Human–Computer Interaction – INTERACT '90
D. Diaper et al. (Editors)
Elsevier Science Publishers B.V. (North-Holland)

For more details on this project see:
Gray, W. D., John, B. E., and Atwood, M. E. (1992). The précis of Project Ernestine or an overview of a validation of GOMS. In CHI '92 Conference on Human Factors in Computing Systems, pages 307–312. ACM Press.
For the definitive version see:
Gray, W. D., John, B. E., and Atwood, M. E. (1993). Project Ernestine: Validating a GOMS analysis for predicting and explaining real-world performance. Human-Computer Interaction, 8(3):237–309.
http://homepages.rpi.edu/~grayw/pubs/papers/1993/GJ&A93_HCIj.pdf

GOMS Meets the Phone Company: Analytic Modeling Applied to Real-World Problems

Wayne D. Gray[1], Bonnie E. John[2], Rory Stuart[1], Deborah Lawrence[1], & Michael E. Atwood[1]

[1]NYNEX Science & Technology Center, 500 Westchester Avenue, White Plains, NY 10604, USA.
[2]School of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15113, USA.

GOMS analyses were used to interpret some perplexing data from a field evaluation of two telephone operator workstations. The new workstation is ergonomically superior to the old and is preferred by all who have used it. Despite these advantages telephone operators who use the new workstation are not faster than those who use the old but are, in fact, significantly slower. This bewildering result makes sense when seen with the aid of GOMS. With GOMS we can see that very few of the eliminated keystrokes or ergonomic advantages affect tasks that determine the operator's work time. Indeed, GOMS shows that some presumed procedural improvements have the contrary effect of increasing the time an operator spends handling a phone call. We conclude that if GOMS had been done early on, then the task, not the workstation, would have been redesigned.

1. INTRODUCTION

In the world of the phone companies, small differences in time per call can result in large savings of dollars. For example, NYNEX estimates that each second reduction per call in work time for its Toll & Assistance Operators (TAO) saves three million US dollars per year. With such an economy of scale, a newly introduced workstation that promised up to 2.5 seconds reduction in average operator work time (AWT) (or $7.5 million US per year) appeared very attractive[1]. However, the potential savings in yearly operating costs must be balanced against a capital cost of $10 to $15 thousand (US) per workstation. Since NYNEX has approximately 1,000 TAO workstations the initial investment in new technology is large and the cost of making a bad buying decision is great.

To evaluate the actual AWT saved by the new workstation with realistic call traffic, our group at the NYNEX Science & Technology Center helped to conduct a six-month field trial. Unfortunately, empirical trials are often long, expensive, and hard to control under real-world conditions. Analytic models, such as GOMS (Card, Moran, & Newell, 1983), have the potential to replace empirical trials but have not been validated with large-scale, complex systems. Therefore, in addition to the empirical component of this project, we have built GOMS models and are testing their predictions against the empirical results.

With its ergonomic enhancements, the new workstation was expected to improve (decrease) AWT. The new workstation had a 1200-baud, graphic display while the old workstation was a line and character-oriented 300-baud display. The new workstation eliminated many keystrokes and those that remained took place on a keyboard specifically designed for the TAO task. In contrast, the old keyboard had evolved through years of functional changes, adding new keys wherever space allowed.

As we framed it, the job was clear cut. We would collect data showing how much better the new workstation was than the old. Of special interest to the phone companies would be data showing for which call-types the new workstation had the greatest advantage and for which it had the least advantage. For the GOMS part we would compare the empirical data with GOMS models for accuracy, reliability, and cost (Gray, et al, 1989).

The empirical data surprised us. No matter how we looked at it, it showed that operators who used the old workstation to be faster than those who used the new workstation. To understand this unexpected result we turned to the GOMS analyses. These analyses show why the old workstation is faster than the new, and, most importantly, they indicate that it would be very difficult for any new workstation to be faster than the current one.

In this paper we provide an overview of the methodology of the study, the ***WHAT*** of the empirical data, and the ***WHY*** provided by GOMS. This is our first published report, more detailed reports will follow.

2. METHODOLOGY

2.1. Design

2.1.1. Task

The task of the TAO is to assist the customer in completing calls and recording the correct billing. Among others, TAOs handle person-to-person calls, collect calls, credit-card calls, and calls billed to a third number.

2.1.2. Office

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 existing workstations were removed and 12 new workstations installed.

2.1.3. Participants

All participants were NET employees who had worked as TAOs for a minimum of two years. Twenty-four participants were selected for the new workstations (the ***NEW*** condition) from a list of approximately 60 volunteers. Each new participant was paired with an ***OLD***, control participant matching for shift worked (that is, time of day), and AWT on the old workstation. The NEW condition was assigned a full-time manager from the management staff of the office.

2.1.4. Training

All NEW participants went through three days of training on the new workstation. The course was conducted on-site, by regular NET trainers. The course taught the techniques and procedures advocated by the manufacturer. The course itself was for ***conversion*** training, not new training; that is, it was intended for TAOs familiar with call handling and billing, who were simply being taught a new workstation. As such, it was very similar to other commonly taught conversion training courses .

2.1.5. Duration

The trial began in April 1989 and was originally scheduled to continue for six months. It was interrupted during the fifth month by a work stoppage (AKA strike) that affected all 60,000 NYNEX employees who belonged to a union. NEW and OLD participants worked their normal shifts during the trial. From the perspective of the NEW participants their tasks and duties as a TAO was identical to their pretrial job in all respects but one; namely, a new workstation was used. For the OLD participants nothing had changed.

2.2. Empirical Data Collection

In GOMS terminology, TAO performance is tracked at the level of the ***unit-task***, where a unit-task is one completed phone call (one completed customer request). For billing purposes, a database is maintained of every completed customer request handled by every operator in each office. Reports generated for each office randomly sample one out of every ten calls that pass through that office. We used this office database to extract data on the calls handled by our 24 NEW and 24 OLD participants.

In the database each completed call is classified as one out of over 250 call-types. Pretrial analysis showed that 19 call-types accounted for over 90% of all completed calls. Both the empirical and GOMS parts of this study included just these 19 most frequent call-types.

2.3. Data for GOMS

The GOMS analyses are based on previous human performance research and task specific information (John, 1990). The task specific information includes commonly used TAO training materials, observations of TAOs handling actual calls with the old workstation, pretrial AWT statistics, and videotapes of TAO handling staged calls with the old workstation. Staged calls, placed by a TAO supervisor and identified for the TAO, are a standard phone company practice used to debug new equipment or software; for this study, they were videotaped. Using these sources we can estimate the knowledge and procedures used by experienced TAOs, and produce estimates of system response times and customer conversation time. GOMS analyses for the new workstation are based only on manufacturer-supplied training materials and performance estimates from the old workstation. No observations or AWT data for the new workstations were used because our goal in using GOMS was to predict performance on the new workstation without empirical evidence.

3. WHAT: EMPIRICAL DATA

3.1. Pretrial

Pretrial matching of participants was based upon data collected by TAO managers at the phone company office. After the trial began we were able to use database information to check whether our two groups showed pretrial equivalence on this measure. The average difference between groups was small, 0.06 seconds, and insignificant, $F(1, 46) < 1$[2,3]. From this we conclude that our two groups, NEW and OLD, were equivalent on pretrial performance on the old workstation.

3.2. Trial

3.2.1. By Month

For the analysis by month we collapsed over call-type and chose the median time for each participant for each month. The data show that median work time (MWT) for the NEW group is 104% that of the OLD[4]; that is, the new workstation requires 4% more time on an average call than does the old workstation. This difference is significant. A two (group) by four (months) ANOVA (with months as a within subject variable) yields $F(1, 44) = 4.17$. The main effect of month is also significant, $F(3, 132) = 12.11$. This main effect reflects seasonal fluctuations in call-mix that affect MWT for both groups of TAOs.

More interesting, as shown by a non-significant interaction ($F(3, 132) = 1.39$, $p > 0.10$), between group differences in work time do not converge over the four month period. (From April through July work times for the NEW group are, respectively, 6%, 3%, 5%, and 4% higher than for the OLD.) This lack of an interaction suggests that the NEW participants master the new workstations very fast and reach asymptotic performance during the first month of the trial.

3.2.2. By Call-Type

For the analysis by call-type we collapsed over months and chose the median time for each participant on each of 18 call-types[5]. The two (group) by 18 (call-type) ANOVA (with call-type as a within subject variable) yielded a significant effect of group, $F(1, 46) = 5.92$, again indicating that the new workstation was slower than the old.

The main effect of call-type was also significant, $F(17, 782) = 101.27$, indicating that different call-types required different amounts of time to process. To our surprise, call-type did not interact with group, $F(17, 782) < 1$.

3.3. Summary of Empirical Results

Not only was the new workstation NOT faster than the old, but it was significantly slower. The approximate 4% difference translates to about a one second loss in MWT. Using the heuristic of $3 million (US) dollars per second per year, the 1 second difference would cost NYNEX $3 million per year.

Clearly something is wrong. How could an ergonomically engineered, modern workstation be slower than ergonomically indifferent, 5 year old technology?

The obvious answers seem to be wrong. First, the NEW participants were very motivated and very interested in "beating" their old work times. Managers reported that the NEW participants enjoyed the new workstations and actively tried to lower their work times. Second, although the study lasted for four rather than six months, all the data we have examined indicates that the NEW group had already reached asymptote and would not have improved their MWT with more practice. Indeed, preliminary analyses suggests that asymptotic performance was reached in the first week of the study.

4. WHY: GOMS

TAO's do several things in parallel when processing a customer's request: they listen to the customer, they perceive information on the CRT screen, they move their hands to appropriate keys and strike them. We represented this parallelism in a PERT chart, displaying all perceptual, cognitive, and motor operators (as boxes) and the dependencies between them (as lines connecting the boxes) according to goal decomposition and operator-placement heuristics (Card, et. al, 1983; John, 1990).

Figures 1 and 2 show the first and last segments of a GOMS analysis for one 15 second phone call. For each figure, the top chart represents the call using the new workstation while the bottom shows the GOMS analysis for the same call using the old workstation. For this report we make a virtue of necessity. A readable version of the full GOMS analysis would require more pages than this proceedings permits; also, a readable form would tend to draw the reader's attention to details which, while interesting in their own right, are tangential to this paper. The reduced version allows us to avoid such details while drawing attention to the overall pattern.

An important concept in analyzing the total task time for complex parallel tasks is "critical path". In project management, the critical path is "the sequence of tasks that determines the soonest the project can finish" (p. 6, CLARIS Corp, 1987); in GOMS analyses, it is the perceptual, cognitive, and motor operators that determine the total time for the task. For example, consider three partially overlapping activities: the TAO saying "New England Telephone, may I help you?", the TAO moving his/her right hand towards specific function keys, and the customer's saying "Operator, bill this to . . ." The TAO must perceive the customer's request before s/he can either press the appropriate key or make the appropriate verbal response, however, experienced TAOs can prepare for the most likely request in advance. In this case, the TAO moves towards likely function keys while still saying the greeting. Since saying the greeting takes longer than moving to the function keys, and customers typically wait for the operator to finish the greeting before they state their request, the movement toward the function keys is said to have "slack time" and is not on the critical path. This slack time is observed in the videotapes as the TAO hovering over some function keys waiting for the customer to give information that will dictate which key is to be pressed. In Figures 1 and 2, the critical path is shown as bold lines and boxes.

Figure 1 has two striking features. First, the analysis for the new (top) workstation has 10 fewer boxes than the analysis for the old (bottom) workstation, representing two fewer keystrokes. Second, none of the deleted boxes were on the critical path so the total task time for this portion of

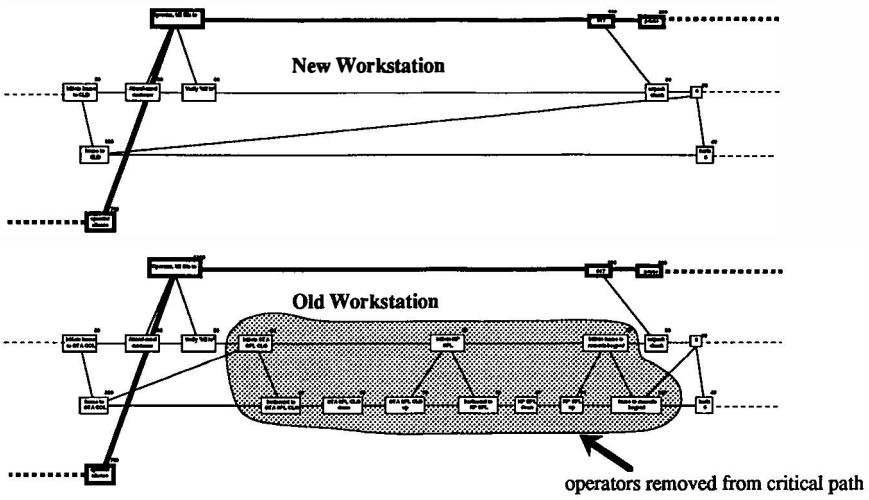

Figure 1. Section of GOMS analysis from near the beginning of the call. Notice that the new workstation (top) has removed two keystrokes (which had required 7 motor and 3 cognitive operators) from this part of the call. However, none of the ten operators removed were along the critical path (shown in bold). (Note: the print in the boxes is intentionally illegible.)

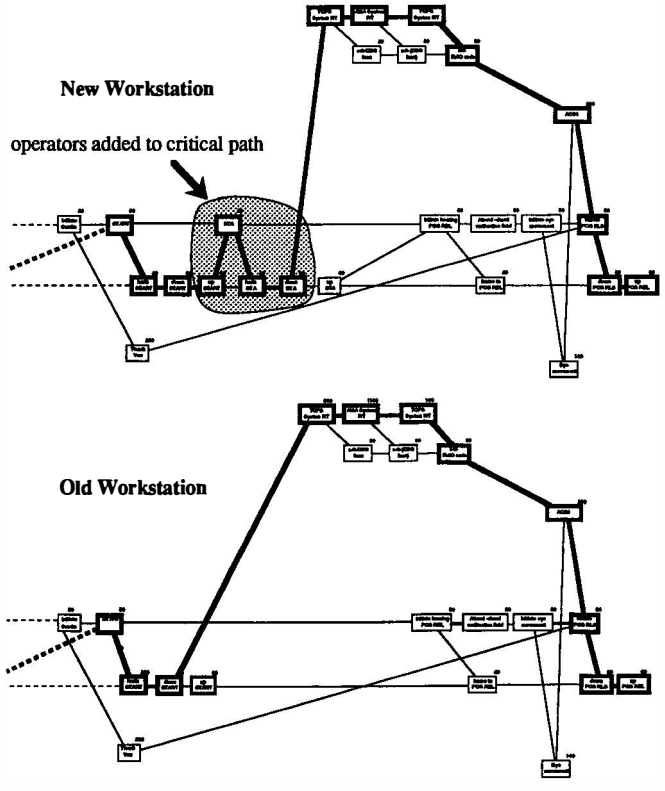

Figure 2. Section of GOMS analysis from the end of the call. Notice that the new workstation (top) has added one keystroke to this part of the call which results in four operators (three motor and one cognitive) being added to the critical path (shown in bold). (Note: the print in the boxes is intentionally illegible.)

the task would not change between workstations. At this point in the task the critical path is determined by the TAO greeting and getting information from the customer. Removing keystrokes here 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.

The missing middle of the analysis (the activities between those shown in Figures 1 and 2) is identical for both workstations and essentially shows the critical path being driven by how fast the customer says the ten-digit number to which the call should be billed. TAOs are taught to "key along" with the customer. While a rapidly speaking customer could force the critical path to be determined by the TAO's keying speed, given that both workstations use the standard numeric keypad, the critical path (and resulting speed of keying in numbers) would be the same for both workstations.

If the new workstation simply eliminated the two keystrokes required by the old workstation in the beginning of the call, then GOMS would predict equivalent performance. However, for the new workstation, the procedure has been changed so that 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 moves from a position off of the critical path to one that is on the critical path. Hence, the cognitive and motor time required for this keystroke now adds to the time required to process this call. Thus, GOMS predicts that the AWT for this call would be slower for the new workstation than it would be for the old workstation. Indeed, the empirical data show that this call is 5% slower on the NEW than OLD workstation (a marginally significant difference, $F(1,46) = 3.14$, $p < 0.10$).

This analysis also speaks to the unexpectedly fast learning curve found in the empirical data. On the old workstation, the critical paths for many call-types are dominated by TAO and customer conversation time, and system response time, with individual keystrokes having as much as several seconds of slack time. Previous research with expert typists suggests that at most, the duration of keystrokes would increase from approximately 100 msec to approximately 1000 msec for the least experienced, hunt-and-peck keying on an unfamiliar keyboard (John & Newell, 1989, Card, et. al., 1983). Thus, it is unlikely that initial difficulty with the new workstation would change the critical path to being dominated by keystrokes. Therefore, most of the learning would be off the critical path and thus unobservable in the MWT.

To date, we have performed GOMS analyses for five call-types. The above example is representative of what we are finding. Generally the new workstation removes keystrokes that are off the critical path. Most of these are eliminated while some are placed onto the critical path. Current analyses are very specific to protocols obtained from individual TAOs. It is our intention to generalize the analysis for each call-type to all TAOs and to then assess the goodness of fit against the empirical data.

5. CONCLUSION & SUMMARY

The new workstation is ergonomically superior to the old and is preferred by all who have used it. Despite these advantages TAOs who use the new workstation are not faster than those who use the old. Indeed, statistical analyses show that the NEW TAOs are significantly slower than the OLD.

This bewildering result makes sense when seen with the aid of GOMS. With GOMS we can see that very few of the eliminated keystrokes or ergonomic advantages affect tasks that are on the critical path of the example call. Indeed, GOMS shows that some of the procedural changes moved previously non-critical keystrokes onto the critical path. Such changes add to the predicted work time even when the net result is fewer keystrokes.

Although the GOMS analyses is ongoing, the emerging conclusion is that there is very little that could be done to a workstation itself that would decrease work time for TAOs. The factors most limiting performance are neither the display nor entry of information, but system response time (other than workstation time) and customer conversation time. Clearly, if GOMS had been done early on, then the task, not the workstation, would have been redesigned.

For the TAO job, GOMS can be used in one of two ways. First, based upon GOMS, we can redesign TAO procedures to reduce the number of bottlenecks along the critical path. Second, and longer term, we can use GOMS to redesign the task itself. With GOMS we can ask whether it is worth while to speed up a component of the system. For example, when calling cards are used, a database is accessed to verify the number. What affect would a 50% decrease in time for database access have on operator work time? (If waiting for verification is not on the critical path then a faster access time would not reduce work time.) As another example, we can ask whether some parts of the customer interaction might best be automated.

Our main conclusion concerns GOMS itself. GOMS is an important and valuable tool that can be applied to tasks other than text-editing or spreadsheets. While it will continue to be useful to academic researchers, the time is ripe to apply GOMS to complex, real-world tasks.

ACKNOWLEDGEMENTS

Thanks are due to Karen O'Brien for her sponsorship, support, and assistance throughout all stages of the trial. Thanks also to Sandy Esch.

FOOTNOTES

[1] Reductions in AWT is just one of the many features of the new workstation. Its other features are such as to make it attractive to NYNEX even if AWT remained constant.

[2]Note that all all comparisons, ANOVAs, and figures are based upon median work time (MWT).

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

[4]Note that the absolute magnitude of work time of NYNEX TAOs is considered proprietary information. Therefore, here and in several other places in this report, we have made an attempt to accurately depict the relative results without compromising corporate information

[5]One of our call-types was chosen not because of its frequency but because of interest in the procedure required to process it. Because many of our participants had such low numbers of recorded calls of this type (< 16) we excluded it from the current analysis.

REFERENCES

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

CLARIS Corporation (1987). ***MacProject II Manual.*** Mountain View, CA: CLARIS Corporation.

Gray, W. D., John, B. E., Lawrence, D., Stuart, R., & Atwood, M. E. (1989). ***GOMS meets the phone company, or, can 8,400,000 unit-tasks be wrong?*** Poster presented at CHI '89 (Austin, Texas, April 30-May 4).

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

John, B. E., & Newell, A. (1989). Cumulating the science of HCI: From S-R compatibility to transcription typing. In ***Proceedings of CHI, 1989*** (Austin, Texas, April 30-May 4 1989). New York: ACM, pp. 109-114.