How are Windows Used? Some Notes on Creating an Empirically-Based Windowing Benchmark Task

Kenneth B. Gaylin

Human Factors Laboratory
Virginia Polytechnic Institute and State University
Blacksburg, Virginia 24061

ABSTRACT

Users of a windowing system were studied for the purpose of creating an empirically based windowing benchmark. Each filled out a paper questionnaire that sampled subjective opinions of windowing commands, and were observed for approximately 22 minutes while performing typical daily activities on the computer. Subjects were also asked to demonstrate a typical log-on procedure and were personally interviewed. Windowing command frequencies, and screen layout characteristics were collected and analyzed. The data revealed a relatively high use of a small number of commands that were primarily concerned with moving between windows. This study enabled the creation of a more accurate windowing benchmark task.

1. INTRODUCTION

Benchmarking has been used to compare computer speed and accuracy for more than 20 years (Lewis and Crews, 1985), but it has only recently been applied to the assessment of the human/computer interface. In this respect, benchmarking is simply the process of determining the performance of an appropriate sample of users on a standard set of tasks. Specifically, one is concerned with how well users perform the required tasks given a particular interface. The most notable application was that of Roberts (1979), and Roberts and Moran (1983) for their use of benchmarking to evaluate the performance and functionality of text editors. Borenstein (1985) provides a useful assessment of the Roberts and Moran methodology and gives some guidelines for suggested improvements. Benchmarks have also been succussfully used to evaluate other aspects of the user interface. For example, Whiteside, Jones, Levy, and Wixon (1985) used a file manipulation task developed by Magers (1983) to evaluate performance on 7 different interfaces.

To be an effective design aid, a benchmark must extrapolate well to the tasks that are being performed in the work setting. The more accurately a benchmark resembles the real world environment, the better its predictions of real world performance. This report describes the collection of windowing usage data that were used in the creation of an empirically-based windowing benchmark task. Due to space limitations, readers are referred to Gaylin (1985) for the actual benchmark tasks and a more detailed discussion of their creation and use.

2. METHOD

2.1 Subjects

Nine experienced computer users participated in the study. All subjects were Digital Equipment Corporation employees using a workstation with windowing capabilities. Four other subjects volunteered to participate but were rejected because they did not use the windowing capabilities of the system. One of the nine subjects stated that he was not a regular user of windows. Subjective ratings of the mean number of hours per day spent actively working on the computer was 5.6 hours (standard deviation = 1.7 hours) with a low of 3.0 and a high of 8.0 hours. Eight of the nine subjects stated that their primary duties centered around computer programming, the one exception being a technical writer involved with the construction of user manuals.

2.2 Equipment

The windowing workstation consisted of a large, high resolution, bit-mapped video display unit (VDU), a mouse, and keyboard. The major windowing capabilities of the system included:

- Creating windows
- Deleting windows
- Moving windows
- Attaching the keyboard to a window
- Popping windows to the front
- Pushing windows to the back
- Resizing windows
- Printing windows

2.3 Procedure

All subjects were videotaped in their own office at a prearranged time. Before starting, subjects were required to read and sign a statement of informed consent, and a photographic release form.

© 1986 ACM 0-89791-180-6/86/0400 - 0096 $00.75

They were then given a questionnaire which asked them to rate a number of windowing commands on several parameters.

The video equipment was set up in back of the subject and to their left or right side to maintain a clear view of the VDU. Videotaping began upon the completion of the paper survey and lasted approximately 22.5 minutes (the length of one videotape). Subjects were asked to do typical daily work at the terminal and ignore the presence of the experimenter as best as possible. Videotaping was stopped during telephone calls or interruptions from other employees.

At the end of the observation period subjects were interviewed for approximately 10 - 15 minutes. Each was asked a similar set of questions regarding their use of windows. Subjects were also asked to demonstrate how they typically log-on to the computer and set up their screen. Total participation time for each subject was approximately one hour. Videotapes were later reviewed to determine the type and frequency of command use.

3. RESULTS

3.1 Observation Period Results

Windowing command frequencies were totaled across all subjects and are depicted in Figure 1. A total of 254 window related commands were observed across the 9 subjects. The most frequently used command was "cycling through the windows" which used a function key to attach the keyboard to a window, or bring a particular window to the forefront. Two other methods could have also been used to achieve these results. These were a "mouse activated pop" which required the user to place a mouse controlled "arrow" icon inside the border of the desired window and depress a mouse key, and a "menu activated pop" which was activated by selecting a menu icon located in the upper left corner of the desired window, and then selecting the option to pop the widow in front of other windows. When these three methods were summed they accounted for 59.5 percent of the total windowing commands used. Switching between split screen edit sessions (or edit buffers) was also logged because of its similarity to switching between windows and accounted for 3.5 percent of the commands used. When added to the above categories, the number of commands used to switch the active work area climbed to 63.0 percent. Thus most of the windowing commands involved switching between windows as opposed to the creation, deletion, or manipulation of window size and location on the VDU.

```
Command                                           Percent  Mean   SD
                    |
Cycle thru windows  |**************************     50.8   14.3  12.2
Menu window select. |*****                          10.6    3.0   5.7
Mouse activated pop |****                            8.3    2.3   3.4
Create window       |****                            7.5    2.1   2.8
Delete window       |***                             6.3    1.8   2.8
Move window         |**                              4.3    1.2   1.6
Dual edit session   |**                              4.3    1.2   1.8
Switch edit buffers |**                              3.5    1.0   1.6
Resize window       |*                               2.0    0.6   1.3
Delete menu window  |*                               2.0    0.6   1.7
Menu activated pop  |                                0.4    0.1   0.3
                    ----+---+---+---+---+---+---+
                      20  40  60  80  100 120 130

                          Command frequency
```

Figure 1. Observed window command frequencies

Figure 1 also shows the mean and standard deviation for commands used in the 22.5 minute period. In all but one instance the standard deviations exceeded the means, indicating a large amount of variation of command use between subjects.

In addition to window commands, data were collected on the amount and type of windows used and is shown in Figure 2. ALL WINDOWS refers to any type of window used, with the exclusion of menu windows. The other categories are a subset of this category. GENERAL PURPOSE windows were regular windows that were used for a variety of purposes. INFORMATIONAL windows were any that were used for status information such as the current directory or file being edited. A CLOCK WINDOW was fairly popular among users and usually contained a small (4 cm. x 4 cm.) clock with an analog display. An INACTIVE WINDOW was any general purpose window that was either reduced to the size of an icon or saved on the screen for use if necessary, but was not being actively used.

```
Window type                                                Mean  SD
                |
All windows     |***********************************       3.7   1.9
                |
General purpose |***********************                   2.4   0.9
Informational   |*****                                     0.7   1.7
Clock window    |***                                       0.3   0.5
Inactive window |**                                        0.2   0.4
                ---------+---------+---------+--------
                         1         2         3

                           Mean windows
```

Figure 2. Mean number of windows used in the 22.5 minute observation period

All frequency counts, which were the basis for these means, were a measure of the maximum number of windows present at one time during the 22.5 minute observation period. In actuality then, these values will be somewhat inflated because not all persons maintained the maximum number of windows on the VDU for their entire observation period. However, most of the subjects maintained relatively stable amounts of windows throughout the observation period.

From the data it can be seen that there was a mean of 3.7 windows maintained on each VDU, with 2.4 being GENERAL PURPOSE windows. Standard deviations for all categories are again high, indicating a large amount of variation of the number of windows used between subjects. The CLOCK WINDOW category is the result of three persons using one clock window each and the INACTIVE WINDOW category results from two persons, each having one inactive window. The INFORMATIONAL window category resulted from one person using five, and a second person using one informational window.

Subjective responses (collected in the paper survey) pertaining to the average, as well as the minimum and maximum number of windows typically used while working on the computer are shown in Figure 3. In general, subjects rated using an average of a little less than 3 windows, with a mean maximum of

```
                                                N    Mean   SD
                |
Max windows     |**************************     8    5.0    2.2
Avg windows     |***************                8    2.9    1.1
Min windows     |**********                     8    1.8    1.4
                -----+----+----+----+----+---
                     1    2    3    4    5

                      Mean windows
```

Figure 3. Subjective assessment of the amounts of windows used

5, and a mean minimum of 1.8 windows.

3.2 Observed Log-on Commands

Observation of log-on procedures revealed the command frequencies shown in Figure 4. The log-on period was defined as that required by users to initially access the computer accounts available to them, and setup their terminal in a typical work configuration. The most frequently used command continued to be "cycle through windows." However, some other commands had much higher percentages of use than in the 22.5 minute observation period. Mean setup time was 122.6 seconds (SD = 121.7 seconds). In general, subjects indicated that their initial screen setup remained unchanged throughout most of the day.

```
Command                                Percent  Mean  SD
                   |
Cycle thru windows |********************* 32.4  4.3   7.4
Menu selection     |************          18.1  2.4   1.6
Move window        |***********           17.1  2.3   1.8
Create window      |***********           17.1  2.3   1.6
Resize window      |********              12.4  1.6   1.9
Delete window      |**                     2.9  0.4   1.1
                   ------+----+----+----+-
                         1    2    3    4

                      Mean logon commands
```

Figure 4. Mean observed log-on command frequencies

3.3 Paper Survey Results

Subjects were asked to rate 17 window-related commands on the basis of six major categories: 1) Frequency of use, 2) Usefulness, 3) Friendliness, 4) Complexity, 5) Naturalness, and 6) Importance. A seven point bipolar scale was used with -3 being the worst rating and +3 the best rating, zero was neutral. The purpose of the paper survey was to obtain data on commands that might be too infrequently used to be sampled in 22.5 minute observation period. In this regard the most pertinent rating categories for the creation of a benchmark were frequency of use, importance, and usefulness. The paper survey results for these categories are shown in Figures 5 - 7 respectively. The ratings for the other three categories can be found in Gaylin (1985). The commands that people were asked to rate are located on the far left of each graph. N refers to the number of persons who responded to that command definition. Caution should be exercised when evaluating some of the ratings because they may be based on a small number of respondents (for example, SET SCREEN DEFAULTS).

```
Command                                             N   Mean  SD

Cycle thru windows                 |**************  4   2.8   0.5
Scroll in window                   |*************   6   2.7   0.8
Create window                      |************    8   2.4   0.7
Paginate in window                 |**********      6   2.0   1.5
Delete window                      |*********       8   1.8   1.8
Put window in front                |*********       8   1.8   1.4
Move window                        |********        8   1.6   1.3
Rescale window                     |*******         3   1.3   0.6
Set screen defaults                |*****           1   1.0    .
Resize window                      |*****           8   1.0   1.1
Jump a windows view                |*****           7   1.0   1.8
Dual edit session                  |*****           8   1.0   0.8
Use a menu window                  |****            5   0.8   2.3
Copy between windows               |****            5   0.8   2.7
Put window in back               **|                7  -0.4   2.4
Change a window label       *****|                  5  -1.0   2.3
Use a help window           ******|                 7  -1.1   2.0
Print an area              *******|                 7  -1.4   1.7
                   -+----+----+----+----+----+----+
                   -3   -2   -1    0    1    2    3
          Extremely                               Extremely
          infrequently                            frequently
```

Figure 5. Frequency ratings

```
Command                                             N   Mean  SD

Cycle thru windows                 |*************** 4   3.0   0.0
Put window front                   |**************  8   2.9   0.4
Move window                        |**************  8   2.8   0.5
Create window                      |*************   8   2.6   0.5
Scroll in a window                 |*************   6   2.5   1.2
Delete window                      |*************   8   2.5   0.8
Paginate in a window               |***********     7   2.3   1.1
Jump a windows view                |***********     7   2.3   1.0
Set screen defaults                |**********      1   2.0    .
Resize a window                    |**********      8   2.0   1.1
Copy between windows               |**********      6   2.0   1.3
Put window in back                 |**********      6   2.0   1.3
Print an area                      |*********       6   1.8   1.6
Use a menu window                  |********        5   1.6   1.1
Rescale window                     |********        4   1.5   1.3
Use a help window                  |******          7   1.3   1.0
Change a window label              |*               5   0.2   1.6
                   -+----+----+----+----+----+----+
                   -3   -2   -1    0    1    2    3
          Extremely                               Extremely
          useless                                 useful
```

Figure 6. Usefulness ratings

```
Command                                             N   Mean  SD

Put window in front                |**************  8   2.9   0.4
Cycle thru windows                 |**************  4   2.8   0.5
Scroll in a window                 |*************   6   2.7   0.8
Move window                        |*************   8   2.6   0.5
Delete window                      |************    8   2.4   0.9
Create window                      |***********     8   2.3   1.8
Use a menu window                  |***********     5   2.2   1.3
Set screen defaults                |**********      1   2.0    .
Jump a windows view                |**********      7   2.0   1.4
Resize window                      |*********       8   1.9   0.8
Paginate in a window               |*********       7   1.9   1.5
Use a help window                  |*********       7   1.9   1.1
Copy between windows               |*********       6   1.8   1.2
Print an area                      |********        6   1.5   1.4
Put window in back                 |******          6   1.2   1.9
Rescale window                     |*****           6   1.0   0.9
Change a window label              |                6   0.0   1.3
                   -+----+----+----+----+----+----+
                   -3   -2   -1    0    1    2    3
          Extremely                               Extremely
          unimportant                             important
```

Figure 7. Importance ratings

Subjects were asked to rate a command even if it was functional only for a specific application within a window. This enabled the capture of commands used in dual edit sessions that were felt to be similar to windowing features currently available in other products. For example, moving text from one edit window to another. If a command was not currently available on the system, subjects were asked to rate how important and useful they thought that command might be.

To determine whether the frequency of use of a command had been influenced by its relative ease or difficulty of use, the three rating scales consisting of friendliness, complexity, and naturalness were analyzed. That is, were some commands infrequently used because they were too unfriendly or complex? Or were they rarely used because they were useless or unimportant? If the former was true then the command might be used more frequently if it was better designed. If the command was rated as being easy to use (friendly, natural, simple) and was still unused then it was probably not important.

To this end, Pearson Product Moment correlations were calculated between all rating categories in the paper survey, and between rating categories and observed command frequencies (see Table 1). Low positive correlations were found between the ease of use categories and both the observed and rated frequency of use of a command. This indicates that poor ease of use probably had little effect on either the observed or rated frequencies of command use. In addition, observed and rated frequency of use was much more highly related to the "usefulness" and "importance" categories.

	Usefulness rating	Friendliness rating	Complexity rating	Naturalness rating	Importance rating	Observed command frequency
Frequency rating	0.71 0.0001	0.22 0.0291	0.32 0.0011	0.39 0.0001	0.61 0.0001	0.23 0.0713
Usefulness rating		0.18 0.0844	0.31 0.0022	0.37 0.0002	0.79 0.0001	0.29 0.0291
Friendliness rating			0.54 0.0001	0.55 0.0001	0.26 0.0089	0.15 0.2623
Complexity rating				0.65 0.0001	0.41 0.0001	0.11 0.4386
Naturalness rating					0.28 0.0050	0.14 0.2949
Importance rating						0.30 0.0225

KEY:
Correlation Coefficient =
Probability of Occurance =

Table 1. Pearson correlation coefficients

Correlation analyses were also used to determine whether the subjective data obtained on the paper survey could be used as an accurate predictor of command frequency of use for those commands not sampled in direct observation of a subject's work, or if collecting subjective ratings would obviate the need to obtain frequency data of command usage. If there was a high degree of correlation between observed, and subjective ratings of command frequencies then subjects could be assumed to be accurate predictors of command usage. The correlation between each subject's observed and rated command frequencies, however, was low (r = 0.23, p = 0.07), and was not statistically significant.

The observed command frequencies summed across subjects was also compared to the mean of the subjective frequency rating scores. The results showed a much higher correlation between these two values (r = 0.64, p = 0.06), indicating that although individual subjects were not very good at predicting their use of windowing commands, mean rating scores did reflect command usage to some degree, and would probably be useful where more accurate information is unavailable.

3.4 Benchmark Construction

Construction of the benchmark was primarily based upon the observed command frequency of use data. Preserving the relative frequencies of the different methods of overlapping windows (mouse, function key, menu activated pops, and menu activated pushes) which could be used interchangeably was accomplished by creating task situations which were difficult to complete without using a specific technique. Although this leaves some room for the frequency of these commands to vary, specifying the use of a particular technique would have made the benchmark inapplicable to products lacking that method, and would introduce bias in the form of aiding less knowledgeable users.

To ensure that the benchmark accurately reflected the relative frequencies of observed commands, two experienced users were tested. The frequency of window commands that they used were found to be highly correlated with the previously collected observational data (r = 0.95). This is a large improvement over the previously used windowing benchmark task which was less highly correlated with observed command use (r = 0.31, p = 0.38). This poor correlation coefficient is primarily due to the large difference in the amount of moving between windows that was actually used (observed), and that expected by expert performance on the previous windowing benchmark. When commands involving moving between windows were removed from the analysis the correlation coefficient became much greater (r = 0.64, p = 0.17).

It is notable that the correlation coefficient of the previous version of the windowing benchmark task with subjective ratings of command frequency of use is higher (r = 0.43, p = 0.29) than that of the previous benchmark task and observed command frequency of use. This may be related to the fact that the command types and frequencies included in the former windowing benchmark were based upon the subjective opinions of several human factors engineering staff. Thus, this data tends to further reinforce the need to apply objective measures when creating benchmark tasks.

4. DISCUSSION

Observational data revealed a relatively high use of a small number of commands that are primarily concerned with moving between windows. In general there were two reasons for moving between windows. Some subjects maintained windows that were dedicated to a specific function such as receiving mail, a calendar program, or a local versus mainframe process, and would move between windows to access these functions. Secondly, waiting for program control to return from activities which are CPU intensive such as compilations, or copying large numbers of files, was usually avoided by continuing work in a new window.

Observation of log-on procedures indicated that creating, resizing, and moving windows is much more common during initial setup as opposed to post setup work. Most resizing involved making a window larger so that more lines of text were visible. There was no general trend regarding window placement except for a tendency to avoid partially overlapping windows due to a technical problem that caused slower scrolling speeds. Some subjects stated that they might use more windows were this problem to be corrected.

In the past, benchmarks have not always been created by using objective techniques. Often benchmark designers feel that they know which functions of the interface are used the most. The low correlations between observed command usage, and ratings of command frequencies (r = 0.22) strongly indicate that subjective opinions are not very accurate, and thus should be avoided where possible. This study seems to indicate that observational data can be a relatively fast and simple method of constructing a realistic set of benchmark tasks. The added initial time and effort should eventually aid the iterative design process by more accurate and valid predictions of interface problems, and evaluation of proposed solutions.

The constructed benchmark has several limitations, however. Foremost, it is based upon the command frequencies of a small number of subjects using a particular piece of hardware/software. No attempt was made to find a random, representative sample from the population of all window users and available windowing products. In addition, the system tested lacked many commands, such as copying between windows, scrolling, or paginating the view of a window, that are currently available on other systems. Future research in this area should be directed at sampling the types of work and commands used on these other systems to validate or modify the current benchmark.

ACKNOWLEDGMENTS

Special thanks to John Whiteside and Robert Williges for their guidance and support of this study. Alana Brassard, and Sandy Jones helped with subject recruitment. Thanks also to Jim Burrows, Charlie Frean, Michael Good, Tom Spine, Dennis Wixon, and other members of the Software Human Engineering Group at DEC for their assistance and advice. This research was made possible through the Human Factors Internship Program which is funded by Digital Equipment Corporation. The views expressed in this article are the authors' and in no way represent the policy of Digital Equipment Corporation.

REFERENCES

Borenstein, N.S. (1985). The evaluation of text editors: A critical review of the Roberts and Moran methodology based on new experiments. In Proc. CHI'85 Human Factors in Computing Systems. (pp. 99-105). New York: ACM.

Gaylin, K. B. (1985). Creating an empirically-based windowing benchmark task. DEC-TR 370, Spitbrook, N.H: Digital Equipment Corporation.

Lewis, B. and Crews, A. (1985). The evolution of benchmarking as a computer performance evaluation technique. MIS Quarterly, March, 7-15.

Magers, C.E. (1983). An experimental evaluation of on-line help for non-programmers. In Proc. CHI'83 Human Factors in Computing Systems. (pp. 277-281). New York: ACM.

Roberts, T.L. (1979). Evaluation of computer text editors (Report SSL-79-9). Palo Alto, California: Xerox PARC.

Roberts, T.L., and Moran, T.P. (1983). The evaluation of text editors: Methodology and empirical results. Communications of the ACM, 26, 4, 265-283.

Whiteside, J., Jones, S., Levy, P., and Wixon, D. (1985). User performance with command, menu, and iconic interfaces. In Proc. CHI'85 Human Factors in Computing Systems. (pp. 185-191). New York: ACM.