ERGONOMICS, 1996, VOL. 39, NO. 9, 1146–1162

Effects of control order, augmented feedback, input device and practice on tracking performance and perceived workload

P. A. HANCOCK

Human Factors Research Laboratory, University of Minnesota, Minneapolis, USA

Keywords: Perceived workload; Skill acquisition; Skill retention; Augmented feedback.

Virtual interfaces to advanced human-machine systems will present operators with a variety of perceptual-motor challenges. To inform the virtual interface design processes, the present experiments examined the effects of track order, level of knowledge of performance, type of control device, and the extent of practice on tracking performance and associated mental workload. Tracking was assessed by root mean square error. Subjective workload was measured using both the NASA Task Load Index (TLX) and the Subjective Workload Assessment Technique (SWAT). Results indicated non-linear effects, where tracking error and subjective workload both increased non-proportionally with track order. Trackball use resulted in more accurate performance and was judged to be of lower subjective workload than input using a mouse. Augmented knowledge of performance had little effect on either performance or workload. There were a number of interactions affecting performance that were replicated in perceived workload. Over acquisition trials, second-order tracking exhibited continuous improvement. This capability was retained even after a 30-day rest interval. Decrease in workload followed performance improvement in both initial acquisition and subsequent retention phases. The two subjective workload scales were essentially equivalent in terms of their sensitivity to task manipulations. These results support the direct association between workload and performance and confirms the use of workload in helping to evaluate the influence of diverse task-related demands. The implications for the design of virtual interfaces to real-world systems are examined in the light of these findings.

1. Introduction

In most contemporary systems, computers mediate between human and machine. In many systems, the computer itself is the machine. This dominance has stimulated a convergence in the evolution of both the physical and cognitive interface. In addition to the 'push' of a common processing medium, there is a 'pull' of inherent human capabilities (Hancock 1995). As progressively more advantage has been taken of human spatial processing, interface displays have developed from alpha-numeric designations to iconic representations (Shneiderman 1983). Given the human facility for action in complex four-dimensional worlds, it appears that virtual environments represent strong candidates to become 'interfaces of the future' as information displays advance along this line of progress (Kozak *et al.* 1993). While these developments may specify *how* future interfaces could operate, exactly what will be displayed in these 'worlds' has yet to be articulated clearly. Smith and Hancock (1995) have proposed that one such representation will be a 'risk space' in which the constraints to safe operation will be directly perceivable as visual boundaries

0014–0139/96 $12·00

(Harwood *et al.* 1994, Rasmussen and Pjetersen 1995). In such circumstances, system operation will be akin to 'navigating' a dynamic phase space whose changes are predicated upon the specific characteristics of the system under control and the external perturbations to which it is subject. Elsewhere, this metaphor of navigation has been explored in more detail, focusing particularly upon situation-specific dynamics (Hancock and Chignell 1995). The main task of the operator in navigating these environments is one with which ergonomists are intimately familiar; that is tracking. The task of the interface designer in these 'worlds' is to present the relevant information in a manner that facilitates the discovery and pursuit of safe and efficient passages. How such displays might be constructed is currently a particularly active area of research (Bennett and Flach 1992, Hansen 1995, Vicente and Rasmussen 1990).

Several practical considerations emanate from this expectation for future human-machine interfaces. How should control be effected in these dynamic spaces? What form of feedback should be presented and what are the effects of extended practice in these conditions? The present experimental sequence addresses such issues through the concurrent examination of both performance and mental workload. This dualistic approach has the added advantage of addressing the current contention over the mapping between performance and workload in different task situations (Hancock and Meshkati 1988). Some researchers have reported a direct link between the two. For example, Warm *et al.* (1991) found that in sustained attention, characteristics of the task directly influenced perceived load so that as the level of objective task difficulty grew, workload increased and performance efficiency diminished accordingly (see also Becker *et al.* 1991). In contrast, others have reported conditions in which workload and performance 'dissociate' so that the direct link between task difficulty and workload level is broken (Derrick 1988, Eggemeier *et al.* 1982). This 'dissociation' question is one of both theoretical and practical concern for workload application. Yeh and Wickens (1988) have provided an approach that attempts to distinguish why and how dissociation occurs. Their framework is grounded strongly in attentional resource theory. They propose that dissociation occurs under a number of circumstances that include: first, when greater 'resources' are invested to improve a resource-limited task (Norman & Bobrow 1975); second, if demands on working memory are increased by time-sharing; and third, when performance is sensitive to some sub-task element while subjective measures are reflecting more global demands.

The present experiments address these issues by evaluating response to different task factors and the link between performance and workload over extended practice. In the first experiment the influence of three task-related variables was investigated. The first of these, tracking order, was chosen as an 'anchor' variable as it had already been demonstrated to exert a strong and consistent influence upon both performance and workload (Eberts and Schneider 1985, Jex 1988, Jex and Clement 1979). The manipulations of input device and augmented feedback were employed to examine whether changes in the physical and cognitive interface, respectively, modified established effects. A critical question was whether any observed effects were matched in both performance change and workload change together. The influence of input device is a practical concern since choice of device can affect interaction dependent upon task context (Card *et al.* 1978). Feedback manipulations are traditional methods of seeking performance improvement (Newell 1981) and a purpose here was to examine whether added feedback would prove useful or distracting to the operator.

What has not been pursued to any experimental depth are the concomitant effects on workload and performance of extended task practice. Reports of association or dissociation are related typically to brief performance periods upon relatively unpracticed tasks. The purpose here is to examine whether association or dissociation between workload and performance changes over time. It has frequently been observed that task practice improves performance efficiency, particularly under consistent mapping conditions (Schneider 1985, Schneider and Shiffrin 1977). It has also been asserted that with this performance improvement there is a reduction in the level of workload experienced. The foundation of this assumption is, in large part, an attentional resource rationale (Kahneman 1973, Navon and Gopher 1979, Wickens 1980, 1984, 1987) which links the growth of response automaticity to reduction in attentional resources utilized. With reduced attention demand, it is argued, perceived workload diminishes. While the experimental evidence concerning automaticity with extended practice is unequivocal, interpretations of such data are still the subject of contention (Logan 1988, Schneider and Detweiler 1988). Assertions concerning workload change are one facet of this argument that has yet to be thoroughly explored. The task practice experiment includes a manipulation that allows for the testing of retention of skills after a rest interval. What happens to workload associated with performance skills when an individual ceases to practice is not known. Therefore, the purpose of the reported experiments was to determine the influence of task characteristics on performance and workload to inform design of future interfaces that are posited to rely heavily upon such perceptual-motor capabilities.

2. Experiment 1

2.1. *Method*

2.1.1. *Participants and apparatus:* Six right-handed subjects (3 females, 3 males) ranging from 19 to 35 years of age volunteered to participate in the experiment. The subjects were members of the staff and student body of the University of Southern California and had either 20/20 or corrected 20/20 vision. None of the subjects was on medication at the time of testing. Subjects were informed on the general nature of the experiment for consent purposes, but were only provided with feedback on their performance following completion of all trials. The experimental task was presented on a Macintosh II computer and was generated by part of the SCORE performance assessment package (Hancock and Winge 1988). The tracking task was displayed on an Apple Color High Resolution Monitor and was contained in a $7{\cdot}5 \times 7{\cdot}5\,cm^2$ area. Control response was either via a trackball device (Kensington Turbo Mouse ADB, Version 3.0), or a (Macintosh) mouse. Responses on the NASA-TLX workload assessment technique were recorded on an IBM PC-XT, which subsequently analysed the collected ratings. The Subjective Workload Assessment Technique (SWAT) ratings were collected using a paper version of the test and were subsequently analysed using a purpose-developed program (Reid and Nygren 1988).

2.1.2. *Task and procedures:* The experimental task was two-dimensional pursuit tracking. The SCORE facility generated a random track and set tracking sensitivity at 0·8, and precise details of possible task manipulations are given in Hancock and Winge (1988). There were three orders of tracking control which were zero (position), first (velocity), and second (acceleration) order (Poulton 1974, Wickens 1986). There were two conditions of feedback. Using the control condition, participants were

aware only of the cursor position with respect to the target track. In the augmented feedback condition, an additional display was given under the track that showed a time history of integrated error with respect to the target track together with a region of acceptable performance. Thus there were 3 (Track Orders) by 2 (Input Devices) by 2 (Feedback Conditions), giving a total of twelve different experimental conditions. Each participant completed all twelve conditions in an order selected by random lot. Each trial was 120 s in duration. The dependent performance variable was RMS error (rmse), which was calculated from the displayed target path.

Mental workload was assessed using the NASA Task Load Index [TLX] (Hart and Staveland 1988) and the Subjective Workload Assessment Technique [SWAT] (Reid and Nygren 1988). In the SWAT procedure the first step was scale generation. The three identified dimensions of workload; time, effort, and stress, each possesses three levels; low, medium, and high. These are each described by brief statements. The complete combination of these statements provide 27 total descriptions of workload, which are each printed on a separate card. The participants sorted these cards in the order from the lowest to the highest level of their own perceived workload. This order is then used to derive a workload scale from 0–100. An individual event, which in the present experiment was a single trial, is scored when the participant responds with either a 1, 2, or 3, on each of the time, effort, and stress dimensions. The combination specifies the SWAT workload value according to the initial scale generation (Reid and Nygren 1988).

The NASA Task-Load Index (TLX) is also composed of a two-step procedure (Hart and Staveland 1988). In the first step, participants make pairwise comparisons of six sources of workload which are mental demand, physical demand, temporal demand, frustration, effort, and self-rated performance. This pairwise comparison of the sources generates a 'weighting' for each source, where the highest possible weighting is 5 and the lowest possible weighting is 0. The total number of weights add up to 15, which is the number of pairwise comparisons. Following each trial, participants rate their perceived workload on each of these individual sources on a scale from 0 to 100. These scores represent the raw ratings. The overall workload score for the condition is obtained by multiplying raw ratings by source weights and dividing the sum by 15 (the number of pairwise comparisons). In the present experiment, the SWAT card sort and the TLX pairwise comparisons were completed prior to the start of the experiment. Participants gave raw SWAT and TLX ratings following each trial. The analysis of tracking and workload response is presented below.

2.2. *Results*

2.2.1. *Performance measures:* Analysis of variance on tracking performance capability indicated a significant main effect of track order on RMS error ($F(2, 10) = 430{\cdot}9$, $p < 0{\cdot}001$). Each pairwise comparison between zero, first, and second order, showed reliable differences where RMS error (rmse) increased with track order. This is not an unexpected pattern and it has been observed by many others who have previously evaluated tracking ability (Poulton 1974). There was also a significant main effect for trial in which rmse decreased sequentially across the 12 imposed conditions, indicating learning irrespective of the order in which other conditions were administered. Analysis of tracking performance also indicated a significant main effect for input device ($F(1, 5) = 53{\cdot}3$, $p < 0{\cdot}001$). Performance with the trackball was clearly superior to that with the mouse. There was no main effect for augmented feedback

alone. This result confirms the complexity and nuances of such additional aiding and indicates a need for further clarification of our understanding of knowledge of performance on concurrent perceptual-motor tasks (Newell 1981). In addition to the significant main effects, there were two significant interactions in performance and both involved the input device. There was a significant interaction between track order and input device ($F(2, 10) = 18{\cdot}67$, $p < 0{\cdot}0001$). At the zero order level of tracking, rmse was essentially equal for the two devices. RMSE was greater for the mouse than the trackball with the first order control condition and this difference was even greater with the second order control. There was also an interaction between input device and the presence or absence of augmented knowledge of performance, ($F(1, 5) = 8{\cdot}75$, $p < 0{\cdot}01$). The presence of augmented knowledge of performance (KP) facilitated capability with the trackball. However, knowledge of performance depressed tracking efficiency when the mouse was used.

2.2.2. *Perceived workload*

(1) *SWAT* (*Subjective Workload Assessment Technique*): With respect to SWAT workload scores, there were significant main effects for trial and track order. As might be expected from the performance findings, there was a massive effect for track order on workload. Subjects reported higher SWAT scores as track order increased ($F(2, 10) = 88{\cdot}5$, $p < 0{\cdot}05$). As with the performance measure, analysis also indicated a significant effect for trial. Subjects reported lower SWAT scores as trials increased in the sequence. In addition to these main effects there were two interactions. The first of these replicated the effect seen in the performance data, namely the interaction between track order and input device ($F(2, 10) = 3{\cdot}32$, $p < 0{\cdot}05$). This interaction followed that for performance at the two lower track orders. Each device was equivalently successful at zero track order and rapidly diverged at the first order level where input via the mouse was perceived as more loading than that with the trackball. This divergence did not continue to second order control where the SWAT scores were equivalent. There may be a number of reasons for this reversal including a ceiling effect for the workload scale under consideration. There was also a three-way interaction between track order, input device and knowledge of performance for the SWAT scores.

(2) *NASA Task Load Index* (*TLX*): For the overall workload measure on the TLX, there were significant main effects for track order, trial, and input device. TLX workload scores increased with track order (i.e. zero order = 36·2, first order = 48·9, and second order = 70·0) and the pattern of increase followed that of both performance and SWAT scores. There was also a sequential decrease in load across trial and a significant effect for input device, where the trackball again resulted in the perception of significantly lower load than the mouse (Trackball = 51·9, Mouse = 56·1). In addition to the above main effects, there were significant interactions, which matched those derived from the performance and SWAT measures. Each included the influence of track order. For the interaction between track order and augmented knowledge of performance, the presence of augmented knowledge of performance had differential effects at the two lower track orders. While TLX overall workload was higher in the presence of augmented knowledge of performance at zero order, this trend was reversed at first order level. At the second order track level, the difference in TLX values disappeared and the presence of augmented knowledge of performance had no differential effect.

For the TLX subscale measures, subjects reported significantly higher mental demand, physical demand, temporal demand, effort, frustration, and a lower perceived performance as track order increased. Subjects perceived significantly lower mental demand, physical demand, and significantly higher own perceived performance when using the trackball compared to the mouse. The main effect of trial and input device configuration were significant only for the perceived performance and frustration scales. The interaction expressed in the subscales replicated those in actual performance and overall workload scales. Specifically, the mental demand scale exhibited the interaction between track order, input device, and the presence of augmented feedback. Temporal demand exhibited the two-way interactions between track order and input device, and track order and augmented feedback while effort showed the former two-way interaction and the three-way interaction between all three factors. The patterns for perceived performance and frustration scale were identical. They each showed both two-way and three-way interactions. In addition each exhibited a two-way interaction between augmented feedback and input device. These results directly follow the observations for the rmse interaction between the same factors.

2.3. *Discussion*

The overall summary of the present findings was a confirmation of the linkage between performance and workload. The increase in workload and decrement in performance with ascending order confirms the earlier results of Wempe and Baty (1968) who used secondary tasks as reflections of workload (Hancock and Meshkati 1988). Even in circumstances that yielded interactive effects, workload predominantly mapped with performance change. The finding of consistent changes in performance efficiency with track order is not unexpected. However, the fact that response on both subjective workload scales followed these changes in pattern and degree is direct confirmatory evidence of association. The sensitivity of the workload measures to the difference in the input devices also strengthens this conclusion.

3. Experiment 2: acquisition phase

The second experiment is based upon an observation derived from the first experiment. As noted, there was evidence of sequential reduction in both performance error and perceived workload across trials. However, there were insufficient trials to confirm this effect with confidence, especially since other significant factors were manipulated. To examine this parallel change in perceived workload and performance further, the tracking task was restricted to second order control only. Previous work has indicated that second order tracking in its early stages of performance is perceived as a highly demanding task (Eberts 1987, Eberts and Schneider 1985). Consequently, workload change could be monitored from an initially highly demanding situation across performance on a number of trials. Therefore, the specific purpose of the second experiment was to examine the relationship between subjective workload response and performance efficiency as each changed across trials. This experiment represents the skill acquisition phase and is contrasted with the skill retention phase, which is examined subsequently in experiment 3.

3.1. *Method*

3.1.1. *Participants and apparatus:* For this experiment, six different, right-handed

male volunteers served as subjects. They were members of the staff and student body of the University of Southern California and ranged in age between 18 and 35 years of age. Subjects had either 20/20 or corrected 20/20 vision, and no subject was on medication at the time of testing. As with the first experiment, the task in this procedure was presented on a Macintosh II and was generated by the SCORE program (Hancock and Winge 1988). In light of the results from the first experiment, the present procedure used only the trackball device and no augmented knowledge of performance was presented beside the intrinsic feedback of the position of the on-screen cursor and target track. With these exceptions the apparatus was exactly the same as that in the first experiment.

3.1.2. *Task and procedure:* The task was a one-dimensional compensatory tracking task with second-order control dynamics (Wickens 1986). Participants were given an auditory warning stimulus to signal the start of each trial. They were instructed to minimize track error throughout each trial, which lasted 120 s each. Tracking error for each trial was recorded and transferred to an analysis file after the completion of a 10-trial block. Participants took approximately 30 min to complete each block. There were 10 blocks in all, giving 100 total trials per subject. The analysis of change in tracking performance and subjective workload response over trials for both individual and group response are presented below. The workload assessment procedures used in the present experiment were the Subjective Workload Assessment Techniques (SWAT), and the NASA (TLX) as previously described. Scale generation and weighting comparisons were completed before performance began. Individual raw scores for each method were taken after the completion of each trial.

3.2. *Results*

The independent variable in the present experiment was performance trial. For the purpose of analysis, trials were condensed into 10 blocks of 10 trials each. Analysis of variance was performed on three dependent variables which were root mean square tracking error (rmse), overall SWAT and overall TLX scores.

3.2.1. *Performance measures:* Analysis of variance indicated a significant decrease in rmse across block, ($F(9, 45) = 11{\cdot}97$; $p < 0{\cdot}001$). As might be expected, there was a large initial improvement followed by progress toward an asymptote (Welford 1968). The data for the mean RMS error versus trial block are illustrated in figure 1. *Post hoc* analysis distinguished differences in a manner consistent with the illustrated data. Scheffe's test distinguished rmse in block 1 as significantly higher than rmse in blocks 4–10. RMSE in block 2 was significantly higher than rmse in either blocks 9 and 10. No other comparisons exhibited significant differences. These data confirm the well-established picture for practice and progressive improvement on such a performance task (Crossman 1959). In keeping with previous findings, variability also systematically decreased across blocks. In confirmation of the findings of the first experiment, there were large reductions in rmse following the first one or two trials of the overall sequence. In the absence of direct kinematic analysis, these changes were observationally related to the reduction of large-scale arm movements on behalf of the participant. This observation is consistent with comparable observations of others who have noted strategy changes with practice on second-order tracking (Goettl 1991).

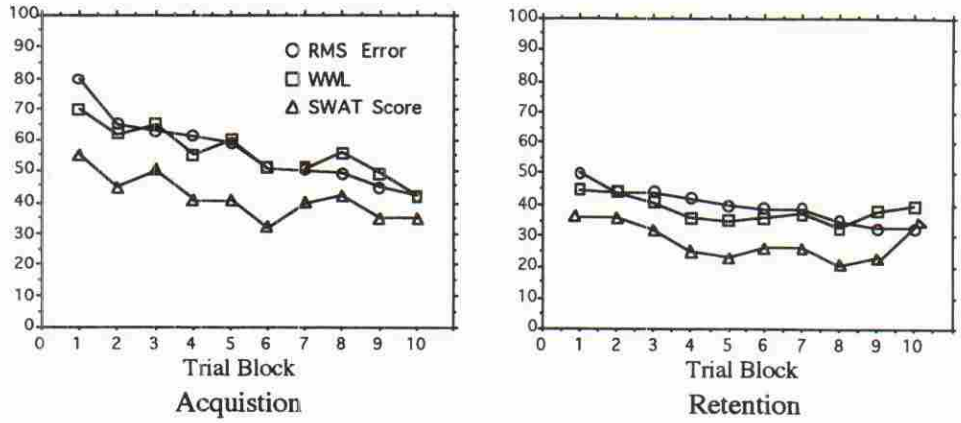

Figure 1. Change in root mean square error (RMSE), overall NASA TLX workload (WWL) and overall SWAT workload (SWAT) versus trial block. The section at left represents the acquisition phase, the section at right represents the retention phase.

3.2.2. *Perceived workload*

(1) *SWAT* (*Subjective Workload Assessment Technique*): Analysis of variance of the SWAT scores revealed a significant effect for trial block. ($F(9,45) = 8{\cdot}01$; $p < 0{\cdot}001$). In this case, *post hoc* analysis distinguished the perceived workload in block 1 as significantly higher than that in blocks 5–10. The only other significant difference was for block 3 in which workload was significantly higher than blocks 6 and 10 respectively. The means of SWAT scores versus trial block are also illustrated in figure 1. Analysis of the standard deviation of SWAT scores resulted in no significant effects across the trial blocks. So, while mean SWAT workload was reduced across trials, variability remained consistent.

(2) *NASA Task Load Index:* For the combined workload measure on the NASA Task Load Index (TLX) there was a significant effect for trial block ($F(9,45) = 11{\cdot}46$; $p < 0{\cdot}001$). As illustrated in figure 1, the mean TLX workload score was reduced across trial block. Again, *post hoc* analysis distinguished workload experienced in block 1 as significantly higher than that in blocks 6–10, respectively. There were similar differences between block 2 and blocks 9–10. Block 3 was significantly higher than blocks 7, 9 and 10, respectively. No other comparisons reached significance. As with the SWAT scores, standard deviation of the overall TLX score did not vary significantly with trial block (figure 1).

3.2.3. *Performance change versus workload change:* A major purpose of the present experiment was to compare the change of perceived workload with change in performance. To assess this relationship, correlations were calculated between the rmse value for each individual trial for each participant and their subjective workload score on that same trial. For the SWAT measures these correlations ranged between (0·13 and 0·65) and for the TLX these correlations ranged between (0·25 and 0·82). With the exception of one subject, all correlations were positive indicating diminishing workload with performance improvement. There were dramatic improvements in performance in the first few trials for each subject. These initial and substantive changes are not apparent in the workload data for either SWAT or TLX.

There is a tendency to score experienced workload as slightly higher on the TLX scale compared to the SWAT scale. This is in part due to the construction of the respective scales. Subjects respond with one of 27 combinations on the SWAT scale. This response is subsequently rescaled into values between 0 and 100. One of these combinations is rated at zero and subjects can produce this score of zero workload on multiple occasions. For the TLX, there is a tendency to rate at least one or more scale above the zero value. In the present experiment, the lowest overall TLX rating was 13. However, several subjects reported zero load on the SWAT scale on several occasions by replying, for example, with a 1, 1, 1, on the time, effort, and stress dimensions. This difference in the lower bound is probably responsible for the absolute difference in SWAT and TLX scores. The iterative jumps on the SWAT scale represents the tendency of subjects to respond in similar categories, which in the present case provide the five bands around zero, 24, 45, 65, and 83 respectively. While the scales exhibit strong qualitative similarities, deducting a floor value of 15 from the TLX scores provides strong quantitative similarities also.

3.2.4. *Performance variability and workload variability:* Separate analyses were performed on the standard deviation of the performance scores and the variability of response on each workload scale. Variability in rmse was sequentially and significantly reduced with trial block. However, there were no equivalent changes in the variability of workload. With respect to the absolute level of variability between the two workload scales, analysis indicated ($t(9) = 3{\cdot}86$; $p < 0{\cdot}005$) that there is a significantly lower variability in TLX scores compared to SWAT scores, a difference consistent with previous findings (Vidulich and Tsang 1986).

3.3. *Discussion*

Performance change in the present acquisition experiment followed a classic learning curve. While no augmenting aids were presented, participants were still able to assimilate the regularity of the dynamics and, upon this basis presumably, improved their tracking performance over trials (Eberts 1987). Both subjective workload assessment methods elicited mean responses that followed the pattern for performance change. The only failure for workload to track performance was in the first one to two trials of the first performance block. It has been suggested (Goettl 1991) that this might reflect a change in tracking strategy on behalf of the subject where one strategy, for example 'bang-bang' control, was substituted for another (Wickens 1986). This suggestion is supported by two observations. First, dissociation was seen in the first trials of this experiment in which the task was new to the performers. However, in a subsequent experiment described below such dissociation did not occur. Presumably this was because having once 'discovered' a preferred strategy, there is no need to 're-discover' that strategy at some later point in time. The second observation was made by experimenters who observed a behavioural change between the first trials where rapid, and sometimes violent, arm motions predominated and subsequent trials where much 'quieter' activity occurred featuring mostly wrist and finger control of the trackball. Exactly how strategy changed in terms of the kinematics of limb response clearly needs further clarification from those more directly involved in motor control. The failure to find a comparable decrease in variability between performance and workload is of interest. Like the absolute difference between the workload scales themselves, the incongruity of variability

might be related to the nature of the construction of the workload scales and thus represent a measurement artefact. However, this may also reflect a veridical tendency, the ramifications of which are discussed in respect of the retention findings.

4. Experiment 2: retention phase

While skill acquisition is an important phase of development with respect to any human-machine interface, an equally important aspect is the retention of those skills over time. This is particularly important in circumstances where an individual cannot practice such skills in the intervening period. Such periods when practice is not possible occur all the time with events such as vacations or temporary duties in military operations. Consequently, a purpose of the retention experiment was to evaluate how well individuals maintained their performance skills after a month of rest.

4.1. *Method*

In the retention experiment, the same subjects returned after a break of 30 days. The task, apparatus, and procedure were identical to that of the previous experiment. Tracking performance and subjective workload were also assessed according to the regimen as specified in the acquisition phase.

4.2. *Results*

4.2.1. *Performance measures:* The participants in the retention experiment continued to improve across the subsequent performance trials. The data for means for this retention experiment are also given in figure 1. As with the results of the acquisition phase, there was a significant effect for trial block ($F(9,45) = 7{\cdot}58$; $p < 0.0001$), and *post hoc* analysis again distinguished block 1 as significantly higher than blocks 6–10. The actual range of rmse values is much smaller for this retention phase and the value of rmse for the first retention block is close to the value for the last block of the acquisition phase.

4.2.2. *Perceived workload:* For the overall SWAT score, there was a significant effect for trial block during the retention phase ($F(9,45) = 3{\cdot}81$, $p < 0{\cdot}0001$). While the SWAT score did not decrease significantly in variability, the mean value did drop across sequential trial blocks. For the NASA TLX scores, trial block also exhibited a significant influence ($F(9,45) = 3{\cdot}07$, $p < 0{\cdot}002$). As was the case for the acquisition phase, variability in the NASA TLX score did not decrease with trial block although mean value was significantly reduced. The significant difference between the absolute level of variability exhibited between each respective scale in the acquisition phase persisted in this retention phase ($t(9) = 11{\cdot}34$; $p < 0{\cdot}0001$). The absolute difference between the scores on each scale for mean response also persisted in this phase of performance.

4.3. *Discussion*

The findings here confirmed those in the previous acquisition experiment. That is, mean workload was reduced as mean performance improved. In this phase, there was no divergence in initial trials between performance and workload. Also, performance

variability sequentially decreased while workload variability remained unchanged on both scales. In the previous discussion it was suggested that this might result in an artefact of the characteristics of the respective assessment scales. However, it may also be a veridical finding. If so, the finding that variability of perceived workload does not track performance variability has important ramifications for the performance-workload linkage on an event-by-event, that is, trial-by-trial basis. As is highlighted below, the time-frame in which respective measurements are taken may very well be a critical element in the association-discussion picture (Yeh and Wickens 1988).

5. Overall discussion and conclusions

5.1. *Workload dissociation*

The present experiments examined the linkage between subjective workload and performance. Pragmatically, if workload response always followed performance variation, then there would be little reason to collect such additional measures. They might provide *post hoc* information, as do Cooper-Harper ratings of aircraft handling qualities (Cooper and Harper 1969); however, they would be of little practical benefit in developing, for example, adaptive human-machine systems since no additional information would be available (Hancock and Chignell 1987, Rouse 1988). As a result, how workload links with performance and the conditions in which workload and performance dissociate are of distinct importance. Figure 2 shows possible combinations that link workload and performance. Across one diagonal are the direct associations, shown by the diagonal cross (X) symbols. In these conditions, better performance is associated with lower workload, poorer performance is associated with higher workload and no change in performance is associated with no change in workload. An example of first form of this association is found in the results for the second experiment reported here.

The illustration also shows conditions in which performance remains constant but workload changes. These circles (○) represent potentially important diagnostic conditions. For example, where performance is stable but workload increases one would suggest that a trade-off is occurring such that the stability is maintained only at the expense of greater effort (Nelson *et al.* 1991). In the obverse condition, performance remains the same while workload decreases. Here, it might be suggested

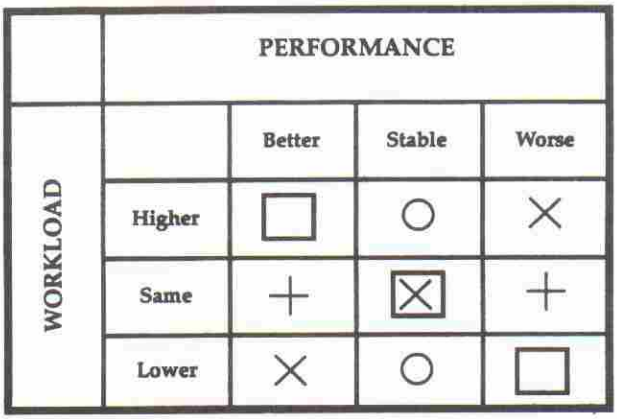

Figure 2. Matrix of performance and workload associations and dissociations.

that the development of task-related skill allows the performer to reduce load while maintaining a constant performance level. Conditions indicated by the vertical crosses (+) are characterized by an insensitivity of the performer to their own output. So, while performance is either increasing or decreasing, workload stays at the same level. These are typically the circumstances that have been described as dissociation (Yeh and Wickens 1988). They may arise from the insensitivity of the performer, but we cannot ignore the possibility that the measurement scales themselves may be limited by artefacts such as floor and ceiling effects.

The square symbols represent direct divergence. In these conditions, workload is increasing as performance improves or is decreasing as performance gets worse. The latter could come from conditions in which the subject is 'giving up' (Hancock 1989). As has been indicated earlier, task-related mental workload can only occur in conditions in which the individuals *believe* that they are able to achieve the goal of the task under consideration (Hancock and Chignell 1988, see also Hancock and Caird 1993). During incipient failure, it may well be that an operator perceives that they can no longer achieve what is being asked and report diminishing workload even as performance grows progressively worse. The obverse situation might well occur in a preceding interval in which difficulty is growing and the operator is responding successfully so that performance is enhanced but only at the cost of considerable workload increase. Whether the latter condition is considered 'dissociation' depends upon what theoretical foundation is used as a rationale for workload in general. In the present experiments, results showed a high degree of concordance. For example, performance was poorer with the mouse compared to the trackball and workload was correspondingly higher with the mouse than the trackball. Performance largely did not change with the presence or absence of augmented knowledge of performance and workload was also not altered by this manipulation alone.

One of the main questions addressed in the present experiments was the relationship between performance and workload under the influence of differing factors. Warm and his colleagues have reported consistent associations between workload and performance (Warm *et al.* 1991). For example, when the event rate of a task was manipulated, workload changed so that higher workload was associated with more frequent events. This means a direct correspondence between psychophysical factors and perceived workload. Their experiments have focused particularly upon conditions that are characterized as vigilance tasks that require the sustenance of attention over a prolonged period. Typically, in such circumstances the individual is required to monitor a display of repeated non-signals for the irregular appearance of critical 'signal' events. It has been shown that in such performance tasks, which are of growing importance as system operator turns system monitor (Lee and Moray 1992, Singleton 1989), workload is mapped directly to the psychophysical characteristics of the task. In contrast, Wickens and his colleagues have reported experiments in which workload 'dissociates' from performance. Under such circumstances, changes in the difficulty of the task no longer yield compatible influences on workload response, as measured by subjective report.

One critical point concerns the time-scale of events and the time-scale of measurement. In the present series of experiments, performance via root mean square tracking error is measured several times per second. The global measure of performance ability is the summation of these momentary errors over a 2-min trial. Subjective workload is taken at the end of the trial and is also, putatively designed to measure the effects of the whole trial. However, equivalence in these terms means that the memory of the

performer accumulates workload units in the same simple additive way that the computer sums the root mean square error. This is problematic for a number of reasons. First, the frequency of collection of root mean square error is a parameter decided in software design. There is no rationale to equate this with either the qualitative or quantitative way in which human observers treat time (Hancock 1993). Second, from all that we know about human memory (e.g. recency, primacy, and salient feature effects, see Lachman *et al.* 1979), simple accumulation of events is an untenable model. One might therefore propose that tasks which present rare events, which put relatively small demand on memory, would be more likely to produce direct associations between performance and perceived workload. In contrast, tasks with many events, such as continuous control or multiple task environments, would be more likely to result in dissociation. This proposal seems to be supported by contrasting existing findings (Warm *et al.* 1991, Yeh and Wickens 1988). The critical phenomenon to distinguish is discordance between objective and subjective reflections of performance compared with discordance of temporal nature of the methods of assessment. Until some clear method of equilibrating these intrinsic and extrinsic time scales is developed, the influence of 'windowing' different facets of response is liable to remain a source of variation between workload and performance.

5.2. *Automated skills, performance and workload*

It has been postulated that automaticity is one critical characteristic of the development of high performance skills (Schneider 1985). A necessary condition for the creation of automated response is the consistent mapping between stimulus and response. Briefly, what this means is that the individual must react to the stimulus in the same way each time it is presented and not in a different manner depending upon conditions. A complete description of this linkage together with the theoretical and practical ramifications of this conception have been the subject of extensive research (Kirlik *et al.* 1994, Shiffrin and Schneider 1977). Eberts and Schneider (1980, 1985) have indicated that automation of second-order tracking is problematic since the moment-to-moment kinematics of the input device are not consistently mapped to the target. They have sought ways to present augmenting cues to promote the assimilation of the higher-order consistency as represented by the deterministic equation describing the dynamics. While participants have difficulty discovering this consistency through perusal of instantaneous response, certain forms of graphic cue can help with skill development. Therefore, the improvement shown in the acquisition and retention experiments can still be accounted for by the automatic and controlled processing theory (Schneider and Shiffrin 1977) and the theoretical link between workload and attention is preserved. The application of this approach to higher-order consistencies in a number of realms has been explored and articulated by Fisk *et al.* (1988). The present evidence therefore, provides indirect support for the automated attention account of skill development. Indeed, on a global level, some general form of the consistency must be true, since the very basis of any perception-action system must rely on some form of regularity in any environment. In addition, at a practical level, the equation generating the track itself is deterministic.

One possible reason for the transitional dissociation observed between performance and perceived workload in the first two trials here, concerns the possibility of a subject strategy change. Upon interview at the termination of the experiment, subjects commented that it took one or two trials to understand the second-order response characteristics. Specifically, subjects initially attempted to minimize error by large

excursions of the trackball using rapid movements of the entire arm. However, after only the initial trials, it was found that accurate control over the cursor could be achieved with small finger movements alone. For most subjects, transition from the arm to finger control strategy was not gradual but abrupt. Evidence for this effect, although consistent with the observation of others (Goettl 1991), is tentative, deriving from observation on behalf of the experimenter and subjective response on behalf of the participant. One potential reason, then, for this transitory dissociation may have resulted from the inability of the workload measures to capture this strategic transition. Embedded in such an account is the confound of anchoring, where initial workload response from any individual subject serves as a reference to all their subsequent trials, but such initial trials are not preceded by any experience on the task. This tendency, together with an adversity to scoring at the extremes of the workload scales, serves to render early workload responses inherently unstable. This may occur whether or not there is a change in performance strategy as suggested here. Consequently, one reason for potential dissociation between subjective workload and performance is the fundamental nature of the measures used to assess each.

Upon retention, after a month without any practice, the subjects continued to improve with a starting level that was close to the point at which they have collectively ceased. As might be expected there were large individual differences between subjects. In general, the workload scales continued to reflect the improvement in the mean level of performance. However, one clear trend in the performance data, i.e. the reduction of variability with sequential trials were clearly not matched by either of the workload scales. In general these findings support the contention that workload is reduced as performance level increases. However, there are a number of important cautions that must be emphasized before such evidence is used for unquestioned practical application.

5.3. *Interface implications*

Two forces were implicated at the beginning of this work as critical in influencing the development of interfaces of the future. The first was the ubiquity of the computer. While complex technical systems obviously feature computer control or computer-mediated control, it is not emphasized as often that simple everyday technologies, such as cars and washing machines, also feature this control characteristic. The presence of computers does not mandate convergent interface evolution and could even be cited as force for diversification. However, the economy and compatibility of a common user interface has a strong draw in a marketplace redolent with positive feedback (Lewin 1992). What is, perhaps, the ultimate stimulus is the intrinsic human ability to operate in complex spatio-temporal worlds. The interface has been seen to evolve from coded alpha-numerics on a two-dimensional display, through two-dimensional icons, toward wrap-around virtual environments. That such interfaces take progressive advantage of operator perception-action abilities is not happenstance. Navigating in these worlds is a tracking task. Although the designer might rarely choose second-order dynamics for such tracking, as evidenced by relatively poor performance with these dynamics in experiment 1, the designer might not have this choice. If the virtual interface is slaved to real-world systems, the inclusion of such response dynamics may be unavoidable. How to facilitate performance of systems such by manipulating the physical characteristics of the interface, the feedback provided, or the extent that practice influences efficiency and workload promise to be central ergonomic issues in these evolving interfaces.

Acknowledgements

This research was supported by Grant NCC 2-379, from the National Aeronautics and Space Administration (NASA), Ames Research Center. Sandy Hart was the technical monitor for the grant. The views expressed are those of the author and do not necessarily represent those of the named agency. The author would like to thank Mr. G. B. Reid for help with the derivation and interpretation of the SWAT scores given in this experiment and Marie Robinson and Andy Chu for help with data collection. I am grateful to my colleague John Flach, whose comments helped to significantly improve my understanding of the dynamics involved in the present work and Max Vercruyssen, Olu Olofinboba, and Erik Arthur for their help with statistical analysis. I am particularly grateful for the additional comments of other reviewers that have helped in the preparation of this paper. Part of this work was presented at the 33rd Annual Meeting of the Human Factors Society and at the International Symposium on Aviation Psychology. The work is dedicated to my former student Darren Hansen who participated in data collection for this work and whose untimely death saddened us all.

References

BECKER, A. B., WARM, J. S., DEMBER, W. N. and HANCOCK, P. A. 1991, Effects of feedback on perceived workload in vigilance performance, *Proceedings of the Human Factors Society*, **35** (Human Factors Society, Santa Monica, CA), 1491–1494.

BENNETT, K.B. and FLACH, J. M. 1992, Graphical displays: implications for divided attention, focused attention, and problem solving, *Human Factors*, **34**, 513–533.

CARD, S. K., ENGLISH, W. K. and BURR, B. J. 1978, Evaluation of mouse, rate-controlled isometric joystick, step keys, and text keys for text selection on a CRT, *Ergonomics*, **21**, 601–613.

COOPER, G. E. and HARPER, R. P. 1969, The use of pilot rating in the evaluation of aircraft handling qualities, NASA-TN-D-5153, National Aeronautics and Space Administration, Washington, DC.

CROSSMAN, E. R. F. W. 1959, A theory of the acquisition of speed skill, *Ergonomics*, **2**, 153–166.

DERRICK, W. L. 1988, Dimensions of operator workload, *Human Factors*, **30**, 95–110.

EBERTS, R. E. 1987, Internal models, tracking strategies, and dual task performance, *Human Factors*, **29**, 407–419.

EBERTS, R. E. and SCHNEIDER, W. 1980, Computer assisted displays enabling internalization and reduction of operator workload in higher order systems, or, pushing the barrier of human control beyond second order systems, *Proceedings of the Human Factors Society*, **24** (Human Factors Society, Santa Monica, CA), 59–62.

EBERTS, R. E. and SCHNEIDER, W. 1985, Internalizing the system dynamics for a second-order system, *Human Factors*, **27**, 371–393.

EGGEMEIER, F. T., CRABTREE, M. S. and REID, G. B. 1982, Subjective workload assessment in a memory update task, *Proceedings of the Human Factors Society*, **26** (Human Factors Society, Santa Monica, CA), 643–647.

FISK, A. D., ORANSKY, N. A. and SKEDSVOLD, P. R. 1988, Examination of the role of 'higher-order' consistency in skill development, *Human Factors*, **30**, 567–581.

GOETTL, B. P. 1991, Tracking strategies and cognitive demands. *Human Factors*, **33**, 169–183.

HANCOCK, P. A. 1989, The effect of performance failure and task demand on the perception of mental workload, *Applied Ergonomics*, **20**, 197–205.

HANCOCK, P. A. 1993, Body temperature influence on time perception, *Journal of General Psychology*, **120**, 197–215.

HANCOCK, P. A. 1995, On convergent technological evolution, *Ergonomics in Design*, **4**, 22–29

HANCOCK, P. A. and CAIRD, J. K. 1993, Experimental evaluation of a model of mental workload, *Human Factors*, **35**, 413–429.

HANCOCK, P. A. and CHIGNELL, M. H. 1987, Adaptive control in human-machine systems, in P. A. Hancock (ed.), *Human Factors Psychology* (North-Holland, Amsterdam), 305–354.

HANCOCK, P. A. and CHIGNELL, M. H. 1988, Mental workload dynamics in adaptive interface design, *IEEE Transactions on Systems, Man, and Cybernetics*, **18**, 647–658.

HANCOCK, P. A. and CHIGNELL, M. H. 1995, On human factors, in J. M. Flach, P. A. Hancock, J. K. Caird and J. K. Vicente (eds), *Global Approaches to the Ecology of Human-Machine Systems* (Lawrence Erlbaum, Hillsdale, NJ), 14–51.

HANCOCK, P. A. and MESHKATI, N. (eds), 1988, *Human Mental Workload* (Elsevier, Amsterdam).

HANCOCK, P. A. and WINGE, B. 1988, Strategic control of response efficiency (SCORE). An integrated dynamic multi-task simulation program for the assessment of human operator response behavior. Technical Report, TRL-NASA-8804, University of Southern California, Los Angeles, CA.

HANSEN, J. P. 1995, Representation of system invariants by optical invariants in configural displays for process control, in P. A. Hancock, J. M. Flach, J. K. Caird and K. Vicente (eds), *Local Applications in the Ecology of Human-Machine Systems* (Lawrence Erlbaum, Hillsdale, NJ).

HART, S. G. and STAVELAND, L. E. 1988, Development of the NASA-Task Load Index (NASA-TLX); Results of empirical and theoretical research, in P. A. Hancock and N. Meshkati (eds), *Human Mental Workload* (Elsevier, Amsterdam), 139–183.

HARWOOD, K., SMITH, K., OLSEN, W. and HANCOCK, P. A. 1994, Shared decision making in the national airspace system: flightdeck-ATC integration, in *Proceedings of the Applied Science Symposium*, US Air Force Academy, Colorado Springs, CO, April.

JEX, H. R. 1988, Measuring mental workload: problems, progress, and promises, in P. A. Hancock and N. Meshkati (eds), *Human Mental Workload* (North-Holland, Amsterdam), 5–39.

JEX, H. R. and CLEMENT, W. F. 1979, Defining and measuring perceptual-motor workload in manual control tasks, in N. Moray (ed.), *Mental Workload: Its theory and Measurement* (Plenum Press, New York), 125–177.

KAHNEMAN, D. 1973, *Attention and Effort* (Prentice-Hall, Englewood Cliffs, NJ).

KANTOWITZ, B. H. and CASPER, P. A. 1988, Human mental workload in aviation, in E. L. Wiener and D. C. Nagel (eds), *Human Factors in Aviation* (Academic Press, San Diego), 157–187.

KIRLIK, A., WALKER, N. and FISK, A. D. 1994, Supporting perception in the service of dynamic decision making. Manuscript in review.

KOZAK, J. J., HANCOCK, P. A., CHRYSLER, S. and ARTHUR, E. 1993, Transfer of training from virtual reality, *Ergonomics*, **36**, 777–784.

LACHMAN, R., LACHMAN, J. L. and BUTTERFIELD, E. C. 1979, *Cognitive Psychology and Information Processing* (Lawrence Erlbaum, Hillsdale, NJ).

LEE, J. and MORAY, N. 1992, Trust control strategies and allocation of function in human-machine systems, *Ergonomics*, **35**, 1243–1270.

LEWIN, R. 1992, *Complexity: Life at the Edge of Chaos* (Macmillan, New York).

LOGAN, G. D. 1988, Automaticity, resources, and memory: theoretical controversies and practical implications, *Human Factors*, **30**, 583–598.

NAVON, D. and GOPHER, D. 1979, On the economy of the human processing system, *Psychological Review*, **84**, 214–255.

NELSON, W. T., Warm, J. S., Dember, W. N. and Parasuraman, R. 1991, Vigilance for real and subjective contours. Paper presented at the Psychonomic Society, San Francisco, CA.

NEWELL, K. M. 1981, Skill learning, in D. H. Holding (ed.), *Human Skills* (Wiley, New York).

NORMAN, D. A. and BOBROW, D. 1975, On data-limited and resource-limited processes, *Cognitive Psychology*, **7**, 44–64.

POULTON, E. C. 1974, *Tracking Skill and Manual Control* (Academic Press, London).

RASMUSSEN, J. and PJETERSEN, A. M. 1995, Virtual ecology of work, in J. Flach, P. A. Hancock, J. Caird and K. Vicente (eds), *Global Perspectives on the Ecology of Human-Machine Systems* (Lawrence Erlbaum, New Jersey).

REID, G. B. and NYGREN, T. E. 1988, The subjective workload assessment technique: a scaling procedure of measuring mental workload, in P. A. Hancock and N. Meshkati (eds), *Human Mental Workload* (Elsevier, Amsterdam), 185–218.

ROUSE, W. B. 1988, Adaptive aiding for human/computer control, *Human Factors*, **30**, 431–443.

SCHNEIDER, W. 1985, Training high-performance skills: fallacies and guidelines, *Human Factors*, **27**, 285–300.

SCHNEIDER, W. and DETWEILER, M. 1988, The role of practice in dual-task performance: toward workload modelling in a connectionist/control architecture, *Human Factors*, **30**, 539–566.

SCHNEIDER, W. A. and SHIFFRIN, R. M. 1977, Controlled and automatic information processing. I. Detection, search, and attention, *Psychological Review*, **84**, 1–66.

SHIFFRIN, R. M. and SCHNEIDER, W. 1977, Controlled and automatic human information processing. II. Perceptual learning automatic attending and a general theory. *Psychological Review*, **87**, 127–190.

SCHNEIDERMAN, B. 1983, Direct manipulation: a step beyond programming languages, *IEEE Computer*, **16**, 57–69.

SINGLETON, W. T. 1989, *The Mind at Work: Psychological Ergonomics* (Cambridge University Press, Cambridge).

SMITH, K. and HANCOCK, P. A. 1995, Situation awareness is adaptive, externally-directed consciousness, *Human Factors*, **37**, 137–148.

VICENTE, K. J. and RASMUSSEN, J. 1990, Ecology of human-machine systems. II: Mediating 'direct perception' in complex work domains, *Ecological Psychology*, **2**, 207–249.

VIDULICH, M. A. and TSANG, P. S. 1986, Techniques of subjective workload assessment: a comparison of SWAT and the NASA-Bipolar methods, *Ergonomics*, **29**, 1385–1398.

WARM, J. S., DEMBER, W. N., GLUCKMAN, J. P. and HANCOCK, P. A. 1991, Vigilance and workload, *Proceedings of the Human Factors Society*, **35** (Human Factors Society, Santa Monica, CA), 980–981.

WELFORD, A. T. 1968, *Fundamentals of Skill* (Methuen, London).

WEMPE, T. E. and BATY, D. L. 1968, Human information processing rates during certain multiaxis tracking tasks with a concurrent auditory task, *IEEE Transactions on Man-Machine Systems*, **MMS-9**, 129–138.

WICKENS, C. D. 1980, The structure of attentional resources, in R. Nickerson (ed.), *Attention and Performance, VIII* (Lawrence Erlbaum, Hillsdale, NJ), 63–102.

WICKENS, C. D. 1984, Processing resources in attention, in R. Parasurman and D. R. Davies (eds), *Varieties of Attention* (Academic Press, New York).

WICKENS, C. D. 1986, The effects of control dynamics on performance, in K. R. Boff, L. Kaufman and J. P. Thomas (eds), *Handbook of Perception and Human Performance*, vol. 39 (Wiley, New York), 1–60.

WICKENS, C. D. 1987, Attention, in P. A. Hancock (ed.), *Human Factors Psychology* (North-Holland, Amsterdam).

YEH, Y. Y. and WICKENS, C. D. 1988, Dissociation of performance and subjective measures of workload, *Human Factors*, **30**, 111–120.

Hancock, P.A. (1996). Effect of control order, augmented feedback, input device and practice on tracking performance and perceived workload. Ergonomics, 39, 1146-1162. by Electronic Intelligence, England.