D4

N82-13669

SUBJECTIVE RATING SCALES AS A WORKLOAD ASSESSMENT TECHNIQUE

Kathleen L. Bird*

Virginia Polytechnic Institute

SUMMARY

Any investigation of the task workload inherent in flying must address many dimensions including cognitive, perceptual, and psychomotor. The present study employs a multidimensional bipolar-adjective rating scale as a subjective measure of operator workload in the performance of a one-axis tracking task. The rating scale addressed several dimensions of workload, including cognitive, physical, and perceptual task loading as well as fatigue and stress effects. Eight subjects performed a one-axis tracking task (with six levels of difficulty) and rated these tasks on several workload dimensions. Performance measures were tracking error RMS (root-mean square) and the standard deviation of control stick output. Significant relationships were observed between these performance measures and skill required, task complexity, attention level, task difficulty, task demands, and stress level.

INTRODUCTION

There is little agreement among scientists in how they conceive workload. To arrive at a functional, accurate, definition of workload several questions must be addressed. Does workload refer to the task demands imposed on the operator, or is workload the operator effort required to satisfy those task demands? What role, if any, does operator fatigue, physical and mental, as well as emotional stress play in the operator's assessment of workload? Is an individual's assessment of workload level really an assessment of a combination of all these factors? Most of the current workload definitions focus on a single facet of this multidimensional area. Jahns (1973) defines workload as "...the extent to which an operator is occupied by a task" (reference 1). Focusing on task performance measurements, Levison, Elkind, and Ward (1973) define workload as "...the fraction of the controller's capacity that is required for him to perform a given task to some specified or criterion level of performance" (reference 2). In an attempt to express the multidimensional aspects of workload, Tennstedt (1973) defines it as "...a summation of such processes as perception, evaluation, decision-making and actions taken to accomodate those needs generated by influences originating within or without the aircraft" (reference 3). While Tennstedt's workload definition addresses several workload dimensions, it falls short of addressing those questions previously posed.

* This research was conducted at NASA-Ames Research Center and was sponsored by NASA grant NAG-217 to Virginia Polytechnic Institute and State Univ.

The multidimensional aspect of workload is demonstrable in the flight duties of a pilot. A pilot's flight duties may encompass facets of cognitive (e.g., in-flight computations, fuel consumption management), perceptual (e.g., monitoring instruments, kinesthetic cues), and psychomotor (e.g., manual control of the yoke, rudder pedals) aspects. The pilot must also encounter, and cope with, the effects of fatigue and stress (mental and physical).

To investigate the multidimensional aspects of flying, workload assessment techniques (both behavioral and physiological) should address the cognitive, perceptual and psychomotor dimensions, as well as measure operator fatigue, and stress. Wierwille (1979) has suggested that a fruitful area of research would combine the best of physiological measures with behavioral measures in a multivariate analysis as a function of workload (reference 4). One or more workload assessment techniques need to be developed that can reliably measure the multiple dimensions of workload.

A subjective rating scale may offer a promising behavioral workload assessment technique. Hicks and Wierwille (1979) compared workload measurements obtained from rating scales with those obtained from primary task performance, secondary task performance, occlusion, and physiological measures (reference 5). Specifically, the rating scale proved to be a sensitive measure of operator workload in the performance of an automobile driving simulation task. Jenney, Older, and Cameron (1972) reported "...encouraging findings as to the usefulness and validity of subjective magnitude estimates" (reference 6). They recorded hourly subjective estimates of fatigue, tension, and task difficulty in assessing workload levels involved in performing an information processing task. Borg (1971) employed a simple rating scale and reported good agreement between perceived exertion, and difficulty, and physiological indicators of effort (stress) (reference 7).

The purpose of this study was to develop and validate a multidimensional rating scale to assess pilot workload. Several dimensions of workload were addressed, including cognitive, physical, and perceptual task loading as well as fatigue and stress effects.

Subjective Rating Scale

The multidimensional rating scale included 15 bipolar adjective pairs, one or more pairs addressing each of the several workload dimensions. These bipolar adjective pairs dichotomized: 1. skill required (no skill-much skill), 2. task complexity (simple - complex), 3. attention level (extremely low - extremely high), 4. monitoring (none - constant), 5. task difficulty (easy - difficult), 6. controlability (easy - difficult), 7. my performance (unsatisfactory - satisfactory), 8. instructions (clear - confusing), 9. task demands (undemanding - demanding), 10. energy level (lazy - energetic), 11. stress level (low stress - high stress), 12. activity level (idle - busy), 13. fatigue (tired - refreshed). 14. task stability (predictable - unpredictable), and 15. interest level (bored - interested).

The scales appeared one at a time on a CRT (cathode ray tube). The adjectives were positioned at opposing ends of a vertical line, with the descriptor (e.g., skill required, attention level) positioned below the scale. Subjects assigned a subjective rating (scale 1 to 100) to the tasks by positioning a cursor along the vertical line.

To validate the multidimensional rating scale, it is necessary to have subjects perform a battery of tasks which concentrate on different aspects of workload and examine whether the rating scale accurately measures these aspects. Future studies will employ a battery of six to eight primary tasks (similiar to the Civil Aeromedical Institutes Multiple Task Performance Battery, MTPB) which will include cognitive, perceptual, and psychomotor components. The primary tasks selected will closely approximate tasks demanded in flying.

The present study examines the psychomotor aspects of workload. Subjects performed a one-axis compensatory tracking task with six levels of difficulty. They rated the six tracking tasks for degree of workload using the multidimensional rating scale

Tracking Task

The task was a one-axis compensatory tracking task with a K/S plant. A random number generator provided a rectangular distribution of frequencies (bandwidth of 1.0, 1.5, 2.0 rad/sec) filtered through a second-order filter to produce the forcing function. The filtered output produced the movement of the cursor.

Difficulty was manipulated by varying the standard deviation (SD of 32, 64) and bandwidth (1.0, 1.5, 2.0). The following tasks were presented: 1. task 1 (bw 1.0, SD 32), 2. task 2 (bw 1.5, SD 32), 3. task 3 (bw 2.0, SD 32), 4. task 4 (bw 1.0, SD 64), 5. task 5 (bw 1.5, SD 64), 6. task 6 (bw 2.0, SD 64). Performance measures were the tracking error RMS (root-mean square) and the standard deviation of the control stick output.

The tracking tasks were presented on a CRT. The six tracking tasks consisted of a vertical line (5.56 cm) which randomly moved in a lateral direction. Maximum displacement of the cursor was 12.70 cm. The subjects task was to keep this cursor centered between two stationary vertical lines (2.11 cm) by means of a control stick right and left.

METHOD

Subjects

Five males and three females (aged 18 to 42) served as paid volunteers. These subjects had been previously screened for tracking ability to gaurantee a minimum amount of psychomotor ability. A pilot study

yielded a criterion score which the subjects were required to achieve before selection. All subjects were right handed.

Apparatus

This study was conducted in a small, sound-attenuated experimental chamber. The subject was seated before a CRT. The control stick was located on the right arm of the chair. The throttle was located on the left arm. Data aquisition was recorded and task presentations were programmed through a Digital Equipment Corporation PDP-12 computer.

Procedure

Subjects were told the purpose of the study, given a description of the required tasks, and instructions for rating the tasks on the various workload parameters. They were told these tasks would vary in degree of difficulty. The importance of maintaining an equally high standard of performance across all tasks was stressed. To familiarize the subjects, tasks were presented (in order of ascending difficulty) and the subjects were permitted to track each task for one minute. During the experimental session the tasks were not presented in order of ascending difficulty but rather in random order. The subjects were given a one-minute practice session prior to each experimental session. The experimental session (for each tracking task) immediately followed the practice session for a duration of four minutes. After completing each tracking task subjects gave a rating for each of the 15 bipolar adjective pairs. As the scales appeared on the CRT subjects would move the throttle to position a cursor along the vertical line to indicate their rating. When they were satisfied with their rating, they pressed a response button. Immediately, a second scale would be displayed.

Subjects were required to perform each tracking task for a duration of four minutes. Standard deviation of the tracking error, output error RMS, standard deviation of control stick output, and stick output RMS were sampled every 30 milliseconds. The following analyses were performed on the data collected during the final two minutes of each experimental tracking session.

RESULTS

Effect of task difficulty on error RMS

A 2 (standard deviation) x 3 (bandwidth) analysis of variance was computed on the error RMS to determine if a significant difference in error RMS would appear as a result of manipulating the bandwidth and standard deviation. Results indicated a significant difference attributable to the bandwidth factor, F (2,14) = 49.30, $p < .01$ and the standard deviation

factor, F (1,7) = 311.59, $p < .01$. In addition, there was a significant interaction between the bandwidth and standard deviation factors, F (2,14) = 10.60, $p < .01$.

Effect of task difficulty on the standard deviation of stick output

A 2 (standard deviation) x 3 (bandwidth) analysis of variance was computed on the standard deviation of stick output to determine whether manipulating the bandwidth or standard deviation would produce a significant difference. There was a significant difference attributable to the bandwidth factor, F (2,14) = 8.25, $p < .01$ and the task standard deviation factor, F (1,7) = 59.81, $p < .01$. There was no significant interaction between these factors, F (2,14) = 2.24, $p > .05$.

Effect of task difficulty on bipolar adjective ratings

Fifteen 2 (standard deviation) x 3 (bandwidth) analyses of variance were computed on the ratings for each of the 15 bipolar adjective scales. There was a significant difference attributable to the bandwidth factor for the following scales: skill required (F (2,14) = 9.62, $p < .01$); monitoring (F (2,14) = 5.05, $p < .05$); task difficulty (F(2,14) = 12.59, $p < .01$); my performance (F (2,14) = 8.71, $p < .01$); task demands (F (2,14) = 4.46, $p < .05$); and stress level (F (2,14) = 5.90, $p < .05$). There was a significant difference in the ratings attributable to the standard deviation factor for the following scales: skill required (F(1,7) = 55.86, $p < .01$); task complexity (F(1,7) = 28.84, $p < .01$); task difficulty (F (1,7) = 18.04, $p < .01$); controlability (F(1,7) = 13.93, $p < .01$); my performance (F(1,7) = 7.67, $p < .01$); task demands (F(1,7) = 6.90, $p < .05$); stress level (F(1,7) = 7.94, $p < .05$); and fatigue (F(1,7) = 20.87, $p < .01$). No significant effect attributable to either the bandwidth or standard deviation factors were found for the following scales: attention level; instructions; energy level; activity level; task stability; or interest level ($p > .05$). There were also no significant interactions between the bandwidth and standard deviation factors for any of these analyses (with the majority of F values less than one, $p > .05$).

Relationship between error RMS and bipolar adjective ratings

To determine if a significant relationship exists between the error RMS and the scale ratings, Pearson product-moment correlations were computed (df = 47). The following significant correlations were derived between scale ratings and error RMS scores: skill required (r = +.55, $p < .01$); task complexity (r = +.40, $p < .01$); attention level (r = +.42, $p < .01$); monitoring (r = +.44, $p < .01$); task difficulty (r = +.57, $p < .01$); controlability (r = +.61, $p < .01$); task demands (r = +.51, $p < .01$); stress (r = +.28, $p < .05$); and task stability (r = +.32, $p < .05$). The remaining scales were not significantly correlated (at or above $p > .05$) with error RMS: my performance, instructions, energy level, activity level, fatigue,

and interest level.

Relationship between stick output standard deviation and scale ratings

To determine if a significant relationship exists between the stick output standard deviations and the scale ratings, Pearson product-moment correlations were computed (df = 47). The following bipolar adjective scales were found to be significantly correlated with stick output standard deviations: skill required ($r = +.55$, $p < .01$); task complexity ($r = +.53$, $p < .01$); attention level ($r = +.29$, $p < .05$); task difficulty ($r = +.55$, $p < .01$); my performance ($r = -.45$, $p < .01$); task demands ($r = +.46$, $p < .01$); stress ($r = +.40$, $p < .05$); and activity level ($r = +.29$, $p < .05$). The following scales were not significantly correlated with stick standard deviation: monitoring, controlability, instructions, energy level, fatigue, task stability, and interest level.

The correlation between error RMS and standard deviation of the stick output was significant ($r = +.44$, $p < .01$).

DISCUSSION

The relationship between increasing task demands and task performance (output error RMS) was examined. As the task demands increase (with an increase in input bandwidth and input standard deviation) subjects' ability to reduce this error decreased. Increasing the input standard deviations from 32 to 64 produces an increase in output error RMS. This result might be expected considering the relative amount of error the subject is asked to reduce. Increasing the bandwidth (1.0, 1.5, 2.0) produced an increase in output error RMS (Mean = 25.8, 36.2, 43.5). Manipulating the task bandwidth and standard deviation had a significant effect on the standard deviation of stick output. Doubling the input standard deviation (from 32 to 64) also doubled the mean standard deviation of the stick output (Mean = 12.3 (SD 32), Mean = 24.9 (SD 64)). A similiar increase in mean stick standard deviation could be attributed to an increase in the bandwidth (1.0 (mean = 15.7), 1.5 (mean = 19.7), 2.0 (mean = 20.6)). In summation, as the task demand increases, a degradation of task performance occurs. As task demand increases, subject effort (as measured by stick output standard deviation) also increases.

Theoretically, an increase in task difficulty should be reflected in the subject's evaluation of task workload level. Several rating scales are strongly related to output error RMS scores. Apparently performance degradation was strongly reflected in evaluation of skill required, task difficulty, controlability, and task demands. An increase in subject effort was strongly reflected in evaluation of skill required, task complexity, task difficulty, and task demands. These scales are most promising as indicators of certain workload dimensions and should be investigated further with other flight related tasks.

REFERENCES

1. Jahns, D.W. Operator workload: What is it and how should it be measured? In K.D. Cross and J.J. McGrath (Eds.) Crew System Design. Santa Barbara, California: Anacapa Sciences, July, 1973.

2. Levison, W.H., Elkind, J.I. and Ward, J.L. Studies of multivariable manual control systems: A model for task interference. NASA-CR-1746, 1971.

3. Wierwille, W.W. Pysiological measures of aircrew mental workload. Human Factors, 1979, 21, pp. 575-593.

4. Hicks, T.G. and Wierwille, W.W. Comparison of five mental workload assessment procedures in a moving-base driving simulator. Human Factors, 1979, 21(2), pp.129-143.

5. Jenney,L.L., Older, H.J., and Cameron, B.J. Measurement of operator workload in an information processing task. NASA-CR-2150, 1972.

6. Borg, G. Psychological and physiological studies of physical work. In Singleton, W.T., et al. (Ed.) Measurement of Man at Work. London: Taylor and Francis, 1971, pp. 121-128.