Thoughts on User Studies: Why, How, and When

Robert Kosara
VRVis Research Center, Vienna, Austria
Kosara@VRVis.at

Christopher G. Healey
Department of Computer Science, North Carolina State University
healey@csc.ncsu.edu

Victoria Interrante
Department of Computer Science and Engineering, University of Minnesota
interran@cs.umn.edu

David H. Laidlaw
Computer Science Department, Brown University
dhl@cs.brown.edu

Colin Ware
Center for Coastal and Ocean Mapping, University of New Hampshire
colinw@cisunix.unh.edu

Introduction

Visualization as currently practiced is mostly a craft. Methods are often designed and evaluated by presenting results informally to potential users. No matter how efficient a visualization technique may be, or how well motivated from theory, if it does not convey information effectively, it is of little use. User studies offer a scientifically sound method to measure a visualization's performance. Although their use has become more widespread, we believe they have the potential for a much broader impact. This article describes our experiences with user studies. We offer some examples of our own studies, talk about the pitfalls and problems we encountered, and show how the results were applied to produce successful visualizations. Although our main goal is to encourage the use of studies in visualization, we recognize that other disciplines also offer important insights into visualization design, for example, the areas of visual design or the visual arts. We conclude by discussing when knowledge from these areas might be preferable to a traditional user study.

Why Conduct User Studies?

There are many reasons to pursue user studies. Studies can be used to evaluate the strengths and weaknesses of different visualization techniques. For example, Laidlaw compared six methods for visualizing 2D vector fields (Figure 1, [7]). His experiments measured user performance on three flow-related tasks for each of the six methods. The results were used to identify what makes a 2D vector field visualization effective.

Studies can show that a new visualization technique is useful in a practical sense, according to some objective criteria, for some specific task. Even more exciting are studies (like Laidlaw's) that show that a new technique is more effective than an existing technique for an important task, where the existing technique was previously considered the *best* technique to use. User studies can objectively establish which method is most appropriate for a given situation.

A more fundamental goal of conducting user studies is to seek insight into *why* a particular technique is effective. This can guide future efforts to improve existing techniques. We want to understand for what types of tasks, and under what conditions, a particular method will give high quality results. This knowledge is critical, since different analysis tasks may be best served by different visualization techniques.

A final use for studies in visualization is to show that an abstract theory applies under certain practical conditions. For example, results from psychophysics or computer vision may or may not extend to a visualization environment. User studies can be run to test this hypothesis. Results can show *when* the theories hold, and *how* they need to be modified to function correctly for real-world data and tasks.

A good starting point in any study is the scientific or visual design question to be examined. This drives the process of experiment design. A poorly designed experiment will only yield results of limited value. Although a comprehensive discussion of experimental design is beyond the scope of

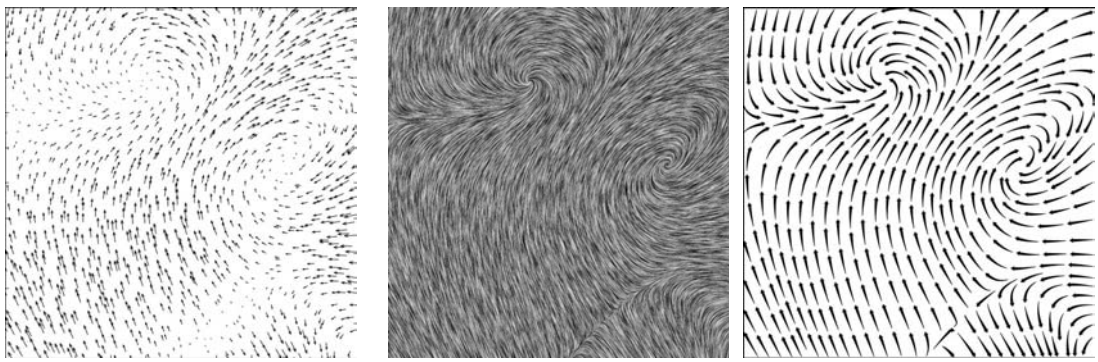


Figure 1. Three of six visualization methods compared with a user study. Each method shows the same vector field. User performance on different tasks provided quantitative comparisons of the methods.

this article, we offer some suggestions and lessons learned in the *Basics of User Study Design* sidebar. We also describe how we designed experiments to answer important questions from our own research.

Color Sequences

One reason for conducting studies is to determine if theoretical principles derived from other disciplines (such as psychophysics) can be applied to visualization design. The theory of human color vision has been studied for more than a century. Results from this work provide a solid foundation for the use of color in visualization. However, choosing colors for a particular visualization problem is normally very different from the extremely simple displays used by experimental psychologists. Experiments are needed to bridge this gap between theory and practice.

Consider the problem of designing pseudo-color sequences for scientific images. We have a continuous data field over a plane (*e.g.,* an energy distribution or a density distribution), and we want to use color to illustrate features in the data. Briefly, the relevant theory from human vision is as follows. Neural signals from the rods and cones in the retina are transformed by neural connections in the visual cortex into three *opponent color channels:* a luminance channel (black-white) and two chromatic channels (red-green and yellow-blue). The luminance channel conveys the most information, enabling us to see form, shape, and detailed patterns to a much greater extent than the chromatic channels. Perception in the chromatic channels tends to be categorical. Colors tend to be placed into categories like red, green, yellow, and blue. Hues such as turquoise or lime green are seen more ambiguously. Another relevant theoretical point is that simultaneous contrast (the phenomena where perceived color is affected by surrounding colors) occurs in all three opponent channels. This can cause large errors when viewers try to "read" values in the data based on color.

We can use these theories to draw a number of conclusions regarding the design of color sequences:

1. If we want our color sequence to reveal form (such as local maxima, minima, and ridges), or if detailed patterns need to be displayed, then a sequence with a substantial luminance component should be used.

2. If we want to display categories of information, for example, the classification of a terrain into regions of different geological type, then a chromatic sequence should be used.

3. If it is important to minimize errors from contrast effects, then a sequence should be arranged to cycle through many colors.

4. A general solution can be constructed that cycles through many colors (to allow for categorization) while continuously increasing luminance.

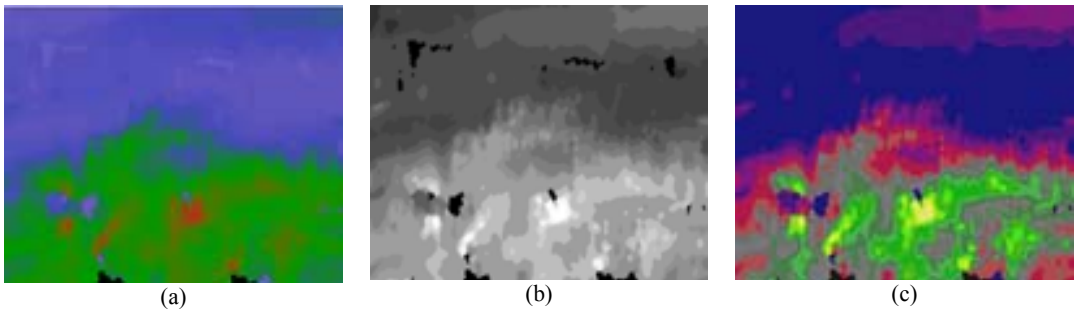
Figure 2 illustrates three different color sequences selected to emphasize a different aspect of the underlying data. Experimental studies have verified that these theoretical predictions apply in the case of color sequences [11]. This demonstrates the use of well established theories to build design guidelines, together with experiments that validate the guidelines in an applied setting.

Shape From Texture

Numerous applications in scientific visualization involve the computation and display of arbitrarily shaped, smoothly curving surfaces. A common case is level surfaces in volume data. By default, the standard practice is to render these surfaces with a smooth, Phong-shaded finish. One important question that arises is, "Can we better convey the 3D shape by rendering the surface as if it were made from a subtly textured material, rather than polished plastic?" There is ample evidence from psychophysics[9] to suggest that certain kinds of surface texture can facilitate shape perception (Figure 3, [4]). Unfortunately, the exact mechanisms by which surface texture affects shape perception, and hence the specific characteristics of texture patterns that best show shape, remain unknown. Complicating any naïve attempt to use texture to enhance shape appearance is the complementary evidence that under many conditions texture can camouflage surface shape features [2].

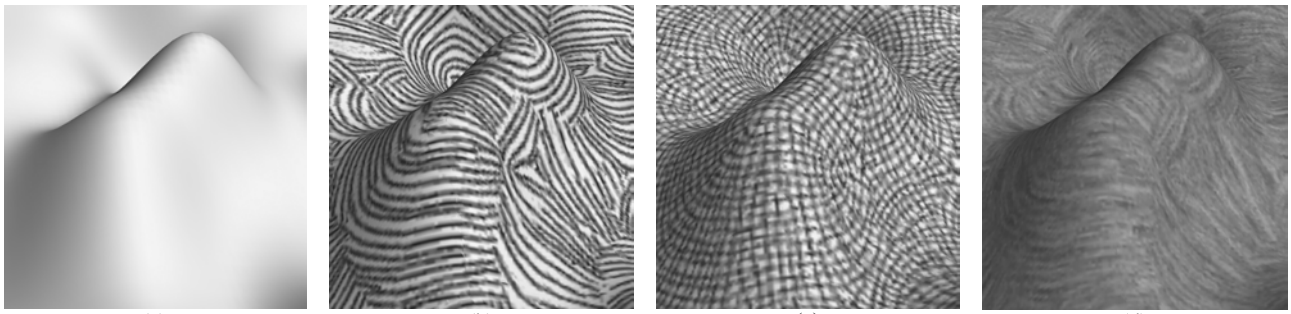
Through carefully designed experiments, it is possible to gain concrete insights into how texture might be used most effectively to support accurate shape perception. More specifically, we can start to answer the question, "If we want to design the ideal texture that best conveys the shape of a smoothly curving surface, what should the characteristics of that texture be?" User studies conducted by visualization researchers are essential to this endeavor for several reasons.

First, traditional vision researchers are primarily concerned with elucidating the neural processes involved in the perception of shape from texture, and their investigations



(a) (b) (c)

Figure 2. Three color sequence: (a) a chromatic sequence, good for representing categories; (b) a luminance sequence, good for representing form; (c) a combine chromatic-luminance sequence, good for representing both categories and form.



(a) (b) (c) (d)

Figure 3. Four examples from a study to test different methods to enhance shape perception: (a) Phong shading (b) one principal direction; (c) two principal directions; (d) line integral convolution (LIC)

do not fully encompass the scope of the questions that we would like to ask.

Second, there is a limit to the depth of understanding we can derive purely from introspection and informal empirical comparison. In the absence of a clear task, viewers may adopt differing opinions about which textures they believe have the greatest potential to be effective. Without concrete experimental evidence, it may be impossible to sort out these differences. Furthermore, it is rarely the case that complex problems yield simple answers. If texturing can help, it is unlikely that any method we initially attempt will turn out to be “the best” in all cases. We expect to discover complicated interactions between surface texture and shading, between texture orientation and surface geometry, and between aesthetics and convention. We may also find numerous task dependencies. This suggests we will need to iterate to achieve progressively more effective methods for different purposes. These goals are best achieved through carefully controlled, quantitative user studies that objectively assess the impact of particular texture pattern characteristics on the accuracy of performance on specific tasks.

Perceptual Textures

One key issue we must address when we design an experiment is which conditions to study. As the number of conditions (and the interactions between conditions) grows, so too does the number of trials needed to properly test each condition. Because of this, experiments are often restricted to only the most important conditions.

As a practical example, consider an area where visualization and experimentation converge: visual perception. Understanding how we "see" the basic properties of an image allows us to create representations that take advantage of the human visual system. An important discovery in psychophysics from the past twenty-five years is that human vision does not resemble the largely passive process of modern photography. A much better metaphor is a dynamic and ongoing construction project, where the products are short-lived models of the external world that are specifically designed for the current visual tasks of the viewer. Harnessing human vision for visualization therefore requires that the images be constructed so as to draw attention to their important parts.

When we design visual perception experiments, we can exploit the fact that participants with normal or corrected-to-normal vision usually exhibit the same relative performance variations during low-level vision tasks. Since we are interested in measuring exactly these differences, we can combine trials across viewers to increase the number of repetitions of each type of trial.

To demonstrate how these kinds of experiments are constructed, we describe a study that investigates the visual properties of texture. Previous work in computer vision and psychophysics has decomposed texture patterns into a number of basic "texture dimensions" like size, contrast, regularity, and directionality. We wondered, "Can individual texture dimensions be used to display multiple attribute values?" Controlled experiments offer a way to answer this question.

Viewers were shown regularly-spaced 20x15 arrays of perceptual texture elements (or pexels, as we call them) that look like upright paper strips. The pexels allow for the variation of multiple texture dimensions including height, density, and regularity of placement. Viewers saw the pexel grid for a short duration, and were then asked whether a group of pexels with a particular target value was present or absent. Our experiment tested five different conditions selected in part from models of human vision, and in part from texture segmentation and classification experiments in computer vision.

We decided to vary *target type* (target pexels were defined by height, density, or spatial regularity), *target-background pairing* (different types of targets were tested, for example, both medium and tall targets), *display duration* (the amount of time the pexel array was shown to the viewer), *target patch size* (the number of pexels used as targets), and *background texture pattern* (whether non-target texture properties were held constant, or varied randomly).

Each condition served a specific function. Target type allowed us to test three different texture dimensions. Target-background pairing searched for differences in performance based on the particular value of the target dimension. Display duration measured the amount of time needed to perform a target detection task. Target patch size asked whether smaller texture patches were harder to identify. Finally, background texture pattern tested for visual interference when secondary texture dimensions vary randomly across the display. Even these basic conditions

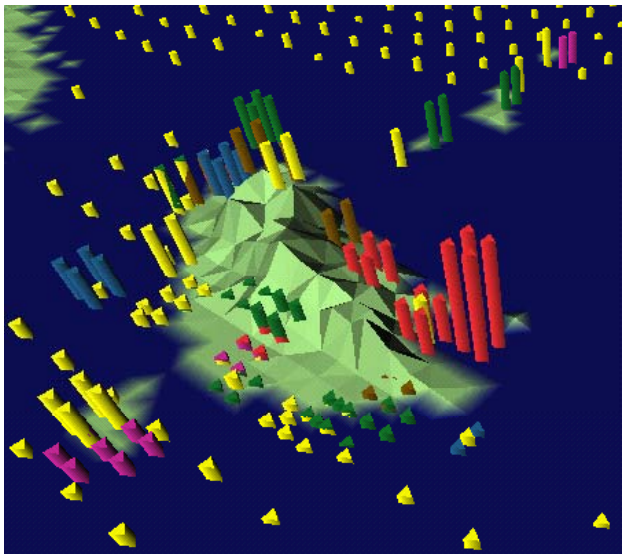


Figure 4: Perceptual texture elements (pexels) used to visualize a typhoon striking the island of Taiwan: pexel height represents wind speed (taller for stronger winds), density represents pressure (denser for lower pressure), and color represents precipitation (blue and green for light rainfall to purple and red for heavy rainfall; yellow indicates an unknown rainfall amount).

produced 108 different display types (three target types by two target-background pairings by three display durations by two patch sizes by three background patterns). Each viewer who participated during the experiment observed 576 trials from one target type (16 repetitions of a target's 36 different display types). Eight trials in each display type were randomly selected to contain a target patch; the remaining eight did not.

Results from the experiment showed a preference for target type (taller targets were easier than shorter, denser, and sparser targets, which were themselves easier than irregular or regular targets). High accuracy was possible for many target types, even for display durations of 150 msec or less. Finally, variations in regularity interfered with the identification of shorter, sparse, and denser targets (but not taller ones). A complete description of the experiment's results is available in [3]. These results have been applied as guidelines on the use of texture for multidimensional visualization. Figure 4 shows an example of using pexels to visualize typhoon activity in Southeast Asia.

Usability Testing and SDOF

Much of the work presented in this article is designed to test basic perceptual features or visualization techniques. We have found, however, that visualization applications have important aspects that need to be studied in the context of the application itself.

The approach for this type of study is quite different from basic perception experiments. Participants must solve a relatively complex task, where there is a greater freedom of action, and also a higher potential for mistakes. Studying a technique in an application setting (as opposed to an artificially simple environment) is critical, because we

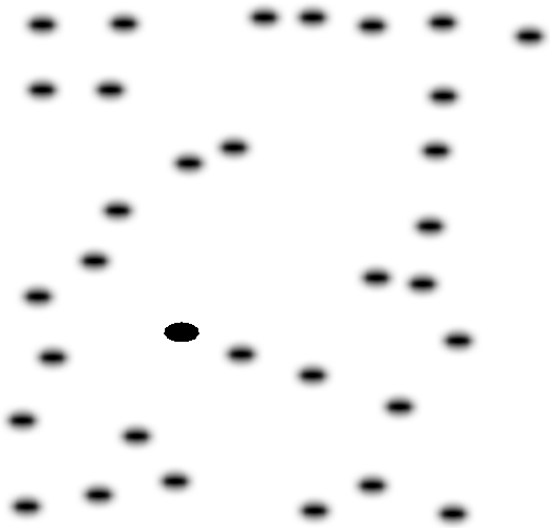


Figure 5: An image from the SDOF study. The image was displayed for 200 msec, after which participants were asked to point at the quadrant with the sharp object.

cannot assume that low-level results automatically apply to more complex displays.

Comments from participants in such a study are often more important than the other data we collect, since they provide valuable hints about what is happening during the experiment. Close observation of the participants can also offer information about experiment details that might not have been part of the original hypotheses.

An example of this type of study is the evaluation of semantic depth-of-field (SDOF) [6], a technique for guiding a viewer to specific information in an image. SDOF is based on the depth-of-field effect from photography, where different parts of a picture are in or out of focus based on their distance from the focal point of the lens. SDOF generalizes this concept. The sharpness of an object depends not on its physical position, but on its relevance. Viewers are immediately drawn to the sharp (*i.e.,* highly relevant) parts of the image, but they can still choose to look at other, out of focus objects (Figure 5). We designed an experiment that contained both basic perception and application components. The perception studies produced results that were significant, and close to what we expected to find. The application findings, however, were much less conclusive.

During the perception part of the study we showed each of our 16 participants 480 separate images (for a total of 7680 trials). We wanted to see if participants could rapidly locate a sharp object in a field of blurred objects. Each image was displayed for 200 msec. Participants were then asked to point to the quadrant that contained the sharp object. It was interesting to see how well participants performed (for one test, we recorded an error rate of less than 0.7%), and how little they used the breaks that were offered.

During the application part of the study, one application was a map viewer which presented the user with a map containing nine layers of information (*e.g.,* roads,

elevations, and cities). They were asked to position a project (*e.g.,* a factory) based on three "very important" and three "somewhat important" factors. Viewers could reorder the layers by selecting which layer was "on top." The layers were displayed in three different ways: opaque, semi-transparent, and SDOF (in SDOF the top layer was sharp, and underlying layers were increasingly blurred). The hypothesis was that SDOF would make it easier to stack the layers in order of importance, and thus answer more quickly and more correctly.

While some useful results were identified during the application study, we did not find statistically significant results in either response time or correctness. We concluded there were two problems with the study. First, the maps we used were visually too simple. Second, the number of tasks was too small; more examples per viewer might lead to significant results. We plan to consider these ideas in future work on SDOF.

Automated Mini-Studies

A final type of user study is something we are calling an *automated mini-study*. An example is the automatic generation of isoluminant colormaps (*i.e.,* colormaps where all the colors appear equally bright) via a face-based method [5]. Images of faces are used to identify colors that are the same perceived brightness as a target gray patch. This is easy to do, because we are highly accurate at interpreting facial characteristics. In essence, the method automatically performs a small user study. Results from the study are accessed by a computer program to generate isoluminant colors.

This automated mini-study offers an elegant solution to the problem of uncalibrated monitors. It also represents a compelling motivation for user studies and their results, since it demonstrates the added value that computer science can provide. To date, we have used methods from psychology to perform tests and analyze results, without thinking about ways to improve this testing procedure. With an automated study the computer generates results automatically, and uses the results directly to improve an application in different ways for different users.

When Do User Studies Help?

While user studies are an important tool for visualization design, they are not the proper choice in every situation. Other techniques are available, and experiments do not always work as expected. We discuss some of these issues in this section.

Beyond User Studies?

While formal, quantitative user studies have a clear role in visualization, they may not always be the right choice. It is important to consider other options before jumping in to design and run a user study. Studies are very time consuming to design, implement, run, and analyze. Typically, they can only be used to answer small questions, and any larger conclusions rely on generalizations that may not be valid. Often, measures that are less precise, quantitative, and objective may provide sufficient insight about a visualization question to allow us to move forward.

In our investigation of virtual reality tools for archaeological analysis [10], we labored long and hard to design a good user study to test the system that we had developed. But the experimental design eluded us. In the end, we videotaped a pair of archaeologists using the system to evaluate some of their scientific hypotheses. They also generated several new ideas, some of which would have been very difficult to generate with other analysis methods. This was sufficient to demonstrate the utility of the visualization application.

In another context we can also, perhaps, transcend the traditional user study. Artists and designers have been creating visualizations for centuries and have evolved very effective methods. User studies come from science; in fact, they embody the scientific method of posing a hypothesis, taking measurements, analyzing them, and iterating to gain insight. For the scientific study of low-level vision, the methodology works, but as we rise up to the level of a scientific visualization application, it may not be possible to use these techniques to answer important questions.

Can we replace some parts of the user testing process with expert visual designers? This is a conjecture we can likely test (not surprisingly) with a user study comparing results of a standard user study with expert visual designer input. Preliminary results suggest that visual designers can replicate some user study results more quickly and with more insight about why differences occur. However, there is still much to learn about the space between perceptual psychology and visual design.

When Things Go Wrong

For some studies, experimental design may lead to results that are not statistically significant. For example, in a recent study we hypothesized that users would perform differently for a visual search task in virtual reality if the virtual environment were different. In fact, we found that there was no statistically significant difference. Perhaps our conjecture was wrong, but it is also possible that our choice of task or other parts of the experimental design misled us. The virtual environment may really matter in some cases. We continue to think about how the virtual environment might make a difference, particularly since visual context has been shown to be important in 2D visual search tasks.

Some studies are not published because of null results, or because the results are inconclusive or not compelling. These are three very different cases that need to be treated accordingly.

Null results are completely natural (albeit usually not intended), because they simply show that the original hypothesis was not supported by the data. This can be because the difference is too small for the amount of data collected, but is most often because the hypothesized difference is not significant. This is why the study was done in the first place, and should therefore not be considered a failure. In visualization, null results cannot be published (at least not on their own) easily. But they can provide valuable information about which directions of research to pursue, and which to abandon.

Inconclusive results are a much more serious problem. They usually mean that there was a design error in the

study, and that it has to be run again. Usually, however, only one part of a study is affected, so the effort is considerably smaller the second time. Also, additional hypotheses can be tested that might have arisen from the successful parts of the study.

Results that are not compelling can be the result of choosing the wrong task or measuring the wrong performance quantity. For example, in the "Great Potato Search" [1], we chose a 3D visual search task. Unfortunately, it was a task that involved looking inward at a relatively small model. We believe a task that involved searching more broadly around the user might have shown important performance differences correlated with changes in the virtual context. While we can (and will) go on to test that new hypothesis, if we had chosen a different task in the first place, we would have been better off. There is a tension between proceeding to execute an experiment quickly and spending time on design. Practice can help to reduce or alleviate these types of mistakes.

Conclusions

In this article we have tried to go beyond the current state of the art in two ways. The first is to promote evaluating visualization methods with user studies. This is being done in certain cases, but it is still far from being a standard practice in our field. The second is to ask where user studies might be useful, and where other techniques might be more appropriate (*e.g.,* ideas from the visual arts).

User studies can improve the quality of our research since we normally strive for effective visualizations. Although it is difficult to design a good experiment, and the relevant skills require substantial study tempered with experience, a well conducted study is usually worth the effort. The results can ultimately have a considerable impact and potentially contribute to the scientific foundations of the discipline.

Even though we advocate more user studies, we recognize that other methods are available that may be more appropriate in certain situations. We should be aware of these methods, so we can select the best tool for the problem at hand. One reason visualization is such a fascinating part of computer science is because there are so many other fields (*e.g.,* psychology and the visual arts) that overlap with our research.

How to do User Studies (Sidebar)

While a complete tutorial on user studies is beyond the scope of a short article, we hope to share some useful lessons we have learned.

The approach we are advocating is a form of applied perception research. Proper use of this technique requires an understanding of how to build experiments that include human participants. It is challenging to design an experiment that will give robust answers to the questions of interest. A typical study might ask, "Which of a set of prospective methods is most promising?" or, "Do any of these methods perform better than the best available alternative?" Unfortunately, there are many problems that can compromise the validity of the study, or make it difficult to draw useful insights from the results. Is the task appropriate? Is it possible that participants were using cues other than the ones being examined to perform the task? Is there a control condition to provide a baseline for comparison between different methods? Do all participants have a correct and equivalent understanding of the task? Are all participants sufficiently willing and able to perform the task? Is there a learning effect?

These problems can be addressed by testing participants for adequate spatial acuity, stereo ability, and absence of color blindness, by randomizing the presentation order of the trials, by using written instructions, by allowing participants to rest during the experiment to avoid becoming fatigued, by devising robust methods to identify when participants are giving "garbage answers", and by asking participants to successfully complete a training task before proceeding to the recorded trials. Due to the significant costs associated with running an experiment, it is often valuable to conduct a pilot study with one or two viewers. This allows testing and refining the experimental design before starting a full-fledged study with numerous participants.

A wide range of experimental methods may be appropriate. At one end of the spectrum is the rigorous application of signal detection methods [8]. These can be used to assess the detectability of a target structure from a background of noise. A more common experiment type is the evaluation of a number of different visual features. For example, a study might address the question of how well motion parallax, stereoscopic depth, and surface texture contribute to the perception of surface shape. Such an experiment calls for a factorial design with analysis of variance (ANOVA) to evaluate the results.

Another concern is the question of how many participants to use. The answer depends critically on what is being studied. For psychophysical experiments that measure low-level visual phenomena it is acceptable to use only a few participants. This is because between-viewer variability is expected to be low. These experiments contain numerous repeated measures (*i.e.,* multiple trials with the same experimental conditions) to ensure a sufficient total number of trials. If cognitive (as opposed to purely perceptual) processes are involved, more participants are normally required. Counterbalancing participants based on characteristics like gender, age, or experience may also be necessary. A detailed description of both participants and methods is an essential component for any publication involving user studies.

Finally, researchers at US universities should be aware that they may be required to obtain prior approval (or exemption) from the Institutional Review Board (IRB) at their institution before conducting any work involving human subjects; in other countries, similar requirements may apply.

In all cases, consulting with an expert on experiments can be invaluable. This will help not only with design, but also in applying appropriate statistical analyses to study the experimental results.

Acknowledgements

We would like to thank Helwig Hauser, who played a key role in proposing the idea for this paper.

Work described in this article was completed in part as a component of the basic research on visualization (http://www.VRVis.at/vis/) at the VRVis Research Center in Vienna, Austria (http://www.VRVis.at/), which is funded by the Austrian research program Kplus. Work was also supported in part by the National Science Foundation (ACI-0083421, CCR-0086065, CCR-0093238).

References

1. C. D. Jackson, D. B. Karelitz, S. A. Cannella, and D. H. Laidlaw (2002). *The Great Potato Search: The Effects of Visual Context on Users Feature Search and Recognition Abilities in an IVR Scene*. IEEE Visualization 2002 Poster Session.

2. J. A. Ferwerda, S. N. Pattanaik, P. Shirley and D. P. Greenberg (1997). *A Model of Visual Masking for Computer Graphics*, Proceedings of ACM SIGGRAPH 97, pp. 143-152.

3. C. G. Healey and J. T. Enns. "Large Datasets at a Glance: Combining Textures and Colors in Scientific Visualization." *IEEE Transactions on Visualization and Computer Graphics 5*, 2, (1999), 145-167.

4. S. Kim, H. Hagh-Shenas and V. Interrante (2003). *Showing Shape with Texture: Two Directions are Better than One*, Human Vision and Electronic Imaging VIII, SPIE 5007, to appear.

5. G. Kindlmann, E. Reinhard, and S. Creem (2002). *Face-based Luminance Matching for Perceptual Colormap Generation*, Proceedings IEEE Visualization 2002, pp. 299-306.

6. R. Kosara, S. Miksch, H. Hauser, J. Schrammel, V. Giller, and M. Tscheligi (2002). *Useful Properties of Semantic Depth of Field for Better F+C Visualization*, Proceedings of the Joint Eurographics–IEEE TCVG Symposium on Visualization (VisSym 2002), pp. 205-210.

7. D. H. Laidlaw, M. Kirby, J. S. Davidson, T. Miller, M. DaSilva, W. H. Warren, and M. Tarr (2001). *Quantitative comparative evaluation of 2D vector field visualization methods*. Proceedings IEEE Visualization 2001, pp. 143-150.

8. J.A. Swets and R.M. Pickett, *Evaluation of diagnostic Systems: Methods from Signal Detection Theory*. Academic Press, New York, New York, 1982.

9. J. T. Todd and F. D. Reichel. "Visual Perception of Smoothly Curved Surfaces from Double-Projected Contour Patterns." *Journal of Experimental Psychology: Human Perception and Performance 16,* 3, (1990), 665-674.

10. E. Vote, D.A. Feliz, D. H. Laidlaw, and M. S. Joukowsky "Discovering Petra: Archaeological Analysis in VR." *IEEE Computer Graphics & Applications 22*, 5, (2002), 38-50.

11. C. Ware. Color Sequences for Univariate Maps: Theory, Experiments and Principles. *IEEE Computer Graphics & Applications 8*, 5, (1998), 41-49.