COMPARISON OF SUBJECTIVE WORKLOAD RATINGS AND PERFORMANCE MEASURES OF A REFERENCE IVIS TASK

Tony Wynn (ESRI, United Kingdom),
John, H. Richardson (ESRI, United Kingdom)

T.wynn@lboro.ac.uk

ABSTRACT: The aim of this study was to establish subjective workload ratings (NASA-TLX) for the ADAM surrogate reference task (SURT) and compare them to ratings of a test battery of real IVIS tasks while participants performed a surrogate of the driving task (Lane Change Task). The results indicated that subjective workload ratings were comparable for both the real and reference tasks, performance measures were significantly correlated with measures of subjective workload and that performance of the secondary reference task declined as task difficulty increased. Reasons for the pattern of results are discussed.

1 Introduction

Traditionally car drivers have operated a number of 'additional' devices in the driving environment including; radios, heating, ventilation and air conditioning systems. Quite often these systems are controlled in addition to the primary task of driving. Technological advances have facilitated the development more sophisticated driver information systems intended to make travel more efficient and less arduous [1]. The major concern with the introduction of new IVIS is that they may introduce driver distraction sufficient to interrupt the primary driving task [2].

Workload is a hypothetical construct [3] that represents the cost incurred by a human operator in achieving a particular level of performance. It can be viewed as the proportion of resources required to meet task demands. If the demands of the task exceed the resources available to the skilled operator performance will inevitably decline [4, 5]. Subjective workload ratings are the most common method of obtaining estimates of the workload associated with a variety of tasks and are an important consideration in the evaluation of performance. If operators consider the workload of a task to be excessive they may behave as though they are overloaded, even though the task demands are objectively low. In such cases participants may report that the task was hard, but then perform just as well as if undertaking a low workload task [6, 7].

A central theme in the development of IVIS is the evaluation of the user's ability to complete two or more tasks concurrently [8]. The primary aim of any evaluation of IVIS is to ascertain whether drivers are able to perform the system tasks in the dynamic and sometimes difficult circumstances that can be expected during performance of the driving task. For these tests the primary task does not necessarily have to be driving a car. It is only necessary to present a primary task which demands an equivalent level of continuous

attention from the participant while operating a secondary task (the interface). The dual task approach purposely divides the attention of the subject in order to define the limits of their processing capabilities.

The rationale behind the use of reference tasks is that an IVIS device will produce a particular pattern of behaviour and if a reference task produces the same pattern of behaviour we can conclude that the reference is a suitable surrogate. Reference tasks are intentionally generic in nature allowing the impact of IVIS tasks to be assessed in a context independent fashion. A good secondary reference task should be analogous to IVIS devices, complex enough to illicit the same responses as a real IVIS yet, simple enough as to be analogous to all IVIS. The 'trade off' in using reference tasks is between the loss of ecological validity (real tasks), compared with the advantages of being able to compare the results of studies more readily (standard tasks).

A robust secondary reference task was developed as part of the ADAM project. The visual search task requires participants to report whether or not a pre-specified target is embedded in a multi-item display. In this instance the specified target is a circle that is distinguishable from other items (also circles) within the display by size. Non-target items are designed to act as distracters.

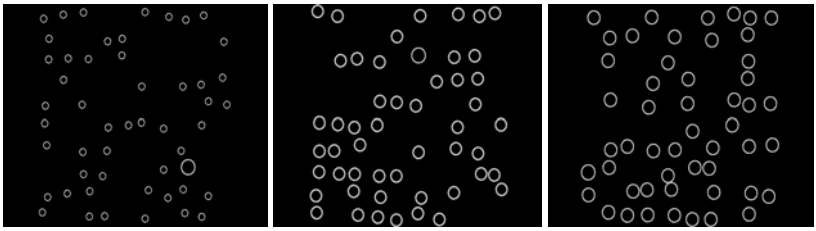

Fig.1. Screen shot of the surrogate reference task (SURT).

Figure 1 illustrates how visual complexity of the task can be manipulated by altering the size of the target, the thickness of the line and the number of distracters present. The size of the target is always 44 arc minutes. In the left image distracters are 22 arc minutes in which the target is easy to find, in the centre image distracters are 35 arc minutes finding the target is moderately difficult and in the right hand image the distracter size is 40 arc minutes, in this instance the target is very difficult to distinguish. Task difficulty can therefore be adjusted to match the difficulty (in terms of visual search) of a wide range of real IVIS tasks.

The key aim of the current study was establish the subjective workload ratings for an established reference task and a test battery of real tasks and then consider these ratings with respect to the participant's performance on the Lane Change Task. This would then establish the utility of reference tasks as a suitable surrogate for real tasks in the investigation of performance.

2 Method

2.1 Participants

16 participants (2 Female, 14 Male) were selected at random from staff and students at Loughborough University. Participants were required to have a full United Kingdom driving licence and normal or corrected vision. They had an average age of 35 years and 6 months and had held a driving license for an average of 15 years and 7 months, self report estimate of annual mileage was an average of 10,464miles.

2.2 Design

A within subjects 'repeated measures' design was used wherein each subject completed each of the conditions. The trial design was counterbalanced so that learning effects could be controlled for in the statistical analysis.

2.3 Lane Change Task

The Lane change Task is a laboratory control and event detection metric based on the dual task paradigm. The dual task paradigm posits that primary task performance will degrade with the introduction of a secondary task. In this case LCT performance can be viewed as the primary task and it is designed to be analogous to the driving task. It was developed as part of the ADAM project (Advanced Driver Assistance Metrics [10].

Fig.2. Screen shot from the LCT. In this instance the driver has to change from the centre lane to the right lane.

The LCT requires participants to negotiate a 3000m long section of three lane highway. Participants are instructed by signs on the roadside (150m apart) to perform a lane change manoeuvre (see figure 2). While completing the LCT participants are required to perform a specific secondary task. To avoid speed confounding the results it is controlled by the program and is kept at a constant 60kmph. The illumination reflects daytime driving with a constant light level. Simulated low level engine noise is also provided. Visual information is

presented using an egocentric (front) view; no visual information is presented regarding side or rear views.

The vehicle dynamics are such that the simulated car will behave as a standard passenger car. Participants are required to change lane when instructed, when not performing a lane change manoeuvre they are required to maintain a central position within the lane. Performance of the lane change task by itself is used as a measure of baseline performance for comparison with performance of the LCT and a secondary task. During a LCT trial the LCT program automatically records data to the computer on which it is running. From this data the LCT analysis program can calculate a number of performance measures. These include; mean deviation from the normative model, standard deviation from the normative model, mean steering angle, as well as time course and distance information to allow for standardisation of experimental runs. The normative model is an 'ideal' vehicle path.

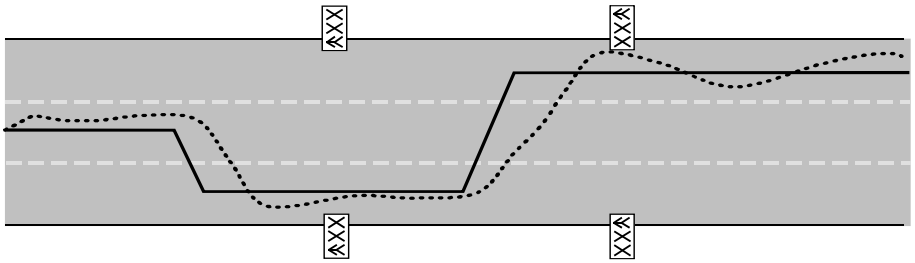

Fig.3. The LCT compares the normative model (solid line) to the participants driving course (broken line)

Mean deviation from the normative model is the major variable of interest. It is a measure of the effect of secondary-task demand and refers to the deviation between the normative model and the actual driving course of the participant along the track (see Figure 3 for a symbolic example of the normative model and driving data). This deviation measure covers important aspects of the driver's performance; namely his/her perception (late perception of the sign or missing a sign), quality of the manoeuvre (slow lane change results in larger deviation) and lane keeping quality, which all result in an increased deviation.

2.4 In-Vehicle Information Systems

The IVIS used in this experiment was a Tom-Tom satellite navigation system running under windows on a HP iPAQ (PDA). The ADAM SURT task was presented on an 8 Inch TFT LCD Monitor produced by LinITX. Using this equipment participants were required to complete seven LCT trials. Each of these trials was dedicated to one of seven IVIS tasks (see Table 1) which have been applied in previous research [10, 11].

Table 1. List of Tasks

Task 1: PDA POI – entering a destination by selecting a "point of interest" using the PDA
Task 2: PDA address – entering a destination by "address" function using the PDA
Task 3: Shares short – searching for a share price from a single scrolling column of text using an LCD screen
Task 4: Shares long – searching for a share price from three scrolling columns of text using an LCD screen
Task 5: SURT circles task easy: - distracters are 22 arc minutes[1] in which the target is easy to find.
Task 6: SURT circles task average: - distracters are 35 arc minutes finding the target is moderately difficult.
Task 7: SURT circles task difficult: - distracter size is 40 arc minutes; in this instance the target is very difficult.

2.5 NASA Task Load Index (NASA-TLX)

The NASA-TLX [12] is a multidimensional tool for the measurement of workload. It identifies a number of subjective factors that are relevant to workload [13]. It has been extensively tested and widely used in the study of human performance [14]. It provides an overall workload score based on a weighted average of ratings on six subscales: mental demand, physical demand, temporal demand, performance, effort, and frustration.

2.6 Procedure

Informed consent was sought from participants prior to commencement of the experiment. Participants were informed of their right to withdraw at any time, for any reason without penalty. Participants were invited to complete a practice session (a block of 5 LCT trials lasting 15 minutes, instructions for the LCT were provided) in order to familiarise themselves with the operation of the LCT. Upon satisfactory performance of the practice task participants completed the experimental trials. In the IVIS condition participants were required to complete 7 LCT trials, one for each task under dual task conditions. The order of these trials was counterbalanced. The seven IVIS tasks are detailed in table 1. After completion of each dual task participants were required to complete the NASA-TLX.

3 Results

Mean deviations from the normative model and participants NASA-TLX ratings for the SURT tasks were significantly correlated ($r(48) = .424$, $p < .01$). There was not a significant correlation between NASA-TLX ratings and tasks 1-4 (Real). A one-way repeated measures ANOVA was calculated for mean deviation from the normative model on the LCT across the seven conditions (PDA POI, PDA Address, Shares short, Shares long, SURT easy, SURT

[1] The size of the target is always 44 arc minutes.

average and SURT difficult). The main effect was not significant [F (6, 15) = 1.029, P>0.05].

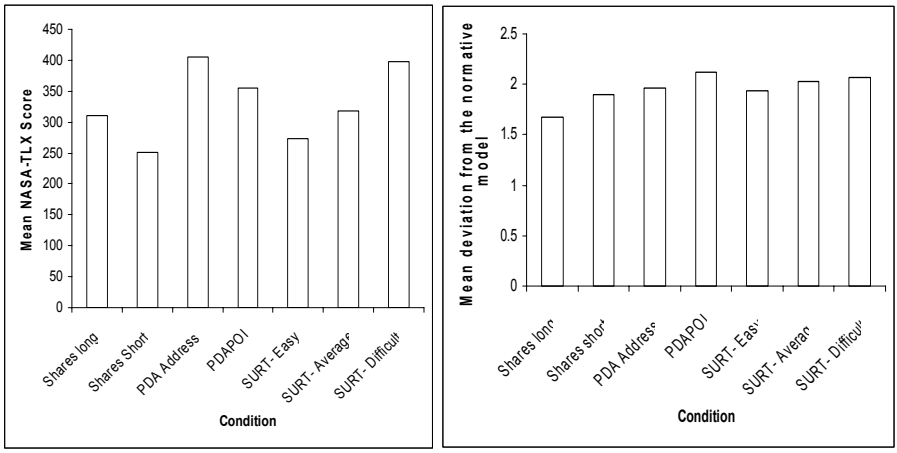

Fig.4. Average NASA-TLX rating and average mean deviation from the normative model by condition

A repeated measures ANOVA was calculated for mean un-weighted NASA-TLX score across the seven conditions (PDA POI, PDA Address, Shares short, Shares long, SURT easy, SURT average and SURT difficult). The main effect was significant [F (6, 15) = 14.421, P<0.05]. The subjective nature of workload is illustrated in a significant test of between-subjects effects [F (1, 15) = 193.206, P<0.005].

3.1 Secondary task performance

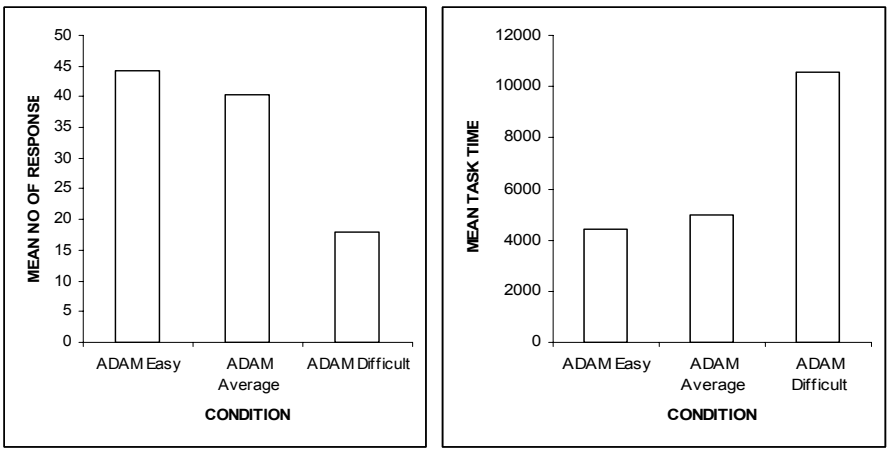

Fig.5. Mean number of responses and task time by SURT condition

A repeated measures ANOVA was calculated for mean number of responses in the three SURT conditions (Easy, Average, and Hard). There was a significant difference between the three conditions [F (2, 47) = 36.150, P< 0.005]. Similar results were observed for total task time [F (2, 47) = 63.467, P< 0.005], mean

reaction time [F (2, 47) = 13.388, P< 0.005], and errors [F (2, 47) = 14.302, P< 0.005]. The difference between number of key presses required to complete the tasks and actual number of presses by participants was not significant [F (2, 47) = 1.755, P> 0.05].

4 Discussion

Performance measures were significantly correlated with measures of subjective workload as evidenced by the significant correlation between NASA-TLX rating and mean deviation from the normative model; this was the case for the SURT tasks, but not the real tasks. Subjective workload ratings were comparable for both the real and reference tasks. There were similarities in the NASA-TLX profiles of the real and reference tasks. Figure 4 shows that the subjective workload ratings obtained on the NASA-TLX for the reference tasks were an approximate replication of the patterns observed in the battery of real tasks. The ratings for the SURT easy task were similar to those for the shares short task, the SURT average task rating was similar to ratings for the shares long and PDA POI conditions and the SURT difficult rating reflected the rating for the address entry condition. This suggests that the amount of workload generated by the SURT tasks is reflective of the workload that would be generated by a real task. The subjective nature of workload is illustrated by a significant test of between-subjects effects. This result strongly suggests that the ratings for each of the seven tasks were different for each participant.

These results reflected the hypothesised difficulty levels of these tasks at their original conception. [11] evaluated the four real tasks (PDA POI, PDA Address, Shares Short, Shares Long) used here using expert opinion and a key stroke analysis. The panel of experts consisted of members of staff from TRL, Nottingham University (UK) and the Chemnitz University of Technology (Germany). The tasks were assessed on four criteria; input (how the driver enters information), task (what needs to be done), display (what information is presented), and output (what results are displayed by the system). The negative, neutral and positive factors of each task were identified. All the tasks were primarily visual however it was hypothesised that the Shares Short task would be the easiest as it contained the least amount of visual information. The Shares Long task contained only a small increment in visual demand (three columns instead of one) therefore we would expect a similar increase in workload rating. The PDA POI task contains an added element in that a physical response is required therefore we would expect a further increase and finally a further increment for the PDA address entry task which is more demanding both visually and physically.

However, the performance measures do not support this. Scores for the mean deviation from the normative model were not significantly different. We can however examine the pattern of results with reference to [11]. What is surprising are the scores for the mean deviation from the normative model for the real tasks. In both the PDA and scrolling shares tasks the mean deviation was higher than expected for the task rated as easier (PDA POI and Shares short) in the expert review. We would have expected a smaller deviation from the normative model in these conditions and therefore a significant difference.

However, there is a plausible explanation for this pattern of results. [15] demonstrate task adaptation in which participants performance in a dual task setting was shown to improve due to participants applying more effort to these tasks and relaxing when performing the driving task alone. A similar effect may be observed here in that participants may be mobilising more resources and increasing effort in the more difficult conditions and/or relaxing and decreasing effort while completing the easier tasks.

This possible explanation is particularly relevant in an examination of the SURT tasks. As the difficulty of the SURT tasks increases so does the subjective workload ratings of participants as illustrated by an increase in the mean NASA-TLX rating. In conjunction with this increase there is a decrease in the mean deviation from the normative model. This is surprising given that this would be indicative of good performance of the LCT. However, this point is clarified in an analysis of the participants' performance on the secondary task.

Data regarding performance of the secondary task suggests that as the difficulty of the task increases participant's performance declines. This is illustrated in Figure 5 which shows that as the difficulty of the SURT task increases the number of responses declines and the time taken to make these responses increases. The results also imply that the SURT difficult task is too difficult for participants and those participants disengage from the task; secondary task performance is sacrificed to maintain what the participant perceives to be an acceptable level of primary task performance. These results are in line with the theory of compensatory effort [16] which states that when faced with a difficult task participants will either mobilise more effort in order to achieve performance goals (with an associated increase in physiological and behavioural costs) or reduce their performance goals to avoid such costs. [6] imply that people are aware of their task performance and that their estimates of subjective workload are based on their perceptions of performance. It is reasonable to suggest that in this instance participants are basing their estimates of workload on their success at completing the SURT task and not on their performance of the LCT. It is important to note that the SURT difficult task is still useful in performance evaluation as it is representative of tasks that should probably not be undertaken while driving.

5 Conclusion

The original aim of this study was to establish the subjective workload ratings (NASA-TLX) for the SURT reference task and compare them to ratings of a test battery of real IVIS tasks while participant's perform a surrogate driving task (LCT). The results of this study support the conclusions that subjective workload ratings are comparable for both the real and reference tasks. Performance measures suggest that participants trade-off secondary task performance in order to maintain LCT performance at their perceived level of acceptable performance.

6 References

[1] Beyreuther, G., and Laidebeur, A.: RTI, State of the Art Review, DRIVE Project V1017 (BERTIE). Changes in Driver Behaviour Due to the Introduction of RTI Systems, No. 28. 1989.

[2] Wierwille, W. W.: Demands on Driver Resources Associated with Introducing Advanced Technology into the Vehicle. Transportation Res (c), 1993, 1, (2), pp. 133-142

[3] Gopher, D. and Donchin, E.: Workload- An examination of the concept, in K. Boff, L. Kaufman and J. Thomas (eds.): Handbook of perception and human performance (Wiley, New York, 1986).

[4] Kahneman, D.: Attention and Effort. Englewood Cliffs, NJ: Prentice Hall. 1973

[5] Wickens, C.D.: Multiple resources and performance prediction. Theor Issues in Ergon Sci., 2002, 3(2), pp. 159-177.

[6] Leggatt, A and Noyes, J.: Workload under complex task conditions in S. A. Robertson (ed.): Contemp Ergon (1996; CRC Press).

[7] Moray, N., Johanssen, J., Pew, R., Rasmussen, J., Sanders, A.F. and Wickens, C.D.: Report of the experimental psychology group, in N. Moray (ed.): Mental workload: its theory and measurement (Plenum, New York, 1979).

[8] van Driel, C.J.G. Hoedemaeker, M. and van Arem B.: Impacts of a Congestion Assistant on driving behaviour and acceptance using a driving simulator, Transportation Res (F), 10, 2007, pp. 139–152

[9] Bischoff, D.: Developing guidelines for managing driver workload and distraction associated with telematic devices, 2007, NHTSA Paper No 07-0082

[10] Mattes, S.: The lane-change-task as a tool for driver distraction evaluation. In H. Strasser, K. Kluth, H. Rausch, and H. Bubb, (Eds.), Quality of work and products in enterprises of the future. Ergonomia, Stuttgart, 2003, 57.

[11] Horberry, T., Stevens, A., Robins, R., Cotter, S. and Burnett, G.: Development of an occlusion protocol with design limits for assessing driver visual demand, 2007; TRL, PPR 256

[12] Hart, S.G. and Staveland, L.E.: Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research, 1988, In P.A. Hancock & N. Meshkati, (eds.), Human Mental Workload, Amsterdam: North Holland Press, pp. 239-250.

[13] Moroney, W.F., Biers, D.W. and Eggemeier, F.T.: Some measurement and methodological considerations in the application of subjective workload and measurement techniques, Int. Jour. of Aviat. Psych., 1995, 5, 87–106.

[14] Jorgensen, A.H., Garde, A.H., Laursen, B. and Jensen, B.R.: Applying the concept of workload to IT work. In Virtual Proceedings of CybErg 1999: The Second International Virtual Conference on Ergonomics, L. Straker and C. Pollock (Eds.) (Perth, Australia: International Ergonomics Association/Curtin University of Technology).

[15] Matthews, G. and Sparkes, T.J. : The role of general attentional resources in simulated driving performance. In Gale, A.G., Brown, I.D., Haslegrave,

C.M., Moorhead, I. & Taylor, S.P. (Eds.), Vision in Vehicles V. Elsevier Science B.V, Amsterdam. 1996

[16] Hockey, G.R.J.: Compensatory control; in the regulation of human performance under stress and high workload: A cognitive-energetical framework, Biol. Psyc., 1997, 45, pp. 73-93.