Behavior Research Methods & Instrumentation
1975, Vol. 7 (2), 175-178

SESSION VIII CONTRIBUTED PAPERS: GENERAL TECHNIQUES AND APPLICATIONS

DAN KEARNS, *Oregon Research Institute, Presider*

The precision of latency measures on real-time computing systems

THOMAS W. CHRISTIAN and PETER G. POLSON
The Computer Laboratory for Instruction in Psychological Research, University of Colorado, Boulder, Colorado 80302

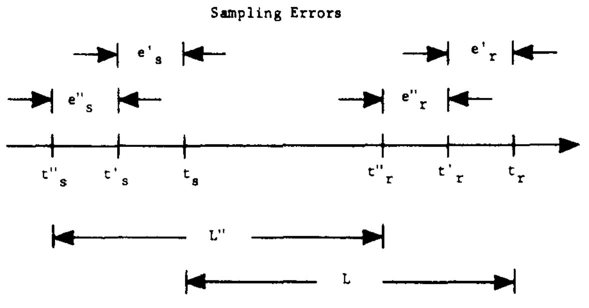
We derive expressions for the bounds on the precision of response latency measures made using a free-running digital clock and discuss other possible sources of measurement errors. In a multitask, real-time environment, there are three possible sources of large measurement errors: (1) the finite resolution of the digital clock, (2) unscheduled delays in recording the time of occurrence of an event, and (3) the uncertainty of the time of stimulus presentation for stimuli presented on a CRT terminal.

This paper presents an analysis of the precision of latency measurements made using a free-running digital clock and discusses problems involved in measuring reaction times. We assume that the program controlling the experiment is executing in some type of multitask environment and that computer time is allocated on a demand-for-service basis using some type of multilevel priority interrupt scheme. The display and response recording hardware are interfaced to the computer. The experiment program controls the stimuli and records latencies and responses. The values of latencies are derived from successive readings of a free-running digital clock. We show that there are four potential sources of error in the latency measures: (1) the finite resolution of the clock; (2) the accuracy of the current clock value; (3) unscheduled delays in recording the time of occurrence of an event, and; (4) in the case of stimuli presented on a CRT terminal, the uncertainty of the time of stimulus presentation.

We will carry out our analysis in the context of a simple reaction time experiment. The experiment program carries out the following operations on each trial by calling subroutines that are part of the operating system. First, the stimulus is presented to the subject and the time of stimulus presentation is recorded. The program then suspends execution and waits for the subject to respond. When the subject responds, the program is reactivated and the time of the subject's response is recorded. We will first derive bounds on the precision of response latency measures from a free-running clock, then discuss several other causes of measurement error, showing, where possible, how to minimize or eliminate these variations.

FORMAL ANALYSIS

Assuming, for the moment, that the samples from the digital clock are taken at the same instant that the stimulus is presented and that the response is received, there are still two potential sources of error–the finite resolution of the digital clock and, if applicable, the error in approximating the digital clock with an internal, interrupt-driven counter. Figure 1 illustrates the potential sampling errors.



t_s	Time of stimulus presentation is recorded.
t'_s	Current value of digital clock at t_s.
t''_s	Current value of internal counter at t_s.
t_r	Time of subject's response is recorded.
t'_r	Current value of digital clock at t_r.
t''_r	Current value of internal counter at t_r.

Figure 1. Possible errors in measuring response latencies with a digital clock and a digital counter.

Suppose t_s represents the time at which the time of stimulus presentation is recorded. Since the digital clock can assume only certain discrete values, t_s must be approximated by t'_s, the current value of the digital clock at t_s where

$$t'_s = t_s - e'_s, 0 \leqslant e'_s < p \qquad (1)$$

where p is the period of the digital clock. This approximation assumes that the digital clock will be synchronized with a continuous clock at integral multiples of its period.

If an internal, interrupt-driven counter is used as the time standard, a second error may be introduced by a lack of synchronization between the digital clock and the internal counter, an error typically caused by a delayed response to the periodic counter update signal. In this case, the current value of the digital clock will be approximated by t''_s, the current value of the internal counter at t_s where

$$t''_s = t'_s - e''_s, e''_s \geqslant 0 \qquad (2)$$

By substituting the expression for t'_s in Equation 1 for t'_s in Equation 2, we have

$$t''_s = t_s - e'_s - e''_s, 0 \leqslant e'_s < p, e''_s \geqslant 0. \qquad (3)$$

If t_r represents the time at which the time of the subject's response is recorded, then we can similarly say that

$$t'_r = t_r - e'_r, 0 \leqslant e'_r < p, \qquad (4)$$

$$t''_r = t'_r - e''_r, e''_r \geqslant 0, \text{ and} \qquad (5)$$

$$t''_r = t_r - e'_r - e''_r, 0 \leqslant e'_r < p, e''_r \geqslant 0. \qquad (6)$$

We now approximate the time interval L, where

$$L = t_r - t_s = t''_r - t''_s + e'_r + e''_r - e'_s - e''_s \qquad (7)$$

with L″, where

$$L'' = t''_r - t''_s \qquad (8)$$

Since all of these error terms are nonnegative, the following bounds can be placed on the error in the approximation:

$$L - e'_r - e''_r \leqslant L'' \leqslant L + e'_s + e''_s \qquad (9)$$

The errors e'_s and e'_r will be uniformly distributed over the range

$$0 \leqslant e'_s < p, 0 \leqslant e'_r < p$$

where p is the period of the digital clock. The errors e''_s and e''_r are entirely dependent on the ability of the computer to keep the internal counter in synchronization with the digital clock. In most cases, the errors e''_s and e''_r will, with high probability, be quite small.

If we could consistently rely on t_s and t_r to represent accurately the time of the onset of the stimulus and the time of the subject's response, respectively, this would be the end of the analysis. Unfortunately, this situation is quite improbable in a multiprogrammed computer system. We must now introduce additional error terms to account for the fact that we cannot record the current time at precisely the same time that an event occurs.

Let $\hat{t}_s$ represent the time at which the stimulus is actually presented, as opposed to t_s, the time at which the time of stimulus presentation is recorded. Similarly, define t_r to be the time at which the subject responds. We can now define the following approximations:

$$\hat{t}_s = t_s - d_s, d_s > 0, \text{ and} \qquad (10)$$

$$\hat{t}_r = t_r - d_r, d_r > 0. \qquad (11)$$

Note that by choosing $d_s > 0$ we have assumed, without loss of generality, that the stimulus is presented before the time of stimulus presentation is recorded. The analysis for the other case is similar. Obviously, there will be no attempt to record the time of the subject's response until the response is actually detected, so $d_r > 0$ represents the only reasonable case.

We can now look at the error in approximating $\hat{L}$, the actual response latency, where

$$\hat{L} = \hat{t}_r - \hat{t}_s, \qquad (12)$$

with L″, as defined by Equation 8. Substituting for t_s and t_r in Equations 10 and 11, we have

$$t''_s = \hat{t}_s + d_s - e'_s - e''_s, \text{ and} \qquad (13)$$

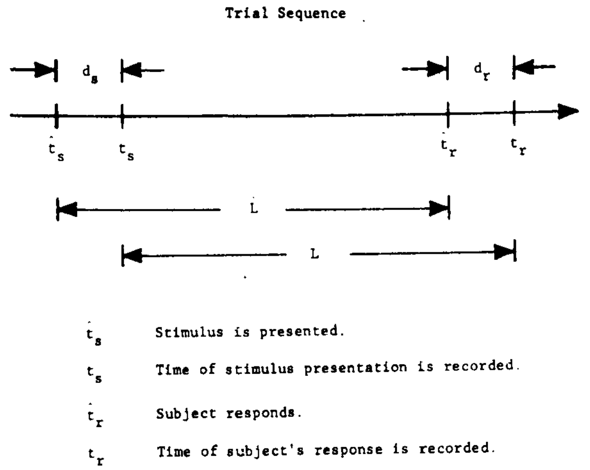

Trial Sequence
d_s
d_r
$\hat{t}_s$
t_s
$\hat{t}_r$
t_r
$\hat{L}$
L
$\hat{t}_s$ Stimulus is presented.
t_s Time of stimulus presentation is recorded.
$\hat{t}_r$ Subject responds.
t_r Time of subject's response is recorded.

Figure 2. Possible errors in measuring the time of occurrence of a stimulus or response.

$$t''_r = \hat{t}_r + d_r - e'_r - e''_r. \qquad (14)$$

Substituting for $\hat{t}_s$ and $\hat{t}_r$ in Equation 12 we have

$$L'' = \hat{L} + d_r - e'_r - e''_r - d_s + e'_s + e''_s \qquad (15)$$

Since the error terms d_s and d_r are nonnegative, the following bounds can be placed on the approximation of the actual response latency:

$$\hat{L} - d_s - e'_r - e''_r \leqslant L'' \leqslant \hat{L} + d_r + e'_s + e''_s. \quad (16)$$

In most computer systems, it will be possible to reduce d_s to a very small value by executing the sequence of instructions required to present a stimulus and to record the current time while in the "lock" or "inhibit" mode, so that no higher priority task will be able to interrupt the sequence. However, the unpredictability of the subject's response makes it impossible to do the same for d_r. Even if a response causes a very high priority interrupt to be activated, it is quite possible for responses from several subjects to arrive in less time than it takes to process them, causing one or more of the response tasks to have to wait to record the time of the corresponding subject's response. Several other factors may also serve to increase d_r, including tasks that must execute at a higher priority than the response tasks and lower priority programs executing in the "lock" or "inhibit" mode.

Ignoring those error terms that either can be made very small or that are very small with high probability, we arrive at the following bounds on the accuracy of the response latency measurement:

$$\hat{L} - e'_r \leqslant L'' \leqslant \hat{L} + d_r + e'_s \qquad (17)$$

This essentially shows that most of the error in measuring response latencies either can be attributed to the resolution of the digital clock or to delays in recording the time of the subject's response.

The preceding analysis is valid and complete for the tachistoscopic presentation of stimuli. An interesting exception occurs when using CRT terminals to display textual material as a stimulus. In this instance, a rather large error term is introduced because the text cannot be made to appear with the rapidity of a tachistoscopic display.

In CRT terminals that use the standard television interlace scan for the display, characters are usually generated by turning the CRT beam on and off according to dot matrix specifications for individual characters, which usually require seven-nine consecutive raster lines for each row of text. Since only half of the raster lines, either the even numbered lines or the odd numbered lines, are swept out during each 1/60 sec cycle, approximately 17-33 msec will be required to display a body of text, depending on the location and length of the text. For displays that scan by text row, and actually write individual characters on the CRT screen, the display time can vary from a few miscroseconds to approximately 33 msec, assuming the usual 1/60 sec refresh period, and again depending on the location and length of the text. It is important to note that the time required to display the text cannot be reduced for either type of display by simply transmitting the text to the terminal at a higher rate, though it can certainly be increased by doing the opposite.

This situation is further compounded by the fact that the text transmitted to the CRT terminal is unlikely to be synchronized with the refresh scan; in fact, it is very debatable whether there can be any meaningful synchronization with the television interlace scan. We now have a stimulus that not only appears very slowly, but that may start appearing at any place in the body of text, perhaps even by character segments. This makes it difficult to state the exact moment of stimulus presentation.

A partial solution to this problem exists for displays in which the CRT beam can be turned on and off under program control. First, the CRT beam is turned off while the text is being transmitted to the CRT terminal. Then, when the blanked CRT beam reaches a fixed point in its refresh scan, it is turned on and generation of the textual display begins. While this does not reduce the time required to generate the display, it does make the display generation predictable and repeatable. In the case of a display where each character is generated as a separate entity, it also makes it possible to time the generation of the display to within a few microseconds. If the time of stimulus presentation is consistently taken to be the time at which the CRT beam was turned on, the experimental results should differ from tachistoscopic presentation of the same material by only a small linear shift in the response latencies.

CASE STUDY

We will now apply the results of the preceding analysis to a specific experiment program in use at the CLIPR laboratory. This program presents stimuli on a Univac Uniscope 100 CRT terminal and records responses on a five-level pushbutton box. The program runs on a Xerox Sigma 3 Computer, and is very dependent on the interrupt structure of the Sigma 3 for accurate measurement of response latencies.

The time base for response latency measurement is a free running counter that is incremented at 1-msec intervals by a high priority, interrupt driven program. The counter update signals are supplied by an internal, crystal controlled clock. Actual measurement of the delays in responding to the counter update signals shows that the synchronization error is always less than 1 msec, and that the probability is less than .005 that the error will exceed 125 microsec. Using Equation 9; the maximum error that will result from using the digital counter to measure response latencies is given by

$$L - 2\ \text{msec} < L'' < L + 2\ \text{msec}$$

or

$$| L - L'' | < 2\ \text{msec}$$

including the worst possible counter synchronization error.

The Univac terminal has been modified by the CLIPR staff to enable the CRT beam to be turned on and off under program control. When the "beam on" signal is transmitted to the terminal, the next center-of-screen condition in the CRT refresh cycle causes the terminal to turn the CRT beam on and signal the computer that the stimulus has been presented. This signal causes a very high priority interrupt to be activated, the same interrupt that is activated by the subject's response, so that the time of stimulus presentation can be recorded. The "beam off" signal causes the CRT beam to be turned off immediately.

By computing the execution time of all tasks that could possibly delay execution of the task that records the time of the stimulus presentation and the time of the subject's response, maximum values for d_s and d_r can be established. The maximum delay at CLIPR is slightly less than 1 msec. From Equation 16, the maximum error in response latency measurement becomes

$$\hat{L} - 3\ \text{msec} < L'' < \hat{L} + 3\ \text{msec}$$

or

$$| \hat{L} - L'' | < 3\ \text{msec}$$

This represents the worst case situation for errors in response latency measurement and includes terms that are very unlikely to appear under normal operating conditions. The program used for this case study was tested while running six subjects, and while the computer was being subjected to the heaviest possible load, and the actual response latencies were measured to the nearest 10 microsec with an electronic counter of known accuracy and were recorded. The absolute measurement errors were approximately uniformally distributed over the range $0 < e < 1$ msec, the period of the counter. As the result of these analyses and tests, we are able to guarantee a response latency measurement accuracy of ± 3 msec and to state that, with high probability, the error is ± 1 msec.