APPENDIX C

The Perceptual Evaluation of Visualization Techniques and Systems

There is a hierarchy of value in research. The ultimate goal of the scientist is to discover an immutable truth that will form the foundation for a new way of understanding the world. Applied researchers must often be satisfied with more humble objectives; sometimes it may be necessary to show that soon-to-be-obsolete interface *A* is better in some small way than soon-to-be-obsolete interface *B*. Between these two scenarios are many graduations. A research-based design guideline is something that can be of enduring value. A rough continuum of value exists, depending on the research goals. The following list starts with those goals that are most valuable.

Research Goals

Uncover fundamental truths and test theories. This is the holy grail of research—a fundamental truth that forever changes how we think of the world. Even small truths are to be prized. Because visualization techniques often produce patterns that do not exist in nature, or rarely do, studies of such techniques can be part of the new discipline of *information psychophysics*. Cognitive modeling of the way people interact with the interfaces to information systems is an important part of cognitive systems theory; and because all human intellectual achievements are ultimately the products of cognitive systems, not individuals alone, lasting truths may be achieved.

Discover the nature of the world. The early stages of science can be like butterfly collecting. It is necessary to get a feeling for the range of phenomena to be encompassed before developing theories. Some areas of perception are still like this. For example, the perception of patterns in motion is still at an early stage. The application of motion in visualization similarly lags.

Ascertain if an existing theory generalizes to practice. Many phenomena that are studied by vision researchers in the simplified conditions of the laboratory may not apply to a more complex data visualization. It can be a useful contribution simply to show that a well known laboratory result generalizes to a common visualization problem.

Make an objective comparison between two or more display methods. Directly comparing two display methods can show which is the more effective. Ideally, the two methods should be tested with a variety of test data to provide some degree of generality.

Make an objective comparison between two or more display systems. Directly comparing two interfaces to an information system has obvious value to someone intending to choose one or the other. However, because system interfaces are typically complex, with usually dozens of differences between them, it is rarely possible to make valid generalizations from such studies.

Measure task performance. Simply measuring the time to perform a task with a particular interface is useful; it is even more useful if the task is elementary and frequently used. Error rates and error magnitudes are other common measurements providing useful guidelines for the designer of information systems.

Ascertain user preferences for different display methods. Occasionally, factors such as the "cool appearance" of a particular interface can be decisive in its adoption. Naturally, the techniques used for research should be suited to the goals of the research. Finding the balance between an attractive display, and an optimal display for the task, should be the goal.

This appendix is intended to provide a preliminary acquaintance with the kinds of empirical research methods that can be applied to visualization. It is not possible in a few thousand words to give a complete cookbook of experimental designs. When studies are looked at in detail, there are almost as many designs as there are research questions, but a number of broad classes stand out. It is generally the case that the methods used for evaluating visualization are borrowed from some other discipline, such as psychophysics or cognitive psychology. Such methods have been continually refined through the mill of peer review. For introductory texts on experimental design and data analysis, see Elmes et al. (1999) or Goodwin (2001). What follows is an introduction to some of the more common methodologies and measurement techniques.

Psychophysics

Psychophysics is a set of techniques based on applying the methods of physics to measurements of human sensation. These techniques have been extremely successful in defining the basic set of limits of the visual system. For example, how rapidly must a light flicker before it is perceived as steady, or what is the smallest relative brightness change that can be detected? Psychophysical techniques are ideal for discovering the important sensory dimensions of color, visual texture, sound, and so on, and more than a century of work already exists. Psychophysicists insist on a precise physical definition of the stimulus pattern. Light levels, temporal characteristics, and spatial characteristics must all be measured and controlled.

Psychophysical techniques are normally used for studies intended to reveal early sensory processes, and it is usually assumed (sometimes wrongly) that instructional biases are not significant in these experiments. Extensive studies are often carried out using only one or two observers, frequently the principal investigator and a lab assistant or student. These results are then generalized to the entire human race, with a presumption that can infuriate social scientists. Nevertheless, for the most part, scientific results—even those obtained with few subjects or as early as the 19th century—have withstood the test of time and dozens of replications. Indeed, because some of the experiments require hundreds of hours of careful observation, experiments with large subject populations are usually out of the question.

If a measured effect is easily altered because of instructional bias, we must question whether psychophysical methods are appropriate. The sensitivity of a measurement to how instructions are given can be used as a method for teasing out what is sensory and what is arbitrary. If a psychophysical measurement is highly sensitive to changes in the instructions given to the subject, it is likely to be measuring something that has higher-level cognitive or cultural involvement.

A few of the studies that have been published in recent years can be understood as a new variant on psychophysics, namely *information psychophysics*. The essence of information psychophysics is to apply methods of classical psychophysics to common information structures, such as elementary flow patterns, surface shapes, or paths in graphs.

When designing studies in information psychophysics, it is important to use meaningful units. For specifying the size of graphical objects there are three possibilities: pixels, centimeters, and visual angle. Each of these can be important. For larger objects, the size in centimeters and the visual angle should be determined. For small objects, pixel size can also be an important variable, and this should be specified. If you want to get really serious about color, then the monitor should be calibrated in some standard way, such as the CIE *XYZ* standard (Wyszecki, 1982). For moving objects, it is also important to know both the refresh rate (the frame rate of the monitor) and how fast your computer graphics are actually changing (update rate). It is worth thinking about how a graphics system actually works to get a better idea of the true precision of measurement. For example, if the update rate and refresh rate of the display are 60 Hz, then the granularity of measurement cannot be better than 16 msec.

Following are some of the common psychophysical methods that may also be applied to information psychophysics.

Detection Methods

There is a range of techniques that rely on how many errors people make when performing a certain task. Sometimes, determining an *error rate* is the goal of the experiment. If, for example,

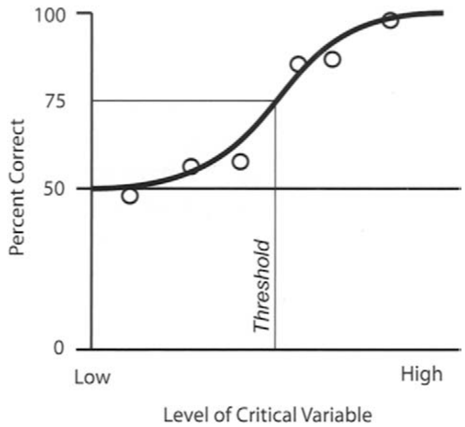

Figure C.1 The threshold in this particular example is defined as a 75% correct rate of responding. Errors are determined at many levels of a critical variable. A curve is fit through the points (heavy line) and this is used to define the threshold.

a visualization is used as part of an aircraft inspection process, then the expected error rate of the inspector is a critical issue.

More commonly, error rates are used as a rigorous way of finding thresholds. The idea is to keep showing subjects a display with some parameter at a range of levels. The percentage of correct detections is measured at each level, and a plot like Figure C.1 is generated. We define the threshold by some error rate; for example, if the chance error rate is 50%, then we might define the threshold as 75%. A problem with this process is that it requires a large number of trials to get a percent error rate for each level of our test parameter, and this can be especially difficult if the region of the threshold is not known. Hundreds of trials can be wasted making measurements that are well above, or below, threshold.

The *staircase procedure* is a technique for speeding up the determination of thresholds using error rates. The subject's responses are used to home in on the region of the threshold. If the subject makes a correct response to a target, the stimulus level is lowered for the next trial. If a subject fails to see a target, the stimulus level is raised. In this way, the program homes in on the threshold (Wetherill and Levitt, 1965).

The most sophisticated way of using error rates in determining thresholds for pattern detection is based on *signal detection theory*. A target pattern is assumed to produce a neural signal with a normal distribution, in the presence of neural noise caused by other factors. A parameter in the model determines whether observers are biased toward positive responses (producing false positives) or negative responses (producing false negatives). One way to represent the results of a study using signal detection theory is the *receiver operating characteristics* (ROC) *curve* (Swets, 1996; Irwin and McCarthy, 1998).

Method of Adjustment

A useful technique for tuning up a visualization is to give application domain experts control over some variable and ask them to adjust it so that it is optimal in some way for them. This is called the *method of adjustment*. Get a population of users to do this, and a useful default setting can be derived from the mean or median setting.

The method of adjustment can also be used to answer questions about perceptual distortion. For example, if we are interested in simultaneous lightness contrast, we can ask subjects to adjust a patch of gray until it matches some other gray and use the difference to estimate the magnitude of the distortion due to contrast.

There are biases associated with method of adjustment. If we are interested in the threshold for just seeing a target, we might turn up the contrast until it is visible or turn down the contrast until it disappears. The threshold will be lower in the latter case. Once we can see something, it is easier to perceive at a lower contrast.

Cognitive Psychology

In cognitive psychology, the brain is treated as a set of interlinked processing modules. A classic example of a cognitive model is the separation of short-term and long-term memory. Short-term memory, also called *working memory*, is the temporary buffer where we hold concepts, recent percepts, and plans for action. Long-term memory is a more or less permanent store of information that we have accumulated over a lifetime.

Methods in cognitive psychology commonly involve *measuring reaction time* or *measuring errors*, always with the goal of testing a hypothesis about a cognitive model. Typical experiments involve very simple, but ideally important tasks, such as determining whether or not a particular object is present in a display. The subject is asked to respond by hitting a key as fast as possible. The resulting time measurement can be used to estimate the time to perform simple cognitive operations, once the time taken to physically move the hand or depress a key is subtracted.

Another common kind of experiment measures *interference* between visual patterns. The increase in errors that results is used as evidence that different channels of information processing converge at some point. For example, if the task of mentally counting down in sevens from 100 were to interfere with short-term memory for the locations of objects in space, it would be taken as evidence that these skills share some common cognitive processing. The fact that there is little or no interference suggests that visual short-term memory and verbal short-term memory are separate (Postma and De Haan, 1996).

Recently, some cognitive theories have gained a tremendous boost because of advances in brain imaging. Functional MRI techniques have been developed that allow researchers actually to see which parts of the brain are active when subjects perform certain tasks. In this way, functional units that had only been previously inferred have actually been pinpointed (Zeki, 1993).

Structural Analysis

In structural analysis, theories of cognitive processing are constructed using direct observation as evidence. Structuralist researchers conduct studies that are more like interviews than formal experiments. Often the subjects are required to carry out certain simple tasks and report at the same time on their understanding and their perceptions. Using these techniques, researchers such as Piaget have been able to open up large areas of knowledge very rapidly and to establish the basic framework of our scientific understanding. However, in some cases, the insights obtained have not been confirmed by subsequent, more careful experiments. In structuralism, emphasis is given to hypothesis formation, which at times may seem more like the description and classification of behavior than a true explanation.

A structural analysis is often especially appropriate to the study of computer interfaces, because it is fast-moving and can take a variety of factors into account. We can quantify judgments to some extent through the use of rating scales. By asking observers to assign numbers to subjective effectiveness, clarity, and so on, we can obtain useful numerical data that compares one representation to another. There are several tools of structural analysis. We can ask domain experts what they need in a visualization (*requirements analysis*). We can try to understand what they are attempting to accomplish at a more elementary level (*task analysis*). Research tools also include testbench applications, *semistructured interviews*, and rating scales.

Testbench Applications for Discovery

At the early, butterfly-collecting stage of science, the goal is to map out the range of phenomena that exist. In visualization research, the goal is to gain an intuitive understanding of diversity, notable phenomena, and what works and does not work from an applied perspective.

The primary early-stage tool for the visualization researcher is the *testbench application*. It gives the researcher the flexibly to try out different ways of mapping the data into a visual representation. Of course, there is no such thing as a universal testbench. The goal should be to build a flexible tool capable of producing a range of visual mappings of the data and a range of interaction possibilities. For example, if the problem is to find the best way to represent the shape of a surface, the testbench application should be able to load different surface shapes, change lighting, change surface texture properties, turn stereoscopic viewing on and off, and provide motion parallax cues.

There is a tendency for programmers to make the user interface for a testbench too sophisticated. This can be self defeating, because it limits flexibility. The objective should be to explore, not to build a polished application for scientists. If the easiest way to explore is to change a constant in the code and recompile, then this is what should be done. Often a good testbench interface is a text file of parameters, setting various aspects of the display. This can be modified in a word processor and reloaded. Sometimes a panel of sliders is useful, allowing a researcher to adjust parameters interactively. For the most part, the quality of the code does not matter for a testbench application, although it is essential that the parts of the program dealing with display parameters are correct.

There are many ways to play with testbenches, and *play* is definitely the operative word. This is a time for creative exploration, forming hypotheses quickly and discarding them easily. Interesting possibilities can be shown to domain experts. It is especially useful to show them the best solutions you have. You may only get one chance to have a physicist, an oceanographer, or a surgeon to take you seriously. Asking their opinions about something that actually looks better than whatever they are currently using is one way to get their interest. Phenomena that may be significant can be shown to other vision researchers.

Once something interesting has been identified with a testbench, a rigorous study can be carried out using the methods of psychophysics or cognitive psychology.

Structured Interviews

One of the most useful tools, both for initial requirements and task analysis and for the evaluation of problem solutions, is the *structured* or *semistructured* interview. The method is to construct an interview with a structured set of questions to elicit information about specific task requirements. It is structured to make sure that the important questions are asked and that the answers come in a somewhat coherent form.

Structured interviews can be excellent tools to evaluate what aspects of a visualization actually are important to potential users. They can also be used to evaluate a number of different solutions for strengths and weaknesses. In many cases, it is useful for structured interviews to be built around the performance of particular tasks. The participant is asked to perform particular tasks with the system or with more than one system and is then asked to comment on suitability, ease of use, clarity of presentation, and so on. The great advantage of structured interviews is that they make it possible to gain information about a wide range of issues with relatively little effort, in comparison with more objective methods, such as reaction time or error rate measurements. Also, you might learn something you did not ask about.

Rating Scales

The Likert scale (also called a *rating scale*) is a method for turning opinions into numbers. Subjects are simply asked to rate some phenomenon by choosing a number on some range, such as the following:

(GOOD) 1 2 3 4 5 (BAD)

For example, if we have six different visual representations of a flow pattern, we might ask subjects to rate how well they can see each pattern on a scale of 1 to 5.

Subjects tend to use rating scales in their own idiosyncratic ways. Some will be biased to the low end of the scale and others to the upper end, but generally they will tend to try to use most to the scale for whatever set of samples they are shown. Because of this, no absolute meaning should ever be given to rating scale data. If 10 inferior visual representations are shown to subjects, they will still differentiate them into good and bad; the same will be true for 10 very good ones. However, rating scales are an excellent tool for measuring relative preferences.

Rating scales can be used to answer broad questions about preferences for two or more different solutions to a problem. Quite often, users will prefer one solution to another, even though no objective differences are measured. In some cases, one interface might even be objectively superior, but another preferred.

Statistical Exploration

Sometimes it may be useful to use statistical discovery techniques to learn about some class of visualization methods. Suppose we wish to carry out an investigation into how many data dimensions can be conveyed by visual texture. The first obvious question is: How many perceptually distinct texture dimensions are there? The next question is: How can we effectively map data dimensions to them? If the answer cannot be found in the research literature, one way to proceed is to use a kind of statistical data-mining strategy to find the answer. First, we might ask people to classify textures in as many different ways as we can think of (e.g., roughness, regularity, elongation, fuzziness). The next step is to apply a statistical method to discover how many dimensions there really are in the subjects' responses. The following sections list the major techniques.

Principal Components Analysis

The goal of *principal components analysis* is to take a set of variables and find a new set of variables (the principal components) that are uncorrelated with each other (Young et al., 1978; Tabachnick and Fidell, 2001; Hotelling, 1933). This might be used to reduce a high-dimensional data set to lower dimensions. In many data sets (think of multiple measurements on the dimensions of parts of beetles, for example), many of the variables are highly correlated, and the first two or three principal components contain most of the variability in the data. If this is the case, then one immediate advantage of the data reduction resulting from PCA is that the data can be mapped into a two- or three-dimensional space and thereby visualized as a scatter plot.

Multidimensional Scaling

Multidimensional scaling (MDS) is a method explicitly designed to reduce the dimensionality of a set of data points to two or three, so that these dimensions can be displayed visually. The method is designed to preserve, as far as possible, metric distances between data points (Young et al., 1978; Wong and Bergeron, 1997).

Clustering

Cluster analysis is a statistical technique designed to find clusters of points in a data space of any dimensionality (Romesburg, 1984). There are two basic kinds: hierarchical and k-means. In *hierarchical clustering*, a tree structure is built, with individual data points at the leaves. These points are combined recursively, most similar first. Hierarchical clustering can provide the basis for hierarchical taxonomy.

K-means clustering requires the user to input a number of clusters (k). A set of k clusters is generated by finding the cluster means that minimize the sum of squared distances between each set of data points and its nearest mean.

Either kind of clustering can be used as a method for data reduction in visualization, because a tight cluster of points can be reduced to a single data glyph.

Multiple Regression

In visualization, *multiple regression* is a statistical technique that can be used to discover whether it is possible to predict some response variable from display properties. For example, the time required to judge the shortest path in a node–link diagram might be predicted from the number of link crossings in the diagram and the bendiness of the path (Ware et al., 2002).

Cross-Cultural Studies

If sensory codes are indeed interpreted easily by all humans, this proposition should be testable by means of *cross-cultural studies*. In a famous study by Berlin and Kay (1969), color naming was compared across more than 100 languages. In this way, the researchers established the universality of certain color terms, equivalent to our red, green, yellow, and blue. This study is supported by neurophysiological and psychophysical evidence that suggests these basic colors are hard-wired into the human brain. Such studies are rare, for obvious reasons, and with the globalization of world culture, meaningful studies of this type are rapidly becoming impossible. Television is bringing about an explosive growth in universal symbols. In the near future, cross-cultural studies aimed at basic questions relating to innate mechanisms in perception may be impossible.

Child Studies

By using the techniques of *behaviorism,* it is possible to discover things about a child's sensory processing even before the child is capable of speech. Presumably, very young children have only minimal exposure to the graphic conventions used in visualization. Thus, the way they respond to simple patterns can reveal basic processing mechanisms. This, of course, is the basis for the Hochberg and Brooks' (1978) study discussed in Chapter 1.

It is also possible to gain useful data from somewhat older children, such as five-year-olds. They presumably have all the basics of sensory processing in place, but they still have a long way to go in learning the graphic conventions of our culture, particularly in those obscure areas that deal with data visualization.

Practical Problems in Conducting User Studies

Experimenter Bias

Researchers' careers depend on what they publish, and it is much easier to publish results that confirm a hypothesis than results showing no effects. There are many opportunities for experimenter bias in both the gathering and the interpretation of results.

As a rule of thumb, if the data being measured relates to some low-level, fundamental aspect of vision, then it will be less subject to bias. For example, if a subject is given a control that allows the setting of what seems to be a "pure" yellow, neither reddish nor greenish, the setting is likely to be extremely consistent and will be relatively robust even if the experimenter makes comments like "Are you sure that's not a little tinged with green?" On the other hand, if the experimenter says, "I want you to rate this system, developed by me to obtain my PhD, in comparison with this other system, developed at the University of Blob," then experimenter bias effects can be extreme.

When considering your own work or that of others, be critical. The great advantage of science is that it is incremental and always open to reasoned criticism. Applied science tends to adopt somewhat looser standards, and replications of experiments are rare. Many of the studies we read are biased. In evaluating a published result, always look to see whether the data actually supports what is being claimed. It is common for claims to be made that go far beyond the results. Often the abstract and title suggest that some method or other has clearly been demonstrated to be superior. An examination of the method may show otherwise. A common example is when a difference that is not statistically significant is claimed to support a hypothesis. Some of most important questions to ask are:

What is the task?

Does the experiment really address the intended problem?

Are the control conditions appropriate?

Does the experiment actually test the stated hypothesis?

Are the results significant?

Are there possible confounding variables?

Confounding variables are variables that change in the different experimental conditions, although they are not the variables that the researchers claim to be responsible for the measured effect.

How Many Subjects to Use?

In vision research, some kinds of studies are run with only two, three, or four subjects. These are studies that purport to be looking at the low-level machinery of vision. It makes sense; humans all have the same visual system, and to measure its properties you do not need a large sample of the population. On the other hand, if you are interested in how color terms are used in the general population, then the general population must be sampled in some way.

Statistically, the number of subjects and the number of observations required depend on the variability of responses with a single subject and the variability from one subject to another.

Most experiments are run with between 12 and 20 subjects, where all of the experimental conditions can be carried out on the same subjects (a *within subjects* design). In some cases, because of learning effects, different subjects must be assigned to different conditions. Such experiments will require more subjects.

Research is always an optimization problem—how to get the most information with the least effort. One reason there have been so many simple reaching experiments presented at the Association for Computing Machinery (ACM) Computer-Human Interaction (CHI) conferences is that a Fitts' law experiment (the standard experimental method) is very easy to carry out; it is possible to gather a data point every two or three seconds. A substantial amount of data can be gathered in half an hour of subject time, making it possible to run large numbers of subjects.

Combinatorial Explosion

One of the major problems in designing a visualization study is deciding on the *independent variables*. Independent variables are set by the experimenter in the design stage. In a study of the effectiveness of flow visualization, independent variables might be line width and line spacing of streamlines. The *dependent variables* are the measured user responses, such as the amount of error in judging the flow direction.

In visualization design problems, there are often many possible independent variables. Let us take the example of flow visualization consisting of streaklets–small, curved line segments showing the direction of flow. Streaklet length, streaklet start width, streaklet end width, streaklet start color, streaklet end color, and background color may all be important. Supposing we would like to have four levels of each variable and we wish to study all possible combinations. The result is $4^6 = 2048$ different conditions. Normally, we would like at least 10 measurements of user performance in each condition. We will require over 20,000 measurements. If it were to take 30 seconds to make each measurement, the result would be more than 160 hours of observation for each subject. We might decide to have 15 subjects in our experiment. This means over a year of work, running subjects 40 hours per week. For most researchers, such a study would be impossibly large.

The brute force approach experimental design is to include all variables of interest at all meaningful levels. Because of the combinatorial explosion that results, this cannot work. The way to obtain more from studies, with less effort, is to develop either theories or descriptive models that can be applied to a range of design problems. Empirical studies can be much simpler and focus on specific aspects of the theory.

Task Identification

A critical element in experimental design is deciding on the *task* the subject is to perform. Ideally, this will be something that is both theoretically interesting and very commonly used in real applications. Even if the exact task is not common, it should be representative of activities that are common in visualization interfaces. For example, if the application domain involves visualizing node–link diagrams, the subject might be asked if there is a path between two highlighted nodes. This task is good, because perceiving links between nodes is likely to be important for almost all of the great variety of node–link diagrams that exist.

In order to provide a useful measure of performance, it is also important that the task can be set up to have a clear and simple user response. For example, the subject might push the right mouse button to indicate *yes* and the left mouse button to indicate *no*.

Controls

In an experimental design, a *control* is a condition that is used to provide some basis for comparison. In a theoretical study, the control is usually some condition that provides a reference for theory testing. A theory might predict that a contrast effect will bias a judgment by 30%; the control measurement would be made without the contrast-causing factor to provide a baseline for comparison.

In evaluating a new visualization method, the most reasonable control is the current best practice display method. Some studies employ the somewhat dishonest practice of using a very poor alternative method as a control, thereby exaggerating the value of their own method. This is one of the reasons that the research literature should be read with a measure of skepticism.

Getting Help

Studies in information visualization are fundamentally multidisciplinary. Usually knowledge of computer science, human visual perception, and some application domain is necessary. Often, the best way to do research is to be part of a collaborative team—a computer scientist who can design and build novel interactive visualization systems, a psychologist who understands the perceptual issues and has experience in perception and cognition research, and a domain expert who understands the potential application. Naturally, everyone has his or her own area of interest, and finding compatible collaborators can be difficult, but it can also be very rewarding.

In reality, a single researcher must take on several roles, although getting help and advice is usually worth the effort. Most academics are willing to provide a certain amount of free advice for no more reward than a line in the acknowledgments section of a published paper.

Finally, many universities operate a statistical consulting service that can provide help in experimental design or data analysis.