for data analysis and SWAT analysis.

The CTS Version 1.Ø tasks are written in the BASIC programming language and then compiled. Additional software was developed during this project to

automate the presentation sequence of the tasks and automatically label and store raw data in disk files. Software was also written to analyze and reduce the raw data, construct summary statistics files, and label and store these files.

Procedure. Subjects were generally scheduled in one of four testing session periods: 8:00-10:00 am, 10:00-12:00 am, 1:00-3:00 pm, and 3:00-5:00 pm. On occasions where sessions were not filled, it was necessary to run an additional session between 5:00 and 7:00 pm. Subjects attended a minimum of ten (10), two-hour sessions -- one per day, Monday through Friday, for two weeks.

Each subject was seated at an individual workstation facing the elevated CRT display. Controller boxes were placed on a table in front of the subject. Subjects were instructed to use their right hands for responding with the controller boxes. For a few subjects an exception was made if the subject was left handed and felt that using the right hand would cause a noticeable decrement in performance. Also on the table were a pencil and SWAT rating recording materials.

On Monday of the first week, each subject was oriented to the project, given an introduction to each of the CTS tasks, and completed a SWAT Sort and a battery of psychometric tests. Approximately two hours of additional individual difference testing was scheduled and completed during the two-week period.

On the second through fifth days of the first week, subjects were given the first four training trials on the entire CTS battery. Monday of the second week was used for the last training trial.

The sixth and eighth trials on Tuesday and Thursday of Week 2 were baseline experimental trials. Data on these days were collected under the same conditions imposed during training. Data from trial seven (Wednesday of Week 2) was collected under a noise stress condition described later.

A fixed sequence of the nine CTS tasks was constructed with the restrictions that no two highly difficult tasks were adjacent and that the input/output tasks were balanced within the sequence. The subsequent task order used for all test sessions was as follows:

(1) Memory Search
(2) Interval Production
(3) Continuous Recall
(4) Linguistic Processing
(5) Probability Monitoring
(6) Grammatical Reasoning
(7) Mathematical Processing
(8) Unstable Tracking
(9) Spatial Processing.

Once the CTS task sequence was determined, the workload levels of each task were presented in ascending order within each task. During each testing session, subjects were thus presented three-minute trials of each of the 25 CTS task-level combinations (three workload levels for eight tasks, plus the Interval Production task).

Following each trial was a brief rest period during which data was stored on the diskette and the next task was prepared for presentation. These rest periods were approximately 1 to 1.5 minutes in length depending on the number of subject responses. During these rest periods each subject recorded a SWAT rating for the previous CTS task trial. Total test session time ranged from one hour and forty-five minutes to two hours depending on the data storage time.

Secondary Study 1 - Noise Condition

Studies of the effects of noise on cognitive performance have shown that noise produces a decrement, an improvement, or no change in performance depending on the nature of both the task and the noise presented. Because the CTS includes tasks that cover a range of information processing demand and complexity, it was essential to obtain knowledge of the effects of noise levels common to operational environments on CTS performance.

Subjects. Subjects in this study consisted of the same fifty subjects who served in the Primary Study.

Method. On Wednesday of the second week of testing, separating the two baseline testing sessions, subjects performed the entire two-hour CTS testing session while listening to 75 dBA background noise. The noise was of typical air traffic control room activity, over-recorded two times to obliterate comprehensible speech patterns. This tape recording was developed for use in research at the Federal Aviation Agency's Civil Aeromedical Institute Research Laboratory (Thackray, 1982).

The noise was presented over Panasonic Model 1501Ø loudspeakers using a Panasonic Model RS6Ø8 cassette tape

deck. The sound level was periodically calibrated by measuring the level at the subject seating location (Realistic Model 33-1028 Sound Level Meter).

Secondary Study 2 - Sleep Loss Condition

Sleep loss is known to be detrimental to performance on a variety of tasks, in particular those involving vigilance. Response times on vigilance and certain central processing tasks have shown a significant increase in their mean, median and variability following sleep deprivation. Because of the serious effect of sleep loss in the operation of complex systems, and the similarity of CTS tasks to components of such complex systems, performance data was collected under a sleep loss condition.

Subjects. Five groups (two-week cycles) of subjects participated in the Sleep Loss Secondary Study. This provided a total of 40 subjects, 19 men and 21 women.

Method. On Friday of the second week of testing, subjects reported to the laboratory at 6:00 pm instead of their normal testing time. Subjects were assigned to one of three testing groups which corresponded ordinally to the normal testing time, i.e., those subjects normally tested in the first groups in the morning were placed in the first group tested in the evening. Groups were then given the normal two-hour testing sessions beginning at 6:00, 8:00, and 10:00 pm.

Following CTS testing, subjects completed additional questionnaires, were involved in some ancillary research testing, and then had a light meal about 12:00 midnight. From this time until 6:00 am, subjects watched prerecorded movies, studied, or played board or card games. Beginning at 6:00, 8:00, and 10:00 am, groups of subjects were again tested on the CTS in the same order as the previous evening.

RESULTS

Primary Study

Performance measures and SWAT ratings on all tasks clearly demonstrated the distinction between the low, medium and high workload levels with four exceptions.

(1) The SWAT ratings for the Linguistic Processing task showed little difference between the medium and high levels, although the response time and accuracy scores clearly indicated a performance difference. This may indicate that the tasks at the medium and high levels (vowel/consonant matching and antonym identification) are in fact tapping different resources with a subjectively equivalent workload.

(2) The accuracy measure for the Mathematical Processing task was consistently high (97%) across all levels, i.e., a ceiling effect existed. However, response times differed substantially for all three levels.

(3) The mean absolute error for the Unstable Tracking task did not distinguish between the medium and high levels, again indicating a ceiling effect. However, the number of edge violations provided a clear distinction between all three levels.

(4) The SWAT ratings for the medium and high levels of Spatial Processing were often identical. This indicates that subjects believed these tasks were of equivalent difficulty despite the clear differences in the performance measures.

With respect to performance over time for the central processing tasks, response time and accuracy stabilized within the five training days for all tasks. With few exceptions accuracy stabilized rapidly while response time continued to improve throughout training. The exceptions are the Continuous Recall and Grammatical Reasoning tasks in which accuracy continued to improve (GR, CR medium and high) while response time was stable (GR, CR medium) or increasing (CR high).

The Interval Production variability score showed no improvement with training and unusually poor performance on Day 2. Due to the perceived simplicity of this task, subjects often do not give it the attention it deserves, despite instructions to the contrary. Reinforcement of these instructions following Day 2 may have produced the observed performance recovery.

The Unstable Tracking mean absolute error showed only slight improvement over time. The number of edge violations was a more sensitive indicator of improvement for this task. For all tasks, the most drastic improvements in performance occurred between the first and second days of testing.

The SWAT ratings decreased over time for all tasks except Linguistic Processing, Grammatical Reasoning, and Mathematical Processing. The ratings showed consistent ordered differences between workload levels for all tasks. A ranking of the ratings provides a comparison of the relative difficulty across tasks (Table 1).

Table 1. Subjective Task Difficulty.

Workload	Rank	Task
Low	1	Interval Production
	2	Memory Search
	3	Spatial Processing
	4	Linguistic Processing
	5	Mathematical Processing
	6	Probability Monitoring
	7	Continuous Recall
	8	Unstable Tracking
High	9	Grammatical Reasoning

Secondary Study 1 - Noise Condition

No differences were observed on any performance criteria under the noise condition. However, the SWAT ratings were higher for Memory Search, Continuous Recall, Linguistic Processing, and Grammatical Reasoning. These are the first four tasks in the presentation sequence and may reflect the subjects' adjustment to the different environmental conditions.

Secondary Study 2 - Sleep Loss Condition

In general, the mean response times for the central processing tasks increased following sleep deprivation. There was little accompanying change in accuracy. Unstable Tracking was affected only at the lower levels where tracking is relatively easy and is more of a vigilance task. Interval Production was very sensitive to the sleep deprivation due to the lack of attentiveness and number of times that tapping ceased entirely. These results agree with the results of previous studies. SWAT ratings on all tasks, at all levels increased substantially following sleep deprivation.

ACKNOWLEDGEMENT

This research was sponsored by the Workload and Ergonomics Branch of the Armstrong Aerospace Medical Research Laboratory, United States Air Force, under Contract F33615-82-D-0627 through the Southeastern Center for Electrical Engineering Education (SCEEE-ARB/85-62). The United States Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright notation hereon. The authors wish to thank Gary B. Reid, Dr. F. Thomas Eggemeier, and Dr. Clark A. Shingledecker for their interest and support.

REFERENCES

Gilliland, K., Schlegel, R.E. and Dannels, S.A. (1986). Individual Differences in Criterion Task Set Performance. In **Proceedings of the Human Factors Society 30th Annual Meeting.** Dayton, Ohio: Human Factors Society.

Reid, G.B. (1985). The Systematic Development of a Subjective Measure of Workload. In **Proceedings of the 9th Congress of the International Ergonomics Association** (pp. 109-111). Bournemouth, England: International Ergonomics Association.

Reid, G.B., Eggemeier, F.T. and Nygren, T.E. (1982). An Individual Differences Approach to SWAT Scale Development. In **Proceedings of the Human Factors Society 26th Annual Meeting** (pp. 639-642). Seattle, Washington: Human Factors Society.

Schlegel, R.E. (1986). **Development of an Optimal Testing Protocol for the USAF Criterion Task Set** (Final Report SCEEE-84 RIP 47). Norman, Oklahoma: The University of Oklahoma.

Schlegel, R.E. and Shingledecker, C.A. (1985). Training Characteristics of the Criterion Task Set Workload Assessment Battery. In **Proceedings of the Human Factors Society 29th Annual Meeting** (pp. 770-773). Baltimore, Maryland: Human Factors Society.

Shingledecker. C.A. (1984). **A Task Battery for Applied Human Performance Assessment Research** (Tech. Report AFAMRL-TR-84-071). Wright-Patterson AFB, Ohio: Medical Research Laboratory.

Thackray, R.I. (1982). Some Effects of Noise on Monitoring Performance and Physiological Response. **Academic Psychology Bulletin, 4,** 73-81.