THE PRECIS OF PROJECT ERNESTINE OR AN OVERVIEW OF A VALIDATION OF GOMS

Wayne D. Gray, [1] *Bonnie E. John,* [2] *& Michael E. Atwood* [1]

1. NYNEX Science & Technology Center

2. Carnegie Mellon University

KEYWORDS: GOMS, analysis methods, empirical studies, user models, cognitive models, methods for analysis/assessment, prototyping, protocol analysis, theory in HCI

INTRODUCTION

Project Ernestine served a pragmatic as well as a scientific goal: to compare the worktimes of telephone company toll and assistance operators on two different workstations, and to test the validity of GOMS[1] models for predicting and explaining real-world performance. Contrary to expectations, GOMS predicted and the data confirmed, that performance with the proposed workstation was slower than with the current one. Pragmaticly, this increase in performance time translates into a cost of $2.4 million dollars a year to NYNEX. Scientificly, the GOMS models predicted performance with exceptional accuracy.

The empirical data provided us with three interesting results: proof that the new workstation was slower than the old, proof that this difference was not constant but varied with type of call, and no evidence of learning in data that spanned four months and 78,240 phone calls. The GOMS models predicted the first two results and explained all three.

It is important to emphasize that the two major parts of Project Ernestine, the field trial and the GOMS analyses, were done separately and during the same time period. It is NOT the case that the GOMS models were built with knowledge of the empirical data. Also, at the time of this writing, we have not observed a single toll & assistance operator (TAO) using the new workstation.

This paper is the final installment in a series of CHI and other presentations summarizing a major test of the applicability of GOMS to real-world design problems (Gray, et al., 1989, 1990a, 1990b; John 1990). We provide an overview of the methodology of the study, the empirical data, and the GOMS models. This report is a prēcis only; the complete report on Project Ernestine is provided by Gray, John, & Atwood (submitted). Additional insights into conducting empirical studies and doing analytic modeling in the real-world are provided in Atwood, Gray, & John (submitted).

THE TASK & WORKSTATIONS

The TAO is the operator you get when you dial 0. Their job is to assist the customer in completing calls and to record the correct billing. Among other tasks, TAOs handle person-to-person calls, collect calls, calling-card calls, and calls billed to a third number. The TAO does not handle Directory Assistance calls.

Two TAO workstations were evaluated - the *current* workstation and a *proposed* workstation. The *current* workstation had been in use for several years and employed a 300-baud, character-oriented display and a keyboard on which functionally-related keys were color coded and spatially grouped. This functional grouping often separated common sequence of keys by large distances on the keyboard.

In contrast, the *proposed* workstation was ergonomically designed with sequential as well as functional considerations. The graphic, high-resolution display operated at 1200-baud, used icons and, in general, is a good example of a graphical user interface whose designers paid careful attention to human-computer interaction issues. For example, when the phone being called is ringing, an icon of a telephone with its receiver on-hook appears next to the called number; when the phone is answered, the icon changes to a telephone with its receiver lying next to it. In the *current* workstation, this

[1] Goals, operators, methods, & selection rules (Card, Moran, & Newell, 1980; 1983).

Permission to copy without fee all or part of this material is granted provided that the copies are not made or distributed for direct commercial advantage, the ACM copyright notice and the title of the publication and its date appear, and notice is given that copying is by permission of the Association for Computing Machinery To copy otherwise, or to republish, requires a fee and/or specific permission.

© 1992 ACM 0-89791-513-5/92/0005-0307 1.50

is indicated by the ASCII characters "CLD 1" (standing for CalLeD line 1) appearing far away from the called number, in the lower part of the screen. Similar care went into the design of the keyboard, where an effort was made to minimize travel distance among the most frequent key sequences and to reduce the number of keystrokes required to complete a call by replacing common two-key sequences with a single function key.

THE FIELD TRIAL

Methodology

Participants

The phone company office used in the study employs over 100 TAOs and handles traffic in the Boston, Massachusetts area. For purposes of the study, 12 *current* workstations were removed and 12 *proposed* workstations installed.

All participants were New England Telephone (NET) employees who had worked as TAOs for a minimum of two years. Twenty-four participants were selected for the *proposed* workstations (the *proposed* condition) from a list of approximately 60 volunteers. Each *proposed* participant was paired with a control participant matching for shift worked (that is, time of day), and average worktime on the *current* workstation (the *current* condition). TAO worktimes were taken from data routinely collected by office managers for the six months prior to the start of the trial (while both groups were using the *current* workstation).

Trial Procedures

Proposed and *current* participants worked their normal shifts during the four month trial. From the perspective of the proposed participants their tasks and duties as a TAO were identical to their pretrial job in all respects but one; namely, a new workstation was used. For the *current* participants nothing had changed. To obtain our data, we extracted the calls handled by our 24 *proposed* and 24 *current* participants from a NYNEX database that routinely samples one out of every ten calls.

Call Categories

To concentrate our effort for both the empirical and analytic comparisons we decided to focus on calls categories that were either high volume or of special interest to NYNEX Operator Services. The final list of 20 call categories accounted for 88.33% of all completed calls. This percentage is based upon one month's frequency data for all calls handled by all NET TAOs.

Results

For the 48 TAOs (24 *proposed* and 24 *current*) over the four months of the study, the 20 call categories sampled a total of 78,240 calls. Five call categories were eliminated for some analyses due to insufficient occurrence of those call categories[2].

Collapsing over call category to look at the median work time per call for each participant, for each month, the data show that the *proposed* group is slower than the *current* group by 4%; that is, the *proposed* workstation requires 0.8 seconds more time on an average call than does the *current* workstation. This 0.8 seconds is both statistically[3] and financially significant. The 0.8 second deficit translates into a cost of $2.4 million a year in additional operating costs if the *proposed* workstation were to be installed across the NYNEX operating area.

Reflecting seasonal variations in call-mix, the main effect of month is significant, but not the interaction of groups by month. This lack of a significant interaction suggests that TAOs using the *proposed* workstation mastered it very quickly, reaching asymptotic performance within the first month of performance.

For the analysis by call category we looked at the 15 sufficiently represented call categories. This analysis yielded significant effects of group, call category, and their interaction. The effect of call category was expected due to the different nature of the calls. The interaction shows that the advantage of the *current* workstation over the *proposed* is not constant for all call categories. For some call categories this difference is small (0.2 seconds) while for others it is quite large (3.7 seconds). This is an interesting result that cannot be explained by the field data. This is a result on which analytic models may shed much light.

ANALYTIC MODELING

Benchmark Tasks

Rather than model every possible procedure executed by the TAOs, for each call category we modeled one common, or important, variation. With the help of NYNEX Operator Services personnel, we wrote a single script for each of the 20 call categories originally chosen for study and used these as benchmarks. These benchmarks were validated against observed calls and were found to be representative of the average work time for these call categories.

The GOMS Models

To model these benchmarks, two different approaches were

[2] Note: additional statistical analysis, as well as details of those reported here are available in Gray et al., submitted.

[3] The level of significance choosen for this report is $p < .05$.

used: observation-based models and specification-based models. The models of the *current* workstation were based upon videotapes of experienced TAOs handling calls for each of the benchmark tasks. In contrast, TAOs were never observed using the *proposed* workstation. Rather, these models were specification-based; that is, they were constructed based upon system response time estimates and TAO procedures provided by the manufacturer.

TAO's do several things in parallel when processing a customer's request: they listen or talk to the customer, they perceive information on the CRT screen, they move their hands to appropriate keys and strike them. To display these parallel activities and calculate total task times, we use the ***critical path method*** (developed for project management). Because this extension distinguishes between **C**ognition, **P**erception, and **M**otor operators and uses the **C**ritical **P**ath **M**ethod, we call it **CPM-GOMS** (John, 1988).

In CPM-GOMS the parallelism of the TAO's task is represented in a *schedule chart* (Figures 1 and 2). Each activity in handling a call is represented as a box with an associated duration. Dependencies between activities are represented as lines connecting the boxes. For example, the TAO cannot hit the *collect-billing* key until s/he hears the customer request a collect call. Therefore, there is a dependency line drawn between the box representing the perception of the word "collect" and the boxes representing the cognitive operators that verify the word "collect" and initiate pressing the *collect-billing* key. The boxes and their dependency lines are drawn according to a detailed understanding of the TAO's task, goal decomposition, and operator-placement heuristics (John, 1990).

An important concept in analyzing the total task time for complex parallel tasks is the *critical path*. When activities occur in parallel, one sequence of activities will take more time than parallel sequences of activities; the critical path is the sequence of activities that takes the longest and determines the total time for the entire task. The critical path is displayed in boldface in Figures 1 & 2.

Each schedule chart is the CPM-GOMS model of the call it depicts. We constructed 30 such models, yielding performance predictions for the 15 benchmarks on the *current* workstation and the 15 benchmarks on the *proposed* workstation, corresponding to the 15 call categories analyzed in the empirical data.

Workstation design features and call handling procedures have an impact on the length of a call which are reflected in the critical path. For example, Figures 1 and 2 show the first and last segments of a CPM-GOMS analysis for one 15 second calling-card call for both the *current* and *proposed* workstations. Figure 1 has two striking features. First, the analysis for the *proposed* workstation has 10 fewer boxes than the analysis for the *current*, representing two fewer keystrokes. Second, none of the deleted boxes were on the critical path, all were performed in slack time. At this point in the task the critical path is determined by the TAO greeting and getting information from the customer. Removing keystrokes that occur during slack time does nothing to affect the TAO's work time; that is, work time is controlled by the conversation, not by the keystrokes and not by the ergonomics of the keyboard.

For the *proposed* workstation one of the keystrokes eliminated at the beginning of the call (Figure 1) now occurs later in the call (Figure 2). In this analysis, the keystroke goes from being performed during slack time, to being performed on the critical path. As a result, the cognitive and motor time required for this keystroke now adds to the time required to process this call and CPM-GOMS predicts that, for this call category, despite requiring one less keystroke than the *current*, the *proposed* will require more time.

CPM-GOMS Predictions versus the Trial Data

After four months of real-world use, during which we sampled 78,240 calls, the trial data showed that the *proposed* workstation was 4% slower that the *current*. Was this result predicted by the CPM-GOMS models? Despite the fact that the trial result was surprising, the answer is *yes*.

When each of the 15 call categories is weighted by its frequency of occurrence in the trial data the CPM-GOMS models predict that the *proposed* will be 3% slower than *current* workstation. Looking at each of the 15 call categories that were analyzed, the correlation between predicted and observed work times was significant for both the *current*, r^2=0.69, and the *proposed* workstation, r^2=0.65, showing that the models adequately reflected the variation in work time as a function of call type.

Models as Explanation

Including the CPM-GOMS modeling effort in Project Ernestine had one welcome, but unanticipated result. The trial data were so counterintuitive that, in the absence of a compelling explanation as to why the *proposed* workstation was slower than the *current*, there was a tendency to try to find fault with the trial rather than with the workstation.

The manufacturer had predicted that the *proposed* workstation would be, on average, 2 seconds faster than the *current* workstation. Given the general expectation that an ergonomically engineered, modern workstation should be faster than a five year old, ergonomically indifferent one, this estimate seemed reasonable. When the trial data began to accumulate, the immediate and widely-held conclusion was that something (training, procedures, or equipment)

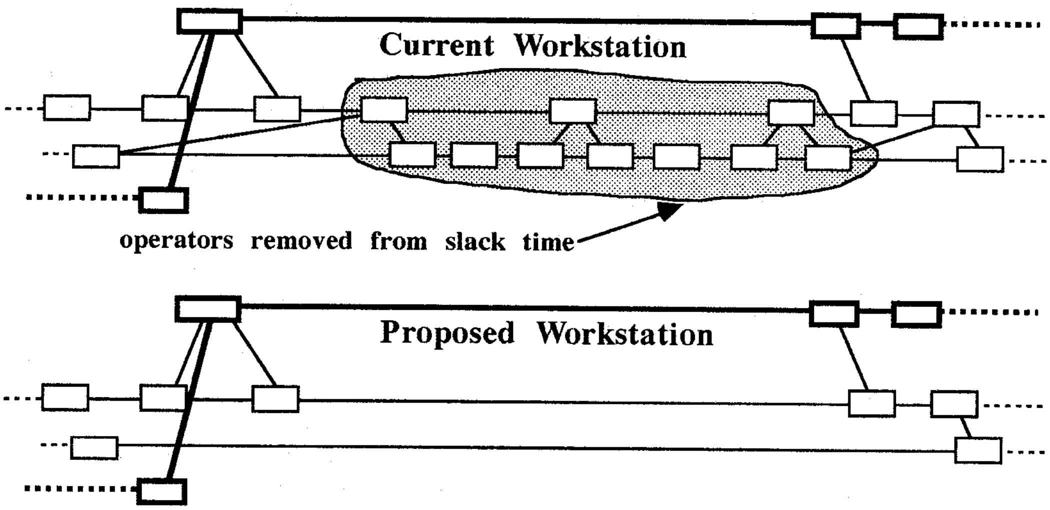

Figure 1. Section of CPM-GOMS analysis from near the beginning of the call. Notice that the proposed workstation (bottom) has removed two keystrokes (which had required 7 motor and 3 cognitive operators) from this part of the call. However, none of the ten operators removed were along the critical path (shown in bold).

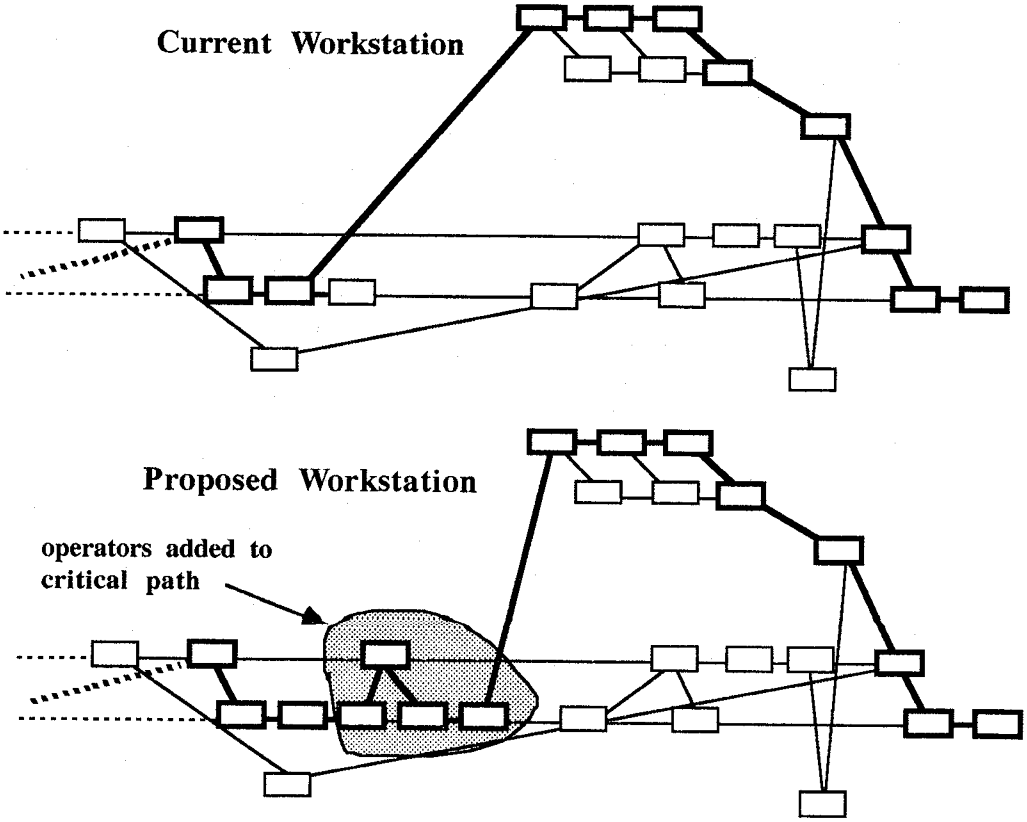

Figure 2. Section of CPM-GOMS analysis from the end of the call. Notice that the proposed workstation (bottom) has added one keystroke to this part of the call which results in four operators (3 motor and 1 cognitive being added to the critical path (shown in bold).

was wrong with the trial.

An initial model of just one call category (Gray et al., 1989) showed us why this *general expectation* was wrong. For example, as discussed above (Figures 1 and 2), although the *proposed* workstation generally had fewer keystrokes than the *current* workstation, the new procedure put more keystrokes on the critical path, rather than in the slack time, increasing the length of the call. In addition, the close spacing of the function keys on the *proposed* keyboard encouraged the use of the right hand for pressing all keys; CPM-GOMS predicted that this would be slower than the old procedure of using the left hand for certain keys that was encouraged by the layout of the old keyboard. Also, while the *proposed* workstation was faster than the *current* workstation in displaying a whole screen of information or in outpulsing large numbers of digits (as in a 14-digit calling card number), the *current* workstation begins displaying information sooner and outpulses a single digit (as for a function key) faster. Since many of these system event are on the critical path, they add to the average worktime of the *proposed* workstation.

Better designed displays provide no advantage to the TAO who knows what information to look for and where it will be displayed. Similarly, a better designed keyboard that reduces the time to move from key to key does not provide a work time advantage when those movements are not on the critical path but provides a deficit when they are.

CONCLUSION

The CPM-GOMS models predicted the empirical field data with remarkable accuracy. In addition, when the empirical data yielded the counter-intuitive result of new technology slower than old technology, GOMS saved the day by explaining why this result occurred. We believe this study validates the usefulness of GOMS models for evaluating real-world systems. Further, this study indicates that GOMS can be used to evaluate design ideas in lieu of empirical studies requiring prototypes or running systems.

The CPM-GOMS models enable us to see the forest rather than the trees. We now understand the TAO's task as a complex interaction involving the TAO, the customer, and various hardware and software. Trying to optimize one component of this interaction, without understanding how it interacts with the others, is unlikely to reduce worktime.

ACKNOWLEDGEMENTS

We thank Sandy Esch, Deborah Lawrence, Jean McKendree, Karen O'Brien, Rory Stuart, Roy Taylor, and Thea Turner for their help and assistance.

Bonnie John's participation was supported, in part, by the Office of Naval Research, Cognitive Science Program, Contract Number N00014-89-J-1975N158. The view and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Office of Naval Research or the U. S. Government.

NOTE

Request for reprints and additional information should be sent to: Wayne D. Gray, Graduate School of Education, Fordham University at Lincoln Center, New York, NY 10023. gray@mary.fordham.edu. (212) 636-6464

REFERENCES

Atwood, M. E., Gray, W. D., & John, B. E. (submitted). ***Project Ernestine: Analytic and Empirical Methods Applied to a Real-World CHI Problem.***

Card, S. K., Moran, T. P., & Newell, A. (1980). Computer text editing: An information processing analysis of a routine cognitive skill. ***Cognitive Psychology, 12***, 32-74.

Card, S. K., Moran, T. P., & Newell, A. (1983). ***The psychology of human-computer interaction.*** Hillsdale, NJ: Erlbaum.

Gray, W. D., John, B. E., & Atwood, M. E. (submitted). ***Project Ernestine: Validating GOMS for Predicting and Explaining Real-World Task Performance***

Gray, W. D., John, B. E., Lawrence, D., Stuart, R., & Atwood, M. E. (May 1989). ***GOMS meets the phone company, or, Can 8,400,000 unit-tasks be wrong?*** Poster presented at: CHI '89, ACM SIGCHI's Conference on Human Factors in Computing Systems. Austin, TX.

Gray, W. D., John, B. E., Stuart, R., Lawrence, D., & Atwood, M. E. (1990a, May). ***GOMS meets the phone company: Part 2, or, Data from the world's first, large-scale application of GOMS.*** Poster presented at: CHI '90, ACM SIGCHI's Conference on Human Factors in Computing Systems. Seattle, WA.

Gray, W. D., John, B. E., Stuart, R., Lawrence, D., & Atwood, M. E. (1990b). GOMS meets the phone company: Analytic modeling applied to real-world problems. In D. Diaper, D. Gilmore, G. Cockton, and B. Shackel (Eds.), ***Human-Computer Interaction -- INTERACT '90.*** North-Holland: Elsevier Science Publishers.

John, B. E. (1988) ***Contributions to engineering*** *models of human-computer interaction.* Doctoral dissertation, Carnegie Mellon University.

John, B. E. (1990). Extensions of GOMS analyses to expert performance requiring perception of dynamic visual and auditory information. ***Proceedings of CHI, 1990*** (Seattle, WA, April 1-5). New York, ACM.