Time Estimation as a Measure of Mental Workload

Mats Lind and Henning Sundvall

Department of Information Science/HCI, Uppsala University,
Box 513,
SE-751 20 Uppsala, Sweden
mats.lind@dis.uu.se

Abstract. Technical systems of different kinds might differ in the mental demands they put on their users while being equally usable in a more conventional sense. Several methods exist that measure mental workload. However, in the everyday practice of usability evaluations, none of these methods seems to be used. This is probably due to the amount of effort needed to use them. The study of our errors in estimation of durations show that errors increase as a function of the amount of attentional resources being needed for other concurrent tasks. This points towards a simple way of estimating mental workload. By asking people to provide estimates of elapsed time after a task, disruptions in their estimates could indicate the mental workload of the task. We conducted a first study aimed at validating this idea with the NASA TLX method. Results show that errors in time estimates correlate significantly with TLX scores.

Keywords: Mental workload, time estimation, usability.

1 Introduction

Mental workload is most often assessed when complex and safety critical tasks are investigated such as the ones found in air traffic control or in the control of nuclear power plants. However, from a usability point of view mental workload could be an important factor to consider in many other areas. It seems reasonable to speculate that high mental workloads could result in worker health problems as well as affecting the quality of work output in many, and more common, work situations (see e.g. [1]). Nevertheless, most usability evaluations of systems intended for use in more ordinary work situations do not, in our experience at least, employ measures of mental workload. One important reason is that funding for usability activities still often is bordering on inadequate. Therefore usability evaluations need to be time and cost effective and usability professionals, consequently, often settle for the bare minimum in terms of measurements. Furthermore, representative users in these types of usability tests are not usually accustomed to the use of special lab procedures or scales and they sometimes give subjective ratings of general usability that are quite inconsistent with the more objectively gathered measurements. These constraints make existing measures of mental workload difficult or impossible to use (for an excellent overview of such measures, see [2] pp. 301-364).

D. Harris (Ed.): Engin. Psychol. and Cog. Ergonomics, HCII 2007, LNAI 4562, pp. 359–365, 2007.

The problem remains, however, and we believe that if a measure could be found that is minimally intrusive and practical enough to use in almost any usability evaluation, a whole new insight into many problems associated with computer use in working life could be found. As a step towards finding such a measure we have looked into the use of errors in time estimation as an indicator of mental workload. Time estimation is easy to administer, a simple "x" on a time line is enough, and has a solid theoretical and empirical foundation in cognitive psychology.

1.1 Time Estimation and Mental Workload

Most of us can produce anecdotal evidence of mental workload affecting our perception of elapsed time. We often literally feel time rushing away from us when we are heavily engaged in trying to meet a deadline for a difficult task whereas the minutes drag on while we are inactively waiting for a delayed airplane. More solid scientific support for this phenomenon can be found in the literature on working memory. Here, the notion of a central executive function with limited capacity is very common and a recent discussion of this particular aspect can be found in [3]. A number of experiments have shown that there is a trade-off between temporal processing of intervals in the range of seconds and simultaneous nontemporal information processing (e.g. [4], [5]). For instance, Brown [4] report interference effects when subjects produced 2 and 5 second intervals from such different task as visual pursuit, visual search and mental arithmetic. On the theoretical side these results are often explained by a so-called "attentional-gate" model [6],[7] where the output of a hypothesized time pulse generator must pass a cognitive gate. This gate is controlled by the amount of attentional resources presently allocated to temporal information processing. When more resources are allocated to such processing, timing estimates are more precise and vice versa.

Recently, longer time periods than a few seconds have also been investigated [8]. In one experiment, subjects were asked to give time estimates of durations as long as 40 seconds. Even for these longer intervals the same pattern of results emerged. When working memory was heavily engaged by memorizing a sequence of letters, the errors in time estimation more than doubled compared to the condition with no working memory load.

In light of this it seems safe to say that the empirical evidence so far point to that disruptions of time estimations are caused by attentional resources being heavily engaged in one or several of other, nontemporal, tasks of a varied nature. Of course, mental workload is a multi-dimensional concept [9]}, but how heavily engaged a hypothesized limited attentional resource is, must in any case be of central importance.

Using time estimation as a measure of mental workload could, perhaps, be regarded as using dual task paradigm. In this case, however, the secondary task is something most of us perform more or less constantly. No training or special instructions are thus needed, at least in principle, although some previous studies (e.g. [8]) have employed a calibration procedure.

An interesting difference between the anecdotal evidence and the formal experiments is that the latter use unsigned error as measure of discrepancy between actual time and perceived time. The anecdotal evidence suggests that the sign of the

discrepancy might be interesting as well. Low memory load could perhaps lead to an overestimation of durations whereas high memory load could lead to an underestimation.

1.2 Usability and Mental Workload

Usability, as defined by ISO [10], is measured in terms of user's effectiveness, efficiency and satisfaction for specified task in a particular context of use. Efficiency is associated with the amount of resources consumed in the process of solving the task. In practice, this usually means how much time a user spends while solving a task, but could easily incorporate also the mental resources spent. In this view, mental workload would be a subscale of efficiency. In this way, it would be quite possible that two computer systems used to solve the same task would results in equal amounts of time spent by the users to perform the task while demanding different amounts of mental resources and, thus, still demonstrate different degrees of efficiency for the task.

2 Method

As pointed out by Jex [9] there are inherent logical and conceptual problems in defining and validating measures of mental workload. He therefore recommends the use of a well-known subjective method for initial validation of a new proposal. Hence we chose the NASA-TLX method [11] as our standard comparison. We also decided to use the main scale of the TLX, the "Total mental workload" scale, mainly because we would use only a limited number of tasks and subjects in this initial evaluation of the duration estimation idea. There would also be a limited time for each subject to learn the TLX method before using it, something that could cast further doubts on the use of the measurements in the TLX subscales.

2.1 Subjects

Thirteen persons, 7 male and 6 female, aged between 19 and 31 and with a mean age of 24.6 years, participated voluntarily in the study. They all received a small compensation for participating.

2.2 Tasks

In a simple pilot test with three subjects we investigated 30 candidate tasks. Our aim was to find a subset of tasks that gave a homogenous range of scores on the TLX main scale and that could be solved within the course of one hour. Nine tasks were selected, two as simple warm-up tasks and seven as the main tasks. The selected tasks included finding specific pieces of information on websites, solving computer based puzzles and solving a simple sudoku game. The seven main tasks were administered in random order to each subject.

2.3 Apparatus

All tasks were performed on a PC running Windows XP and was equipped with a 17" TFT monitor. The PC was connected to the Internet. No time related information was visible on the monitor and all subjects were asked to remove any time keeping device they carried before the session began. Each session was recorded using a screen capture program and timings of real elapsed time were later obtained from these recordings. The subjects' duration judgments were performed by drawing a line on a time scale printed on paper.

2.4 Procedure

When a subject came to the lab s/he was first instructed in the use of the TLX method for about 5 to 10 minutes. Both oral and written instructions were used for this tutoring. The instructions explicitly told the subjects that they would be asked to use the TLX forms and to estimate elapsed time after each completed task. Elapsed time was always estimated before the TLX form was filled in.

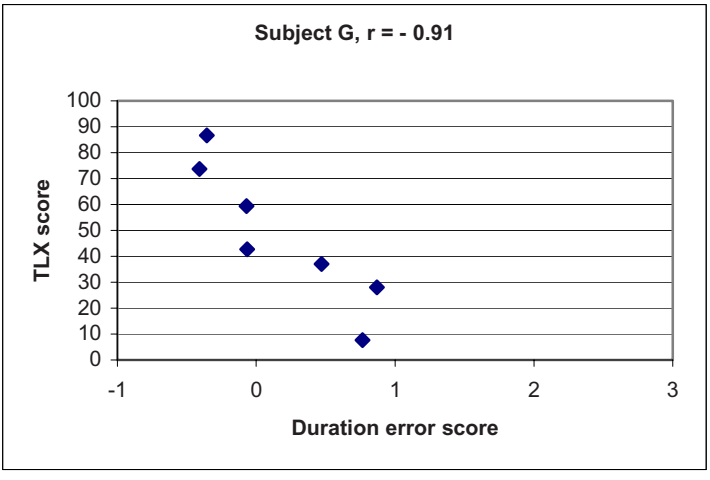
3 Results

Interestingly, several subjects showed underestimation of duration for the tasks with a high TLX score. Because of this we decided to use signed errors as our dependent variable for the duration estimations. We simply calculated the following score for each trial and each subject:

$$\text{Duration error score} = (\text{judged duration} - \text{actual duration}) / \text{actual duration} \quad (1)$$

There was a large variation in these scores, both within and between subjects, ranging from –0.4 to 7.6. Values above 4 were found in two persons only. The TLX scores also showed a large variation. They ranged from 5 to 85 with a clear majority of subjects producing values between 20 and 70. The mean TLX values over all subjects for the seven tasks were [22 33 39 52 60 64 69] indicating a reasonably well-balanced set of tasks in terms of mental load.

The main analysis was the correlation for each subject between the TLX scores and the duration error score. Each such correlation is, of course, only based on seven pairs of values making the correlations error prone. However, the correlations were surprisingly stable. With the exception of two subjects who showed correlations close to zero (0.08 and –0.18), the values ranged between –0.67 and –0.91. For the statistical analysis all thirteen correlations were transformed by Fishers' r to Z procedure [20]. These values were then tested by means of a t-test to see if they were significantly different from zero. They were ($t=-7.25$, $N=13$, $p< 0.00001$). The mean value of these transformed values was –0.9682 which corresponds to $r = -0.75$. A 95% confidence interval around this mean was also calculated. It was [-1.26 –0.68] corresponding to r=[-0.85 –0.59]. The two scatter plots in Figure 1 illustrate the correlation results.



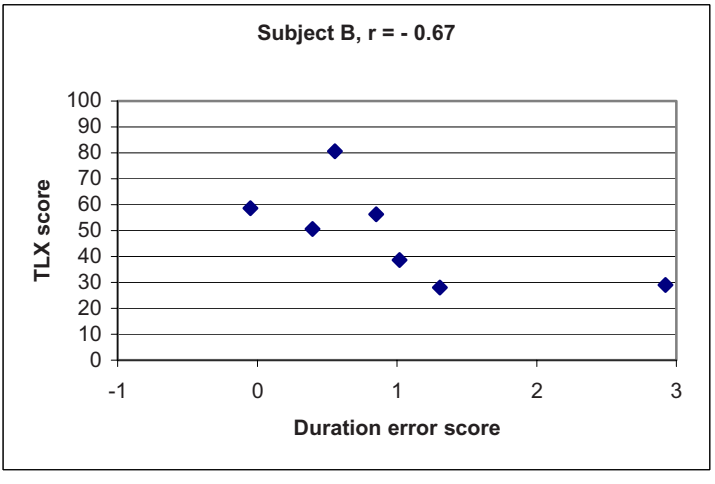


Fig. 1. Scatter plots of duration error scores versus TLX scores for two subjects whose results exhibited the highest and the lowest measured correlation, respectively (not including the two subjects who exhibited correlations close to zero)

The fact these correlations are negative is perhaps surprising but it reflects that the duration of tasks with a low TLX value are systematically more overestimated than duration of tasks with a high TLX value.

4 Discussion

This is of course only a first validation of the basic idea and there are a number of issues that need to be addressed. First of all, there exists a co-variation in our data

between actual task duration and TLX score for most subjects. According to previous results, for instance experiment 3 in [8], errors increase as durations increase regardless of load. However, we got the opposite result; duration estimates are closer to actual duration when task durations are longer. This apparent contradiction could perhaps be explained by the co-variation between mental load, as measured by TLX, and task duration. Larger loads also increase the error in duration estimations, so, given that the sign of these error are opposite, this could result in the errors averaging out for longer durations/higher loads. Since previous studies only report the unsigned error this must remain a speculation until further evidence is gathered. In any case it points to the necessity of conducting more studies where the differential effects of task duration and mental load can be studied.

Another result that needs to be further investigated is the large between subject variation we found in duration estimates. In most previous experiments subjects have been calibrated, for instance by being shown a regularly flashing light before each trial. Doing this would ruin the whole idea of a simple measure of mental workload to use in everyday usability studies. Of course, within subjects designs where the same subject does the same task with two or more interfaces are not affected by this problem since each subject is only compared to herself. Also the possibility of administering a simple calibration task to each subject before the evaluation proper begins could solve the problem, not by calibrating the subject, but by calibrating the measure. Assuming, of course, that each subject shows the same pattern of results over time without any major drift.

Given the small number of subjects and tasks in this study, more studies are needed just to validate the basic results. However, the results we obtained are very promising and it seems to be a worthwhile effort to continue the investigation into the use of time estimation as a measure of mental workload.

In summary it seems appropriate to use the five criteria on a useful measure of mental workload put up by Jex [9]. A good measure of mental workload should be: Relevant, Sensitive, Concordant, Reliable and Convenient.

- The relevance criterion seems fulfilled. The psychological evidence point directly to disruptions in duration estimation as being caused by attentional load.
- Whether or not the measure is sensitive remains to be investigated. Our results seem to indicate that there is at least hope and the procedure can definitely be improved upon.
- The concordance criterion means that the measure should reflect “ubiquitous trends in target population” ([9], p.13). Again, the psychological studies provides hope but more studies are necessary in more applied settings and with more representative samples than the usual university students.
- To find out how reliable this measure is should be a fairly straightforward task and is one of the objectives of planned, forthcoming studies.
- That the measure it is convenient, that it is easy to learn and administer, is portable for use in field trials and evaluations and involves a low cost seems clear already at this point.

References

1. Arnetz, B.B.: Techno-stress: a prospective psychophysiological study of the impact of a controlled stress-reduction program in advanced telecommunication systems design work. J. Occup. Environ Med. 38, 53–65 (1996)
2. Stanton, N.A., Salmon, P.M., Walker, G.H., Baber, C., Jenkins, D.P.: Human Facotrs Methods, A Practical Guide for Engineering and Design. Ashgate Publishing Limited, Aldershot (2005)
3. Baddeley, A.D.: Is working memory still working? Am Psychol 56, 851–864 (2001)
4. Brown, S.W.: Attentional resources in timing: interference effects in concurrent temporal and nontemporal working memory tasks. Percept Psychophys 59, 1118–1140 (1997)
5. Fortin, C., Breton, R.: Temporal interval production and processing in working memory. Percept Psychophys 57, 203–215 (1995)
6. Zakay, D., Block, R.A.: Temporal cognition. Current Directions in Psychological Science 6, 12–16 (1997)
7. Zakay, D.: Gating or switching? Gating is a better model of prospective timing (a response to 'switching or gating?' by Lejeune)(1). Behav Processes 52, 63–69 (2000)
8. Venneri, A., Pestell, S., Nichelli, P.: A preliminary study of the cognitive mechanisms supporting time estimation. Percept Mot. Skills 96, 1093–1106 (2003)
9. Jex, H.R.: Measuring mental workload: Problems, progress, and promises. In: Hancock, P.A., Meshkati, N. (eds.) Human Mental Workload, vol. 52, Elsevier science publishers, Amsterdam (1988)
10. ISO: ISO 9241-11:1998 Ergonomic requirements for office work with visual display terminals (VDTs) – Part 11: Guidance on usability. ISO (1998)
11. Hart, S.G., Staveland, L.E.: Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. In: Hancock, P.A., Meshkati, N. (eds.) Human Mental Workload, vol. 52, Elsevier science publishers, Amsterdam (1988)
12. Hays, W.L.: Statistics for the social sciences. Holt, Rinehart and Winston, New York (1973)