

# VISUAL RATING SYSTEM FOR HFES GRAPHICS: DESIGN AND ANALYSIS

Paul Aumer-Ryan  
The University of Texas at Austin  
Austin, Texas

This paper introduces research that evolved out of an analysis of Gillan, Wickens, Hollands, & Carswell's (1998) guidelines for the graphical presentation of quantitative data in HFES publications. The impetus for this research stemmed from a concern about substandard illustrations accompanying publications and the difficulty in communicating how to improve the quality of those illustrations. The aim in this paper is twofold: (1) to offer a pragmatic method for operationalizing and measuring the adherence of a specific graphic to the aforementioned guidelines; and (2) to present the analysis of preliminary data gathered using the metric developed. I will first present the methodology used to develop a Web-based rating interface, then examine the implementation of that interface, and finally discuss the results of a comparative analysis performed on samples of graphics published before and after the original guidelines. Counterintuitively, the only evidence we found of a change between graphs published before the guidelines and those published after was a decrease in the readability of text.

## INTRODUCTION

In some fashion, the research herein relies on the oft-repeated truism that a picture is worth a thousand words (e.g., Larkin & Simon, 1987). Especially in the presentation of scientific research and quantitative data, a few annotated lines or bars can communicate a concept (say, exponential growth) much better than a fully textual attempt can. In that vein, most authors published in HFES venues elect to include at least a few graphical descriptions of their data, whether it is an advanced statistical routine or a basic comparison of variables. However, as is often the case, it is easy for us to take the truism at face value, and simply believe that *any* picture is worth a thousand words. Such are the vagaries of life: a truism strikes us as true because of its depth of meaning, but after repeated use the meaning gets lost in compulsory behavior. The point here is that well-designed pictures (in our case, graphs) are adept at communicating what words alone cannot. The work here is dedicated to reviving the above truism by encouraging diligence and respect for the process of creating understandable and readable graphs in scientific publications.

Building upon research on the human perception and comprehension of different presentations of quantitative data, most notably graphs (e.g., Simkin & Hastie, 1987; Pinker, 1990; Lohse, 1993; Gillan & Lewis, 1994), Gillan, Wickens, Hollands, & Carswell (1998) presented a set of guidelines for authors submitting graphics with their publications to the various HFES venues. The ostensible purpose of these guidelines was to improve the quality of the accompanying graphics by presenting authors with a convenient resource they could use during the production of their graphics. The authors communicated this intent via illustrated examples and 26 numbered guidelines containing a mixture of directives (e.g., textual labels should be large and discriminable), factors to take into consideration (e.g., readers tend to perceive line slopes as nearer to 45 degrees than is actually the case), and techniques to avoid (e.g., exercise caution in using 3D graphs).

## DESIGN

In evaluating these guidelines I counted over 90 specific messages for authors, which struck me as a large enough number to cause difficulty in referring to them during the creation of graphs. Additionally,

HFES Graphics Guidelines - Visual Rating System

**Graphic**  
Now evaluating number 2 of 50 (user: eval00).

Target Density	Mean Percentage Task Error
1	~2
2	~4
3	~5
4	~7
5	~10
6	~13
7	~15
8	~16
9	~17

Figure 2. Mean percentage task error as a function of target density for all participants. Error bars represent one standard error of the mean.

Investigators: Paul Aumer-Ryan, Dr. Randolph Bias

**Instructions and Definitions**  
Welcome and thank you for taking the time to help evaluate these graphics. Your participation in this exercise will likely lead to world peace. Be proud of yourself for making this monumental effort. I know I am. (Seriously, I can't thank you enough.)  
Hint: Press [F11] to make the browser window fullscreen (this will give you more space to see the graph and questions).  
Glossary (click on the term to see an example):  
Axes: y axis (vertical line at the left of the graph) and x axis (horizontal line at the bottom of the graph).  
Graph types: Bar graph, Line graph, Pie chart, Scatter plot.  
Indicator: Geometric elements of the graph that express the value of the dependent variable (e.g., lines and shapes in a line graph and bars in a bar graph).  
Label: Descriptive words for the axes, indicators, and legend.  
Quantitative label: Numbers on the axes or by indicators (to express their value).  
Tick marks: Lines on an axis that indicate even divisions in the axis.

**Graphic Type**  
This section is concerned with the selection of the appropriate graph type.

	Strongly disagree	Strongly agree	N/A
This line graph is used to show rate of increase, interactions between independent variables, or relative/absolute amounts.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
This line graph has an independent variable that is continuous or ordinal.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

**Shapes and Patterns**  
This section is concerned with the geometric aspects of the graph.

	Strongly disagree	Strongly agree	N/A
The styles of lines are varied to make them distinct from one another. [Example]	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
The styles of indicator shapes are varied to make them distinct from one another. [Example]	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Large geometric shapes are used as the indicator shapes. [Example]	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Dark, filled plotting symbols or dark, thick lines are used to draw attention to an important line (if there is one that is more important than the rest).	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Error bars (if they exist) are thin and light relative to the indicator.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

**Axes**  
This section is concerned with the graph's axes.

	Strongly disagree	Strongly agree	N/A
If an axis obscures graph data, the authors have offset the axis slightly. (Example)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
The axis labels are placed near their axes.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Tick marks are only used if there are gaps between the scale values on the axis. (Example)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

**Text and Labels**  
This section is concerned with the text accompanying the graph.

	Strongly disagree	Strongly agree	N/A
The labels of indicators that will be compared are placed close together and	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

**Figure 1. A sample graph evaluation page of the Visual Rating System (VRS). The graph being evaluated appears on the left, while the instructions and guideline statements appear on the right.**

since I intended to create a tool that allowed an inexperienced research assistant to evaluate a specific graphic's adherence to the guidelines, I felt that that number would diminish the effectiveness of the final rating. Accordingly, I proceeded to perform a specialized sorting and culling process to reduce the number of guidelines to a manageable amount. I utilized a heuristic approach accentuated with focused discussions with novice and expert readers of graphs to reduce the number of guidelines to 33 statements, all designed to be measured with a 5-point Likert scale ranging from strongly disagree to strongly agree (with the option to mark any statement as not applicable). These items were separated into five new categories based on their purpose: selecting the appropriate graph

type (line graph, bar graph, etc.), presentation of the shapes and patterns in the graph, formatting of the axes, formatting of the text and labels, and a general purpose category. A full listing of these statements can be found in the VRS on the Web (Aumer-Ryan, 2005).

The 33 statements were not fundamentally different from the originally published guidelines, however they differed in the following ways: they used language that was less technical (i.e., wording that was more understandable to novice or intermediate readers of graphs); they standardized the use of technical terms that refer to elements of the graph; they combined guidelines with similar intent into one statement; they attempted to discard guidelines that were too subjective in scope (based

on results from a focus session, discussed below); and they discarded guidelines that evaluated the relationship between the graph and the content of the paper (since raters would only be evaluating the graphs, not the entire paper).

In order to help ensure that the new statements maintained integrity with the originally published guidelines, the initial focus group discussion focused on anomalies between the versions until everyone was satisfied that their breadth of coverage and intended meanings were equivalent. Additionally, a general purpose statement—“Overall, I like this graph”—was added to the list as a subjective measurement of effective graph design, and as a way to recover some of the subjective intent of the original guidelines.

These 33 statements formed the basis for the Visual Rating System (VRS) that was to be developed as an evaluation tool for rating a graph's adherence to the original guidelines published in Gillan, Wickens, Hollands, & Carswell (1998). Early in the development process I decided against a paper method of evaluation (because of the added organizational difficulty in distributing paper copies of the statements and each sample graph to any number of designated raters, and the increased workload of inputting the resulting data by hand into a statistical package). A “Web survey” format seemed most hospitable for my purposes, and thus the Visual Rating System was developed (see Figure 1).

The VRS was built around several components: (1) an authentication page that allowed individual raters to log in so they could be presented with a subset of graphics that had been assigned to them to rate; (2) a back-end database that managed the available graphics to rate, and stored information about which graphics were assigned to specific raters and whether or not they had completed an evaluation of each graphic; (3) a front-end PHP-based Web page (Figure 1) that displayed an image of a graph, the 33 statements it was to be evaluated against, and several instructional guides (such as common word definitions and pictorial examples of graphs that adhered to or violated the guidelines); and (4) a database that stored the graph evaluations from each rater.

## METHOD

The completed Visual Rating System was now used in a preliminary research study that aimed to answer the following question: did the original guidelines published by Gillan, Wickens, Hollands, & Carswell (1998) have a noticeable effect on graphs published in the *Human Factors* journal? I set out to answer this question by comparing graphs in papers published one year before the guidelines were published, in 1997 (one year was chosen because it seemed sufficiently close enough to the publication date of the guidelines, but not so close as to expect any authors to have seen an early draft of the guidelines), and three years afterwards, in 2001 (three years was expected to be a good balance between full dispersal of the paper amongst authors and the expected attrition of reference as the guidelines paper aged).

A total of over 900 illustrations were present in papers published over the course of these two selected years, around 500 of which were classified as bar graphs, line graphs, scatter plots, or pie charts (the focus of the guidelines). Because of the large amount of available graphics, I chose to sample 50 graphs from 1997 and 50 graphs from 2001 (for a total of 100 graphs) to be evaluated using the Visual Rating System. The sample graphs were randomly selected from the total number and accurately represented the percentage distribution of the four graph types. For the 1997 sample, the proportion of the four graph types were as follows: 28 line graphs, 19 bar graphs, three scatter plots, and zero pie charts. For the 2001 sample, the proportions were as follows: 25 line graphs, 21 bar graphs, 4 scatter plots, and zero pie charts.

Six research assistants were recruited to evaluate the chosen sample of graphs using the Visual Rating System. In order to enhance intercoder reliability and adherence to the intent of the originally published guidelines, a focus session was convened between the author and the six assistants where we discussed the meaning and intent of each of the statements in the Visual Rating System, and proceeded to evaluate, as a group, several representative graphs (which were not included in the subsequent study). We discussed our responses to the statements for each graph until

a consensus was reached, and continued this process until our individual responses to the statements were similar.

Rather than have each research assistant evaluate all 100 graphs (which was determined to be too taxing on the assistants' time and ability to adequately evaluate each graph), a subset of 30 graphs was chosen for each research assistant to evaluate. These subsets of the sample were chosen in such a fashion as to ensure that at least three assistants evaluated each graph in the sample to further our aim of intercoder reliability. The order in which each assistant evaluated their selected graphs was randomized and any reference to the date of publication was removed from the graph images, so each research assistant would be unable to tell which year a certain graph was published. Each assistant was assigned a login ID for the Visual Rating System that would allow them to evaluate their assigned graphs on their own time on a computer of their choosing.

The data were analyzed by performing 33 t-tests (one for each statement) with year of publication as the independent variable and the values of the Likert-5 scales as the dependent variable. To offset the likelihood of getting false significance values because of the large number of analyses run, a Bonferroni adjustment was performed that reduced the necessary p value for significance to 0.0015 (0.05 / 33).

## RESULTS

Two of the statements were found to be statistically significant, and three more had near-significance values (see Table 1 for a listing of these results). The two significant statements ( $p < 0.000$ ) were: "Words are large and readable" and "Numbers (quantitative labels) are large enough to read comfortably." Interestingly, these responses showed a downward trend from 1997 to 2001 (i.e., the results show that words and numbers were easier to read in articles from 1997 than in articles from 2001). One of the near-significant statements—"The graph is visible at small sizes (e.g., 2-inches wide)"—also showed a downward trend, while the remaining two near-significant statements showed a slight upward trend.

Statement	Significance	1997	2001
<i>Words are large and readable.</i>	$t(180)=3.656,$ $p < .000$	$M=3.72,$ $SD=1.386$	$M=2.94,$ $SD=1.480$
<i>Numbers (quantitative labels) are large enough to read comfortably.</i>	$t(174)=3.581,$ $p < .000$	$M=3.90,$ $SD=1.187$	$M=3.18,$ $SD=1.451$
<i>The axis labels are placed near their axes.</i>	$t(180)=2.776,$ $p < .006$	$M=3.95,$ $SD=1.246$	$M=4.37,$ $SD=0.744$
<i>The styles of lines are varied to make them distinct from one another.</i>	$t(139)=2.639,$ $p < .009$	$M=2.28,$ $SD=1.312$	$M=2.88,$ $SD=1.386$
<i>The graph is visible at small sizes (e.g., 2-inches wide).</i>	$t(180)=2.347,$ $p < .020$	$M=3.42,$ $SD=1.432$	$M=2.90,$ $SD=1.560$

**Table 1. Guideline statements with a significant or near-significant difference between the two years.**

## DISCUSSION

The conclusion drawn from these preliminary findings is that the originally published guidelines had no positive effect on the quality of graphs published in *Human Factors*, and, if anything, graphs published since the guidelines are harder to read than those published before the guidelines. This was not an unexpected result, and further research is in progress to augment this sample.

Several troubling questions arise out of this research, the first of which appeared during the focus session aimed at increasing reliability among the graph raters. We ran into notable difficulty when trying to individually express graph readability—while some raters found a graph to be clear and readable, others found it difficult and poorly designed. Could the concept of graph readability be so personalized and subjective that there is no common ground upon which to build a set of guidelines? There is a good deal of research that implicitly contests this claim (e.g., Lindsay & Norman, 1977; Wickens, 1992; Loftus, 1993), but the creation and acceptance of a set of graphic guidelines has not met with widespread acceptance. It seems too cavalier to attribute this to apathy on behalf of authors; could it be that some authors actually prefer their peculiar graphs?

Secondly, since the vast majority of graphs for publication are produced by statistical software packages (e.g., SAS, SPSS, Excel), it seems like the best way to implement these guidelines would be by addressing them via improved software design. Many graphs that are difficult to understand are left in the default format provided by the statistical software, and often these designs leave much to be desired. While textual matters (such as lengthy sentences used as axis labels or indecipherable or missing graph titles) are solely the liability of the author, much of the underlying graph design can be improved by providing better graph templates for use with the popular statistical packages. We are approaching such companies to broach this subject.

### CONCLUSION

I contend that the Visual Rating System developed in this research project has intrinsic value in answering the above questions and others like them. Not only can it be used to search for trends in graph design according to the guidelines it is based upon, it can be used to examine the efficacy of the guidelines themselves (e.g., if raters cannot agree on how to answer certain statements, then those statements may not convey a solid meaning). The ideal outcome of this research is an increased understanding of graph perception, and the communication of graph design that most effectively conveys the intended information. A picture may be worth a thousand words, but an ill-conceived picture may be worth no words at all—or worse, may require two thousand words of explanation.

### ACKNOWLEDGEMENTS

I would like to thank Dr. Randolph Bias (UT Austin), Dr. Douglas Gillan (NMSU), Zachary Golden (UT Austin), Crystal Meeks (UT Austin), Elizabeth Grinton (UT Austin), Nicole Ryan (UT Austin), Jamie Ratliff (UT Austin), Katherine Aumer-Ryan (UT Austin), and Robert Montalvo (NMSU) for their research efforts and collegiality in pursuit of this project. I could not have done it without you.

### REFERENCES

- Aumer-Ryan, P. R. (2005). *HFES Graphic Guidelines – Visual Rating System*. Retrieved February 20, 2006, from <http://sentra.ischool.utexas.edu/~paul/hfes/>
- Gillan, D. J., & Lewis, R. (1994). A componential model of human interaction with graphs: I. Linear regression modeling. *Human Factors*, 36, 419-440.
- Gillan, D. J., Wickens, C. D., Hollands, J. G., & Carswell, C. M. (1998). Guidelines for presenting quantitative data in HFES publications. *Human Factors*, 40(1), 28-41.
- Krippendorff, K. (2004). Intrinsic motivation and human-centred design. *Theoretical Issues in Ergonomics Science*, 5(1), 43-72.
- Larkin, J., & Simon, H. (1987). Why a diagram is (sometimes) worth 10,000 words. *Cognitive Science*, 4, 317-345.
- Lindsay, P. H., & Norman, D. A. (1977). *Human Information Processing, 2nd ed.* New York, NY: Academic Press.
- Loftus, G. R. (1993). Visual data representation and hypothesis testing in the microcomputer age. *Behavior Research Methods, Instruments, & Computers*, 25, 250-256.
- Lohse, G. L. (1993). A cognitive model for understanding graphical perception. *Human-Computer Interaction*, 8, 313-334.
- Morkes, J., Kernal, H., & Nass, C. (1999). Effects of humor in task-oriented human-computer interaction and computer-mediated communication: A direct test of SCRT theory. *Human-Computer Interaction*, 14(4), 395-435.
- Pinker, S. (1990). A theory of graph comprehension. In R. Freedle (Ed.), *Artificial intelligence and the future of testing* (pp. 73-126). Hillsdale, NJ: Erlbaum.
- Simkin, D., & Hastie, R. (1987). An information processing analysis of graph perception. *Journal of the American Statistical Association*, 82, 454-465.
- Wickens, C. D. (1992). The human factors of graphs at HFS annual meetings. *Human Factors and Ergonomics Society Bulletin*, 35(7), 1-5.