

Evaluating models of visual comprehension

Mary Ellen Foster (M.E.Foster@ed.ac.uk)
Institute for Communicating and Collaborative Systems
School of Informatics, University of Edinburgh
2 Buccleuch Place, Edinburgh EH8 9LW
<http://www.iccs.informatics.ed.ac.uk/~mef/>

Abstract

This paper describes the results of a study evaluating the predictions of two existing models of graph comprehension: BOZ (Casner, 1990) and UCIE (Lohse, 1993). Each model was implemented in Java and was used to make predictions about the relative efficiency of different graphical presentations of numerical data for use in different tasks. These predictions were then compared with the results of human subjects performing the same tasks using the same presentations.

The results of the human study do not correspond to the predictions of either model. In particular, while both models predict that tabular presentations would have the worst results in practice the tables actually proved to be the best presentation type. A possible explanation for this result is that the models capture optimal, expert performance, while the subjects used less efficient techniques. Updating both the models and the experimental setting should make it possible to bring the predicted and actual results more in line with one another.

Overview

Intuitively, one way of determining the usefulness of any presentation of data is by measuring the ease with which tasks involving that data can be performed using that presentation. Several studies have shown that the effectiveness of a graph can vary from task to task—see, for example, Zhang (1996). Selecting a graph to use in a particular context therefore depends not only on the data to be presented, but also on the task that the user is intended to perform using that data.

In the context of a system that aims to generate presentations automatically, one possible method of choosing among multiple presentation techniques is to simulate the user task on each of the presentation alternatives, and to choose the presentation that allows for the most efficient achievement of the task.

The goal of the current study is to evaluate two existing models of how users comprehend and make use of “information graphics” (bar charts, line graphs, and tables). If a model does a good job of predicting human performance, it could be used as part of a generation system as described above. The models were evaluated by comparing their predictions on a range of graphs and tasks to actual human performance on the same graphs and tasks.

Task type	Example
COMPARE ₁	Was the exchange rate of IEP greater in 1993 than in 1992?
COMPARE ₂	Was the price of Sulphur less than the price of of Copper in 1986?
READ	Was the market share of Toyota in 1983 less than 10.0?
TREND	Was the exchange rate of USD generally increasing between 1991 and 1997?

Table 1: Sample task of each type

This paper is arranged as follows. First, the tasks and presentation types that were used in the simulation and the human experiment are described, along with the graph-comprehension models that were used. Next, the design and results of the simulation study are described, followed by those of the experimental study. Then, the predicted results are compared with the actual results, and an explanation is proposed for the differences between the two. Finally, some possible future studies that could lead to models that more accurately capture human performance are proposed.

Tasks, presentation types, and models

This section describes the tasks and presentation types that were used in both the simulation and the experiment. The tasks and the presentation types were based on those used by Lohse (1993).

Tasks

All of the tasks for the simulation and the experiments consisted of answering a *yes-no* question about the data presented on the graph. The task was of one of four types, as follows:

COMPARE₁ Comparing adjacent values within a series

COMPARE₂ Comparing values in two adjacent series

READ Reading a single value from a series

TREND Reading a sequence of values from a series

An example of each task type is given in Table 1.

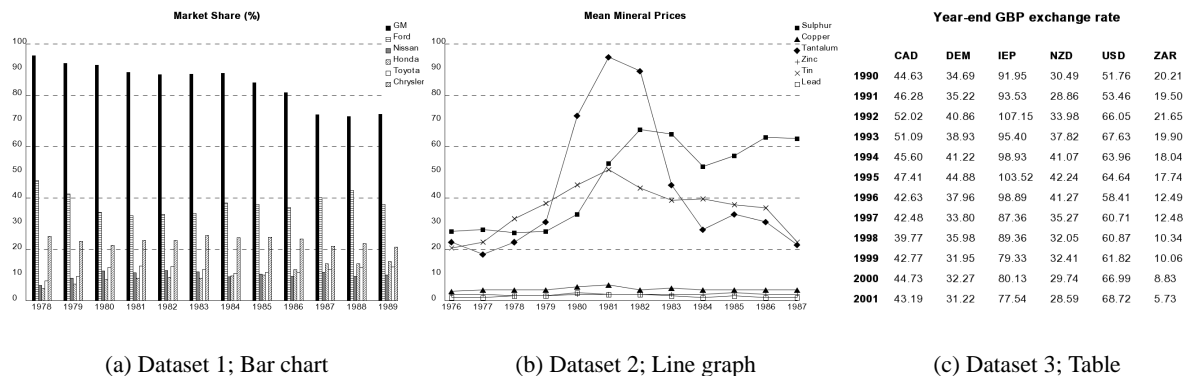


Figure 1: Datasets and sample presentations

Presentation types

There were three types of presentations used: bar charts, line graphs, and tables. An example of each type of presentation is shown in Figure 1. Each presentation was produced in black and white, and grid lines were added to the bar charts and line graphs as shown in the figure.

The data used to produce the graphics consisted of three sets of time-series data. Each set was made up of six data series, and each series had twelve points. The sample presentations in Figure 1 also show the three datasets used. The first two datasets were adapted from the examples used by Lohse (1993), while the third was created using a historical currency-exchange website.

Models

The following two models of graph comprehension were implemented for this study.

BOZ (Casner, 1990) This model begins with an abstract *logical procedure* that a user would employ to perform a task. The logical procedure is then translated into a *perceptual procedure*, which is a sequence of perceptual operators that can be used to accomplish the task using a particular presentation. For example, on a graph where the height of the bars encodes the cost of flights, the logical operator *determine-cost* might be translated into the perceptual operator *read-height*. Under certain circumstances, BOZ also permits steps in a perceptual procedure to be skipped when it is run on the graph. The efficiency of a perceptual procedure is assessed by counting the number of search and lookup operations required to run it on a graph. In this model, a more effective graph for a task is one that permits a more efficient perceptual procedure to implement the task's logical procedure.

UCIE (Lohse, 1993) This model predicts the sequence of eye fixations that will answer a question posed to a graphics display. That is, it predicts where a person doing the task would look, and the operations that the person would perform at each location. It then determines

the time required to process the information during each fixation, using assumptions based on results from previous human-factors studies. It sums the component times to predict the total time it would take to answer the question. Using this model, a more effective graph is one that requires less predicted time to answer the question.

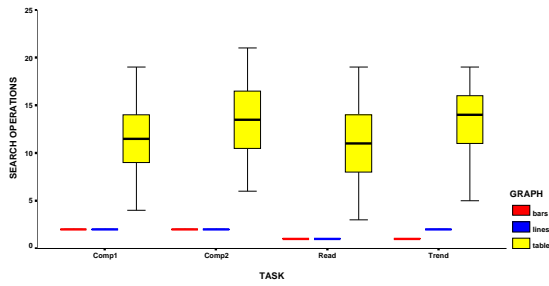
Simulation study

All of the tasks described in the preceding section were simulated on all of the graph types and all of the data series, using both models. This section describes the design and results of this simulation study.

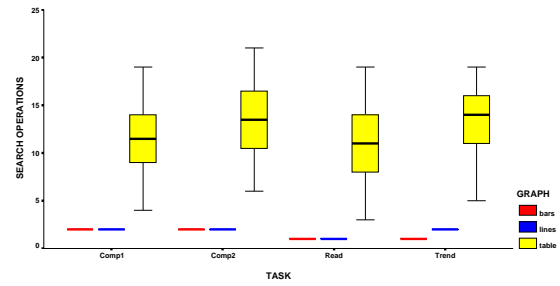
Simulating the models

Each model was implemented in Java. For the BOZ model, a logical procedure was created by hand for each of the task types. Casner's original tasks were primarily iterative search tasks such as *Find the cheapest flight from Pittsburgh to Mexico City*, so these procedures were created in the spirit of the original model. Each logical procedure was then translated—also by hand—into a perceptual procedure for each graph type. In Casner's original implementation, the pre-written logical procedures were automatically translated into perceptual procedures as part of the process of generating the graphs. Since the purpose of the current study was just to compare the model's predictions on different graphs, rather than to generate the graphs themselves, this step was not necessary here.

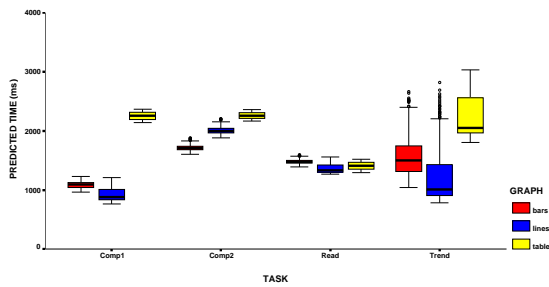
For the UCIE model, the sequences of high-level moves that would be required to perform each of the tasks on each of the graphs were created by hand in advance of the simulation. Examples of such moves are *find the value 1986 on the x-axis* or *read the value for Copper at the current axis position*. A predicted time was assigned to each fixation at run time, based on the method described in the Lohse paper. In Lohse's model, the sequence of movements was automatically determined from the question description; however, as with BOZ, since this was not the focus of the current study, these hard-coded task descriptions were sufficient.



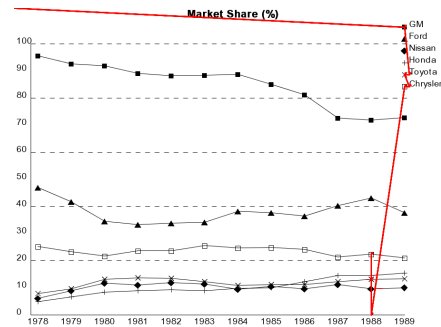
(a) BOZ: Search operations



(b) BOZ: Lookup operations



(c) UCIE: Predicted time (ms)



(d) UCIE: Sample simulated eye trace

Figure 2: Simulation results

The implementations of the two models were used to make predictions about every possible instantiation of each of the tasks, on each of the graph types, using each of the data series. This made a total of 7344 observations for each model. For the BOZ model, an observation consisted of the predicted number of search and lookup operations required to perform the task instance; on the other hand, the UCIE simulation produced a prediction of the total time required and a simulated eye trace.

Results

For both models, the results were broadly similar to those presented in the original papers. By the nature of the models, the specific results produced by each are very different; each set of data will therefore be discussed on its own initially. The results of the two models will then be compared qualitatively at the end of this section.

BOZ The overall results of the BOZ simulation are shown in Figure 2(a–b).¹ Notice that the counts for the bar chart and line graph are nearly identical for each task, with no variation in either. Neither the similarity nor the lack of variation is surprising: the operators used in

¹In Figures 2 and 4, the graphs should be read as follows: The “boxes” represent the 25–75 percentile range; the line inside the box is the mean of the observations; and the “whiskers” represent the highest and lowest values, excluding outliers (which are represented as dots).

BOZ’s procedures are so high-level that they are not sensitive to variations in graph types, datasets, or individual data items.

In all cases, the table required many more search operations than the other two graph types, with more variation in the individual values. The variation comes solely from the row and column in the table that contain the object or objects to be read; cells that are further from the top and left margins require more operations to locate them. The prediction that tables are worse than the other types is also not surprising, as the focus of Casner’s work was to design presentations that were better than tables; that is, tables were essentially a baseline. His simulation method—particularly the mechanism by which steps may be skipped in a perceptual procedure—greatly penalises operations on tabular presentations.

UCIE The results of the UCIE simulation are shown in Figure 2(c), while Figure 2(d) shows the simulated eye trace for answering the question *Was the market share of Toyota less than the market share of Chrysler in 1988?* using a line graph.

The UCIE results show more variation than those of the BOZ model. In almost all cases, the table has the longest predicted time, particularly for COMPARE₁ tasks (comparing adjacent values within a single data series). The line graph, on the other hand, has the shortest predicted time in all cases except for READ. The TREND

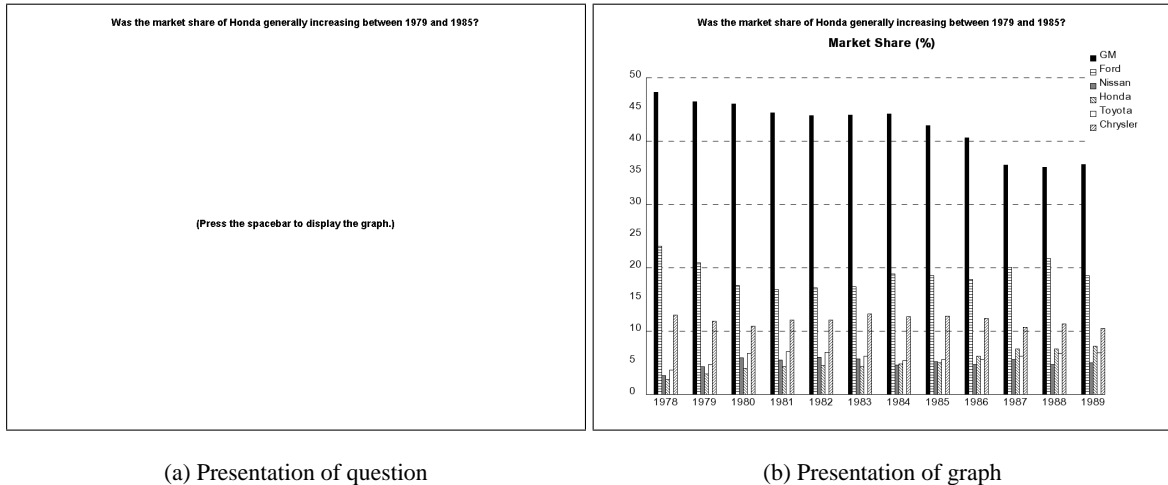


Figure 3: Experimental environment

task shows the most variation in predicted times; the ordering of the per-graph means is the same as in the other cases, but there is a larger amount of overlap between the predicted times here than for other tasks.

Summary The main results of the simulation study can be summarised as follows:

- The predictions of both models were broadly in line with the results presented in the original papers from which the models were adapted.
- In almost all cases, the table is predicted by both models to be the worst presentation type.
- Bar charts and line graphs have similar predicted results, especially in BOZ; when there is a difference, the line graph is almost always better.
- UCIE shows much more variation in its predictions than BOZ. This is because it uses lower-level operations that are more sensitive to differences between the graphs. For example, in the sample eye-trace in Figure 2(d), discriminating the mark for Toyota from the surrounding marks will take longer than performing the same operation on the Chrysler mark, since there are more graphemes near the Toyota mark.

Experimental study

The predictions from the simulation were compared with the results of human subjects performing the same tasks using the same set of materials. This section outlines the design of this experiment and its results.

Experiment design

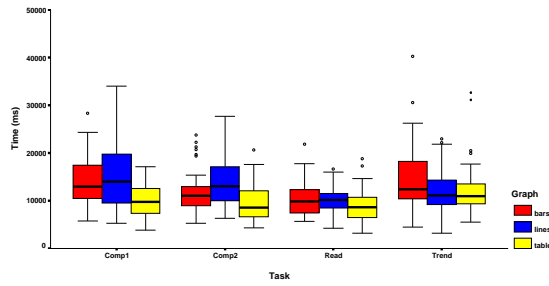
Subjects The participants in this study were 33 post-graduate students at the University of Edinburgh. All had normal or corrected-to-normal vision. Twenty-one of the

subjects were native speakers of English, while 12 spoke English as a second language.

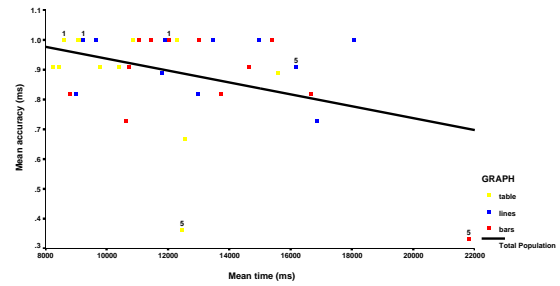
Tasks Each subject answered the same twelve *yes-no* questions: one question per type on each presentation type. Samples of the questions used are shown in Table 1 on the first page. The type of graph used for each individual question was balanced across the subjects, so that there were the same number of responses for each combination of question and graph type.

Procedure The experiment was administered using computers to present the questions and the graphs, as follows. The subject was instructed to press the space bar to display the question to be answered; the question was then displayed near the top of the screen (Figure 3(a)). When ready, the subject then pressed the space bar again to display the graphical presentation; the question remained on the screen along with the presentation (Figure 3(b)). The subject then answered *yes* or *no* to the question by pressing the appropriate key on the keyboard—the **A** and **'** keys on a standard keyboard were labelled with *Yes* and *No* stickers respectively. As soon as the question was answered, the question and graph disappeared and the subject was then prompted to press the space bar to display the next trial.

Each subject first answered a warm-up round of four questions, one question of each type. The warm-up questions were the same for all subjects. Once the warm-up round was completed, the subjects answered the twelve experimental questions, which were presented in a different random order for each subject. For each trial, the program recorded the following data: the time between the appearance of the graph and the subject's response, and whether the question was answered correctly. No feedback was provided to the subjects on the correctness of their answers.



(a) Response time (all answers, all subjects)



(b) Speed vs accuracy in experimental results

Figure 4: Results from experimental study

Results

Figure 4(a) shows a graph of the response times from all users on all tasks. This section discusses various aspects of these results.

Speed vs accuracy Figure 4(b) plots response time against accuracy for all trials, for all subjects. The mean accuracy over all of the trials was 0.94, and there was a small but significant negative correlation between reaction time and accuracy (-0.1183 , $p = 0.019$). In other words, questions that were answered incorrectly also took more time to answer. The data is thus free from any concern of a speed-accuracy trade-off.

Graphs and tasks² Qualitatively, for most of the tasks, the table was the presentation type that permitted the fastest responses, and the line graph the slowest. Only for TREND was the pattern different: for that task, the table was actually the slowest and the line graph the fastest.

An ANOVA was performed on the times of the correct answers. On the initial data, the response times of the native speakers (mean = 11.7s) were significantly faster than those of the second-language speakers (mean = 13.0s). An ANOVA was therefore performed on each of the two groups separately.

For both the native and the second-language speakers, the graph, dataset, and task had a significant effect on the response time (all $p < 0.001$). However, the interaction of interest—graph \times task—was significant ($p < 0.005$) only for the group of native speakers. The trends were similar for both groups, so this lack of significance for the second-language speakers was probably due simply to the smaller size of that group (12 vs 21).

Individual items The individual task instances were generated randomly, which meant that some of the questions proved much more difficult to answer than others. The two extremes were items 1 and 5, both of which were based on Dataset 2 (Figure 1(b)).

²The analysis in this section was performed only on the correct answers; the analysis on the entire set of observations produces very similar results.

Item 1 Was the price of Tantalum in 1979 greater than 100.0?

Item 5 Was the price of Copper generally increasing between 1976 and 1985?

The points corresponding to these tasks are labelled in Figure 4(b).

The mean response time on item 1 was 9.97 seconds, with a mean accuracy of 1.00—that is, *every* subject got this question right, whatever the presentation type. This question was especially easy to answer because the values in all of the series are significantly less than 100.0, so it is not even necessary to examine the graph in any detail.

In contrast, the mean response time for item 5 was 16.8 seconds, with a mean accuracy of 0.54. This task was particularly difficult because the value for Copper remains almost constant during the period in question, and it is hard to tell with most presentations whether it is increasing or decreasing at all. However, with the line graph, the accuracy on this question was 0.91—so, for this task, the line graph was actually the best alternative.

Summary The main results of the experimental study can be summarised as follows:

- There were no signs of a speed-accuracy trade-off in the results; the incorrect responses were generally given more slowly than the correct ones.
- The table was the presentation type that permitted the fastest answers for every type of question except for TREND; for this task, the line graph was the fastest.
- Some of the questions (e.g., Item 1) were *much* easier to answer than others (e.g., Item 5).

Discussion

Neither of the implemented models does a very good job of predicting the actual data. In particular, both models predicted that the table would be the worst presentation type for all tasks; however, for the human subjects, the

table was actually the fastest and most accurate presentation in almost all cases. Similarly, the line graph, which was the worst in almost all cases for the humans, was predicted by both models to be the best presentation type.

Another difference between predicted and actual performance is in the actual time taken to answer the questions, which were much longer than the times predicted by the UCIE model. The times predicted by UCIE are all less than four seconds, while the mean of the actual times is well over ten seconds.

Expert vs novice performance

The differences summarised above between predicted and actual performance indicate that the processes simulated by the models are, in fact, not the same processes that the subjects used. A likely explanation for the difference is that the models simulate *expert* performance—that is, they predict the *minimum* sequence of operators or fixations required to answer the questions. It is likely that the subjects did not make use of such minimal procedures.

This hypothesis is supported by the work of Carpenter and Shah (1998), who describe an experiment in which they used eye-tracking equipment to follow the gaze of subjects as they performed various tasks using line graphs. They found that subjects did not use minimal paths such as the UCIE predicted trace in Figure 2(d); rather, they constantly switched their gaze between different parts of the graph: the title, the legend, the axes, and the pattern itself.

Although the tasks used in the current study were somewhat different than those used by Carpenter and Shah, it is likely that the subjects used a similar iterative technique to answer the questions, which could explain the discrepancy between the predicted and actual results. Some subjects also commented informally after the experiment that they felt they had indeed been constantly looking back and forth between different parts of the graph as they answered the questions.

Next steps

To meet the goal of a visual-comprehension model that accurately captures human performance, there are two possible sorts of modifications to the current study; both sorts of modifications can proceed in tandem.

On one hand, the models can be updated so that they more accurately capture the task. On the other hand, further experiments can be performed to get a better handle on the exact nature of the task, and attempt to induce more expert performance in the subjects. This section describes some possible next steps in both of these directions.

Revised models The predictions of both models are generally in line with the results presented in the original papers. Given the original intention of BOZ and its use of high-level operators, it is unlikely that its predictions can be greatly different than those presented here; however, it is worth verifying the UCIE model to ensure

that it truly represents the one used by Lohse (1993).

However, both of these initial models still ultimately provide simulations of expert performance. A more accurate and useful model would be one that incorporates the iterative-comprehension model of Carpenter and Shah (1998). This model is not as straightforward to implement as the two that were studied here; however, ideas from that work and results from the further experiments described below can be used to update the existing implementations.

Further experiments In the study described here, the specific instances of the tasks were chosen randomly. This led to some tasks that were either extremely easy (Item 1) or extremely difficult (Item 5). The tasks should be chosen in a more systematic manner to make sure that the “easy” and “hard” cases are covered equally well.

As well, some other features of the graphs could be varied, such as: the inclusion of more or fewer data series or series with more or fewer points; the use of colour in the presentations; and possibly other types of presentations such as pie charts or horizontal bar charts.

The subjects in this study were presented with an assortment of graphs in a random order, which probably contributed to the less-than-minimal strategies that they adopted. A possible way to obtain more expert-like, minimal performance is to use familiarity within the experiment. For example, subjects could answer a number of questions in a row using the same presentation type, or even the same presentation instance. As the presentations become more familiar, the subjects should converge towards the minimal strategies predicted by the models.

Conclusion

This study compared the predictions of two existing models of how humans answer questions based on graphical presentations with the results of people actually performing the modelled tasks. The predictions of both models were somewhat different from the actual results. One possible explanation for this difference is that the models were of idealised expert performance, while the subjects used less efficient, iterative techniques such as those described by Carpenter and Shah (1998). The next step in this work is to modify both the simulation models and the experimental set-up in various ways that should bring the predicted and actual results more in line with one another.

References

- Carpenter, P. A. and Shah, P. (1998). A model of the perceptual and conceptual processes in graph comprehension. *Journal of Experimental Psychology: Applied*, 4(2):75–100.
- Casner, S. (1990). *Task-Analytic Design of Graphic Presentations*. PhD thesis, Department of Computer Science, University of Pittsburgh.
- Lohse, G. L. (1993). A cognitive model for understanding graphical perception. *Human-Computer Interaction*, 8:353–388.
- Zhang, J. (1996). A representational analysis of relational information displays. *International Journal of Human-Computer Studies*, 45:59–74.