Evaluating Information Visualizations

Sheelagh Carpendale

Department of Computer Science, University of Calgary,
2500 University Dr. NW, Calgary, AB, Canada T2N 1N4
sheelagh@ucalgary.ca

1 Introduction

Information visualization research is becoming more established, and as a result, it is becoming increasingly important that research in this field is validated. With the general increase in information visualization research there has also been an increase, albeit disproportionately small, in the amount of empirical work directly focused on information visualization. The purpose of this paper is to increase awareness of empirical research in general, of its relationship to information visualization in particular; to emphasize its importance; and to encourage thoughtful application of a greater variety of evaluative research methodologies in information visualization.

One reason that it may be important to discuss the evaluation of information visualization, in general, is that it has been suggested that current evaluations are not convincing enough to encourage widespread adoption of information visualization tools [57]. Reasons given include that information visualizations are often evaluated using small datasets, with university student participants, and using simple tasks. To encourage interest by potential adopters, information visualizations need to be tested with real users, real tasks, and also with large and complex datasets. For instance, it is not sufficient to know that an information visualization is usable with 100 data items if 20,000 is more likely to be the real-world case. Running evaluations with full data sets, domain specific tasks, and domain experts as participants will help develop much more concrete and realistic evidence of the effectiveness of a given information visualization. However, choosing such a realistic setting will make it difficult to get a large enough participant sample, to control for extraneous variables, or to get precise measurements. This makes it difficult to make definite statements or generalize from the results. Rather than looking to a single methodology to provide an answer, it will probably will take a variety of evaluative methodologies that together may start to approach the kind of answers sought.

The paper is organized as follows. Section 2 discusses the challenges in evaluating information visualizations. Section 3 outlines different types of evaluations and discusses the advantages and disadvantages of different empirical methodologies and the trade-offs among them. Section 4 focuses on empirical laboratory experiments and the generation of quantitative results. Section 5 discusses qualitative approaches and the different kinds of advantages offered by pursuing this type of empirical research. Section 6 concludes the paper.

A. Kerren et al. (Eds.): Information Visualization, LNCS 4950, pp. 19 – 45, 2008.

2 Challenges in Evaluating Information Visualizations

Much has already been written about the challenges facing empirical research in information visualization [2, 12, 53, 57]. Many of these challenges are common to all empirical research. For example, in all empirical research it is difficult to pick the right focus and to ask the right questions. Given interesting questions, it is difficult to choose the right methodology and to be sufficiently rigorous in procedure and data collection. Given all of the above, appropriate data analysis is still difficult and perhaps most difficult of all is relating a new set of results to previous research and to existing theory. However, information visualization research is not alone in these difficulties; the many other research fields that also face these challenges can offer a wealth of pertinent experience and advice.

In particular, empirical research in information visualization relates to human computer interaction (HCI) empirical research, perceptual psychology empirical research, and cognitive reasoning empirical research. The relationship to empirical research in HCI is evident in that many of the tasks of interest are interface interaction tasks, such as zooming, filtering, and accessing data details [66]. The aspects of these interactive tasks that provide access to the visual representation and its underlying dataset often relate to the usability of a system. Other challenges that are shared with HCI empirical research include the difficulty of obtaining an appropriate sample of participants. If the visualization is intended for domain experts it can be hard to obtain their time. Also, when evaluating complex visualization software, it may not be clear whether the results are due to a particular underlying technique or the overall system solution. If an existing piece of software is to be used as a benchmark against which to compare an interactive visualization technique, it is likely that participants may be much more familiar with the existing software and that this may skew the results. This problem becomes more extreme the more novel a given visualization technique is. Research prototypes are not normally as smooth to operate as well established software, creating further possibilities for affecting study results and leading to controversy about testing research prototypes against the best competitive solution. Greenberg and Buxton [27] discuss this problem in terms of interaction sketches, encouraging caution when thinking about conducting usability testing on new ideas and new interface sketches in order to avoid interfering with the development of new ideas. In addition, research software does not often reach a stage in which it can support a full set of possible tasks or be fully deployable in real-world scenarios [57].

In addition to usability questions, perceptual and comprehensibility questions such as those considered in perceptual psychology are important in assessing the appropriateness of a representational encoding and the readability of visuals [30, 79]. Also, in information visualization, there are a great variety of cognitive reasoning tasks that vary with data type and character, from low-level detailed tasks to complex high-level tasks. Some of these tasks are not clearly defined, particularly those that hold some aspect of gaining new insight into the data, and may be more challenging to test empirically. Examples of low-level detailed tasks include such tasks as compare, contrast, associate, distinguish, rank, cluster, correlate, or categorize [57]; higher-level and more complex cognitive tasks include developing an understanding of data trends, uncertainties, causal relationships, predicting the future, or learning a domain [1]. Many important tasks can require weeks or months to complete. The success of information

visualization is often an interplay between an expert's meta-knowledge and knowledge of other sources as well as information from the visualization in use.

While all of the above are important, a question that lies at the heart of the success of a given information visualization is whether it sheds light on or promotes insight into the data [55, 63]. Often, the information processing and analysis tasks are complex and ill-defined, such as discovering the unexpected, and are often long term or on-going. What exactly insight is probably varies from person to person and instance to instance; thus it is hard to define, and consequently hard to measure. Plaisant [57] describes this challenge as "answering questions you didn't know you had." While it is possible to ask participants what they have learned about a dataset after use of an information visualization tool, it strongly depends on the participants' motivation, their previous knowledge about the domain, and their interest in the dataset [55, 63]. Development of insight is difficult to measure because in a realistic work setting it is not always possible to trace whether a successful discovery was made through the use of an information visualization since many factors might have played a role in the discovery. Insight is also temporally elusive in that insight triggered by a given visualization may occur hours, days, or even weeks after the actual interaction with the visualization. In addition, these information processing tasks frequently involve teamwork and include social factors, political considerations and external pressures such as in emergency response scenarios. However, there are other fields of research that are also grappling with doing empirical research in complex situations. In particular, ecologists are faced with conducting research towards increasing our understanding of complex adaptive systems. Considering the defining factors of complex adaptive systems may help to shed some light on the difficulties facing empirical research in information visualization. These factors include non-linearity, holoarchy and internal causality [37, 49]. When a system is non-linear, the system behaviour comes only from the whole system. That is, the system can not be understood by decomposing it into its component parts which are then reunited in some definitive way. When a system is holoarchical it is composed of holons which are both a whole and a part. That is, the system is mutually inter-nested. While it is not yet common to discuss information analysis processes in terms of mutual nesting, in practice many information analysis processes are mutually nested. For instance, consider the processes of search and verification: when in the midst of searching, one may well stop to verify a find; and during verification of a set of results, one may well need to revert to search again. Internal causality indicates that the system is self-organizing and can be characterized by goals, positive and negative feedback, emergent properties and surprise. Considering that it is likely that a team of information workers using a suite of visualization and other software tools is some type of complex adaptive system suggests that more holistic approaches to evaluation may be needed.

Already from this brief overview, one can see that useful research advice on the evaluation of information visualization can be gathered from perceptual psychology, cognitive reasoning research, as well as human computer interaction research. Many, but not enough, information visualization researchers are already actively engaged in this pursuit. The purpose of this paper is to applaud them, to encourage more such research, and to suggest that the research community to be more welcoming of a greater variety of these types of research results.

3 Choosing an Evaluation Approach

A recent call for papers from the information visualization workshop, Beyond Time and Errors (BELIV06) held at Advanced Visual Interfaces 2006, stated that "*Controlled experiments remain the workhorse of evaluation but there is a growing sense that information visualization systems need new methods of evaluation, from longitudinal field studies, insight based evaluation and other metrics adapted to the perceptual aspects of visualization as well as the exploratory nature of discovery*" [7]. The purpose of this section is to encourage people to consider more broadly what might be the most appropriate research methods for their purposes. To further this purpose a variety of types of empirical research that can be usefully conducted are briefly outlined and these differing types are discussed in terms of their strengths and weaknesses. This discussion draws heavily from McGrath's paper Methodology Matters [50] that was initially written for social scientists. However, while social scientists work towards understanding humans as individuals, groups, societies and cultures, in information visualization – similarly to HCI – we are looking to learn about how information visualizations do or do not support people in their information tasks and/or how people conduct their information related tasks so that visualization can be better designed to support them. To gain this understanding we sometimes study people using information visualization software and sometimes it may be important to study people independently of that software, to better understand the processes we are trying to support.

There are some commonalities to all studies. They all must start with some question or questions that will benefit from further study. Also, they all must relate their research questions to the realm of existing ideas, theories and findings. These ideas, theories, and concepts are needed to relate the new study to existing research. For example, the results from a new study might be in contrast to existing ideas, in agreement with existing ideas, or offer an extension of or variation to existing ideas. A study must also have a method. This is what this section is about – possible types of empirical methodologies.

All methods offer both advantages and disadvantages. One important part of empirical research is choosing the most appropriate research methods for your content, your ideas, and your situation. The fact that methods both provide and limit evidence suggests that making use of a wide variety of methodologies will, in time, strengthen our understandings. Thus, both conducting a greater variety of studies and encouraging this by publishing research that employs a greater variety of methodologies will help to develop a better understanding of the value of information visualization and its potential in our communities.

When conducting a study there are three particularly desirable factors: generalizability, precision, and realism [50]. Ideally, one would like all of these factors in one's results. However, existing methodologies do not support the actualization of all three simultaneously. Each methodology favours one or two of these factors, often at the expense of the others; therefore the choice of a methodology for a particular goal is important. To define these terms (as used in McGrath [50]):

- **Generalizability**: a result is generalizable to the extent to which it can apply to other people (than those directly in the study) and perhaps even extend to other situations.

- **Precision:** a result is precise to the degree to which one can be definite about the measurements that were taken and about the control of the factors that were not intended to be studied.
- **Realism:** a result is considered realistic to the extent to which the context in which it was studied is like the context in which it will be used.

Figure 1 (adapted and simplified from McGrath [50]) shows the span of common methodologies currently in practice in the social sciences. They are positioned around the circle according to the labels: most precision, most generalizability and most realism. The closer a methodology is placed to a particular label, the more that label applies to that methodology. Next, these methodologies are briefly described. For fuller descriptions see McGrath 1995.

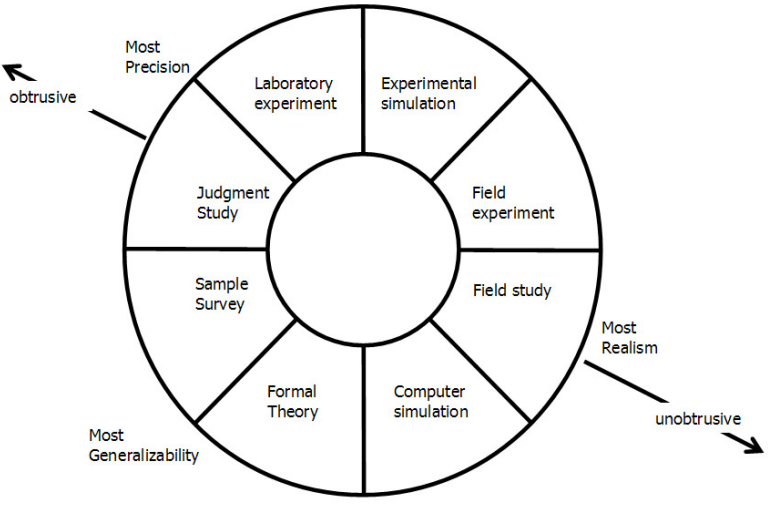

Fig. 1. Types of methodologies organized to show relationships to precision, generalizability and realism. (adapted, simplified from McGrath 1995)

Field Study: A field study is typically conducted in the actual situation, and the observer tries as much as possible to be unobtrusive. That is, the ideal is that the presence of the observer does not affect what is being observed. While one can put considerable effort into minimizing the impact of the presence of an observer, this is not completely possible [50]. Examples of this type of research include ethnographic work in cultural anthropology, field studies in sociology, and case studies in industry. In this type of study the realism is high but the results are not particularly precise and likely not particularly generalizable. These studies typically generate a focused but rich description of the situation being studied.

Field Experiment: A field experiment is usually also conducted in a realistic setting; however, an experimenter trades some degree of unobtrusiveness in order to obtain more precision in observations. For instance, the experimenter may ask the participants to perform a specific task while the experimenter is present. While realism is

still high, it has been reduced slightly by experimental manipulation. However, the necessity of long observations may be shortened and results may be more readily interpretable and specific questions are more likely to be answered.

Laboratory Experiment: In a laboratory experiment the experimenters fully design the study. They establish what the setting will be, how the study will be conducted, what tasks the participants will do, and thus plan the whole study procedure. Then the experimenter gets people to participate as fully as possible following the rules of the procedure within the set situation. Carefully done, this can provide for considerable precision. In addition, non-realistic behaviour that provides the experimenter more information can be requested such as a 'think aloud' protocol [43]. Behaviour can be measured, partly because it is reasonably well known when and where the behaviour of interest may happen. However, realism is largely lost and the degree to which the experimenter introduces aspects of realism will likely reduce the possible precision.

Experimental Simulation: With an experimental simulation the experimenter tries to keep as much of the precision as possible while introducing some realism via simulation. There are examples where this approach is essential such as studying driving while using a cell phone or under some substance's influence by using a driving simulator. Use of simulation can avoid risky or un-ethical situations. Similarly although less dramatically, non-existent computer programs can be studied using the 'Wizard of Oz' approach in which a hidden experimenter simulates a computer program. This type of study can provide us with considerable information while reducing the dangers and costs of a more realistic experiment.

Judgment Study: In a judgment study the purpose is to gather a person's response to a set of stimuli in a situation where the setting is made irrelevant. Much attention is paid to creating 'neutral conditions'. Ideally, the environment would not affect the result. Perceptual studies often use this approach. Examples of this type of research include the series of studies that examine what types of surface textures best support the perception of 3D shape (e.g. [34, 38]), and the earlier related work about the perception of shape from shading [39]. However, in assessing information visualizations this idea of setting a study in neutral conditions must be considered carefully, as witnessed by Reilly and Inkpen's [62] study which showed that the necessity for an interactive technique developed to support a person's mental model during transition from viewing one map to another (subway map to surface map) was dependent on the distractions in the setting. This transition technique relates to ideas of morphing and distortion in that aspects of the map remain visible while shifting. These studies in a more neutral experiment setting showed little benefit, while the same tasks in a noisy, distracting setting showed considerable benefit.

Sample Survey: In a sample survey the experimenter is interested in discovering relationships between a set of variables in a given population. Examples of these types of questions include: of those people who discover web information visualization tools how many return frequently and are their activities social or work related? Of those people who have information visualization software available at work what is the frequency of use? Considering the increased examples of information visualization results and software on the web, is the general population's awareness of and/or use of information visualization increasing? In these types of studies proper sampling

of the population can lead to considerable generalizability. However, while choosing the population carefully is extremely important, often it is difficult to control. For example in a web-based survey, all returned answers are from those types of people who are willing to take the time, fill out the questionnaire, etc. This is a type of bias and thus reduces generalizability. Also, responses are hard to calibrate. For instance, a particular paper reviewer may never give high scores and the meta-reviewer may know this and calibrate accordingly or may not know this. Despite these difficulties, much useful information can be gathered this way. We as a community must simply be aware of the caveats involved.

Formal Theory: Formal theory is not a separate experimental methodology but an important aspect of all empirical research that can easily be overlooked. As such, it does not involve the gathering of new empirical evidence and as a result is low in both precision and realism. Here, existing empirical evidence is examined to consider the theoretical implications. For example, the results of several studies can be considered as a whole to provide a higher-level or meta-understanding or the results can be considered in light of existing theories to extend, adjust or refute them. Currently this type of research is particularly difficult to publish in that there are no new information visualizations and no new empirical results. Instead, the contribution moves towards the development of theories about the use of and practicality of information visualizations.

Computer Simulation: It is also possible to develop a computer simulation that has been designed as logically complete. This method is used in battle simulation, research and rescue simulation, etc. This type of strategy can be used to assess some visualizations. For instance, a visualization of landscape vegetation that includes models of plant growth and models of fire starts and spread can be set to simulate passage of several hundred years. If the resulting vegetation patterns are comparable to existing satellite imagery this provides considerable support for the usefulness of models [22]. Since this type of research strategy does not involve participants, discussion of generalizability over populations is not applicable. Also, since the models are by definition incomplete, notions of precision in measurement are often replaced with stochastic results. On the other hand it does provide a method of validation and offers a parallel with which we can study realistic situations, such as explosions, turbulence in wind tunnels, etc.

4 Focus on Quantitative Evaluation

Quantitative evaluations, most well known as laboratory experiments or studies, are those methodologies in which precision is relatively high and in which some declaration can be made about the possible generalization to a larger population. These declarations can include information about the characterization of this larger population and how likely it is that the generalization will hold. These types of experiments or studies are part of the traditional empirical scientific experimental approach and have evolved and been refined through the centuries of scientific research. Science does and has depended on these methods. Slowly, through careful and rigorous application of the experimental process, knowledge has been built up, usually one piece at a time.

The experiments or studies involve a rigorous process of hypothesis development, identification and control of the independent variables, observation and measurement of the dependent variables, and application of statistics which enable the declaration of the confidence with which the results can be taken. In these formal studies or controlled evaluations, the experimenter controls the environment or setting, manipulates chosen factor(s) or variable(s) – the independent variable(s) - in order to be able to measure and observe the affect this manipulation has on one or more other factors – the dependent variable(s). Ideally no other factors change during the experiment. Once the changes to the dependent variables have been measured, statistical methods can be applied to understand the relative importance of the results. Done with sufficient thoroughness, this process can arrive at facts about which we can be relatively certain. The application of this scientific process will to try to reduce the overall complexity by fine tuning particular questions or hypotheses, using these hypotheses to allow one to cull some of the complexity by trying to eliminate as many of the extraneous variables as possible. Traditionally experiments of this type are used to shed light on cause and effect relationships; that is, to discover whether changes in some factor result in changes to another factor.

This idea that we can observe simpler, more manageable subsets of the full complex process is appealing, and it is clear from centuries of experiments that much can be learnt in this manner.

4.1 Quantitative Methodology

Since quantitative empirical evaluations have evolved over the centuries the methodology has become relatively established (Figure 2). This brief overview is included for completeness; the interested reader should refer to the many good books on this subject [15, 17, 33]. This methodology includes:

- **Hypothesis Development:** Much of the success of a study depends on asking an interesting and relevant question. This question should ideally be of interest to the broader research community, and hopefully answering it will lead to a deeper or new understanding of open research questions. Commonly the importance of the study findings results from a well thought through hypothesis, and formulating this question precisely will help the development of the study.
- **Identification of the Independent Variables:** The independent variables are the factors to be studied which may (or may not) affect the hypothesis. Ideally the number of independent variables is kept low to provide more clarity and precision in the results.
- **Control of the Independent Variables:** In designing the experiment the experimenter decides the manner in which the independent variables will be changed.
- **Elimination of Complexity:** In order to be clear that it is actually the change in the independent variable that caused the study's result, it is often the case that other factors in the environment need to be controlled.
- **Measurement of the Dependent Variables:** Observations and measurements are focused on the dependent variables as they change or do not change in

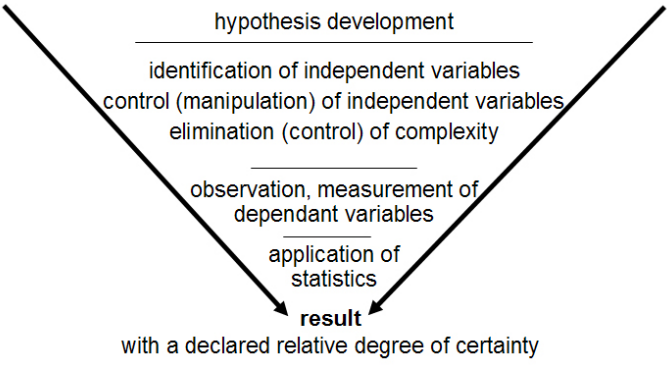

Fig. 2. A simple schematic of the traditional experimental process.

response to the manipulation of the independent variable. The aspects to be measured are often called metrics. Common metrics include: speed, accuracy, error rate, satisfaction, etc.

- **Application of Statistics:** The results collected can then be analysed through the application of appropriate statistics. It is important to remember that statistics tell us how sure we can be that these results could (or could not) have happened by chance. This gives a result with a relative degree of certainty. There are many good references such as Huck [33].

These steps sound deceptively simple but doing them well requires careful and rigorous work. For instance, it is important that the study participants are valued, that they are not over-stressed, and that they are given appropriate breaks, etc. Also, exactly what they are being asked to do must be clear and consistent across all participants in your study. Since small inconsistencies such as changes in the order of the instructions can affect the results, the common recommendation is that one scripts the explanations. Perhaps most importantly, to eliminate surprises and work out the details, it is best to pilot – run through the experiment in full – repeatedly.

4.2 Quantitative Challenges

Even though these types of experiments have been long and effectively used across all branches of science, there remain many challenges to conducting a useful study. We mention different types of commonly-discussed errors and validity concerns and relate these to the McGrath's discussion as outlined in Section 3. In this discussion we will use a simple, abstract example of an experiment that looks at the effect of two visualization techniques, VisA and VisB, on performance in search. There are several widely discussed issues that can interfere with the validity of a study.

Conclusion Validity: Is there a relationship? This concept asks whether within the study there is a relationship between the independent and the dependent variables. Important factors in conclusion validity are finding a relationship when one does not exist (type I error) and not finding a relationship when one does exist (type II error).

Table 1. Type I and Type II Errors

		Reality	
		H_0 TRUE	H_0 FALSE
Experimental decision	H_0 TRUE	ok	Type II
	H_0 FALSE	Type I	ok

Type I and Type II Errors: If one is interested in which visualization technique VisA or VisB helps people conduct a particular task faster one might formulate a null hypothesis (H_0) – *there is no difference in speed of search between VisA and VisB*. The possible type I, false negative, and type II, false positive, errors are specified in Table 1. The columns represent whether the null hypothesis is true or false in reality and the rows show the decision made based on the results of the experiment. Ideally the results of the experiment reflect reality and that if the hypothesis is false (or true) in reality it will show as false (or true) in the experiment. However, it is possible that the hypothesis is true in reality – VisA does support faster search than VisB – but that this fails to be revealed by the experiment. This is a type II error. A type I error occurs if the null hypothesis is true in reality (there is no difference) and one concludes that there *is* a difference. Type I errors are considered more serious. That is, it is considered worse to claim that VisA improves search when it does not, than to say there was no measurable difference.

Internal Validity: Is the relationship causal? This concept is important when an experiment is intended to reveal something about causal relationships. Thus, internal validity will be important in our simple example because the study is looking at what effect VisA and VisB have on search. The key issue here is whether the results of one's study can properly be attributed to what happened within the experiment. That is, that no other factors influenced or contributed to the results seen in the study. Another way of asking this question is: are there possible alternate causes for the results seen in the study?

Construct Validity: Can we generalize to the constructs (ideas) the study is based on? This concept considers whether the experiment has been designed and run in a manner that answers the intended questions. This is an issue about whether the right factors are being measured or whether the factors the experimenter intends to measure are actually those being measured. For instance, does the experiment measure the difference due to the techniques VisA and VisB or the difference in participant's familiarity with VisA and VisB. For instance, if the construct is that a person will have higher satisfaction when using VisB, does measuring error rates and completion times provide answers for this construct? An important part of this concept of construct validity is **measurement validity**. Measurement validity is concerned with questions such as is one measuring what one intends to measure and is the method of measurement reliable and consistent. That is, will the same measurement process provide the same results when repeated?

External Validity: Can we generalize the study results to other people/places/times? External validity is concerned with the extent to which the result of a study can be generalized. If a study has good internal and construct validity the results apply to the setting, time, and participants being studied. The extent to which the results apply beyond the immediate setting, time and participants depends, for participants, on the participant sample and the population from which it was drawn. For instance, in practice it is common to draw participants from the geographic region in which the study is run. Does this mean that the results only apply to people from that region? If culture has a possible impact on the results, they may not generalize. If one addresses the need to include cultural variation by recruiting participants from different cultures from a university's foreign students, one might have at least partially addressed the need to run the study across cultural variations but now have limited the demographic to university students which may introduce its own skew. Understanding the population to which one would like to be able to generalize the study results and successfully obtaining an appropriate participant sample is a difficult issue. This does not mean we can not learn from more specific participant samples. It does mean that reporting the demographics of the sample and being cautious about generalizations is important. Participant sample choice is just one factor influencing external validity. Setting includes other factors such as noise, interruption, and distractions. Possible temporal factors include events that occurred before or are anticipated after the experiment.

Ecological Validity: Ecological validity discussions focus on the degree to which the experimental situation reflects the type of environment in which the results will be applied. This concept relates strongly to McGrath's concept of realism. It is distinct from the idea of external validity, in that external validity is concerned with whether the experimental results generalize to other situations, while ecological validity is concerned with how closely the experimental settings matches the real setting in which the results might be applied. Thus it is possible to have good ecological validity; the study is conducted on site, but that the results are applicable only to that site. This would indicate poor external validity in that the results do not generalize beyond the specific setting.

4.3 Quantitative Studies Summary Remarks

The number of quantitative studies in information visualization is increasing. Early examples include the series of studies done by Purchase and her collaborators that examine the impact of graph drawing aesthetics on comprehension and usability [58, 59, 60, 61]. Dumais et al. [16] explored use of context techniques in web search. Forlines et al. [23] looked at the effect of display configuration on relationship between visual search and information visualization tasks. Recently, Willet et al. [81] studied embedding information visualizations in widgets.

Quantitative experiments have formed the backbone of experimental science and it is to be expected that they will continue to do so. However, it is relatively easy to find fault in any given experiment because all factors can not usually be completely controlled. If they are completely controlled, external and ecological validity can be impacted. This is particularly true for studies involving humans. Designing and working with experiments is often a matter of making choices about what factors are important and understanding the strengths and limitations of any given study and its results. As

a community it is important that we recognise that we are working towards a larger understanding and that any given study will not present the bigger answer. It instead will contribute to a gradual building of a bigger understanding. For this bigger understanding we need to encourage authors to openly discuss the limitations of their studies, because both the results and the limitations are important. This is also true for negative results. It can be just as important to understand when there are no differences among techniques and when these differences exist.

5 Focus on Qualitative Evaluation

Qualitative inquiry works toward achieving a richer understanding by using a more holistic approach that considers the interplay among factors that influence visualizations, their development, and their use [56]. Qualitative techniques lend themselves to being more grounded in more realistic settings and can also be incorporated into all types of studies. This includes qualitative studies conducted as part of the design process [64, 73], in situ interviews [83], field studies [72], and use of observational studies to create design and evaluative criteria that are derived from observed data [71]. These types of studies offer potential for improved understanding of existing practices, analysis environments, and cognitive task constraints as they occur in real or realistic settings. In providing a brief overview of a variety of qualitative methods, we hope to spark further research and application of qualitative methods in information visualization; to expand our empirical approaches to include the application of qualitative methods to design and evaluation; and to encourage a wider acceptance of these types of research methodologies in our field.

5.1 Qualitative Methods

At the heart of qualitative methods is the skill and sensitivity with which data is gathered. Whether the records of the data gathered are collected as field notes, artefacts, video tapes, audio tapes, computer records and logs, or all of these, in qualitative empirical approaches there are really only two primary methods for gathering data: observations and interviews. Observation and interview records are usually kept continually as they occur, as field notes, as regular journal entries as well as often being recorded as video or audio tapes. Artefacts are collected when appropriate. These can be documents, drawings, sketches, diagrams, and other objects of use in the process being observed. These artefacts are sometimes annotated as part of use practices or in explanation. Also, since the communities we are observing are often technology users, technology-based records can also include logs, traces, screen captures, etc. Both observation and interviewing are skills and as such develop with practice and can, at least to some extent, be learnt. For full discussions on these skills there are many useful books such as Seidman [65] and Lofland and Lofland [45].

5.1.1 Observation Techniques

The following basic factors have been phrased in terms of developing observational records but implicitly also offer advice on what to observe:

- Try to keep jotting down notes unobtrusively. Ideally, notes are taken as observations occur; however, if one becomes aware that one's note taking is having an impact on the observations, consider writing notes during breaks, when shielded, or at the end of the day.
- Minimize the time gap from observations to note taking. Memory can be quite good for a few hours but does tend to drop off rapidly.
- Include in observations the setting, a description of the physical setup, the time, who is present, etc. Drawing maps of layouts and activities can be very useful.
- Remember to include both the overt and covert in activities and communications. For example, that which is communicated in body language and gestures, especially if it gets understood and acted upon, is just as important as spoken communications. But be careful of that grey area where one is not sure to what extent a communication occurred.
- Remember to include both the positive and negative. Observed frustrations and difficulties can be extremely important in developing a fuller understanding.
- Do not write notes on both sides of a paper. This may seem trivial but experienced observers say this is a must [6]. You can search for hours, passing over many times that important note that is on the back of another note.
- Be concrete whenever possible.
- Distinguish between word-for-word or verbatim accounts and those you have paraphrased and/or remembered.

5.1.2 Interview Techniques

These are a few brief points of advice about interviewing. Do remember that while sorting out the right questions to ask is important, actively listening to what the participant says is the most important of all interviewing skills.

- Make sure that you understand what they are telling you and that the descriptions, explanations they are giving you are complete enough. However, when asking for clarification, try to avoid implying that their explanations are poor because one does not want to make one's participants defensive. Ask instead for what they meant by particular word usage or if they would explain again. The use of the word *again* implies that the interviewer did not catch it all rather than the explanation was incomplete.
- Limit your inclination to talk. Allow for pauses in the conversation, sometimes note taking can be useful here. The participant will expect you to be taking notes. In this situation note taking can actually express respect for what the participant has said.
- Remember that the default is that the participant will regard the interview to some extent as public and thus will tell you the public version. Do listen for and encourage the less formal, less guarded expression of their thoughts. One example, from Seidman [65], is the use of the word 'challenge'. Challenge is an excepted term for a problem. The details of the problem might be explained more fully if one asks what is meant in the given situation by the word challenge.

- Follow up on what the participant says. Do allow the interview to be shaped by the information your participant is sharing.
- Avoid leading questions. An important part of minimizing experimenter bias is wording questions carefully so as to avoid implying any types of answers. For example, asking a participant what a given experience was like for them, leaves space for their personal explanations.
- Ask open ended questions. This can involve asking for a temporal reconstruction of an event or perhaps a working a day or asking for a subjective interpretation of an event.
- Ask for concrete details. These can help trigger memories.
- With all the above do remember that one of the most important pluses of an interview process is the humanity of interviewer. Being present, aware and sensitive to the process is your biggest asset. These guidelines are just that; guidelines to be used when useful and ignored when not.

5.2 Types of Qualitative Methodologies

This section is not intended to be a complete collection of all types of qualitative inquiry. Rather it is meant to give an overview of some of the variations possible, set in a discussion about when and where they have proven useful. This overview is divided into three sections. First, the type of qualitative methodologies often used in conjunction with or as part of more quantitative methodologies is discussed. Then, we mention the approaches taken in the area of heuristic, or, as they are sometimes referred to 'discount', inspection methodologies. The last section will cover some study methodologies that are intentionally primarily qualitative.

5.2.1 Nested Qualitative Methods

While qualitative methodologies can be at the core of some types of studies, some aspects of qualitative inquiry are used in most studies. For instance, data gathered by asking participants for their opinions or preferences is qualitative. Gorard [26] argues that quantitative methods can not ignore the qualitative factors of the social context of the study and that these factors are, of necessity, involved in developing an interpretation of the study results. There are many methods used as part of studies such as laboratory experiments that provide us with qualitative data. The following are simply a few examples to illustrate how common this mixed approach is.

Experimenter Observations: An important part of most studies is that the experimenter keeps notes of what they observe as it is happening. The observations themselves can help add some degree of realism to the data and the practice of logging these observations as they happen during the study helps make them more reliable than mere memory. However, they are experimenter observations and as such are naturally subjective. They do record occurrences that were not expected or are not measurable so that they will also form part of the experimental record. These observations can be helpful during interpretation of the results in that they may offer explanations for outliers, point towards important experimental re-design, and suggest future directions for study. Here, experimenter observations augment and enrich the primar-

ily quantitative results of a laboratory experiment and in this they play an important but secondary role.

Think-Aloud Protocol: This technique, which involves encouraging participants to speak their thoughts as they progress through the experiment, was introduced to the human-computer-interaction community by [43]. Discussions about this protocol in psychology date back to 1980 [19, 20, 21]. Like most methodologies, this one also involves tradeoffs. While it gives the experimenter/observer the possibility of being aware of the participants' thoughts, it is not natural for most people and can make a participant feel awkward; thus, think aloud provides additional insight while also reducing the realism of the study. However, the advantage for hearing about a participant's thoughts, plans, and frustrations frequently out-weigh the disadvantages and this is a commonly used technique. Several variations have been introduced such as 'talk aloud' which asks a participant to more simply announce their actions rather than their thoughts [21].

Collecting Participant Opinions: Most laboratory experiments include some method by which participant opinions and preferences are collected. This may take the form of a simple questionnaire or perhaps semi-structured interviews. Most largely quantitative studies such as laboratory experiments do ask these types of questions, often partially quantifying the participant's response by such methods as using a Likert scale [44]. A Likert scale asks a participant to rate their attitude according to degree. For instance, instead of simply asking a participant, 'did you like it?' A Likert scale might ask the participant to choose one of a range of answers 'strongly disliked,' 'disliked,' 'neutral,' 'liked,' or 'strongly liked.'

Summary of Nested Qualitative Methods: The nested qualitative methods mentioned in this section may be commonplace to many readers. The point to be made here is that in the small, that is as part of a laboratory experiment, inclusion of some qualitative methods is not only commonplace, its value is well recognized. This type of inclusion of qualitative approaches adds insight, explanations and new questions. It also can help confirm results. For instance, if participants' opinions are in line with quantitative measures – such as the fastest techniques being the most liked – this confirms the interpretation of the fastest technique being the right one to chose. However, if they contradict – such as the fastest techniques not being preferred – interesting questions are raised including questioning the notion that fastest is always best.

5.2.2 Inspection Evaluation Methods

We include a discussion of inspection methods because, while they are not studies per se, they are useful, readily available, and relatively inexpensive evaluation approaches. The common approach is to use a set of heuristics as a method of focusing attention on important aspects of the software – interface or visualization – which need to be considered [54]. These heuristics or guidelines can be developed by experts or from the writings of experts. Ideally, such an inspection would be conducted by individual experts or even a group of experts. However, it has been shown that in practice, that a good set of heuristics can still be effective in application if a few, such as three or four, different people apply them [54]. For information visualization it is important to consider exactly what visualization aspects a given set of heuristics will shed light on.

Usability Heuristics: These heuristics, as introduced and developed by Nielson and Mack [1994], focus on the usability of the interface and are designed to be applied to any application, thus are obviously of use to information visualizations. They will help make sure that general usability issues are considered. These heuristics are distilled down to ten items – visibility of system status, match between system and real world, personal control and freedom, consistency and standards, error prevention, recognition rather than recall, flexibility and efficiency, aesthetic and minimalist design, errors handling, and help and documentation.

Collaboration Heuristics: When interfaces are designed for collaboration, two additional major categories arise in importance: communication and coordination. Baker et al. [4] developed a set of heuristics that explore these issues based on the Mechanics of Collaboration [29]. As information visualizations start to be designed for collaborative purposes, both distributed [31, 78] and co-located [35], these heuristics will also be important.

Information Visualization Heuristics: While the usability heuristics apply to all infovis software and the collaboration heuristics apply to the growing body of collaborative information visualizations, there are areas of an information visualization that these at best gloss over. In response, the Information Visualization research community has proposed a variety of specific heuristics. Some pertain to given data domains such as ambient displays [46] and multiple view visualizations [5]. Others focus on a specific cognitive level, for instance knowledge and task [1], or task and usability [66]. Tory and Möller [74] propose the use of heuristics based on both visualization guidelines and usability. As explored by Zuk and Carpendale [84], we can also consider developing heuristics based on the advice from respected experts such as design advice collected from Tufte's writings [75, 76, 77], semiotic considerations as expressed by Bertin [8] and/or research in cognitive and perceptual science as collected by Ware [79]. Alternatively, we can start from information visualization basics such as presentation, representation and interaction [68]. However, a concept such as presentation cuts across design and perception, while representation advice, such as what types of visuals might best represent what types of data, might be distilled from the guidelines put forth by Bertin [8] and from an increasing body of cognitive science as gathered in Ware [79]. Sorting out how to best condense these is a task in itself [52, 85]. "At this stage of development of heuristics for information visualization we have reached a similar problem as described by Nielson and Mack [54]. It is a difficult problem to assess which list(s) are better for what reasons and under what conditions. This leads to the challenges of developing an optimal list that comprises the most important or common Information Visualization problems" (page 55, [85]).

Summary of Inspection Evaluation Methods: While experience in the human computer interaction communities and the growing body of information visualization specific research indicates that heuristics may prove a valuable tool for improving the quality of information visualizations, there is considerable research yet to be conducted in the development of appropriate taxonomies and application processes for heuristics in information visualization.

The currently recommended application approach for usability heuristics is that evaluators apply the heuristics in a two pass method. The first pass is done to gain an overview and second is used to asses in more detail each interface component with

each heuristic [54]. The original use indicated that in most situations three evaluators would be cost effective and find most usability problems [54]. However, subsequent use of heuristics for web site analysis appears to sometimes need more evaluators [9, 69]. Further, this may depend on the product. While application of heuristics has not yet been formally studied in terms of web sites, it does introduce the possibility that information visualization heuristics may also need to be data, task or purpose specific.

Heuristics are akin to the design term *guidelines* in that both provide a list of advice. Design guidelines are often usefully applied in a relatively ad hoc manner as factors to keep in mind during the design process and heuristic lists can definitely be similarly used. While there are definitely benefits that accrue in the use of guidelines and heuristics, it is important to bear in mind that they are based on what is known to be successful and thus tend not to favour the unusual and the inventive. In the design world, common advice is that while working without knowledge of guidelines is foolish, following them completely is even worse.

5.2.3 Qualitative Methods as Primary

A common reason for using qualitative inquiry is to develop a richer understanding of a situation by using a more holistic approach. Commonly, the qualitative research method's goal is to collect data that enables full, rich descriptions rather than to make statistical inferences [3, 14]. There are a wealth of qualitative research methods that can help us to gain a better understanding of the factors that influence information visualization use and design. Just as we have pointed out how qualitative methods can be effectively used within quantitative research, qualitative research can also include some quantitative results. For instance, there may be factors that can be numerically recorded. These factors can then be presented in combination with qualitative data. For example, if a questionnaire includes both fixed-choice questions and open ended questions, quantitative measurement and qualitative inquiry are being combined [56].

Qualitative methods can be used at any time in the development life cycle. A finished or near to finished product can be assessed via case studies or field studies. Also, there is a growing use of these methods as a preliminary step in the design process. The HCI and particularly the computer-supported cooperative work (CSCW) research communities have successfully been using qualitative methods to gain insight that can inform the initial design. CSCW researchers have learned a lot about how to support people working together with technology through pre-design observation and qualitative analysis of how people work together without technology. The basic idea is that through observations of participants' interactions with physical artefacts, a richer understanding of basic activities can be gained and that this understanding can be used to inform interface design. This approach generally relies on observation of people, inductive derivation of hypotheses via iterative data collection, analysis, and provisional verification [14]. For example, Tang's study of group design activities around shared workspaces revealed the importance of gestures and the workspace itself in mediating and coordinating collaborative work [73]. Similarly, Scott et al. [64] studied traditional tabletop gameplay and collaborative design, specifically focusing on the use of tabletop space, and the sharing of items on the table. Both studies are an example of how early evaluation can inform the design of digital systems. In both cases, the authors studied traditional, physical contexts first, to understand participants' interactions with the workspace, the items in the workspace, and

within the group. The results of these experiments are regarded as providing important information about what group processes to support and some indication about how this might be done. This type of research can be particularly important in complex or sensitive scenarios such as health care situations [72]. Brereton and McGarry [11] observed groups of engineering students and professional designers using physical objects to prototype designs. They found that the interpretation and use of physical objects depended greatly on the context of its placement, indicating that the context of people's work is important and is difficult to capture quantitatively. Their goal was to determine implications for the design of tangible interfaces. Other examples include Saraiya et al. [63] who used domain expert assessments of insight to evaluate bioinformatics visualizations, while Mazza and Berre [48] used focus groups and semi-structured interviews in their analysis of visualization approaches to support instructors in web-based distance education.

The following are simply examples of empirical methods in which gathering of qualitative data is primary. There are many others; for instance, Moggridge [51] mentions that his group makes active use of fifty-one qualitative methods in their design processes.

In Situ Observational Studies: These studies are at the heart of field studies. Here, the experimenter gets permission to observe activities as they take place in situ. In these studies the observer does their best to remain unobtrusive during the observations. The ideal in Moggridge's terms is to become as a 'fly on the wall' that no one notices [51]. This can be hard to achieve in an actual setting. However, over time a good observer does usually fade into the background. Sometimes observations can be collected via video and audio tapes to avoid the more obvious presence of a person as observer but sometimes making such recordings is not appropriate as in medical situations. In these studies the intention is usually to gather a rich description of the situation being observed. However, there is both a difference and an overlap in the type of observations to be gathered when the intention is (a) to better understand the particular activities in a given of setting, or (b) to use these observations to inform technology design. Thus, because different details are of prime interest it is important that our research community conducts these types of observational studies to better inform initial design as well as to better understand the effectiveness of new technology in use. These studies have high realism, result in rich context explicit data and are time and labour intensive when it comes to both data collection and data analysis.

Participatory Observation: This practice is the opposite of participatory design. Here an information visualization expert becomes part of the application expert's team to experience the work practices first hand rather than application experts becoming part of the information visualization design team. In participatory observation, additional insights can be gained through first-hand observer experience of the tasks and processes of interest in the context of the real world situation. Here, rather than endeavouring to be unobtrusive, the observer works towards becoming an accepted part of the community. Participatory observation is demonstrably an effective approach since as trust and rapport develop, an increasingly in-depth understanding is possible. Our research community is interested in being able to better understand the work practices of many different types of knowledge workers. These workers are usually highly trained, highly paid, and often under considerable time pressures. Not

surprisingly, they are seldom willing to accept an untrained observer as part of their team. Since information visualization researchers are of necessity highly trained themselves, it is rare that an information visualization researcher will have the necessary additional training to become accepted as a participatory observer. However, domain expertise is not always essential for successful participatory observation. Expert study participants can train an observer on typical data analysis tasks – a process which may take several hours, and then "put them to work" on data analysis using their existing tools and techniques. The observer keeps a journal of the experience and the outcomes of the analysis were reviewed with the domain experts for validity. Even as a peripheral participant, valuable understandings of domain, tasks, and work culture can be developed which help clarify values and assumptions about data, visualizations, decision making and data insights important to the application domain. These understandings and constructs can be important to the information visualization community in the development of realistic tools.

Laboratory Observational Studies: These studies use observational methodologies in a laboratory setting. A disadvantage of in situ observations is that they often require lengthy observations. For instance, if the observer is interested in how an analyst uses visual data, they will have to wait patiently until the analyst does this task. Since an analyst may have many other tasks – meetings, conference calls, reports, etc. – this may take hours or even days. One alternative to the lengthy in situation wait is to design an observational experiment in which, similarly to a laboratory experiment, the experimenter designs a setting, a procedure and perhaps even a set of tasks. Consider, for example, developing information visualizations to support co-located collaboration. Some design advice on co-located collaborative aspects is available in the computer supported cooperative work literature [35]. However, while this advice is useful, it does not inform us specifically about how teams engage in collaborative tasks when using visual information. Details such as how and when visualizations will be shared and what types of analysis processes need to be specifically supported in collaborative information visualization systems were missing. Here, an observational approach is appropriate because the purpose is to better understand the flow and nature of the collaboration among participants, rather than answering quantifiable lower-level questions. In order to avoid temporal biases in existing software, pencil and paper based visualizations were used. This allowed for the observation of free arrangement of data, annotation practices, and collaborative processes unconstrained by any particular visualization software [36].

Contextual Interviews: As noted in Section 5.1, interviewing in itself is core to qualitative research. Conducting an interview about a task, setting, or application of interest within the context in which this work usually takes place is just one method that can enrich the interview process. Here the realism of the setting helps provide the context that can bring to mind the day-to-day realities during the interview process (for further discussion see Holtzblatt and Beyer 1998). For example, to study how best to support the challenging problem of medical diagnosis, observing and interviewing physicians in their current work environment might help to provide insights into their thought processes that would be difficult to capture with other methodologies. A major benefit of qualitative study can be seeing the big picture – the context in which a new visualization support may be used. The participants' motives, misgiv-

ings, and opinions shed light on how they relate to existing support, and can effectively guide the development of new support. This type of knowledge can be very important at the early stage of determining what types of information visualizations may be of value.

Summary of qualitative methods as primary: These four methods are just examples of a huge variety of possibilities. Other methods include action research [42], focus groups [48], and many more. All these types of qualitative methods have the potential to lessen the task and data comprehension divide between ourselves as visualization experts and the domain experts for whom we are creating visualizations. That is, while we can not become analysts, doctors, or linguists, we can gain a deeper understanding of how they work and think. These methods can open up the design space, revealing new possibilities for information visualizations, as well as additional criteria on which to measure success.

5.3 Challenges for Qualitative Methods

A considerable challenge to qualitative methods is that they are particularly labour intensive. Gathering data is a slow process and rich note taking is an intensive undertaking, as are transcribing and subsequent analysis.

5.3.1 Sample Sizes

Sample sizes for qualitative research are determined differently than for quantitative research. Since qualitative research is not concerned with making statistically significant statements about a phenomenon, the sample sizes are often lower than required for quantitative research. Often, sample sizes are determined during the study. For instance, a qualitative inquiry may be continued until one no longer appears to be gaining new data through observation [3]. There is no guideline to say when this 'saturation' may occur [70]. Sample sizes may vary greatly depending on the scope of the research problem but also the experience of the investigator. An experienced investigator may reach a theoretical saturation earlier than a novice investigator. Also, because each interview and/or observation can result in a large amount of data, sometimes compromises in sample size have to be made due to considerations about the amount of data that can be effectively processed.

5.3.2 Subjectivity

Experimenter subjectivity can be seen as an asset because of the sensitivity that can be brought to the observation process. The quality of the data gathering and analysis is dependent on the experience of the investigator [56]. However, the process of gathering any data must be concerned with obtaining representative data. The questions circle about whether the observer has heard or understood fully and whether these observations are reported accurately. Considerations include:

- Is this a first person direct report? Otherwise normal common sense about 2nd, 3rd, and 4th hand reports needs to be considered.
- Does the spatial location of the observer provide an adequate vantage point from which to observe, or might it have led to omissions?

- Are the social relationships of the observer free from associations that might induce bias?
- Does the report appear to be self-serving? Does it benefit the experimenter to the extent that it should be questioned?
- Is the report internally consistent? Do the facts within the report support each other?
- Is the report externally consistent? Do the facts in the report agree with other independent reports?

As a result it is important to be explicit about data collection methods, the position of the researcher with respect to the subject matter, analysis processes, and codes. These details make it possible for other researchers to verify results.

In qualitative research it is acknowledged that the researcher's views, research context, and interpretations are an essential part of the qualitative research method as long as they are grounded in the collected data [3]. This does not, however, mean that qualitative evaluations are less trustworthy compared to quantitative research. Auerbach suggests using the concept of 'transferability' rather than 'generalizability' when thinking about the concepts of reliability and validity in qualitative research [3]. It is more important that the theoretical understanding we have gained can also be found in other research situations or systems and can be extended and developed further when applied to other scenarios. This stands in contrast to the concept of generalizability in quantitative research that wants to prove statistically that the results are universally applicable within the population under study.

Sometimes the point has been raised that if results do not generalize how can they be of use when designing software for general use. For example, qualitative methods might be used to obtain a rich description of a particular situation perhaps only observing the processes of two or three people. The results of a study like this may or may not generalize and the study itself provides no proof that they do. What we have is existence proof: that such processes are in use in at least two or three instances. Consider the worst case; that is that this rich description is an outlier that occurs only rarely. For design purposes, outliers are also important and sensitive design for outliers has been often shown to create better designs for all. For example, motion sensors to open doors may have been designed for wheelchairs but actually are useful features for all.

5.3.3 Analyzing Qualitative Data

Qualitative data may be analyzed using qualitative, quantitative, or a combination of both methods. Mixed methods research includes a qualitative phase and a quantitative phase in the overall research study in order to triangulate results from different methods, to complement results from one method with another, or to increase the breadth and range of inquiry by using different methods [28].

Many of the qualitative analysis methods can be grouped as types of thematic analysis, in which analysis starts from observations, then themes are sensed through review of the data, and finally coded [10]. Coding is the process of subdividing and labeling raw data, then reintegrating collected codes to form a theory [70]. Moving from the raw data into themes and a code set may proceed using one of three ap-

proaches: data-driven, motivated from previous research, or theory-driven, each with respectively decreasing levels of sensitivity to the data [10]. In the first style, data-driven, commonly called open coding [14]; themes and a code set are derived directly from the data and nothing else. If the analysis is motivated by previous research, the questions and perhaps codes from the earlier research can be applied to the new data to verify, extend or contrast the previous results. With theory-driven coding one may think using a given theory, such as grounded theory [13], or ethno-methodology [24], as a lens through which to view the data.

In either case the coded data may then be interpreted in more generalized terms. Qualitatively coded data may then be used with quantitative or statistical measures to try and distinguish themes or sampling groups.

5.4 Qualitative Summary

Qualitative studies can be a powerful methodology by which one can capture salient aspects of a problem that may provide useful design and evaluation criteria. Quantitative evaluation is naturally precision-oriented, but a shift from high precision to high fidelity may be made with the addition of qualitative evaluations. In particular, while qualitative evaluations can be used throughout the entire development life cycle in other research areas such as CSCW [41, 52, 64, 73], observational studies have been found to be especially useful for informing design. Yet these techniques are under-used and under-reported in the information visualization literature. Broader approaches to evaluation, different units of analysis and sensitivity to context are important when complex issues such as insight, discovery, confidence and collaboration need to be assessed. In more general terms, we would like to draw attention to qualitative research approaches which may help to address difficult types of evaluation questions. As noted by Isenberg el al. [36], a sign in Albert Einstein's office which read, '*Everything that can be counted does not necessarily count; everything that counts cannot necessarily be counted*' is particularly salient to this discussion in reminding us to include empirical research about important data that can not necessarily be counted.

6 Conclusions

In this paper we have made a two-pronged call: one for more evaluations in general and one for a broader appreciation of the variety of and importance of many different types of empirical methodologies. To achieve this, we as a research community need to both conduct more empirical research and to be more welcoming of this research in our publication venues. As noted in Section 4, even empirical laboratory experiments, as our most known type of empirical methodology, are often difficult to publish. One factor in this is that no empirical method is perfect. That is, there is always a trade-off between generalizability, precision, and realism. An inexperienced reviewer may recommend rejection based on the fact that one of these factors is not present, while realistically at least one will always be compromised. Empirical research is a slow, labour-intensive process in which understanding and insight can develop through time. That said, there are several important factors to consider when publishing empirical research. These include:

- That the empirical methodology was sensitively chosen. The methodology should be a good fit to the research question, the situation and the research goals.
- That the study was conducted with appropriate rigor. All methodologies have their own requirements for rigor and these should be followed. However, while trying to fit the rigor from one methodology onto another is not appropriate, developing hybrid methodologies that better fit a given research situation and benefit from two or more methodologies should be encouraged.
- That sufficient details are published so that the reader can fully understand the processes and if appropriate, reproduce them.
- That the claims should be made appropriately according to the strengths of the chosen methodology. For instance, if a given methodology does not generalize well, then generalizations should not be drawn from the results.

While there is growing recognition in our research community that evaluation information visualization is difficult [55, 57, 67], the recognition of this difficulty has not in itself provided immediate answers of how to approach this problem. Two positive recent trends of note are: one, that more evaluative papers in the form of usability studies have been published [25, 40, 47, 63, 80, 82], and two, that there are several papers that have made a call for more qualitative evaluations and complementary qualitative and quantitative approaches [18, 36, 48, 74].

This paper is intended merely as a pointer to a greater variety of empirical methodologies and encouragement towards their appreciation and even better their active use. There are many more such techniques and these types of techniques are being developed and improved continuously. There are good benefits to be had through active borrowing from ethnographic and sociological research methods, and applying them to our information visualization needs. In this paper we have argued for an increased awareness of empirical research. We have discussed the relationship of empirical research to information visualization and have made a call for a more sensitive application of this type of research [27]. In particular, we encourage thoughtful application of a greater variety of evaluative research methodologies in information visualization.

Acknowledgments. The ideas presented in this paper have evolved out of many discussions with many people. In particular this includes: Christopher Collins, Marian Dörk, Saul Greenberg, Carl Gutwin, Mark S. Hancock, Uta Hinrichs, Petra Isenberg, Stacey Scott, Amy Voida, and Torre Zuk.

References

1. Amar, R.A., Stasko, J.T.: Knowledge Precepts for Design and Evaluation of Information Visualizations. IEEE Transactions on Visualization and Computer Graphics 11(4), 432–442 (2005)
2. Andrews, K.: Evaluating Information Visualisations. In: Proceedings of the 2006 AVI Workshop on BEyond Time and Errors: Novel Evaluation Methods for Information Visualization, pp. 1–5 (2006)
3. Auerbach, C.: Qualitative Data: An Introduction to Coding and Analysis. University Press, New York (2003)

4. Baker, K., Greenberg, S., Gutwin, C.: Empirical Development of a Heuristic Evaluation Methodology for Shared Workspace Groupware. In: Proceedings of the ACM Conference on Computer Supported Cooperative Work, pp. 96–105. ACM Press, New York (2002)
5. Baldonado, M., Woodruff, A., Kuchinsky, A.: Guidelines for Using Multiple Views in Information Visualization. In: Proeedings of the Conference on Advanced Visual Interfaces (AVI), pp. 110–119. ACM Press, New York (2000)
6. Barzun, J., Graff, H.: The Modern Researcher, 3rd edn. Harcourt Brace Jovanvich, New York (1977)
7. BELIV 2006, accessed `http://www.dis.uniroma1.it/~beliv06/` (February 4, 2008)
8. Bertin, J.: Semiology of Graphics (Translation: William J. Berg). University of Wisconsin Press (1983)
9. Bevan, N., Barnum, C., Cockton, G., Nielsen, J., Spool, J., Wixon, W.: The "Magic Number 5": Is It Enough for Web Testing? In: CHI Extended Abstracts, pp. 698–699. ACM Press, New York (2003)
10. Boyatzis, R.: Transforming Qualitative Information: Thematic Analysis and Code Development. Sage Publications, London (1998)
11. Brereton, M., McGarry, B.: An Observational Study of How Objects Support Engineering Design Thinking and Communication: Implications for the Design of Tangible Media. In: Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI'00), pp. 217–224. ACM Press, New York (2000)
12. Chen, C., Czerwinski, M.: Introduction to the Special Issue on Empirical Evaluation of Information Visualizations. International Journal of Human-Computer Studies 53(5), 631–635 (2000)
13. Corbin, J., Strauss, A.: Basics of Qualitative Research: Techniques and Procedures for Developing Grounded Theory, 3rd edn. Sage Publications, Los Angeles (2008)
14. Creswell, J.: Qualitative Inquiry and Research Design: Choosing Among Five Traditions. Sage Publications, London (1998)
15. Dix, A., Finlay, J., Abowd, G., Beale, R.: Human Computer Interaction, 2nd edn. Prentice-Hall, Englewood Cliffs (1998)
16. Dumais, S., Cutrell, E., Chen, H.: Optimizing Search by Showing Results In Context. In: Proc. CHI'01, pp. 277–284. ACM Press, New York (2001)
17. Eberts, R.E.: User Interface Design. Prentice-Hall, Englewood Cliffs (1994)
18. Ellis, E., Dix, A.: An Explorative Analysis of User Evaluation Studies in Information Visualization. In: Proceedings of the Workshop on Beyond Time and Errors: Novel Evaluation Methods for Information Visualization, BELIV (2006)
19. Ericsson, K., Simon, H.: Verbal Reports as Data. Psychological Review 87(3), 215–251 (1980)
20. Ericsson, K., Simon, H.: Verbal Reports on Thinking. In: Faerch, C., Kasper, G. (eds.) Introspection in Second Language Research, pp. 24–54. Multilingual Matters, Clevedon, Avon (1987)
21. Ericsson, K., Simon, H.: Protocol Analysis: Verbal Reports as Data, 2nd edn. MIT Press, Boston (1993)
22. Fall, J., Fall, A.: SELES: A Spatially Explicit Landscape Event Simulator. In: Proceedings of GIS and Environmental Modeling, pp. 104–112. National Center for Geographic Information and Analysis (1996)
23. Forlines, C., Shen, C., Wigdor, D., Balakrishnan, R.: Exploring the effects of group size and display configuration on visual search. In: Computer Supported Cooperative Work 2006 Conference Proceedings, pp. 11–20 (2006)
24. Garfinkel, H.: Studies in Ethnomethodology. Polity Press, Cambridge (1967)

25. Gonzalez, V., Kobsa, A.: A Workplace Study of the Adoption of Information Visualization systems. In: Proceedings of the International Conference on Knowledge Management, pp. 92–102 (2003)
26. Gorard, S.: Combining Methods in Educational Research. McGraw-Hill, New York (2004)
27. Greenberg, S., Buxton, B.: Usability Evaluation Considered Harmful (Some of the Time). In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (2008)
28. Greene, J., Caracelli, V., Graham, W.: Toward a Conceptual Framework for Mixed-Method Evaluation Design. Educational Evaluation and Policy Analysis 11(3), 255–274 (1989)
29. Gutwin, C., Greenberg, S.: The Mechanics of Collaboration: Developing Low Cost Usability Evaluation Methods for Shared Workspaces. In: Proceedings WETICE, pp. 98–103. IEEE Computer Society Press, Los Alamitos (2000)
30. Healey, C.G.: On the Use of Perceptual Cues and Data Mining for Effective Visualization of Scientific Datasets. In: Proceedings of Graphics Interface, pp. 177–184 (1998)
31. Heer, J., Viegas, F., Wattenberg, M.: Voyagers and Voyeurs: Supporting Asynchronous Collaborative Information Visualization. In: Proceedings of the Conference on Human Factors in Computing Systems (CHI'07), pp. 1029–1038. ACM Press, New York (2007)
32. Holtzblatt, K., Beyer, H.: Contextual Design: Defining Customer-Centered Systems. Morgan Kaufmann, San Francisco (1998)
33. Huck, S.W.: Reading Statistics and Research, 4th edn. Pearson Education Inc., Boston (2004)
34. Interrante, V.: Illustrating Surface Shape in Volume Data via Principal Direction-Driven 3D Line Integral Convolution. Computer Graphics, Annual Conference Series, pp. 109–116 (1997)
35. Isenberg, P., Carpendale, S.: Interactive Tree Comparison for Co-located Collaborative Information Visualization. IEEE Transactions on Visualization and Computer Graphics 12(5) (2007)
36. Isenberg, P., Tang, A., Carpendale, S.: An Exploratory Study of Visual Information Analysis. In: Proceedings of the Conference on Human Factors in Computing Systems (CHI'08), ACM Press, New York (to appear, 2008)
37. Kay, J., Reiger, H., Boyle, M., Francis, G.: An Ecosystem Approach for Sustainability: Addressing the Challenge of Complexity. Futures 31(7), 721–742 (1999)
38. Kim, S., Hagh-Shenas, H., Interrante, V.: Conveying Shape with Texture: Experimental Investigations of Texture's Effects on Shape Categorization Judgments. IEEE Transactions on Visualization and Computer Graphics 10(4), 471–483 (2004)
39. Kleffner, D.A., Ramachandran, V.S.: On the Perception of Shape from Shading. Perception and Psychophysics 52(1), 18–36 (1992)
40. Kobsa, A.: User Experiments with Tree Visualization Systems. In: Proceedings of the IEEE Symposium on Information Visualization, pp. 9–26 (2004)
41. Kruger, R., Carpendale, S., Scott, S.D., Greenberg, S.: Roles of Orientation in Tabletop Collaboration: Comprehension, Coordination and Communication. Journal of Computer Supported Collaborative Work 13(5–6), 501–537 (2004)
42. Lewin, C. (ed.): Research Methods in the Social Sciences. Sage Publications, London (2004)
43. Lewis, C., Rieman, J.: Task-Centered User Interface Design: A Practical Introduction (1993)
44. Likert, R.: A Technique for the Measurement of Attitudes. Archives of Psychology 140, 1–55 (1932)
45. Lofland, J., Lofland, L.: Analyzing Social Settings: A Guide to Qualitative Observation and Analysis. Wadsworth Publishing Company, CA, USA (1995)
46. Mankoff, J., Dey, A., Hsieh, G., Kientz, J., Lederer, S., Ames, A.: Heuristic Evaluation of Ambient Displays. In: Proceedings of CHI '03, pp. 169–176. ACM Press, New York (2003)
47. Mark, G., Kobsa, A., Gonzalez, V.: Do Four Eyes See Better Than Two? Collaborative Versus Individual Discovery in Data Visualization Systems. In: Proceedings of the IEEE Conference on Information Visualization (IV'02), July 2002, pp. 249–255. IEEE Press, Los Alamitos (2002)

48. Mazza, R., Berre, A.: Focus Group Methodology for Evaluating Information Visualization Techniques and Tools. In: Proceedings of the International Conference on Information Visualization IV (2007)
49. McCarthy, D.: Normal Science and Post-Normal Inquiry: A Context for Methodology (2004)
50. McGrath, J.: Methodology Matters: Doing Research in the Social and Behavioural Sciences. In: Readings in Human-Computer Interaction: Toward the Year 2000, Morgan Kaufmann, San Francisco (1995)
51. Moggridge, B.: Design Interactions. MIT Press, Cambridge (2006)
52. Morris, M.R., Ryall, K., Shen, C., Forlines, C., Vernier, F.: Beyond "Social Protocols": Multi-User Coordination Policies for Co-located Groupware. In: Proceedings of the ACM Conference on Computer-Supported Cooperative Work (CSCW, Chicago, IL, USA), CHI Letters, November 6-10, 2004, pp. 262–265. ACM Press, New York (2004)
53. Morse, E., Lewis, M., Olsen, K.: Evaluating Visualizations: Using a Taxonomic Guide. Int. J. Human-Computer Studies 53, 637–662 (2000)
54. Nielsen, J., Mack, R.: Usability Inspection Methods. John Wiley & Sons, Chichester (1994)
55. North, C.: Toward Measuring Visualization Insight. IEEE Computer Graphics and Applications 26(3), 6–9 (2006)
56. Patton, M.Q.: Qualitative Research and Evaluation Methods, 3rd edn. Sage Publications, London (2001)
57. Plaisant, C.: The Challenge of Information Visualization Evaluation. In: Proceedings of the Working Conference on Advanced Visual Interfaces, pp. 109–116 (2004)
58. Purchase, H.C., Hoggan, E., Görg, C.: How Important Is the "Mental Map"? – An Empirical Investigation of a Dynamic Graph Layout Algorithm. In: Kaufmann, M., Wagner, D. (eds.) GD 2006. LNCS, vol. 4372, pp. 184–195. Springer, Heidelberg (2007)
59. Purchase, H.C.: Effective Information Visualisation: A Study of Graph Drawing Aesthetics and Algorithms. Interacting with Computers 13(2), 477–506 (2000)
60. Purchase, H.C.: Performance of Layout Algorithms: Comprehension, Not Computation. Journal of Visual Languages and Computing 9, 647–657 (1998)
61. Brandenburg, F.J. (ed.): GD 1995. LNCS, vol. 1027. Springer, Heidelberg (1996)
62. Reilly, D., Inkpen, K.: White Rooms and Morphing Don't Mix: Setting and the Evaluation of Visualization Techniques. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 111–120 (2007)
63. Saraiya, P., North, C., Duca, K.: An Insight-Based Methodology for Evaluating Bioinformatics Visualizations. IEEE Transactions on Visualization and Computer Graphics 11(4), 443–456 (2005)
64. Scott, S.D., Carpendale, S., Inkpen, K.: Territoriality in Collaborative Tabletop Workspaces. In: Proceedings of the ACM Conference on Computer-Supported Cooperative Work (CSCW, Chicago, IL, USA), CHI Letters, November 6-10, 2004, pp. 294–303. ACM Press, New York (2004)
65. Seidman, I.: Interviewing as Qualitative Research: A Guide for Researchers in Education and the Social Sciences. Teachers' College Press, New York (1998)
66. Shneiderman, B.: The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations. In: Proceedings of the IEEE Symposium on Visual Languages, pp. 336–343. IEEE Computer Society Press, Los Alamitos (1996)
67. Shneiderman, B., Plaisant, C.: Strategies for Evaluating Information Visualization Tools: Multi-Dimensional In-Depth Long-Term Case Studies. In: Proceedings of the Workshop on BEyond Time and Errors: Novel Evaluation Methods for Information Visualization, BELIV (2006)
68. Spence, R.: Information Visualization, 2nd edn. Addison-Wesley, Reading (2007)
69. Spool, J., Schroeder, W.: Testing Web Sites: Five Users is Nowhere Near Enough. In: CHI '01 Extended Abstracts, pp. 285–286. ACM Press, New York (2001)

70. Strauss, A.L., Corbin, J.: Basics of Qualitative Research: Techniques and Procedures for Developing Grounded Theory. Sage Publications, London (1998)
71. Tang, A., Tory, M., Po, B., Neumann, P., Carpendale, S.: Collaborative Coupling over Tabletop Displays. In: Proceedings of the Conference on Human Factors in Computing Systems (CHI'06), pp. 1181–1290. ACM Press, New York (2006)
72. Tang, A., Carpendale, S.: An observational study on information flow during nurses' shift change. In: Proc. of the ACM Conf. on Human Factors in Computing Systems (CHI), pp. 219–228. ACM Press, New York (2007)
73. Tang, J.C.: Findings from observational studies of collaborative work. International Journal of Man-Machine Studies 34(2), 143–160 (1991)
74. Tory, M., Möller, T.: Evaluating Visualizations: Do Expert Reviews Work. IEEE Computer Graphics and Applications 25(5), 8–11 (2005)
75. Tufte, E.: The Visual Display of Quantitative Information. Graphics Press, Cheshire (1986)
76. Tufte, E.: Envisioning Information. Graphics Press, Cheshire (1990)
77. Tufte, E.: Visual Explanations. Images and Quantities, Evidence and Narrative. Graphics Press, Cheshire (1997)
78. Viegas, F.B., Wattenberg, M., van Ham, F., Kriss, J., McKeon, M.: Many Eyes: A Site for Visualization at Internet Scale. IEEE Transactions on Visualization and Computer Graphics (Proceedings Visualization / Information Visualization 2007) 12(5), 1121–1128 (2007)
79. Ware, C.: Information Visualization: Perception for Design, 2nd edn. Morgan Kaufmann, San Francisco (2004)
80. Wigdor, D., Shen, C., Forlines, C., Balakrishnan, R.: Perception of Elementary Graphical Elements in Tabletop and Multi-surface Environments. In: Proceedings of the Conference on Human Factors in Computing Systems (CHI'07), pp. 473–482. ACM Press, New York (2007)
81. Willett, W., Heer, J., Agrawala, M.: Scented Widgets: Improving Navigation Cues with Embedded Visualizations. In: INFOVIS 2007. IEEE Symposium on Information Visualization (2007)
82. Yost, B., North, C.: The Perceptual Scalability of Visualization. IEEE Transactions on Visualization and Computer Graphics 12(5), 837–844 (2006)
83. Zuk, T.: Uncertainty Visualizations. PhD thesis. Department of Compute Science, University of Calgary (2007)
84. Zuk, T., Carpendale, S.: Theoretical Analysis of Uncertainty Visualizations. In: Proceedings of SPIE Conference Electronic Imaging, Vol. 6060: Visualization and Data Analysis (2006)
85. Zuk, T., Schlesier, L., Neumann, P., Hancock, M.S., Carpendale, S.: Heuristics for Information Visualization Evaluation. In: Proceedings of the Workshop BEyond Time and Errors: Novel Evaluation Methods for Information Visualization (BELIV 2006), held in conjunction with the Working Conference on Advanced Visual Interfaces (AVI 2006), ACM Press, New York (2006)