Cognitive Load During CT Interpretation

David V Beard, Brad M Hemminger, Kevin M Denelsbeck,
Peter H Brown, R Eugene Johnston

Departments of Radiology and Computer Science
University of North Carolina Chapel Hill, 27599-7510
voice: (919)-966-5467 fax: (919)966-5934 email: beard@cs.unc.edu

1. INTRODUCTION

In 1990 Foley et al evaluated a commercial CT workstation which displayed 8 images in from 7.5 to 26 seconds using two 1024x1024 monitors[1]. While the workstation displayed images of acceptable quality, the workstation's interpretation quality was unacceptable, i.e., the dictated reports were of lesser quality than those from film interpretations. Why did Foley find unacceptable dictated-report accuracy despite the acceptable quality of displayed images?

We argue that an image display method — be it film or monitor-based — imposes a "cognitive load" or "mental workload" on the radiologist during interpretation in general, and on human short-term or working memory in particular; Miller [2] indicates that human working memory is limited to 7±2 "chunks", though many currently think the limit may be closer to 5±2. If mental workload becomes excessive, the radiologist may forget to view images or anatomy, or may forget to dictate viewed findings in the interpretation report.

In this paper, we consider the nature of mental workload and how one might determine whether mental workload effects the potential accuracy of a particular medical image display method. Then we detail observer experiments we have conducted evaluating electronic display of CT images. While detailing these results, we will focus on the evidence we have accumulated which suggests that mental workload has a significant influence on workstation accuracy and interpretation speed. These results suggest that experiments involving interpretation tasks that are very similar to those in the clinic are needed in addition to conventional ROC analysis when evaluating the effectiveness of a new display device or method.

2. MENTAL WORKLOAD FOR THE RADIOLOGIST

We expect that human working memory is used in a number of different ways by radiologists during interpretation of medical images. First radiologists must both remember their image-interpretation plan or task list, as well as their current place in that plan or task list. Second, they must remember details about the anatomy being viewed, normal metabolic function in and around that anatomy, and the very large number of abnormal anatomical features or functions that could occur with the anatomy in question. Third, they must remember the various normal and abnormal findings they have discovered during the course of the interpretation. Fourth, they must compose the dictated report "in their heads" again taking up limited human working memory. And fifth, they must remember how to use the alternator or computer workstation, the dictation device, their resident's name, etc. All these items form the radiologist's mental workload during image interpretation. [e.g., 3]

Obviously, there are far more than 5±2 "things" in the above list. Humans in general, and radiologists in particular cope with this problem by combining several memory items into a single "chunk" or item. Experience is, in part, the process of developing more and more useful chunks for a given domain of discourse so experienced individuals are likely to have more and better "chunks" and thus a greater mental workload limit for a given domain. Radiologists develop these various chunks during residency training as

0-8194-1461-1/94/$6.00

well as throughout their career. They also learn mental "boilerplate" text which greatly reduces the mental workload during dictation. In fact, for highly trained radiologists operating in their subspecialty, it is very possible that dictating a report may generate less mental workload than using a computerized dictation system.

Augmenting Human Working Memory. Radiologists have a number of behaviors that reduce mental workload and thus help them keep track of what they are doing during an interpretation. First, radiologists often use a grease pen to place a mark near relevant findings on the images. Such a grease mark is used as a memory aid during subsequent dictation. Such marks also are used by radiologists and other physicians who view the images at a later point in time. Second, radiologists often use written notes to keep track of findings or other points to be included in a dictation. Third, a report is often dictated "on the fly" throughout the interpretation rather than as a single step at the end of the interpretation; the process of dictating the various findings has a tendency of reinforcing them in working memory and thus may also serve as a memory aid. Fourth, radiologists use the spatial organization of the images on the lightboxes as a memory aid; they remember that an interesting image is "down over there." Spatial memory seems to be separate from the more textual human working memory, and thus one can use spatial memory to offset working memory and correspondingly reduce mental workload.

3. DETERMINING EXCESSIVE MENTAL WORKLOAD

Disrupter Experiments. Determining whether mental workload is excessive should be an essential step in the evaluation of a candidate medical image display method. Cognitive psychologists use a disrupter task to determine working memory usage[e.g., 4], which might be a reasonable measure of mental workload. In a disrupter task experiment, the subject is given a memory task such as memorizing a list of numbers. Then a disrupter task is given to the subject, such as counting backwards from one hundred or adding together several four digit numbers. Finally, the subject is asked to recall the items in the original memory task. By varying the type and number of items in the memory task, the time allowed for memorization, the type, extent, and duration of the disrupter task, the delay until the recall task, and other parameters, the experimenter can measure the capacity of human working memory to handle the memory items in question, and thus quantify the effect of the disrupter task.

Disrupter tasks are the definitive method for determining working memory usage and thus mental workload. But while such cognitive psychology experiments can be precise, they are very expensive, time consuming, and of very limited scope. Further, measuring working memory usage is not enough; we also need to determine how much is too much. Thus, we must find other methods for evaluating mental workload during the display of radiologic images.

Cognitive Models. Cognitive models are another method for estimating mental workload. One of the more influential models or family of models in computer-human interaction (CHI) is the GOMS model [5,6]. The GOMS framework describes the knowledge that a user needs when performing a given computer task. This knowledge contains *goals* that reflect the user's intentions; *methods* that reflect the user's plans for accomplishing the goals; *operators* that are the individual actions carried out in a method and *selection rules* that present conditions under which certain methods apply. GOMS recursively decomposes a user's task into subtasks until atomic tasks of known duration are encountered. Then by recursive summation, the duration of the original user's task can be estimated. This general framework, initially applied to the domain of text-editing, has also been used successfully to model other computer-based tasks such as spreadsheet applications [7] and video games [8].

In addition to extending the domains to which this work applies, there have also been extensions to the mechanisms that drive the framework. CPM-GOMS [9] extended GOMS to include a critical path component that allows designers to see how system changes can influence the user's "preferred" sequence of operators. BROWSER-SOAR [10] takes GOMS modeling into more interactive environments and tasks.

Bovair, Kieras, and Polson [11-13] proposed the *cognitive complexity* analysis. This approach combines the production system approach with the GOMS model and extends the model so that it can make predictions about speed of learning for different systems and transfer between system designs.

Cognitive models can be used to estimate relative mental workload between various alternative computer human interactions; these estimates are, however, fairly rough. If the designer is going to develop a model for other purposes, such as predicting the duration of a particular task with a particular human computer interaction design, then the designer might as well go ahead and use that model to estimate mental workload. But building such a model from scratch is very time consuming, especially considering the limited accuracy of the resulting mental workload metric. Further, while we might be able to determine that one interpretation interaction requires more mental workload than another, we will not be able to determine whether either approach required too much.

Observations and Verbal Protocol. We believe the best way of determining whether an image display method imposes excess mental workload is to watch and listen to radiologists while they interpret images with the display system in question. The term "verbal protocol" denotes the verbal comments and other indicatings of what the subject is thing about during the observation session. While such an observational or anthropological approach is not nearly as precise as a laboratory-conducted cognitive psychological experiment, it is tractable within the development time of a medical image workstation, and does provide a considerable amount of useful information.

The following section provides a number of case studies showing signs of excess mental workload we have observed while building and experimentally evaluating seven medical image workstations during the last eight years.

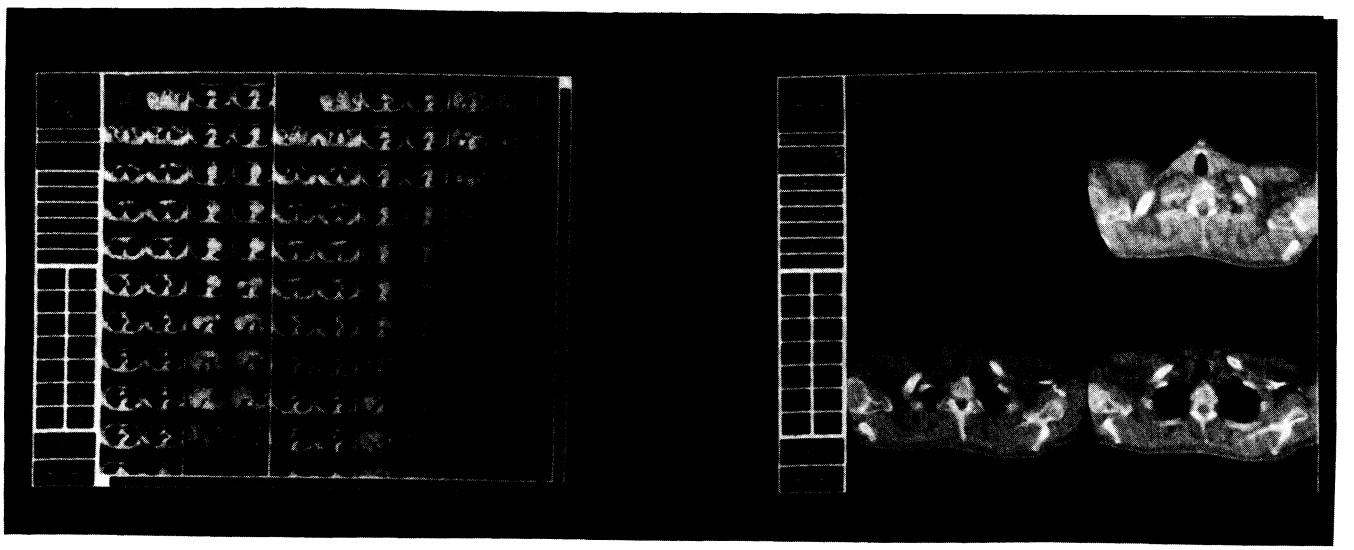

Figure 1
FilmPlane prototype CT workstation. FilmPlane uses one 900x1100 pixel monitor with a 2 second image display time.

4."EVIDENCE"

FilmPlane. In the late 1980s, we developed and evaluated the FilmPlane prototype CT workstation implemented on a Sun 3/180 with one or two 900x1100 pixel greyscale 8-bit monitors [14]. FilmPlane is

shown in figure 1. Images were displayed in about 2 seconds. A navigational view image was provided as a mental model [e.g., 15,16] in which the images of a single CT study were arrayed in a vertical two-image-wide column and multiple study columns were organized horizontally across the screen. A single screen display could be toggled between displaying the navigational view and a detail view showing four images from one of the image columns. Both dynamic and preset intensity windowing were available.

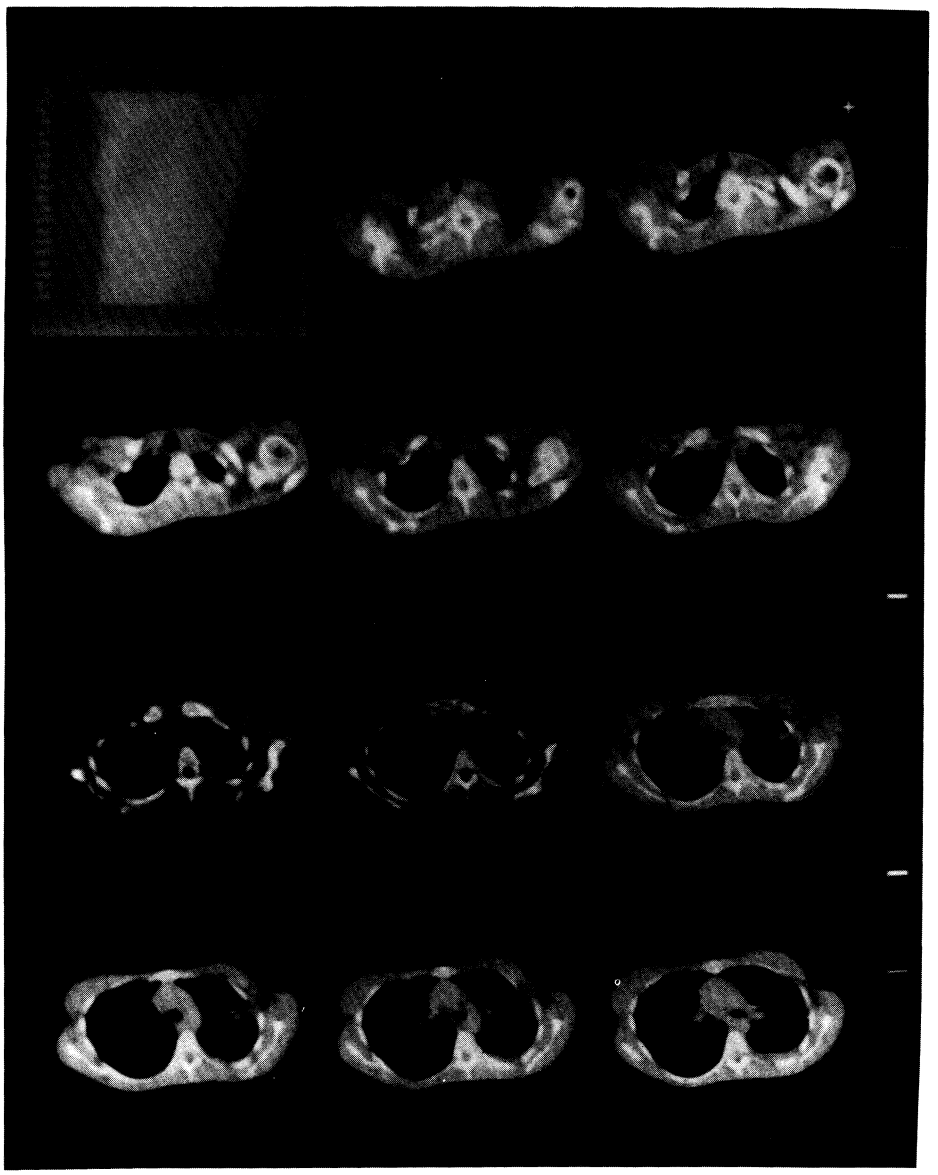

Figure 2
FilmStrip prototype CT workstation. FilmStrip uses one 2500x2000 pixel monitor with 0.11 second image display time.

An observer experiment determined that FilmPlane was about twice as slow as film for the interpretation of current and previous CT studies. even worse, the radiologists felt that the system "imposed itself" on them during the interpretation and that they found themselves thinking about manipulating the workstation rather than thinking about interpreting the CT studies; they felt this never or rarely happened with film and alter-

nator. During one trial, a radiologist with a great deal of CT experience could not remember whether he had looked at the kidney at one point during the interpretation; something that never happened with film. Further, several radiologists forgot critical findings until interpreting subsequent examinations. Radiologists said they would not be happy using such a system in the clinic.

These observations and verbal comments are all indications of excessive mental workload. Part of this additional mental workload is likely to be caused by the unfamiliar display method — even if FilmPlane was as effective as film and alternator we would expect radiologists to have less mental workload with film and alternator than with FilmPlane simply because of the additional experience and familiarity they had with the traditional display method. Nevertheless, the above comments, combined with the vast difference in interpretation speed rather strongly suggests the presence of excessive mental workload and then need for an improved approach to computer display of CT images.

FilmStrip. Subsequently, we developed the FilmStrip prototype CT workstation in 1991 using a single 2500x2000 pixel monitor [17]. FilmStrip, shown in figure 2, used the mental model of a three-image wide strip of film that could be scrolled vertically at 0.11 seconds pre scroll using up and down buttons. A third button cycled through three preset intensity windows, also in 0.11 seconds. A controlled observer experiment found FilmStrip to be "as fast as film and alternator." No excessive mental workload indications were observed. Radiologists thought such a system would be effective in the clinic.

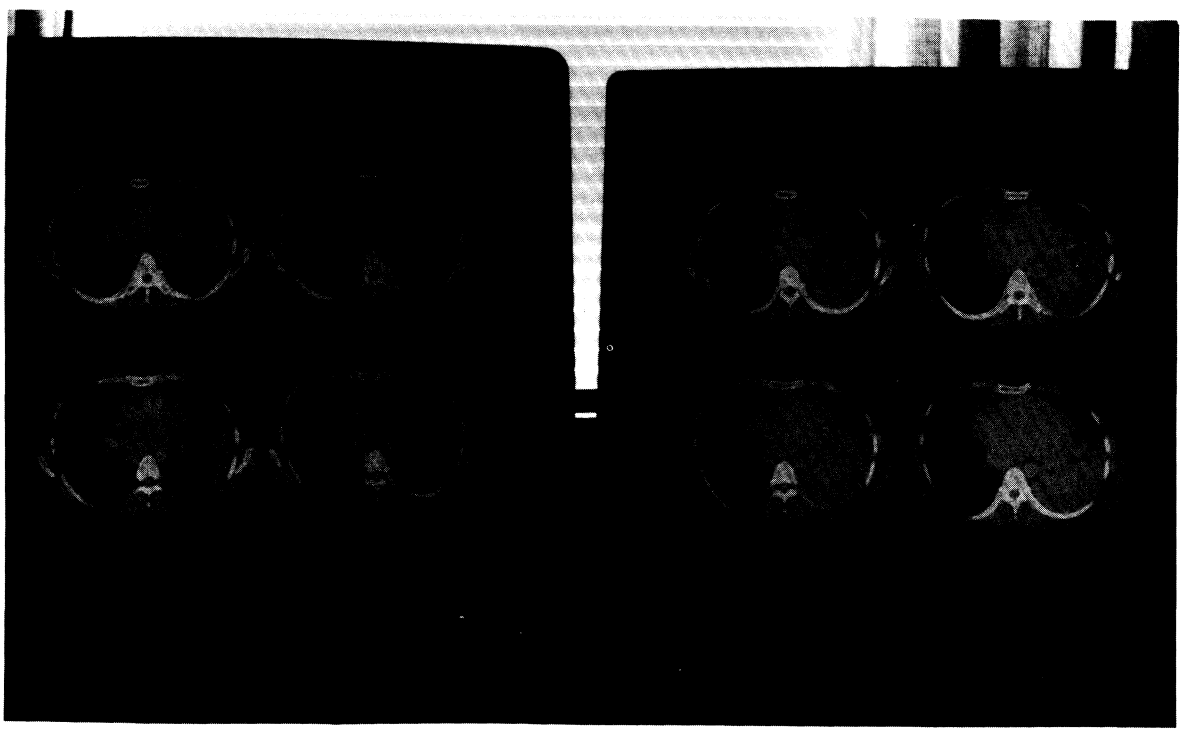

Figure 3.
FilmStripLet prototype CT workstation. Two 900x1100 pixel monitors with 0.15 second image display time.

FilmStripLet. FilmStrip is an effective method for interpretation of CT studies. But the hardware is very expensive. Thus in 1992, we developed a low cost version of FilmStrip called FilmStripLet [18]. FilmStripLet uses the same button organization and mental model as FilmStrip, but is implemented using a

Sun SPARC workstation and two 900x1000 pixel 8-bit monitors. Improvements in hardware and software provide for 0.1 to 0.2 second image display times. A controlled observer experiment found that FilmStripLet had a interpretation time average that was within 30 second of that of film and alternator. Further, the radiologists felt that FilmStripLet would provide acceptable CT interpretations in the clinic. But they expressed feelings of being less comfortable with FilmStripLet than with film.

We had not provided a marking or "grease pen" feature allowing the radiologist to denote CT images constraining interesting anatomical features for subsequent description in the interpretation report. While the lack of a "grease pen" did not appear to be a problem for the frequent CT readers, it was mentioned several times by infrequent readers who instead had to use paper and pencil as a memory aid for their findings. Additional signs of excessive mental workload included omissions in the dictated reports which were remembered later during the dictation.

Dictation Techniques. As usual with observer studies, differences between individuals seemed to vary even more than differences between display methods. for example, several radiologists kept all their observations about the CT study in their head and then dictated the entire report at the end of the interpretation. Others dictated the findings "on the fly" and interpreted the results at the end. Dictation behavior also varied with the complexity of the CT study with more mental-workload-intensive approach being used primarily with simpler interpretations.

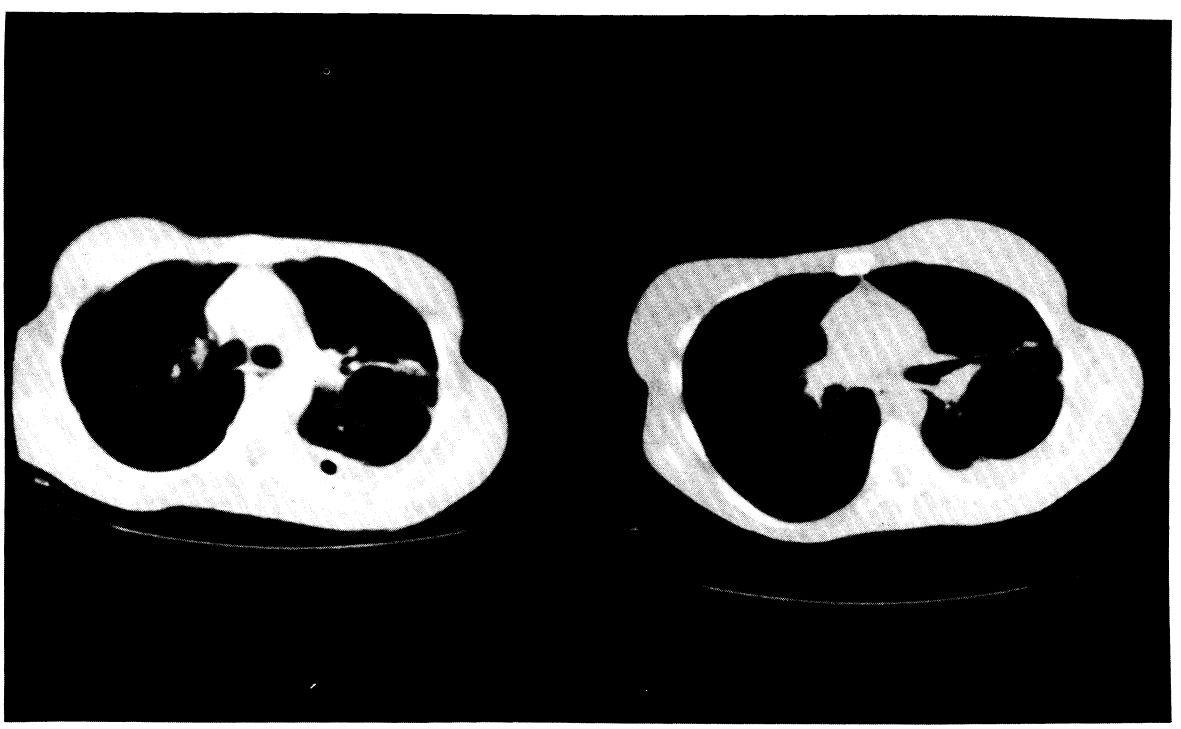

Figure 4
FilmStack prototype CT Workstation. One 900x1100 pixel monitor with 0.15 second image display time.

FilmStack. During 1993 and 1994, we have built and are in the process of experimentally evaluating another low-cost workstation for display of multiple CT or MR studies. Unlike the previous workstation prototypes FilmStack does not use the mental model of a FilmStrip. Rather, it uses the mental model of a

stack of films, with the axis of the stack normal to the surface of the monitor. The effect is that of a "cine" display with user controlled scrolling rather than a continuous scroll approach. FilmStack is implemented using a Sun SPARC or other UNIX/ X Windows workstation and provides for the display of two CT studies with a 0.1 to 0.2 second image display time.

Observation and verbal protocol during initial pilot studies indicated problems with radiologists obtaining a "big" picture or gestalt of various organs and other 3D structures, particularly in the abdomen where interpretation requires noting when one organ is pushing against another organ and distorting it shape in 3-space. To aid in this, we added a mosaic view containing 128x128 miniatures of all the images in each study. By pressing a button, the radiologist can toggle from the stacked displays showing both studies to the mosaic miniatures displays showing all the images at the same time.

A pilot study has been completed generating an average interpretation time for FilmStack that is similar to that of film and alternator. No excess mental workload indications have been observed at this point.

5. ROC ANALYSIS FOR EVALUATING INTERPRETATION QUALITY

With a dictated report, radiologists must keep many observations in their human working memory including what portions of the anatomy they have viewed, what findings they have seen, the spatial shape of the object, etc. If a workstation is sufficiently disruptive, the radiologist may miss or forget critical findings resulting in inadequate interpretation reports even if the image quality is acceptable and all needed information is available on the images. The findings form typically used with ROC analysis not only acts as a working memory aid, reducing the number of items radiologists must keep in their heads during interpretation, but also affects the interpretation process. Since interpretation quality may be affected by excess cognitive load of an inadequate CT workstation, a findings-form study may show adequate accuracy with a display method that otherwise would result in unacceptable interpretation reports.

Cognitive load may also effect interpretation speed, for a radiologist may work more slowly, may use a grease pen or written notes as a working memory aid, or may have to go back in a dictation tape to correct a forgotten finding. Further, filling in the findings form may take a considerably differing amount of time than dictating an interpretation report, so task time results from ROC studies using findings forms, while a useful metric, are not a definitive measure of the relative interpretations times of the candidate methods. Further still, maximal experimental power often demands that the cases used in an ROC experiment consist of a very narrow range of abnormal conditions, while an experiment focused on interpretation speed would more likely consist of a broader range of cases and, most likely, a higher percentage of normals.

ROC analysis would be overly complicated if not impossible if a dictated report were used instead of a fixed findings form, so the ROC study with the dictated report is still a necessary step in evaluating image quality. However, it is not sufficient for determining the clinical accuracy or the interpretation speed that can be expected during clinical use of a new display method. In addition to an ROC study with a findings form, another study should be conducted that has sufficient power to determine interpretation speed equivalence to within a confidence interval of say 5% of the average interpretation time; this would amount to 30 seconds for a five minute CT interpretation. Such a study would use a clinically realistic interpretation task including access to previous interpretation reports, the requisition form, and resulting in whatever reporting mechanism is used in the clinic, most often, the dictated report.

6. ACKNOWLEDGMENTS

This work was funded by NIH R01 CA44060.

7. REFERENCES

1. Foley WD, Jacobson DR, Taylor AJ, Goodman LR, Steward ET, Gurney JW, & Stroka D. Display of CT Studies on a Two-Screen Electronic Workstation versus a Film Panel Alternator: Sensitivity and Efficiency among Radiologists, *Radiology* 1990;174:769-773.
2. Miller GA The magical number seven plus or minus two: Some limits on our capacity for processing information. Psychological Review 1956:63, 81-97.
3. Hancock PA and Caird JK Experimental Evaluation of a Model of Mental Workload, J. Human Factors 1993: 35(3) 413-430.
4. Reitman JS, Mechanisms of forgetting in short-term memory, J. Cognitive Psych. 1971:2 185-195.
5. Card SK, Moran TP, & Newell A The keystroke-level model for user performance time with interactive systems. *Communications of the ACM*, 1980, 23, 396-410.
6. Card SK, Moran TP, & Newell A *The psychology of human-computer interaction*, Hillsdale, NJ: Lawrence Erlbaum Associates 1983.
7. Olson JR & Nilsen E, Analysis of Cognation involved in Spreadsheet software interaction, *Human Computer Interaction*, 1988, 3, 309-350.
8. John BE, Vera AH, & Newell A, *Toward Real-time GOMS*, Carnegie Mellon Univ,. School of Comp. Science Technical Rpt. 1990, CMU-CS-90-195.
9. Grey WD, John BE, & Atwood ME The Precision of Project Ernestine or An Overview of a Validation of GOMS *Proc. CHI '92 Human Factors in Computing Systems* 1992.
10. Peck & John 1992 "Browser-Soar: A computational model of a highly interactive task" School of Computer Science, Carnegie Mellon University.
11. Bovair S, Kieras DE, & Polson PG The acquisition and performance of text editing skill: A cognitive complexity analysis. *Human-Computer Interaction*, 1990 5, 1-48.
12. Polson PG, A quantative model of Human Computer Interaction, in Carroll JM (Ed.) *Interfacing Thought: Cognatative Aspects of Human Computer Interaction*, 1987, (pp.184-235) Cambridge, MA, MIT Press.
13. Kieras DE & Polson PG An approach to the formal analysis of user complexity *International Journal of Man-Machine Studies* 2(2) 1985 365-394
14. Beard DV Designing a Radiology Workstation: A Focus on Navigation During the Interpretation Task, *J. Digital Imaging* 1990;3(3):152-163
15. Rumelhard D & Norman D, Analogical processes in learning, In JR Anderson (Ed.), Cognitive skills and their acquisition, 1981, pp 335-359, Hillsdale, NJ: Erbaum.
16. Young R The machine inside the machines: User's models of pocket calculators. International *Journal of Man-Machine Studies* 1981 15:51-85.
17. Beard DV, Hemminger BM, Perry JR Mauro M, Muller K, Warshauer D , Zito A, & Smith M Single-Screen Workstation vs. Film Alternator for fast CT Interpretation, *J. Radiology*, 1993; 187(2):1-6.
18. Beard DV, Hemminger BM, Pisano ED, Denelsbeck KM, Warshauer D, Mauro M, Keefe B, McCartney W, & Wilcox C, CT Interpretations with a Low Cost Workstation: A Timing Study. Accepted by J. Digital Imaging, 1994.