Using Interval Productions to Measure Mental Workload: Lessons Learned from Human Factors Assessments of Surgical Visualization Technologies

Russell Grant, C. Melody Carswell, Cindy H. Lio,
Matt Field, Duncan Clarke, and W. Brent Seales

University of Kentucky Dept. of Psychology, Dept. of Computer Science, and Center for Visualization and Virtual Environments, Lexington, KY

Feature at a Glance: The use of secondary task performance to assess mental workload in a primary task is appealing because the method clearly reflects a central goal of workload assessment – to determine what other functions an operator can undertake while satisfactorily performing the ongoing (primary) technical challenges of a job. For example, does a surgeon performing a suturing task have the cognitive reserves to maintain situation awareness, deal with unanticipated events, or coordinate the efforts of other team members? Unfortunately, secondary task measures have a reputation for being intrusive, artificial, and difficult to use. The current article describes procedures to minimize these concerns, specifically when using an interval production secondary task. Although our suggestions for implementing interval production are based on experience in surgical training environments, the method is grounded in workload assessment research from a variety of other contexts over the past two decades. The methodology appears to be highly adaptable.

The secondary task technique is an intuitively appealing approach to the measurement of mental workload. It assumes that if a person is simultaneously performing two tasks of different priorities, a reduction in the workload of the more important (or primary) task will make possible an increase in performance of the remaining (or secondary) task. Thus, changes in secondary task performance can, in principle, be used to infer changes in primary task workload.

To be clear, the purpose of this article is not to review the extensive human factors literature comparing performance-based measures of workload with physiological and subjective measurement strategies. Nor is our goal to introduce a totally unique secondary task. Instead, our goal is to describe the issues we encountered when trying to use one seemingly simple secondary task to measure workload in one work domain – surgery. We hope that the problems we encountered, and the many methodical considerations we addressed in adapting a well-known secondary task to a new environment will help others who may 1) want to use this same secondary task in their own projects, or 2) adapt other secondary tasks to new work domains. For those still wanting a primer on the pros and cons of the major approaches to workload assessment, secondary tasks included, Box 1 lists several accessible resources.

Box 1. Suggested Readings on the Theory and Practice of Workload Assessment

Gawron, V. J. (2000*). Human performance measures handbook.* Lawrence Erlbaum Associates. Mahwah, NJ.

Tsang, P. S., & Vidulich, M. A. (2006). Mental Workload and Situation Awareness. In G. Salvendy (Ed.), *Handbook of Human Factors and Ergonomics* (3rd Edition ed.). Hoboken, New Jersey: John Wiley & Sons.

Young, M. S., & Stanton, N. A. (2005). Mental Workload. In N. Stanton, A. Hedge, K. Brookhuis, E. Salas & H. Hendrick (Eds.), *Handbook of Human Factors and Ergonomics Methods*. New York: CRC Press..

Wickens, C. D. **&** Hollands, J. G. (2000)**.** *Engineering Psychology and Human Performance* (3rd ed.). Upper Saddle River, NJ: Prentice Hall.

Mental Workload in Minimally Invasive Surgery

We began our own search for a suitable secondary task workload assessment method several years ago when we were asked to evaluate new display systems for minimally invasive surgery (MIS). At the time, usability tests of new surgical technologies mostly focused on the speed, accuracy, and economy with which surgeons could perform such fundamental skills as knot-tying, suturing, and cutting. However, we felt that new technologies should also reduce the cognitive load imposed by these primary tasks. In aviation, tremendous attention has been devoted to ensuring that pilots have the capacity to cope with a variety of secondary tasks while maintaining primary flight control. In MIS, likewise, it was clear that surgeons needed the cognitive reserves to perform a variety of additional tasks. For example, while executing the delicate movements needed to partition a tumor from healthy tissue, surgeons should also be able to maintain situation awareness, anticipate problems, generate plans, coordinate the surgical team, and cope with a variety of stressors.

Although the rationale for measuring surgeons' cognitive reserves seemed compelling, there were few surgical human factors evaluations that had actually used a secondary task criterion. After excluding our own studies, we found 28 surgical studies that assessed mental workload between 1994 and 2009. Eighty-six percent of these studies employed subjective assessment techniques, but only 21% employed a secondary task. Although we cannot know for sure why secondary tasks were relatively rare, we suspect that it had something to do with their reputation as intrusive, artificial, and difficult to design and administer. We clearly needed to find a secondary task that would minimize these concerns.

We decided to focus our efforts on a secondary task – interval production -- with a relatively long history of successful use in a variety of work settings.(e.g., Baldauf, Burgard, Wittman, 2009, for driving; Liu and Wickens, 1992, for monitoring). Research participants are simply asked to make a response each time they believe a target interval has elapsed. They continue to make the response, that is, to produce the interval, until the experimenter tells them to stop or until the trial ends. In general, as primary task workload increases, the intervals generated by participants become less accurate and more variable. Although our goal was to standardize the application of interval production to the surgical domain, we believe that our administration protocol, nicknamed "Vis Tempo," may increase the method's applicability to other settings as well.

Why Interval Production?

Anyone who has seriously considered assessing mental workload by measuring secondary task performance will have quickly learned that there are many secondary tasks to choose from, including auditory, visual, and vibrotactile detection, simple and complex classifications, mental arithmetic, memory search, manual control, and a variety of time estimation tasks. Given these options, why did we pursue interval production?

Interval production fared well when we considered the criteria by which workload measures are often judged (e.g., Wierwille and Eggemeier, 1993; Wickens and Hollands, 2000). Most obviously, we needed a task that was likely to be sensitive to primary task demands. Interval production was a good bet because it engages a variety of mental resources, including those involved in goal maintenance, phonological processing (e.g., subvocal counts), and response execution, to name a few. This may explain why it has been used successfully to assess workload in so many tasks, from piloting an aircraft to process monitoring.

We knew that acceptability to surgeons would be especially critical, because the effectiveness of secondary tasks strongly depend on the motivation of research participants. Participants must be motivated to 1) maximize joint performance on primary and secondary tasks, and 2) adhere to instructions about task priority. We consulted available MIS task analyses for possible secondary tasks that would have

face validity for both surgeons and surgical trainees, and we found a study by Cynthia Dominguez (2001) that provided inspiration. Surgeons described the importance of their internal time keeping, their "mental clocks," for assessing the progress of surgical procedures and deciding when to change strategies. With the time pressures that characterize many other workplaces, the choice of a temporal secondary task would probably make sense to workers in other jobs as well.

Finally, we needed a secondary task that could be easily implemented and was unlikely to alter typical primary task strategies. A characteristic of interval production, unusual among secondary tasks, is that it is discretionary. That is, it involves no ongoing presentation of stimuli that may intrude upon normal primary task performance, even when the participant wants or needs to ignore the secondary task. Another benefit of having a task that requires no external stimuli is that the need for specialized displays and presentation software is eliminated. Finally, the task can be easily modified to require pedal, manual, or vocal responses, further reducing unnecessary disruptions of the primary task.

Putting Interval Production into Practice

To understand how we actually used interval production in our research, consider the following instructions we gave to participants. Note that participants received these instructions after having already learned to perform their primary task, which involved manipulating small objects with surgical graspers.

"On the next trial, you will perform both the surgical task and a time-keeping task. To mark the passing of time, you should say the word 'time' each time you think that 21 seconds have elapsed. Keep doing this until I tell you to stop. You should start both your surgical task and your time keeping when I say 'go'. Try to complete your surgical task as quickly and accurately as possible, and try to make your intervals in the time-keeping task as accurate and consistent as possible. Although you should try to do both tasks well, keep in mind that it is more important for you to maintain good performance in the surgical task."

Although these sample instructions are sufficient for describing the task to research participants, they will probably leave researchers interested in actually using interval production with many questions. Some of the most important questions are summarized in Figure 1, where the general steps in a typical experimental trial are presented in a vertical flow chart on the left, and associated methodological questions are presented, in parallel, on the right. When we first adopted the interval production task, we could find little practical advice about how to answer many of these questions, for example, how to select a target interval, instruct participants, or summarize time-keeping performance. Now, after having used the technique in a number of validation studies, our file drawers include both successes and failures, with the successes eventually coming to dominate (e.g., Grant, 2010; Grant, Carswell, Lio et al., 2009; Lio, Bailey, Carswell et al., 2007; Lio, Carswell, Seales et al., 2006). Based on a critical review of the sometimes minor changes in procedures that characterized our

experiments, we offer the following suggestions for implementing an interval production secondary task, with an eye toward “keeping it simple.”

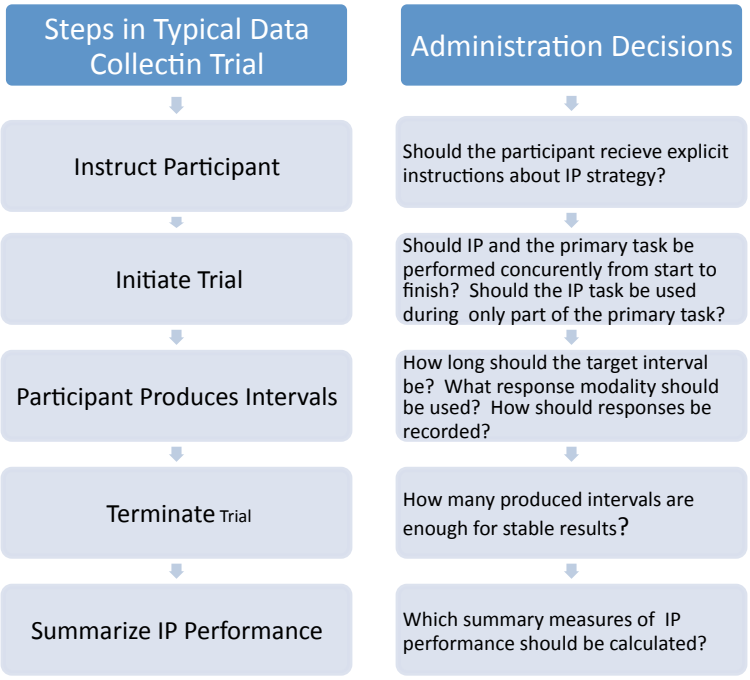

Figure 1. The steps in a typical dual-task trial with interval production used as a secondary task (left) and the associated procedural questions relevant to each step (right).

Vocal vs. Manuel Responses. Most studies using interval production ask participants to mark intervals by pressing a key or pedal. Because surgical primary tasks require the use of both hands, and sometimes require foot controls as well, vocal responses were more appropriate for our application. Although we have not conducted a head-to-head comparison, both vocal and manual interval productions have been successfully employed as secondary tasks. The choice should be driven by the constraints imposed by the primary task of interest to the researcher.

If primary task demands permit the use of multiple response modalities, vocal responses have the advantage of lending themselves to very “low tech” data collection solutions. In the simplest case, a researcher needs nothing more than a good audio recorder and a stop watch. With a little more time and money, a voice key and supporting software can automate the collection of voice responses; however, in practice, it may be difficult to achieve acceptable response detection accuracy amidst the noise of the training environment and participants’ irrelevant vocalizations.

A hybrid (semi-automated) solution for collecting data from vocal responses is to have an experimenter press a key after each of the participant’s responses, allowing close-to-real-time data entry and analysis. Although this procedure inevitably adds noise to the measurement process, we have found correlations between intervals

recorded in this way and those based on visual analysis of the individual vocal waveforms to range from r = .96 to r = .99. Those interested in automating their data collection and analysis can download "Vis Tempo," a free application for use on iPhones and iPads, from the Apple app store. The application collects manual responses (generated either directly by the participant or indirectly by the researcher) by using the entire face of the mobile device as a single response key. Alternatively, there are a variety of time-keeping applications that are available for download to other mobile devices for less than five dollars. We have found rowing timers to be particularly easy to repurpose.

Performance metrics. There are three types of measures used to summarize interval production performance -- measures of bias, absolute error, and consistency. Bias metrics describe the direction of interval production errors, revealing whether a participant is under- or overestimating elapsed time. Although bias metrics sometimes provide useful insights about the cognitive processes involved in subjective time keeping, they are not particularly sensitive to primary task workload. Instead, we would suggest using the Percent Absolute Error (PAE) to describe a participant's overall accuracy and the Coefficient of Variation (CV) to describe the stability of their responses (see concluding summary for formal definitions).

Target Intervals. Intervals from a few seconds to a minute have been successfully used to detect shifts in concurrent task workload. Hart (1975) recommended using target intervals less than 30 seconds and, practically speaking, shorter intervals permit the collection of more data per experimental trial. However, it is possible to make target intervals too short. Psychophysical studies suggest that very short intervals cross a "breakpoint" where interval production is reduced to a sensory memory task with little cognitive demand (e.g., Madison, 2001), but in practice we have achieved acceptable sensitivity with target intervals as low as 3 seconds. Although it appears that a variety of target intervals can be used, we caution researchers to avoid using intervals that are close to the duration of a salient performance outcome in the primary task, such as the time it takes to make a single suture in a suturing task. Participants sometimes report using the completion of such milestones as a response cue for the interval production task.

Number of produced intervals. In order to establish such basic experimental parameters as target interval and trial length, we must know how many produced intervals are necessary per trial to obtain stable estimates of interval production performance. Across a range of target intervals, we find that the effect sizes associated with manipulations of primary task demand approach their maximum when 5 – 8 intervals contribute to summary performance measures. Our rule of thumb is to select a combination of target interval and trial length that will make it likely that at least 5 intervals will be produced by research participants.

Strategy instructions. When we began using interval production, we told participants to avoid counting. We reasoned that counting was so over-learned that its use as a time-keeping strategy would make interval production performance relatively

impervious to even drastic changes in primary task demand. But we were wrong. We found smooth tradeoffs in performance between the primary surgical task and interval production performance as a function of task emphasis instructions even when participants reported counting. In fact, insistence that participants avoid counting actually reduced the sensitivity of interval production to changes in primary task workload. Participants adopted so many different strategies (with some admitting to counting despite our instructions) that performance variability led our statistical tests to suffer under the weight of bloated error terms. We currently give participants no strategy instructions.

Practice and feedback. We usually give participants no more than 3 90-second practice trials producing the target interval. After each trial, we provide them with feedback about the mean duration of their produced intervals and the duration of their longest and shortest intervals. We also allow participants to perform interval production concurrently with the primary task for at least one trial before beginning data collection. One advantage of using the coefficient of variation (CV) to summarize interval production performance is that it makes it largely unnecessary to calibrate participants' duration estimates through practice. Because CV uses the participant's mean produced intervals rather than the actual target interval as a baseline, it is relatively unaffected by idiosyncratic tendencies to over- or underestimate interval lengths.

Summary of Suggested Procedures

As with any workload assessment procedure, the exact administration technique will depend on the nature of the primary task and work environment. The following procedural summary, therefore, is offered as a starting point for those wanting to adopt interval production and modify it for their own evaluation projects.

1. Select a target interval between 3 and 30 seconds. Specific considerations:

    - If the primary task involves salient, repetitive events, choose an interval that is well under or over the average length of the event.

    - If recording is not automated, longer intervals will be easier for the experimenter to collect.

    - If recording is automated (or semi-automated), use shorter (3 – 10 sec) intervals.

2. Choose the response modality (manual, pedal, or voice) that best fits experimental demands.

    - If the primary task requires vocalizations or depends on auditory displays, opt for manual or pedal-based responses.

    - If the primary task requires intensive use of both hands, use vocal or pedal responses.

3. Calculate both the percent absolute accuracy (PAE) and coefficient of variability (CV) as performance metrics.

$$\text{Percent Absolute Error (PAE)} = \frac{M_D}{TI}$$

$$\text{Coefficient of Variation (CV)} = \frac{SD_{PI}}{M_{PI}}$$

Where TI is the target interval, M_D is the mean absolute difference of the actual produced intervals and the TI, M_{PI} is the mean of the produced intervals, and SD_{PI} is the standard deviation of the produced intervals. The Vis Tempo app will automatically calculate these values for each experimental trial.

4. Collect a minimum of 5 intervals per experimental trial.

5. Provide 3 interval production practice trials with feedback about the mean and range of produced intervals. Allow at least one practice trial in which interval production is performed concurrently with the primary task.

6. Neither encourage nor discourage particular interval production strategies.

7. Remind participants to perform both the primary and secondary task as well as they can together while remembering that performance on the primary task is more critical.

References

Baldauf, D., Burgard, E., Wittmann, M. (2009). Time perception as a workload measure in simulated car driving. *Applied Ergonomics, 40,* 929-935.

Dominguez, C. (2001). Expertise and metacognition in laparoscopic surgery: a field study. *Proceedings of the Human Factors and Ergonomics Society 45th Annual Meeting*, Minneapolis, MN.

Grant, R.C. (2010). Time estimation errors as an index of task demand during laparoscopic skills training: Effects of target duration and attention allocation. Thesis presented in partial fulfillment of the M.A. degree, University of Kentucky, Lexington, KY.

Grant, R.C., Carswell, C.M., Lio, C.H., Seales, W.B., and Clark, D. (2009). Verbal Time Production as a Secondary Task: Which Metrics and Target Intervals are Most Sensitive to Workload for Fine Motor Laparoscopic Training Tasks*? Human Factors and Ergonomics Society Annual Meeting Proceedings, 18,* 1191-1195.

Hart, S. G. (1975) Time estimation as a secondary task to measure workload, *Proceedings, 11th Annual Conference on Manual Control* (Washington, DC: US Government Printing) 64 – 77

Lio, C. H., Bailey, K., Carswell, C.M., Seales, W.B., Clarke, D., and Payton, G.M. (2006). Time Estimation as a Measure of Mental Workload During the Training of Laparoscopic *Skills Human Factors and Ergonomics Society 50th Annual Meeting Proceedings,* 1910-1913.

Lio, C.H., Carswell, C.M., Seales, W.B., Clarke, D., Kurs, Y., and DeCuir, J. (2007). Using Global Implicit Measurement Strategies to Assess Situation Awareness during the Training of Laparoscopic Surgical Skills. *Human Factors and Ergonomics Society 51st Annual Meeting Proceedings,* 1280-1282.

Liu, Y. & Wickens, C. D. (1994). Mental workload and cognitive task automaticity: an evaluation of subjective and time estimation metrics*. Ergonomics, 37,* (11), 1843-1854.

Wickens, C. D., & Holland, J. G. (1999). *Engineering psychology and human performance*. (3rd ed.) Prentice Hall.

Wierwille, W.W., Eggemeier, F. T. (1993). Recommendation for mental workload measurement in a test and evaluation environment. *Human Factors, 35* (2), 263-281.