TRAINING AND SUBJECTIVE WORKLOAD IN A CATEGORY SEARCH TASK

Michael A. Vidulich & Parimal Pandit
NASA - Ames Research Center, Moffett Field, California
San Jose State University, San Jose, California

ABSTRACT

This study examined automaticity as a means by which training influences mental workload. Two groups were trained in a category search task. One group received a training paradigm designed to promote the development of automaticity; the other group received a training paradigm designed to prohibit it. Resultant performance data showed the expected improvement as a result of the development of automaticity. Subjective workload assessments mirrored the performance results in most respects. The results supported the position that subjective mental workload assessments may be sensitive to the effect of training when it produces a lower level of cognitive load.

INTRODUCTION

Workload assessment is becoming progressively more important to system evaluation. For example, cognitive overload has been cited as a major problem for many existing and proposed high technology military aviation systems (Cordes, 1985). This trend makes understanding the variables that influence mental workload a crucial need.

One variable that has inexplicably received little attention, but that potentially exerts considerable influence on workload, is training. Perhaps this is because there may appear to be an obvious answer. Most people have experienced a task that at first seems appallingly difficult, but with practice becomes manageable. For many people learning to read or to drive would be likely examples. Generalizing from experiences such as these could lead to an implicit hypothesis that training (or practice) reduces workload. However, little reflection is required to see that this hypothesis, though plausible and descriptive of many situations, is probably incomplete. For example, previous research (e.g., Schneider and Shiffrin, 1977) has shown that even extensive practice does not invariably improve performance. Schneider and Shiffrin demonstrated that if the same information was always associated with the same responses (i.e., there were consistent stimulus-response (S-R) mappings), then performance was considerably improved by practice. However, if during practice the same information led to different responses (i.e., varied S-R mappings), then performance was only marginally improved by practice.

The present study examined the effect of consistency of training on mental workload. Consistency was manipulated in the same fashion as in Schneider and Shiffrin's (1977) research. The consistently mapped (CM) condition had a set of stimuli which the subject always responded to as "targets," and another set of stimuli always as "distractors." The varied mapping (VM) condition used the same basic task, but on any given trial the stimuli used as targets might have been previously used as distractors and vice versa. The task used in the present study was a category search task. This task has been used in automaticity research and its performance has been demonstrated to be influenced by the consistency manipulation (e.g., Fisk and Schneider, 1983). One group of subjects was trained with a CM procedure, and the other group was trained with a VM procedure. Workload was assessed via a subjective rating procedure. Subjective workload ratings were selected for two main reasons: (1) They are the most commonly used technique in applied settings (Sheridan, 1980). Therefore, the need to understand their strengths and weaknesses is most urgent. (2) They are particularly sensitive to the demands placed on working memory (Vidulich, in press), and this is an expected locus of the changes in cognitive resource demand as a result of CM training (Schneider and Shiffrin, 1977). The experimental hypothesis was that CM training of the category search task would produce dramatically improved performance by reducing working memory load and thereby greatly reduce rated workload relative to the effect of the VM training.

METHOD

Subjects

Sixteen right-handed college students (9 males and 7 females) participated as paid subjects in this study. All subjects had normal vision and hearing. Subjects were divided into two equal groups. One group practiced the category search task with a CM procedure, the other group practiced with a VM procedure.

Apparatus

Subjects were seated in a sound and light attenuated booth. Approximately 90 cm in front of the subject and slightly below eye level was a 22 cm Sanyo DM 5109CX CRT display. The right armrest of the subjects' chair had two micro-switches positioned under the subjects' right index and middle fingers. The left armrest had a slide switch and button within easy reach of the left hand. The sequence of stimuli was controlled by a program implemented on an Apple II+ computer equipped with a monitor, a Corvus Systems 20 Megabyte hard disk, and an Epson MX 80 F/T printer. The subjects' input and output (I/O) was interfaced to the computer via an ISAAC Cyborg Model 91A interface unit. The experimenter communicated with the subject through an intercom.

Task

The experimental task was a category search task. Each trial block began with the specification of a memory set. Memory set size was either 1-category or 4-category. Following the memory set a sequence of 36 trials was run. Each trial consisted of the presentation of a probe word in the center of the CRT. If the probe word was a target (i.e., it belonged to a memory set category), the correct response was to press the switch under the right index finger. If the word was a distractor (i.e., it did not belong to a memory set category) the correct response was to press the other button. The inter-trial interval was approximately 500 ms. The reaction time (RT) and accuracy were recorded for each trial.

For the CM group the memory set items were always selected from a specified set of categories; Animals, Vehicles, Body parts, and Weapons. Exemplars from these categories were never employed as distractors for the CM group. The CM group's distractors were always exemplars from the following categories; Trees, Colors, Clothes, and Furniture. In the varied mapping (VM) group, memory set categories for each trial block. were selected randomly from all eight categories listed above. In any given trial block, distractors were selected from any category not in the memory set.

Procedure

Subjects began by reading instructions describing the task and rating scales. The subjects were then run through a series of 40 trial blocks; 20 trial blocks had 1-category memory sets, and the other 20 had 4-category memory sets. The sequence of the trial blocks with the two memory set sizes was randomly determined. Following each trial block, subjects were informed of their average RT and number of correct responses. Short breaks (10-15 minutes) were provided following trial blocks 8 and 16. Subjects were exhorted to respond as quickly as possible without making errors. The session lasted approximately 3 hours.

Subjective Workload Ratings

The NASA-Bipolar technique was used to assess subjective mental workload (Hart, Battiste, and Lester, 1984). Following trial blocks numbers 1, 2, 7, 8, 13, 14, 19, and 20 of both memory set sizes, all subjects rated their mental workload. A set of nine bipolar scales were used: Task Difficulty, Time Pressure, Performance, Mental/Sensory Effort, Physical Effort, Frustration, Stress, Fatigue, and Activity Type. Following each to-be-rated trial block the individual scales were presented to the subject, one at a time, on the CRT. The subject then used a slide switch to position an arrow along a line marked with two endpoint descriptors (e.g., Very Low, Very High).

These individual ratings on the scales were weighted and combined into a single workload score for each trial block. The weights were provided by a "workload parameters evaluation" that was conducted following the completion of all of the trial blocks. This evaluation consisted of a sequence of all 36 pairings of the 9 rating scales being presented to the subject in a forced-choice paradigm. The subject identified the member of each pair that was more important to his/her conception of workload. The number of times the subject picked each scale served as a weight for that scale's ratings for calculating the final workload rating. The magnitude of the independent scales were thus adjusted according to individual subject biases concerning workload (Hart *et al.* 1984).

RESULTS

RT data of the two groups were first compared to test for the development of automaticity. The subjective ratings were then examined in relation to the performance.

Reaction Time Data

The RT data were analyzed with a Consistency (CM or VM) x Memory Set Size (1-category or 4-category) x Trial Block (1 through 20) x Probe Type (target or distractor) analysis of variance (ANOVA). All four variables had significant main effects. The CM group was almost 200 ms faster than the VM group (575 ms vs. 773 ms, $F(1,14)= 22.21$, $p< 0.003$). Four-category memory sets took longer to respond to than 1-category (785 ms vs. 563 ms, $F(1,14)= 97.68$, $p< 0.001$). Subjects' RTs decreased across trial blocks (from 855 ms to 614 ms, $F(19,266)= 17.24$, $p< 0.0001$). There was a RT advantage for identifying targets as opposed to rejecting distractors (641 ms vs. 708 ms, $F(1,14)= 48.67$, $p< 0.0001$).

The Consistency x Memory Set Size x Trial Block interaction was significant ($F(19,266)= 4.42$, $p< 0.0001$). This interaction, illustrated in Figure 1, shows the effect of consistency on the potency of training. The VM group showed only mild changes in performance as a result of the practice. In contrast, the CM group's performance in the 4-category task quickly reached a level comparable to 1-category task.

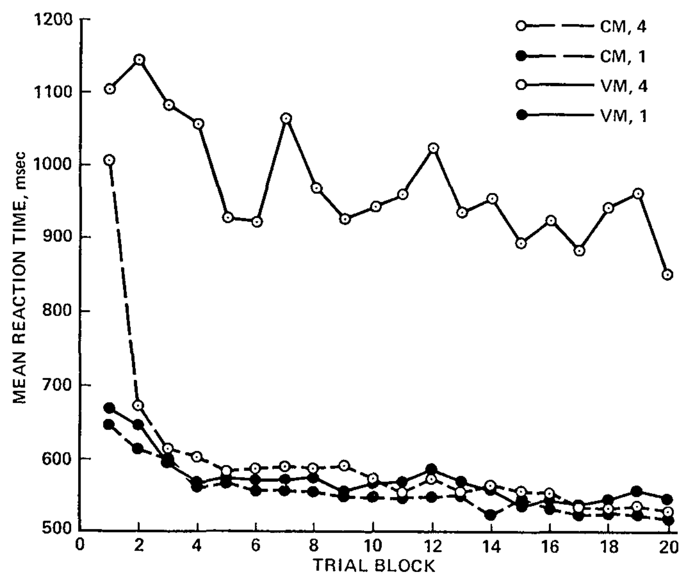

Figure 1 - Mean RT across trial blocks as a function of training consistency and memory set size.

The Consistency x Memory Set Size x Probe Type interaction ($F(1,14) = 10.73$, $p < 0.006$), is shown in Figure 2. For the VM group, the difference between of identifying a target and rejecting a distractor increased from a 54 ms target advantage for the 1-category condition to 140 ms for the 4-category condition; while for the CM group the advantage only increased from 37 ms to 42 ms.

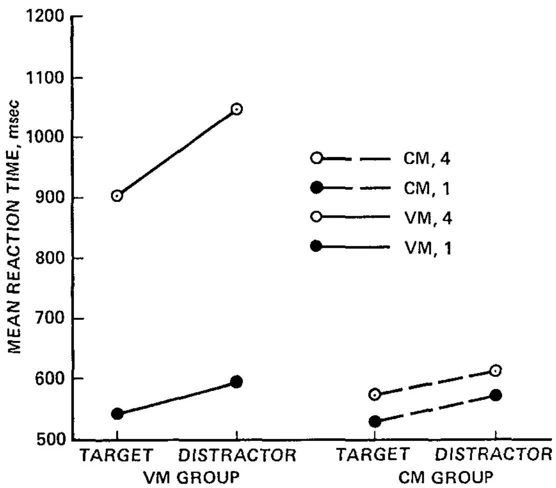

Figure 2 - Mean RT as a function of training consistency, memory set size, and probe type.

The Memory Set Size x Trial Block x Probe Type interaction was also significant ($F(19,266) = 2.91$, $p < 0.0001$). This reflected the fact that the target advantage was relatively stable across trials when the memory set size was 1-category (ranging from 30 to 50 ms), but when the memory set size was 4 the target advantage was highly variable (ranging from near 0 to 120 ms).

Finally, Consistency x Trial Block x Probe Type also interacted ($F(19,266) = 1.94$, $p < 0.02$). For the VM group the difference between the probe types remained relatively large (often more than 100 ms) and very unstable across trial blocks. For the CM group the difference between the probe types rapidly decreased and stayed relatively stable (usually between 30 and 50 ms).

Overall, the interactions in the RT data indicate that, compared to the VM group, the CM group became insensitive to the effects of increasing memory set size and the difference between identifying a target and rejecting a distractor as a result of training. The effects in the RT data appear to confirm that the CM group developed a level of automaticity which resulted in superior performance.

However, before the RT results can be accepted at face value it is essential to examine the accuracy data for any indication of a speed-accuracy trade-off. That a speed-accuracy trade-off was not operative in the present experiment appeared to be indicated by three findings: (1) All subjects, as encouraged by the instructions, made very few errors. Overall, in the experiment, subjects averaged about 1 error per trial block (36 trials). (2) The average accuracy for the two consistency groups were essentially identical; 97.2% for the VM group, 97.1% for the CM group. (3) The correlation between average RT and average percentage of correct responses across task conditions for the two groups was statistically significant ($r(158) = -.40$, $p < 0.01$), but indicated that the fastest RTs were associated with the highest accuracies. Consequently, there was no evidence that a speed-accuracy trade-off was responsible for the trends observed in the RT data.

Subjective Workload Ratings

The subjective ratings data were analyzed with a Consistency x Memory Set Size x Trial Block ANOVA. Two main effects were significant. As the memory set size was increased from 1 to 4, the average rating increased from 24 to 30 ($F(1,14) = 32.98$, $p < 0.0001$). The main effect of trial ($F(7,98) = 2.63$, $p < 0.02$) may have reflected only the variability of the mean rating over trial blocks although there was a slight downward trend.

The Consistency x Memory Set Size interaction was significant ($F(1,14) = 11.16$, $p < 0.005$). As displayed in Table 1, the memory set size manipulation had a relatively minor effect on the CM group compared to the large effect present in the VM group.

Table 1

Workload Ratings as a Function of Memory Set Size and Consistency

	Group	
Memory Set Size	CM	VM
1-category	20	27
4-category	23	37

DISCUSSION

The expected performance differences between the CM and VM groups were obtained. The VM group's performance was consistent with the operation of a serial, self-terminating search strategy. The serial aspect of such a strategy would account for the difference between the memory set conditions. Since, each comparison is done one-at-a-time in a serial strategy the RT is predicted to slow with increases in memory set size. The self-terminating aspect would explain the target identification advantage. Responses to targets would be faster because targets could often be identified without exhaustively searching the entire memory set. However, rejecting a distractor would always require an exhaustive item-by-item comparison with every memory set member. Also, as the memory set size increases, the target advantage will increase. For example, if the memory set is increased from

2 members to 4 members; the number of comparisons to reject a distractor increases from 2 to 4, the average number of comparisons to identify a target increases from 1.5 to 2.5, and the expected target advantage increases from 0.5 to 1.5 comparisons. Overall, the serial, self-terminating search strategy provides an adequate description of the VM group's performance in the present experiment, and has been suggested by previous research with similar tasks (e.g., Naus, Glucksberg, and Ornstein, 1972).

However, the CM group's performance was not consistent with serial self-terminating search strategy. As illustrated in Figures 1 and 2, the effects of memory set size on overall RT and the target advantage were negligible for the CM condition. A model for the CM group's performance would, therefore, seem to require either parallel processing to allow multiple comparisons to occur concurrently or an "integration" mechanism for combining all of the target categories into a single "metacategory." In either case, it is evident that the CM group's task-related cognitive processing was far more efficient than the VM group's processing.

The interaction between consistency and memory set size in the workload data indicates that the subjective assessments were effective in detecting the different levels of cognitive load. In the present experiment, the subjective workload assessments were demonstrated to be sensitive to the difference between the relatively demanding VM condition and the more efficient CM condition. The present experimental design did not permit testing the sensitivity of subjective workload assessments to the difference between target identification and distractor rejection since the ratings were only collected at the end of trial blocks and probe type was manipulated within trial blocks.

In general, the results suggest that training will reduce mental workload only when it reduces the overall demand for cognitive resources. One method of accomplishing this appears to be to develop CM training regimes to promote automaticity. Consequently, in addition to the fact that practice does not always make perfect (Schneider, 1985), the present data would suggest that it does not always reduce workload either.

Further research suggested by the present study includes: (1) A better demonstration of workload changes over practice is needed. The fact that the workload data had no three-variable interaction involving consistency, memory set size and trial block comparable to the performance data displayed in Figure 1, is troublesome. The onset of automaticity in the present experiment was apparently so rapid that, combined with the variability of workload assessments, it precluded detection of the development of automaticity. Extension of the present results with a task that produces more gradual automation would greatly strengthen the present findings. (2) Examination of the changes in the workload associated with a complex task as a result of automatization of its' components is needed. It is not at all clear that practice inevitably reduces workload in such circumstances. It seems likely that practice in complex tasks would develop automaticity at lower levels (for example, learning the proper form for strokes in tennis), but that the cognitive resources thus saved would be reinvested at higher levels (e.g., developing a more global game strategy). An evaluation of changes both in rated workload and verbal reports on the contents of working memory as a result of training in such complex tasks should be very informative. Such work could possibly point the way to using subjective workload assessments and verbal protocols as measures of training program success.

REFERENCES

Cordes, C. (1985). Military waste: The human factor. *American Psychological Association Monitor, 16*(7), 1, 17-18.

Fisk, A. D., and Schneider, W. (1983). Category and word search: Generalizing search principles to complex processing. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 9,* 177-195.

Naus, M. J., Glucksberg, S. and Ornstein, P. A. (1972). Taxonomic word categories and memory search. *Cognitive Psychology, 3,* 643-654.

Hart, S. G., Battiste, V., and Lester, P. T. (1984). Popcorn: A supervisory control simulation for workload and performance research. In *Twentieth Annual Conference on Manual Control* (pp. 431-454). Washington D.C.: NASA Conference Publication 2341.

Schneider, W. (1985). Training high-performance skills: Fallacies and guidelines. *Human Factors, 27,* 285-300.

Schneider, W. and Shiffrin, R.M. (1977). Controlled and automatic human information processing: I. Detection, search, and attention. *Psychological Review, 84*, 1-66.

Sheridan, T. B. (1980). Mental workload: What is it? Why bother with it? *Human Factor Society Bulletin, 23*, 1-2.

Vidulich, M. A. (in press). The cognitive psychology of subjective mental workload. In P. A. Hancock and N. Meshkati (Eds.), *Human Mental Workload*, Amsterdam, The Netherlands: Elsevier.