Viewing personal history records:
A comparison of Tabular format and graphical presentation using LifeLines

Diane Lindwarm Alonso[2], Anne Rose[1], Catherine Plaisant[1], Kent L. Norman[2]

[1]Human-Computer Interaction Laboratory

University of Maryland Institute of Advanced Computer Studies

[2]Department of Psychology

University of Maryland, College Park, MD 20742-3255

http://www.cs.umd.edu/projects/hcil

lindwarm@wam.umd.edu, {rose,plaisant}@cs.umd.edu, kn8@umail.umd.edu

Revised December 1997

ABSTRACT

Thirty-six participants used a static version of either LifeLines, a graphical interface, or a Tabular representation to answer questions about a database of temporal personal history information. Results suggest that overall the LifeLines representation led to much faster response times, primarily for questions which involved interval comparisons and making intercategorical connections. A "first impression" test showed that LifeLines can reduce some of the biases of the tabular record summary. A post-experimental memory test led to significantly ($p<.004$) higher recall for LifeLines. Finally, simple interaction techniques are proposed to compensate for the problems of the static LifeLines display's ability to deal with precise dates, attribute coding and overlaps.

INTRODUCTION

The way in which temporal data is represented has a dramatic effect on the way we interpret and use those data. Metaphors and analogies have been used quite effectively to aid the user and provide a mental model of the system (Carroll & Mack, 1985). In order for a graphical interface (visual, as opposed to textual or numeric) to be the most effective, though, it is useful to "use real-world analogies as much as possible" (Hix & Hartson, 1993, p.89) and to establish "good mappings between the computer display of information and the user's conceptual model of the information" (Nielsen, 1993, p. 126). Shneiderman (1992) notes the benefits of visual displays as compared to textual displays because of this mapping to our three-dimensional world. By using consistent, visual displays we can utilize the cues with which we are familiar -- proximity, containment, color, coding, etc. LifeLines, a graphical interface designed by the Human Computer Interaction Laboratory (HCIL) at the University of Maryland, College Park, MD, attempts to meet these ideals. It uses the metaphor of a timeline to represent chronological data, with the use of color coding and proximity to specify and relate important events and actions (see Figure 1).

A Practical Application

HCIL produced prototypes for the Maryland Department of Juvenile Justice (DJJ) who is redesigning their information system. To better understand DJJ's problems, HCIL performed an extensive evaluation of the existing system (Rose, Shneiderman, & Plaisant, 1995; Slaughter, Norman, & Shneiderman, 1995; Plaisant, Rose, Shneiderman, & Vanniamparampil, to appear). One problem is how difficult and time consuming it is to get an overview of a youth's history with the current system. Case workers must use cryptic codes to navigate through dozens of tabular screens. As an alternative, HCIL proposed LifeLines, a general visualization technique that uses multiple timelines (e.g., cases, workers assigned, placements and reports) to present a youth record overview in one screen (Plaisant, et. al., 1996). Line color is used to indicate the depth of penetration into the system (e.g., before court, after court) and thickness is used to indicate severity. The timeline metaphor allows users to quickly get an overview of the record and see relationships among the events. It is believed that LifeLines is a general method of presenting personal history records and can be used in a variety of applications (e.g. insurance records, financial records, student records, or medical records (Plaisant and Rose, 1996; Plaisant and Shneiderman, 1997)

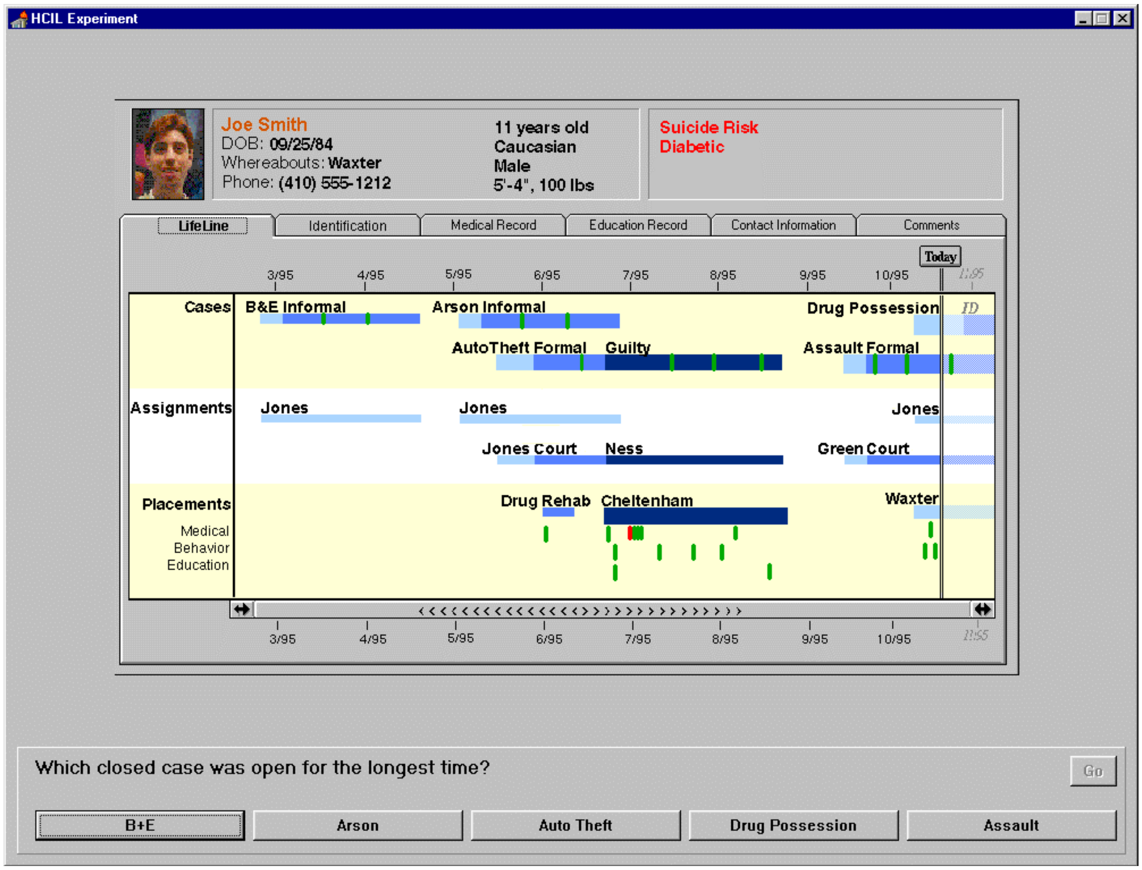


Figure 1. The LifeLines format

THE EXPERIMENT

This experiment examines the effects of the format in which temporal data is represented. Subjects were shown one of two formats, LifeLines or Tabular (Figures 1 and 2), and asked to answer questions based on the information given. It was predicted that participants in the LifeLines condition would do better:

- comparing time intervals,
- making intercategorical (i.e. across tables) connections relating one area of information to another, and
- gaining an appropriate impression of the record.

The participants in the Tabular condition were expected to do better in tasks requiring precise pieces of information (e.g., a specific date or rating).

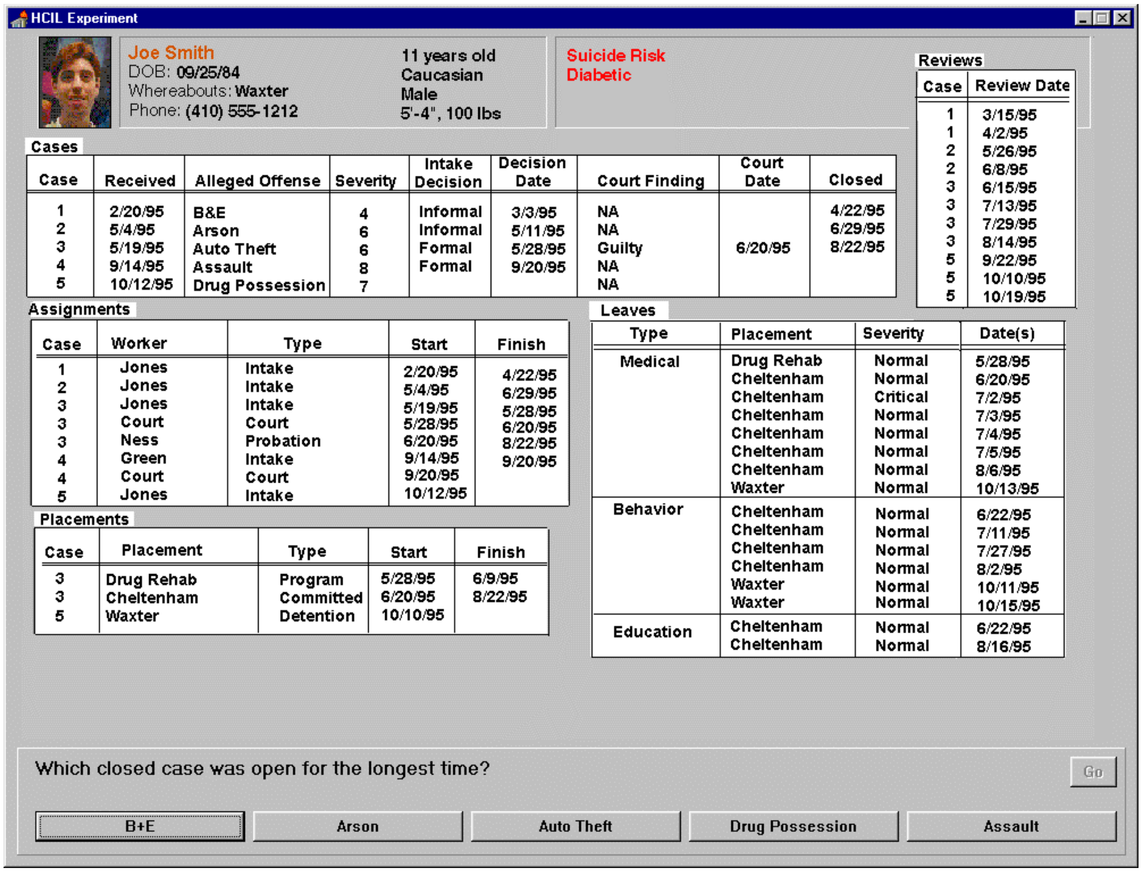


Figure 2. The Tabular format.

By collecting speed and accuracy data as well as user satisfaction ratings and recall data, our goal was to compare the two static displays of information to understand and measure the benefits and pitfalls of the LifeLines display. Because the LifeLines display was always intended to be part of an interactive information system a secondary goal for this experiment was to identify and measure the need for the interactive features implemented in the application to augment the LifeLines display (e.g., active cursor / ballon help or dynamic highlighting of related information).

While this experiment attempts to understand the difference between these two formats, for practical applications, the best solution might be a combination of the two. As Paivio's (1986) Dual Coding Theory predicts, best performance would be found for a combination of textual and spatial pictorial representations because it offers the most information via the two codes (verbal and spatial). Also, there is a benefit of redundancy of information which should help encoding. Multiple Resource Theory (Wickens, 1992) also supports the use of different resources for verbal (textual) and for spatial (pictorial/navigational) processing. For this experiment, we will only look at the benefits of the LifeLines representation as compared to the Tabular representation, but our prototype for DJJ offers both a graphical and tabular view of the data.

Hypothesis

As mentioned earlier, the primary purpose of this experiment is to observe the strengths and weaknesses of both the LifeLines and Tabular formats so as to develop an interface which incorporates the best features of both. In order to do this we created 31 questions that would be used to study speed and accuracy of user interaction with each of the interfaces. Prior to testing we categorized each question by whether we thought that user performance would be best in terms of the LifeLines interface, the Tabular interface, or that both would provide an equal level of performance (See Appendix 2 for the text of the questions). From these predictions we arrived at the following hypotheses (stated here in order of presentation in the experiment):

H1: First Impression Test

It is predicted that more subjects in the LifeLines condition will accurately indicate the that the Complex record is actually less severe than the Simple record.

H2: Main Quiz

It is predicted that subjects in the LifeLines condition will perform with fewer errors and with a faster response time for those questions requiring: a) date/interval comparisons, b) approximate dates estimations with good clue location, c) multiple table lookups, and d) multiple column lookup in single table.

Likewise, it is predicted that subjects in the Tabular condition will perform with fewer errors and with a faster response time for those questions in which:

a) exact dates are requested, b) LifeLines provides ambiguous line overlap for the same information, c) single table lookup where LifeLines uses coding (LifeLines uses color and line thickness coding, whereas Tabular gives text value), Finally, it is predicted that there will be no difference in terms of number of errors and response times for those questions in which: a) approximate dates are requested but no location clues are given, b) exact intervals are requested., c) interval comparison with good clue location in the table/LifeLine display are needed, d) single table, single column lookup is needed, and e) exact date with a multiple table lookup is needed.

H3: Subjective Questionnaire

It is predicted that subjects in the LifeLines condition will have a higher level of user interface satisfaction than subjects in the Tabular condition.

H4: Recall test

It is predicted that subjects in the LifeLines condition will have a higher rate of recall than subjects in the Tabular condition.

A SECONDARY STUDY:
SPATIAL VISUALIZATION ABILITY

A secondary, but related study also investigated individual differences in terms of Spatial Visualization Ability (SVA). That is, is there a difference between high SVA and low SVA individuals in terms of performance?

Research suggests that we may find some differences due to SVA level. SVA has been shown to be closely tied to an individual's ability to successfully navigate through a hierarchical database (Butler, 1990, Norman & Butler, 1989, and Vincente, Hayes, & Williges, 1987). It is heavily dependent on the way in which the user represents the mental image -- whether the user has a pictorial or a verbal representation. Lohman (1989) observes that people use different methods for storing and manipulating mental images. He states that, "Some subjects solve items on such (paper folding) tests by generating mental images that they then transform holistically" (p.346) while other subjects use less visual means to solve these problems. He refers to the former group as high SVA and the latter group as low SVA. In this experiment, we used the VZ-2 (Ekstrom, 1976) to evaluate users and then looked at their performance on each interface.

This leads us to our final hypothesis:

H5: *SVA*

We expect that the high SVA individuals would perform better in the LifeLines condition and that the low SVA individuals might perform better in the Tabular condition thus supporting the need to use both representations in the actual interface.

METHOD

Participants

Thirty-six individuals from the University of Maryland participated in this experiment. The 20 male and 16 female subjects ranged in age from 18 to 49 years old. Each participant was paid $4.00 for taking part in the 40 minute experiment and was told that there would be an extra $4.00 incentive reward for the best performance (highest score in the shortest amount of time) in each condition.

Design

An independent groups design was used to look at subject's performance in terms of the format used for data representation. The independent variable, format, was defined as either LifeLines or Tabular. A series of unpaired t-tests were used to look at differences in terms of the dependent variable, response time for each of the 31 questions, and an error count was used to examine the dependent variable, number wrong. For the Subjective Questionnaire unpaired t-tests were used to look at the differences in ratings between participants in the two groups. For the first impression test a simple summary count was used to see the differences between the groups and finally, for the recall test, a single unpaired t-test was run to describe the difference in terms of number correct.

Finally, a secondary issue, SVA level versus format, was studied independently. Participants were divided into high versus low SVA as based on a median split (with the median score of 12.5, those who scored 0-12 were categorized as low SVA and those that scored 13-20 were categorized as high SVA). This created a 2x2 design for investigating whether there is an interaction between format and SVA Level. For this study, a 2x2 ANOVA was used to look at just the main effect of SVA level and the interaction. The main effect of format was not investigated here , as it was addressed in the previous part of the experiment.

Materials

Adobe Photoshop™ was used to create the two versions of the youth record (Figures 1 and 2) and a Borland Delphi program was created to run the experiment. A computerized version of the VZ-2 test of Spatial Visualization (Ekstrom, 1976) was used to determine SVA level, and the final questionnaire was based on the Questionnaire for User Satisfaction (QUIS) developed by Chin, Diehl, and Norman (1988). The subjects each ran the experiment on the same IBM PC machine running Windows 95.

Procedure

Participants were scheduled one at a time and were run individually at the computer. During the entire 40 minute session, the experimenter was seated nearby to answer questions and to provide the appropriate materials.

Spatial visualization ability test - After filling out the Informed Consent form, the subject was seated at the computer and asked to begin the VZ-2 portion of the experiment. Each subject had six minutes to complete this test.

Reading/training - When the VZ-2 was completed, the experimenter recorded the scores and gave the subject the proper reference sheet and the training hard copy for their condition (LifeLines or Tabular). Each subject was given plenty of time to fully understand the information and when done, notified the experimenter.

First impression test - Subjects were asked to look briefly (approx. 5 seconds) at hard copies of two youth records (see Appendix 1), and asked to answer the following question: "You have to place each youth in one of two facilities. One of the facilities is more secure than the other. Which youth needs to be put into the more secure facility?" Their answer was recorded and then they were given another 15 seconds (or more time, if needed) to look more carefully at the youth records and answer the same above question.

We were concerned that the LifeLines or Tabular representation might be misleading at first glance. A youth record may appear worse than the youth's actual behavior. In particular we knew from our DJJ user study that a record including many minor offenses but few convictions may appear to be “worse” than a record containing fewer but more severe offenses. We refer to the former type of record as Complex (more offenses, but less severe and no convictions, therefore a "better" record) and the latter as Severe (fewer offenses, but more severe and more convictions). Similar situations can be found in other types of records as well (e.g. for school records: a student with more classes but poor grades. For insurance records: a car driver with fewer but more severe accidents).

Main quiz - The main portion of the experiment, which was completely self-paced and on the computer, included a brief background questionnaire, five training questions, the actual experiment which consisted of 31 questions, and a questionnaire of user interface satisfaction. All participant's questions were answered prior to the actual experiment. At this point, participants were informed of a special bonus for the best score in each condition (highest score in the shortest amount of time).

The experiment consisted of 31 questions that were presented in the following manner: The question by itself was presented to the participant at the bottom of the screen. The participant read the question and then pressed the "Go" button when ready. The display appeared (a LifeLines or Tabular representation, depending on the condition) with the question and the possible answers visible at the bottom of the screen. The participant selected an answer, after which the text of the next question appeared. The completion time was recorded (i.e. the time between pressing the "Go" button and selecting an answer) as well as whether or not the answer was correct. Subjects had to complete each question in order to go on to the next question and they were not able to go back to previous questions.

The 31 questions are listed in Appendix 2. The questions were chosen to represent the diversity of possible tasks. We hoped to show the benefits of both the LifeLines and the Tabular display. For each question we tried to predict which format would perform better (see Appendix 2).

Subjective questionnaire - After completing the main quiz, subjects were asked to complete an eleven item subjective questionnaire which rated their experience during the experiment (see Appendix 3). This questionnaire consisted of a selected set of items from the Questionnaire of User Interface Satisfaction (Chin, Diehl, & Norman, 1988) which has a high reliability, Cronbach's alpha = .94. Responses were collected and then the subjects were debriefed.

Recall test - After debriefing the subjects were given one last six question hard copy post experimental memory questionnaire. The recall questions are listed in Appendix 4.

RESULTS

First impression test (H4)

Results (see Figure 3) indicate that at first glance, of the 18 participants in the Tabular condition, 6 thought the Complex record was more severe, 10 thought the Severe record was more severe, and 2 couldn't decide. For the LifeLines condition, only 2 thought the Complex record was more severe and 16 thought the Severe record was more severe. After more study, the results showed that, for the Tabular condition, 3 still thought the Complex record was more severe, and 15 thought the Severe record was more severe. For the LifeLines condition, nobody thought the Complex record was more severe, 17 thought the Severe record was more severe and 1 individual was undecided (that person had originally thought the Complex record was more severe).

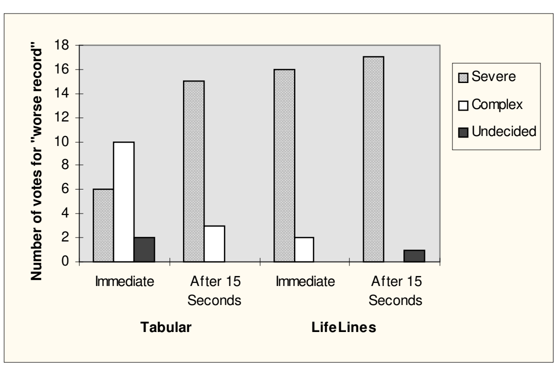


Figure 3. Seriousness Rating -- Number of subjects viewing that record as the more serious record (i.e. the worse record)

A Chi-square test of Independence comparing results for Tabular versus LifeLines produced the following results: Chi^2 (2) = 39.485, p<.01, which indicates that there is a relationship between whether a Tabular or LifeLines representation is used, and the perceived severity of each type of record (Complex or Severe).

Main quiz (H1)

Prior to the experiment, each of the 31 test questions had been categorized as: Tabular, LifeLines, or Both to indicate the condition in which performance was expected to be superior (See Appendix 2).

Twelve questions seemed better suited to a LifeLines representation. These involved:

- interval comparison
- multiple lookup table
- multiple column lookup

Nine questions seemed more suited to the Tabular representation and involved:

- exact dates
- exact values (coded in the LifeLines)
- information hidden by overlaps

Ten questions seemed equally suited for both (e.g., single table-single column lookup or approximate date questions.)

A t-test was performed for all the combined questions that were predicted to favor LifeLines. Results confirming our prediction were significant for t(34)=4.79, p<.0001. A mean comparison shows $M_{Tabular}$ =210.86 sec. and $M_{LifeLines}$ =106.85 sec.

Another t-test (Fig 4) was performed for all the combined questions that were predicted to favor the Tabular condition, results in this case were not significant for t(34)=-.04, p>.05. A mean comparison shows $M_{Tabular}$ =100.70 sec. and $M_{LifeLines}$ =101.22 sec.

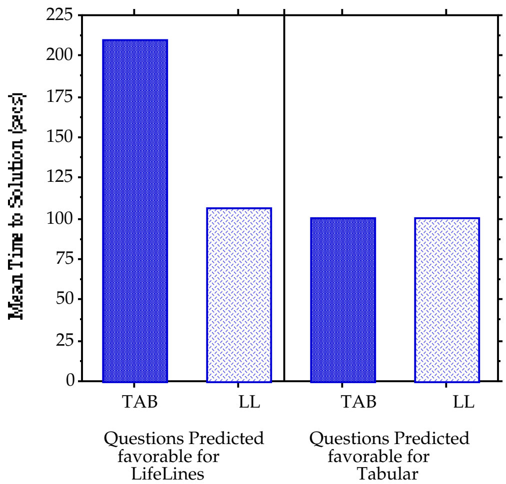


Fig 4: Time to complete combined tasks.

In addition a series of unpaired t-tests were used to determine the actual outcomes of individual test questions. For the most part, the data confirmed our predictions. A Bonferroni adjustment set the alpha level at .0016 (alpha = .05/31) to evaluate the 31 questions considered. The significant results are summarized in Table 1. Five of these scores were significant in the direction of LifeLines. These tasks included interval comparisons and tasks requiring Tabular subjects to look at two tables or two columns in the same table. The mean completion times were dramatically different, showing participants in the LifeLines condition performing twice as fast as Tabular.

The one question that was significant for the Tabular condition was question #26 with a means comparison score of: $M_{Tabular} = 5.41$, $M_{LifeLines} = 13.13$. This question involved a simple table lookup but required the interpretion of a color code on the LifeLines display. Since subjects were all novices, the color codes and names of the facility types were probably confusing and required most users to consult the printed training materials.
Although only these six items showed significant differences, it is beneficial to consider the mean differences for all the questions (Figures 4, 5 and 6).
These values are given in Appendix 2.

Question		$M_{Tabular}$	$M_{LifeLines}$	t-test, p<.0016
1	Which closed case was open for the longest time? *(interval comparison)*	8.94	**.67**	t(34) = 3.49
14	Which case(s) did Jones handle alone (for the entire case from beginning to end)? *(multiple column lookup in single table)*	8.56	**7.43**	t(34) = 3.44
17	As of today (10/16/95) at what facility did Joe Smith stay the longest? *(interval comparison)*	4.08	**.39**	t(34) = 4.52
19	Who was in charge of Joe Smith while he was in Cheltenham? *(multiple table lookup)*	1.45	**.34**	t(34) = 4.58
26	What type of a placement is Waxter? *(single table lookup - LL using color coding*	**.41**	3.13	t(34) = -3.52
27	During which case did Joe Smith have a critical medical event? *(multiple table lookup)*	3.61	**.72**	t(34) = 4.92

Table 1. Questions with Significant Differences in Mean Completion Times (seconds)

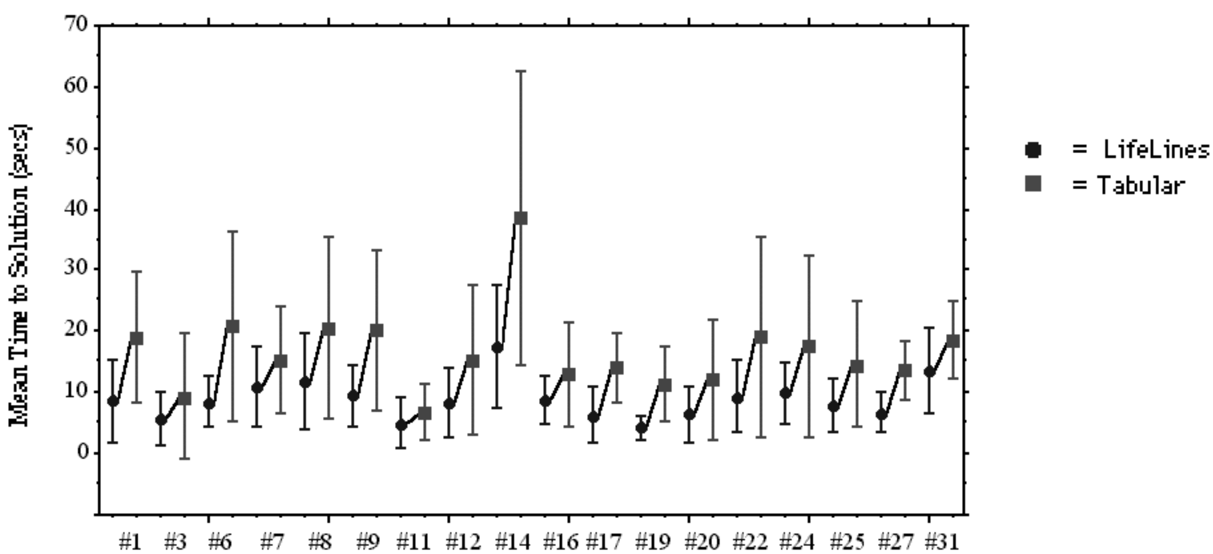


Figure 5. Mean Comparisons of Questions for which LifeLines was faster

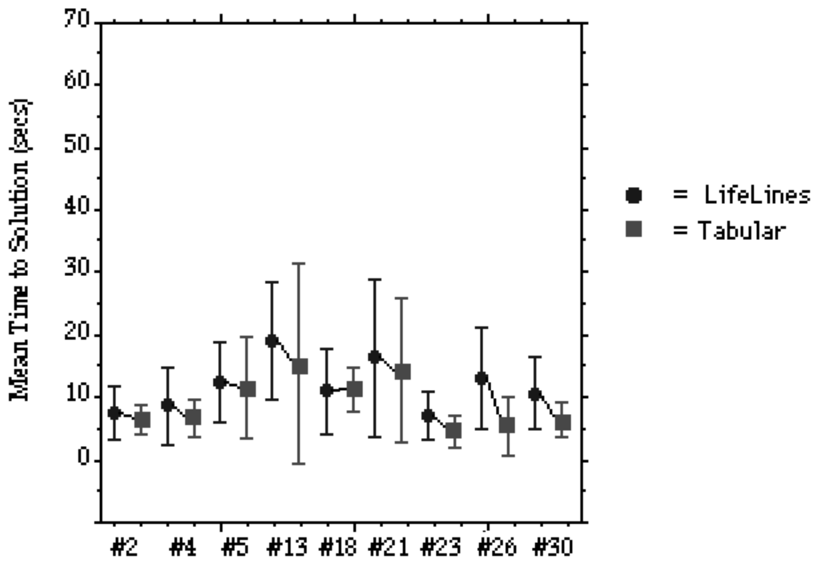


Figure 6. Mean Comparisons of Questions for which Tabular was faster

A look at the mean time differences of the non-significant questions reveals a pattern in favor of LifeLines. Most of the tasks show that users answered the question much faster on average with LifeLines than with the tabular display (e.g., twice as fast in #6, 12, 20, 22, 25 with similar errors rate).

None of the questions for which the Tabular condition had the faster mean time (i.e. "actual" Tabulars) were predicted to be faster in the LifeLines condition. However, two questions which performed better with LifeLines had been predicted to be better for Tabular. More of the questions predicted to be equivalent for Both performed better in LifeLines (5 out of 10) than in the Tabular (2 out of 10).

The following overall summary information is given:

(1) Total Mean Time (seconds)

$TMT_{Tabular} = 429.54$

$TMT_{LifeLines} = 302.02$

(2) Total Errors

$Errors_{Tabular} = 97$

$Errors_{LifeLines} = 135$

(3) Predicted versus Actual

$Predicted_{Tabular} = 9$, $Actual_{Tabular} = 9$

$Predicted_{LifeLines} = 12$, $Actual_{LifeLines} = 18$

$Predicted_{Both} = 10$, $Actual_{Both} = 4$

An overall unpaired t-test comparison between the two display types for the dependent variable of total time was significant ($t(34) = 2.96$, $p<.01$), and was faster for the LifeLines condition ($M_{Tabular} = 429.54$ and $M_{LifeLines} = 302.02$). In some cases there were more errors for the faster condition. If the difference between the two groups was 2 or more errors for the faster group, the question was marked as “Both” (i.e. no “winner”). The total number of errors for the LifeLines condition was higher than for the Tabular condition which we expected. Appendix 2 shows most of the errors occurred for questions in which the Tabular condition had faster response times. They were questions that did not provide sufficient information in the LifeLines condition and therefore were not expected to be answered accurately. These questions required the user to determine the answer based on either: a specific date (#4, #5, #9), overlapping events (#18, #21), or the decoding of color and thickness codes (#23, #26). To confirm that the source of the errors was these questions an additional t-test was run with those questions removed so as to look at performance scores -- time to completion and error rate -- for each item. Results were as follows: $t(34) = 3.67$, $p<.001$ with average completion times $M_{Tabular}= 14.06$ and $M_{LifeLines}= 9.19$, and similar error rates: $Errors_{Tabular}= 64$ and $Errors_{LifeLines}= 65$, confirming the origin of the errors.

Subjective Questionnaire (H3)

For the subjective questionnaire, each question was considered independently in a series of unpaired t-tests. A Bonferroni adjustment of alpha <.004 was used to evaluate the eleven questions considered. None of the eleven question were significant at this alpha level, which is not uncommon for a between-subjects experiment, however, trends indicated better (higher) scores for nine out of the eleven questions in terms of user satisfaction. In addition, an overall t-test was run on the mean score for all eleven questions, however, the result was not significant $t(33) = -.3$, $p>.05$. The questions and the results are shown in the Appendix 3.

The participants in the Tabular condition did say that their overall reaction to the display was better than did the participants in the LifeLines condition. Also, the participants in the Tabular condition said that they understood the terms better that the people in the LifeLines condition. Other than those two items, however, participants in the LifeLines condition said that their display was more: satisfying, stimulating, and clear, and that the characters were easier to read, the screen layout made the task easier, that there was adequate information on the screen, that learning to use the display was easier, and that learning to interpret the information was easier.

Recall test (H2)

Following the experiment, each participant was given a pop quiz -- a post experimental questionnaire to see what, if any, information had been retained. From the six questions asked, participants in the Tabular group only correctly recalled, on the average, 2.83 questions, while participants in the LifeLines group correctly recalled a better average of 4.33 questions. The results of an unpaired t-test were $t(34) = 3.82$, $p<.001$.

Spatial Ability (H5)

As a secondary issue, we looked at whether there was an interaction between SVA and format (LifeLines versus Tabular). Only the twenty five questions which resulted in a difference between the two groups were considered (no “Both” questions). Although the data did not show any interactions in an overall 2x2 ANOVA ($F(1,32) = .002$, $P>.05$) nor did it show any of the interactions for 2x2 ANOVAs using a Bonferroni adjustment of alpha $<.003$, it was interesting to note that from the means comparison, for the questions better suited to the LifeLines condition, response times were all faster for the high SVA individuals. Even more interesting is that for two out of the three questions better suited to the Tabular condition, the low SVA individuals had faster response times. We also noticed that the subject who performed most poorly using the LifeLines format also had a very low SVA.

DISCUSSION

The purpose of this research was to determine how well the LifeLines graphical data representation compared to the Tabular data representation which is commonly used in computer applications.

The first impression test confirmed that the representation of the data can have a strong influence on the first impression users have of a record. This small test seems to indicate that the LifeLines representation can give a better overall summary of the record than the Tabular representation. Designers have to carefully chose display parameters such as color, thickness, character size or style as they can lead to potential biases. But this test showed that even an ordinary tabular display can induce bias in users' first impression. Of course neither LifeLines nor the Tabular display contain all the information in the record but merely a summary and can only provide a subjective impression of the record.

As Norman (1993) notes, the type of format which is most appropriate for a particular task depends upon the nature of the task. Some tasks will undoubtedly benefit from graphical representations, but there are other tasks for which a Tabular representation may be more useful.

In the main quiz, we did find faster response times for the LifeLines condition, not only for those questions where we had predicted faster times, but also for some questions where we had predicted either the Tabular condition would be faster or where there would be no difference. Significant differences were found for tasks involving time interval comparisons and multiple table lookup. The speed gains were dramatic. However, the Tabular condition did have fewer errors than LifeLines. Many of the errors were linked to needing to guess at exact dates, overlapping events and understanding the graphical coding (color and thickness).

No significant differences were found for subjective user satisfaction but for 9 out of the 11 items on the subjective questionnaire, the participants in the LifeLines condition rated their system higher.

Another interesting finding was that there were fewer memory errors for the participants in the LifeLines condition -- it seems that those individuals were able to successfully recall more information that those in the Tabular condition.

From these results, it would seem that the LifeLines graphical interface does provide a good representation of the data. The timeline metaphor does seem to work since, for the most part, the performance of the users did show favorable results.

This experiment was run using static displays but an application using LifeLine can use interactive features to clarify and expand important pieces of information. Active cursors or balloon help can provide exact dates for events and exact values for coded attributes (either at the cursor or in a dedicated area of the screen). This simple technique helps compensate for one of the main weaknesses of LifeLines. A more serious weakness is related to the overlapping of events. In question #18 LifeLines users could not count how many medical events were in the record, but adequate display rules can be devised to spread events vertically, or make the important/critical events always visible or to provide special coding (e.g. a special color) to indicate overlapping event which can be revealed interactively. Zooming also provides a nice way to focus the overview of the record on areas of interest while increasing the resolution of the display. Another method consists of reserving a part of the screen for a small tabular display which can display the details of several related events (e.g., all the medical events).

As for the question regarding the interaction between SVA and format, there does seem to be some implication that individuals with low SVA may prefer to use the Tabular representation for those items which target more specific information (those questions for which the Tabular condition had faster response times). This also provides support for including textual information with the graphical representation for those low SVA users.

Regardless, this study points out many of the advantages of the LifeLines graphical representation. These results show that there are many benefits to using this type of graphical representation to display chronological records such as the DJJ youth record. Hopefully, the

LifeLines prototype display will help provide easy access to large databases of temporal personal history information and make the acquisition such data quicker and more effective.

In conclusion our study indicates that overall, LifeLines provides a useful summary of the record that users are more likely to remember. Tasks requiring interval comparisons and intercategorical information will be performed much faster than they would be from a tabular display. Finally, simple interaction techniques can augment LifeLines ability to deal with precise dates, attribute coding and overlaps.

REFERENCES

Butler, S. A. (1990) . *The effect of method of instruction and spatial visualization ability on the subsequent navigation of a hierarchical data base.* (CAR-TR-488 and CS-TR-2398) Department of Psychology and the Human/Computer Interaction Laboratory, University of Maryland, College Park, MD.

Carroll, J. M., & Mack, R. L. (1985). Metaphor, computing systems , and active learning. *International Journal of Man-Machine Studies,* 22, 39-57.

Chin, J. Diehl, V., and Norman, K. (1988). Development of an instrument measuring user satisfaction of the human-computer interface. *Proceedings of CHI'88 -- Human Factors in Computing Systems,* ACM, New York, 213-218.

Ekstrom, R. B., French, J. W., Harmon, H. H. (1976). *Manual for kit off factor-referenced cognitive tests.* Princeton, NJ: Educational Testing Service.

Hix, D., Hartson, H. (1993). *Developing User Interfaces: Ensuring Usability through Product & Process,* John Wiley & Sons.

Lohman, D. F. (1989). Human intelligence: An introduction to advances in theory and research. *Review of Educational Research, 59* , 333-373.

Nielsen, J. (1993) *Usability Engineering,* Academic Press, New York.

Norman, D. A. (1993). *Things that make us smart.* Addison Wesley Publishing Company, Reading, MA.

Norman, K. L., & Butler, S. (1989). *Search by uncertainty: Menu selection by target probability.* (CAR-TR-432 and CS-TR-2230). University of Maryland, Center for Automation Research and the Department of Computer Science, College Park, MD.

Paivio, A. (1986) *Mental Representations: A dual coding approach.* Oxford, England: Oxford University Press.

Plaisant, C., Milash, B., Rose, A., Widoff, S., Shneiderman, B. (1996). LifeLines: Visualizing Personal Histories, in *Proceedings of CHI '96 (Vancouver, BC, April 13-18, 1996),* ACM, NY, 221-227.

Plaisant, C., Rose, A., Shneiderman, B., Vanniamparampil, A. J., (1997) Low Effort, High Payoff User Interface Reengineering. *IEEE Software*, July/August 1997 67-72.

Plaisant, C., Shneiderman, B., An information architecture to support the visualization of personal histories. HCIL Technical Report, University of Maryland (December 1997).

Rose, A., Shneiderman, B., Plaisant, C. (1995). An Applied Ethnographic Method for Redesigning User Interfaces, in *Proceedings of DIS ’95 (Ann Arbor, Michigan, August 1995),* ACM, NY, 115-122.

Shneiderman, B. (1998). *Designing the User Interface: Strategies for Effective Human-Computer Interaction* (3rd ed.). Addison-Wesley Publishing Co., Reading, Massachusetts.

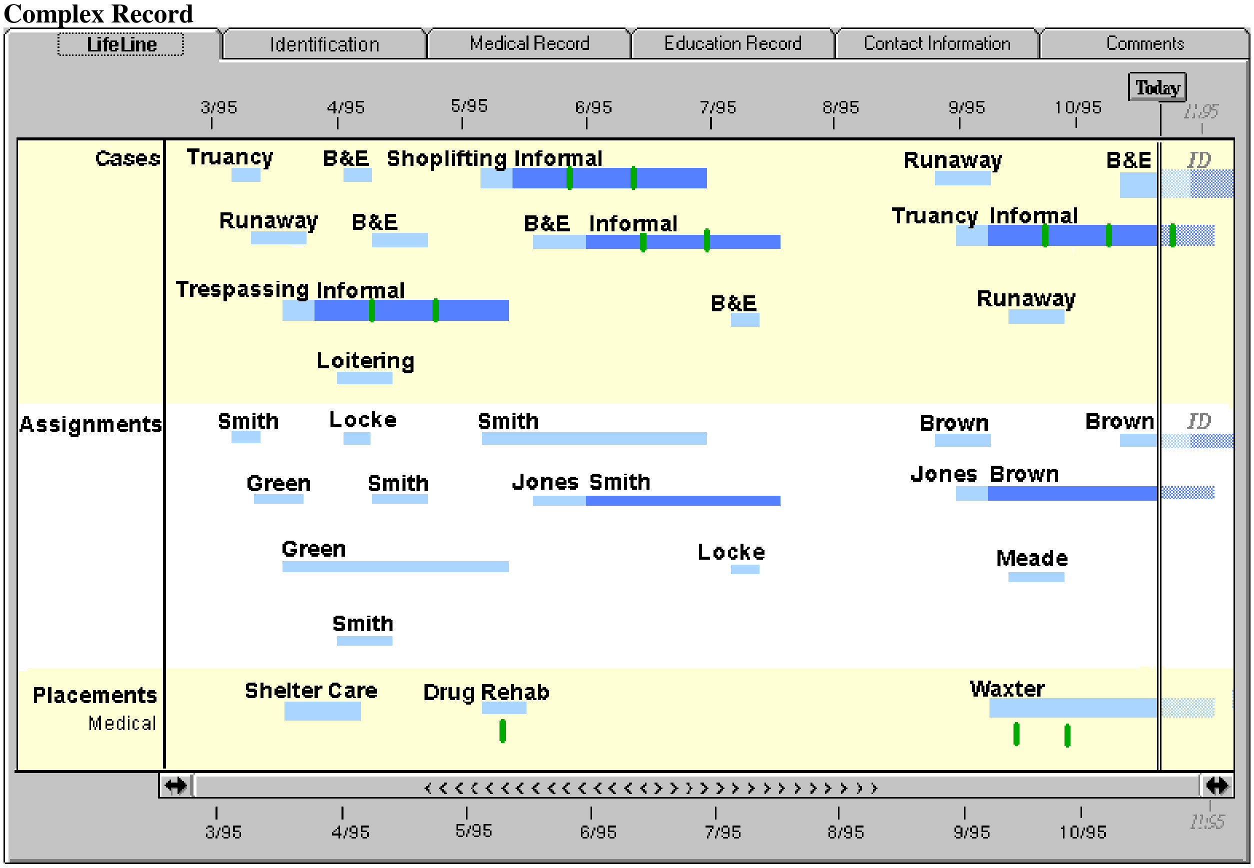
Slaughter, L., Norman, K.L., Shneiderman, B. (1995). Assessing Users' Subjective Satisfaction with the Information System for Youth Services (ISYS), in *Proceedings of Third Annual Mid-Atlantic Human Factors Conference (Blacksburg, VA, March 26-28, 1995),* 164-170.

Vincente, K. J., Hayes, B. C., Williges, R. C. (1987). Assaying and isolating individual differences in searching a hierarchical file system. *Human Factors, 29*, 349-359.

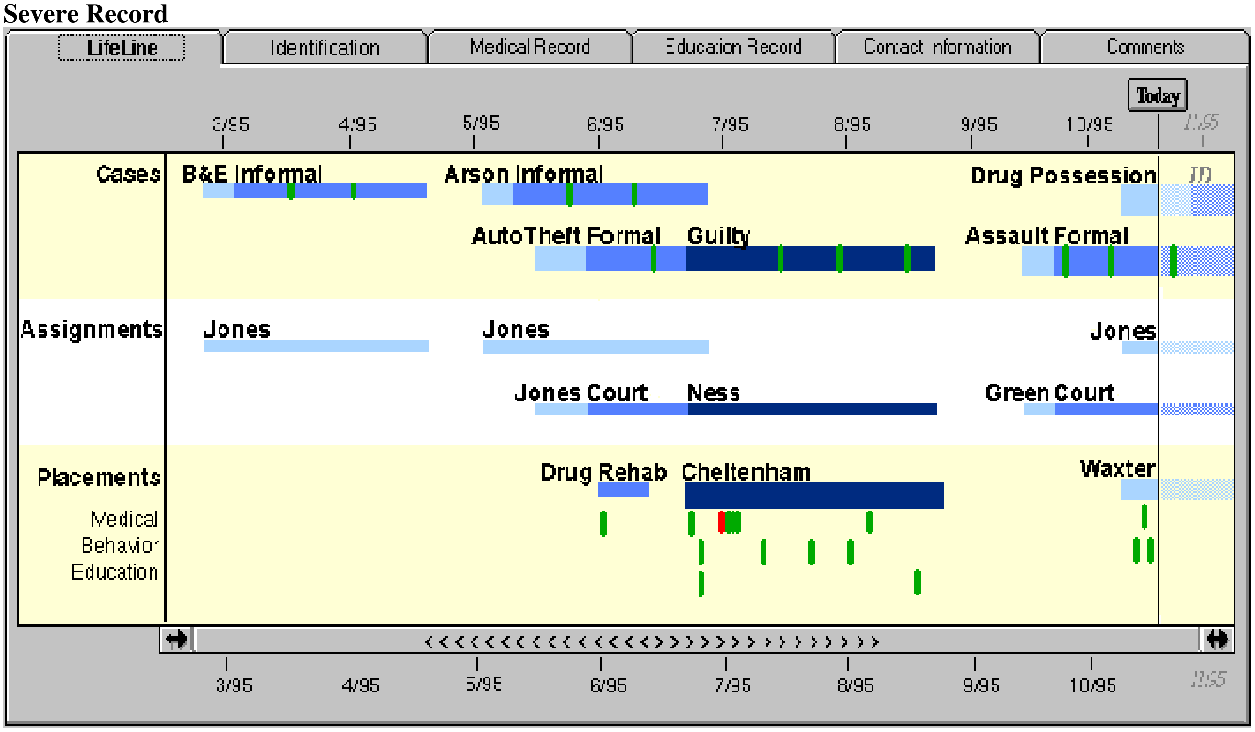
Wickens, C. D. (1992). *Engineering psychology and human performance* (2nd ed.). Champagne-Urbana, IL: HarperCollins Publishers Inc.

APPENDIX 1 - First impression test for LifeLines condition

Complex Record



Severe Record



APPENDIX 2 - Quiz Questions

* indicates that the results were statistically significant **(lines shown in bold)**

Note 1: The questions were predicted to favor either Tabular (T), LifeLines (LL), or Both (no real difference expected). The actual results are based, per question, on the mean time to answer and the number of incorrect answers . The faster time to solution is underlined . Actual results are rated as LL (advantage LifeLine) or T (advantage Tabular) or Both (no winner). Significant results appear in bold **(LL or T).** For the other questions we still report on the LifeLine vs. Tabular advantage trend. Some of the mean time differences are very large (up to 100 or 150% difference) and are marked in upper case (LL or T), while small advantages are maked in lower case (ll and t).

Note 2: Questions #10 and #28 are rated as “Both” for actual results because the LifeLine format had the fastest time but too many errors (a difference of more than 1).

Note 3: Question #18 and #21 are rated as actual “Tabular” because both formats had similar times but LifeLine had far more errors.

					Tabular		LifeLines	
#	Question	Predicted (note 1)	Actual	*	Avg. time, in sec	# wrong	Avg. time, in sec	# wrong
1	**Which closed case was open for the longest time?** *(interval comparison)*	**LL**	**LL**	*	**18.97**	**4**	**8.67**	**1**
2	In what month was the case of Arson closed? *(approximate date)*	Both	T		6.47	1	7.57	4
3	How many cases are still open as of today 10/16/95? *(interval comparison with good location clue in the table/Lifeline)*	Both	LL		9.43	1	5.75	1
4	What case started on 5/4/95 and ended on 6/29/95? *(exact date)*	T	t		6.55	0	8.75	3
5	On what date did the crime of Assault occur? *(exact date)*	T	t		11.70	4	12.72	4
6	Which 2 cases overlapped during June 1995 (were both active at the same time)? *(interval comparison)*	LL	LL		20.90	0	8.49	0
7	In which month did Joe Smith have his last review for Arson? *(approximate date with good location clue)*	LL	LL		15.31	1	10.84	1
8	A call was received on 10/9/95 referring to an active case. To which case would this call be associated? *(exact date - multiple table lookup)*	T	LL		20.62	10	11.95	10
9	A letter was received on 7/13/95. Which caseworker received that letter? *(exact date - multiple table lookup)*	T	LL		20.12	1	9.48	1
10	Which caseworker has never been assigned to Joe Smith's cases? *(single table, single column lookup)*	Both	Both (note 2)		14.48	3	12.98	5
11	Who has handled the majority of Joe Smith's cases? (*single table, single column lookup)*	Both	ll		6.78	0	5.03	0
12	Who was working with Joe Smith after he was found guilty of Auto Theft? *(multiple table lookup)*	LL	LL		15.23	8	8.38	7
13	Which 2 cases went to Court? *(multiple table lookup - misleading location clue)*	Both	T		15.35	3	19.03	8

14	**Which case(s) did Jones handle alone (for the entire case – from beginning to end)?** *(multiple column lookup in single table)*	**LL**	**LL**	*	**38.56**	**9**	**17.43**	**1**
15	Which case did Green handle? *(multiple table lookup)*	LL	Both		7.17	0	6.25	1
16	How many months did Joe Smith spend at Cheltenham? *(exact interval)*	Both	LL		13.09	0	8.78	2
17	**As of today (10/16/95) where did Joe Smith stay the longest?** *(interval comparison)*	**LL**	**LL**	*	**14.08**	**1**	**6.39**	**0**
18	How many times did Joe Smith leave Cheltenham for Medical Reasons? *(multiple table lookup but exact count with overlapping dates)*	T	T (note3)		11.04	1	11.02	6
19	**Who was in charge of Joe Smith while he was in Cheltenham?** *(multiple tables lookup)*	**LL**	**LL**	*	**11.44**	**1**	**4.34**	**1**
20	For what reason was Joe Smith sent to Cheltenham? *(multiple tables lookup)*	LL	LL		12.13	1	6.40	0
21	For what reason was Joe Smith sent to a Drug Rehabilitation Program? *(multiple tables lookup - with ambiguous line overlap)*	T	T (note 3)		14.44	6	16.35	10
22	How many of the cases that Jones handled have gone to Court? *(multiple column, single table lookup)*	LL	LL		19.20	8	9.37	7
23	Which alleged offense has the highest severity rating? *(single table lookup - LL using thickness coding)*	T	T		4.48	6	7.18	14
24	What decision has been made about Joe Smith's innocence in reference to the case of Drug Possession? *(multiple column, single table lookup)*	Both	ll		13.82	3	9.91	4
25	How long will Joe Smith be staying at Waxter? *(single table lookup)*	Both	LL		14.65	3	7.91	3
26	**What type of a placement is Waxter?** *(single table lookup - LL using color coding)*	**T**	**T**	*	**5.41**	**2**	**13.13**	**11**
27	**During which case did Joe Smith have a critical Medical event?** *(multiple table lookup - date/interval comparison)*	**LL**	**LL**	*	**13.60**	**0**	**6.72**	**0**
28	How many new cases, placements, and assignments are there between 10/2/95 and Today 10/16/95? *(exact dates, multiple table lookup)*	Both	Both (note2)		25.59	11	13.78	14
29	Where is Joe Smith currently placed? *(single table lookup)*	Both	Both		3.99	0	3.31	0
30	For which case was there a Review on 6/15/95? *(exact date - ambiguous on LL)*	T	T		6.33	0	10.64	7
31	From the information given on the display, would you say that Joe Smith's behavior: improved over time, worsened over time, stayed the same, worsened, then improved, or cannot determine	LL	LL		18.59	9	13.45	9
	Total:	LL=12 T= 9 Both =10	LL = 18 T= 9 Both= 4	6	429.54	97	302.02	135

APPENDIX 3 - Subjective Questionnaire Items and Scores

SUBJECTIVE QUESTIONNAIRE:

PART A. Overall Reactions:

1. Overall Reactions to the display: terrible — wonderful
   1 2 3 4 5 6 7 8 9 NA
2. frustrating — satisfying
   1 2 3 4 5 6 7 8 9 NA
3. dull — stimulating
   1 2 3 4 5 6 7 8 9 NA
4. confusing — clear
   1 2 3 4 5 6 7 8 9 NA

PART B. Screen

5. Characters on the computer screen were: hard to read — easy to read
   1 2 3 4 5 6 7 8 9 NA
6. Screen layout makes the task: harder — easier
   1 2 3 4 5 6 7 8 9 NA
7. Amount of information that can be displayed on the screen
   inadequate — adequate
   1 2 3 4 5 6 7 8 9 NA
8. Arrangement of information on screen: illogical — logical
   1 2 3 4 5 6 7 8 9 NA

PART C. Terminology and Learning

9. Terms were: confusing — clear
   1 2 3 4 5 6 7 8 9 NA
10. Learning to use the display was: difficult — easy
    1 2 3 4 5 6 7 8 9 NA
11. Learning to interpret the information was difficult — easy
    1 2 3 4 5 6 7 8 9 NA

Subjective Questionnaire Scores

The following results are based, per question, on the (1) total mean score on a 9-point Likert-type scale for each question in each group (T or LL), and (2) a comparison between the expected and observed results. A higher score indicates a better subjective reaction to the group.

Question	T	L	"Better" (not significant)
	5.83	5.28	T
2	4.78	5.56	LL
3	4.5	6.33	LL
4	4.56	5.11	LL
5	5.67	6.33	LL
6	5.11	6.39	LL
7	6.22	6.72	LL
8	5.06	6.39	LL
9	6.22	5.39	T
10	6.06	6.67	LL
11	5.28	6.17	LL

APPENDIX 4 - Recall Questionnaire Items

1. How many different youth records were displayed during the real experiment (not including the practice session)? *(asked by the experimentor)*
2. How many cases were there for Joe Smith?
3. Which was the longest case for Joe Smith?
4. How long was Joe Smith in Drug Rehabilitation?
5. Approximately how many months ago was the last critical event?
6. Approximately how long was the entire youth record?