Increasing the Utility of Quantitative Empirical Studies for Meta-analysis

Heidi Lam
University of British Columbia
hllam@cs.ubc.ca

Tamara Munzner
University of British Columbia
tmm@cs.ubc.ca

ABSTRACT

Despite the long history and consistent use of quantitative empirical methods to evaluate information visualization techniques and systems, our understanding of interface use remains incomplete. While there are inherent limitations to the method, such as the choice of task and data, we believe the utility of study results can be enhanced if they were amenable to meta-analysis. Based on our experience in extracting design guidelines from existing quantitative studies, we recommend improvements to both study design and reporting to promote meta-analysis: (1) Use comparable interfaces in terms of visual elements, information content and amount displayed, levels of data organization displayed, and interaction complexity; (2) Capture usage patterns in addition to overall performance measurements to better identify design tradeoffs; (3) Isolate and study interface factors instead of overall interface performance; and (4) Report more study details, either within the publications, or as supplementary materials.

Categories and Subject Descriptors

H.5 [**Information Interfaces and Presentation (e.g., HCI)**]: Miscellaneous

General Terms

information visualization evaluation, meta-analysis

1. INTRODUCTION

As the fields of information visualization and human computer interaction mature, both communities put more emphasis on interface evaluation. Indeed, a 2007 study finds that over 90 % of the papers that were accepted to the ACM SIGCHI Conference on Human Factors in Computing Systems (CHI) in 2006 included formal evaluation, and only about half of the papers did in 1983 [1].

The type of evaluation has also changed. In the same study, Barkhuus and Rode found a steady increase in qualitative empirical evaluations, from 7% in 1983 to 14% in 2006 [1]. Indeed, researchers have expressed discontent in quantitative empirical evaluation methods, and have suggested looking into more qualitative, exploratory, and long-term evaluations (e.g., [22, 31]).

For system evaluation, system use is heavily influenced by context of use, such as working environment, user characteristics (domain expertise, engagement/incentive, individual differences), task, and data. Given the many potentially important factors involved in interface use, and some of them surprising and only identified after extensive piloting (e.g., how environments affect visualization use [25]), contextual evaluation at the workplace is the most ecologically valid approach for system evaluation. For the same reason, longer term studies provide more complete pictures of system use, as true usage pattern can only emerge over time, as found by González and Kobsa in their evaluations of InfoZoom [12, 11]. Also, using the user's own data and task, rather than synthetic ones, provides a more realistic study setting and ensures participant motivation and engagement, as found in Saraiya *et al.*'s studies on analyzing micro-array data with various visualizations [27, 28]

Despite these advantages, it is often difficult to generalize results obtained from in-depth exploratory studies, since they tend to show specific system effectiveness under specific settings with a handful of participants. We therefore believe that quantitative empirical studies still have their place in visualization evaluation, as such studies often require abstraction of tasks and data, and therefore tend to produce more generalizable results by looking at specific aspects of the system with larger numbers of participants.

Our own struggle in evaluating visualizations using quantitative empirical methods led us to heartily agree on all reported and inherent constraints in the method, such as the difficulties in creating meaningful yet testable study tasks and the difficulties in ensuring sufficient training for effective interface use [22]. We also acknowledge that many important questions in information visualization remain unanswered, despite persistent evaluation efforts using predominantly quantitative empirical approaches. For example, even though focus+context visualization techniques have been around for over 20 years, we do not know when, how, or even if they are useful [9]. Our lack of knowledge is not because of lack of effort: many studies have looked at systems using the focus+context technique, or directly studied the technique itself. We partially attribute this lack of clarity to the design of most studies as head-to-head comparisons between novel systems or techniques with state-of-the-art

BELIV 2008, April 5, 2008, Florence, Italy.

counterparts, with a focus on overall performance. While head-to-head comparisons are useful tests for novel interface effectiveness, we argue for more studies designed to understand interface use by identifying the factors at play and the design tradeoffs to better guide further visualization developments, since overall performance results can be sensitive to specific interface implementation, study tasks, or study data details.

We believe quantitative empirical evaluations can be more useful if general conclusions can be extracted from a large number of study results, for example, via meta-analysis. However, diverse study designs and reporting styles in existing publications make meta-analysis extremely difficult, as illustrated by Chen and Yu's 2000 meta-analysis on visualization systems where only six out of the 35 coded studies were eventually included in the meta-analysis [7]. We believe this difficulty can be alleviated by a few modifications in our current evaluation practices.

We compile our list of methodology modifications based on our experience in extracting design guidelines for multiple visual information resolution (VIR) interface from 19 quantitative empirical user studies [18]. Given the diversity in our reviewed studies in their implementations of the various multiple-VIR techniques, study tasks, and data, and in some cases, experimental design and measurements, we did not attempt to compare between studies. Instead, we performed pairwise interface comparisons within each study to abstract generalizable usage patterns based on task, data characteristics, and interface differences. By doing so, we bypassed many of the difficulties encountered in typical meta-analysis, such as requiring straight adherence to selection criteria which may exclude a majority of available studies (e.g., in [7]). Despite being able to use most of the study results in our analysis, we encountered difficulties in interpreting selective study results and had to exclude them from our analysis, since we focused on teasing out factors that affect visualization use. The four main scenarios that led to result exclusion are:

1. Study interfaces are not comparable at the individual factor level, such as visual elements, information content and amount displayed, level of organization displayed, or interaction complexity;

2. Measurements are not sensitive enough to capture usage patterns, which is needed to understand the factors at play in visualization use;

3. Studies investigate multiple interface-use factors, making it difficult to isolate the effect of each;

4. Studies did not report sufficient details for our analysis, since we wish to extract effects of selective design factors in interface use, instead of overall system or technique effectiveness.

We therefore argue that by using comparable study interfaces, capturing usage patterns in addition to overall performance measures, isolating interface-use factors, and by reporting more study details, we can increase consistency among studies and increase their utility, since study results would be more amenable to meta-analysis.

In the remainder of this paper, we illustrate each of our suggestions based on our experience [18][1]. In our discussion section (Section 6), we reflect on challenges in adapting our suggestions and propose possible solutions such as emphasizing the need for follow-up studies. We also postulate potential benefits of our suggestions in advancing information visualization evaluations.

2. USE COMPARABLE INTERFACES

In order to understand factors influencing study interface use, studies should identify possible factors at play, and if possible, use comparable factors between interfaces. For visual design, some factors include the interfaces' basic visual elements such as the number of views and the use of image distortion, the amount and type of information displayed, and the number of levels in the displayed data. For interaction, study designers should consider the required number of input devices, the types of action required, and the number of displays on which the action is applied.

2.1 Basic visual elements

While it is understandable that the study interfaces may be dramatically different in appearance, they should be comparable in their basic visual elements whenever possible to allow for direct comparison. For example, in Baudisch *et al.*'s 2004 study on visual searches on webpages, they included two interfaces that show web documents at two levels of detail simultaneously. The Overview interface had a scrollable detail page and an overview that showed the entire webpage by compressing all elements equally. The Fisheye interface was a non-scrollable browser that showed the entire webpage by differentially compressing pertinent versus peripheral content in order to keep the pertinent text readable [3]. On the surface, the two interfaces are ideal candidates for studying the effects of spatial arrangement of the focus/detail and context/overview components: in the Overview interface, the two components were arranged as separate views; in the Fisheye interface, they were embedded into a single view.

However, there is another factor at play that affected performance results. Since their interfaces displayed readable words pertinent to their study tasks as highlighted popouts, the spatial association between the original web documents and these popout words becomes important. Unfortunately, association by row is more difficult in their Fisheye interface than by column, as their focus+context implementation selectively distorts in the vertical dimension. On the other hand, their Overview implementation proportionally reduces both vertical (row) and horizontal (column) dimensions. Their study results reflect the interfaces' ability to associate popouts with document rows and columns: their Fisheye interface better supported a task that does not require row-specific information (the *Product Choice* task), but not for row-dependent tasks (e.g., the *Co-occurrence* task). The Overview interface results show opposite trends.

Since Baudisch *et al.*'s study aimed to evaluate the overall effectiveness of their novel Fisheye interface relative to two existing techniques, both the visual components' spatial arrangement and the row-column association with the highlighted popouts are part of their interface design and should be evaluated together. However, when we tried to tease out

[1] Interfaces of our surveyed studies are listed at `http://www.cs.ubc.ca/~hllam/res_ss_interfaces.htm`

the effect of spatial arrangement to extract general design guidelines, we failed to isolate the effect and therefore could not include their study results in our analysis.

We encountered similar problems in analyzing Bederson *et al.*'s 2004 study on PDA-size calendar use [5]. Their study looked at two interfaces: the Pocket PC calendar that provided a single level of detail per view (day, week, month, or year), and the DateLens interface that used a Table Lens-like distortion technique to show multiple levels of details simultaneously. Again, the study seemed to compare the effects of providing separate views one at a time, or embedding them in a single view.

Their study looked at a variety of calendar tasks that involved searching for appointments, navigation and counting scheduled events, and scheduling given constraints. While their study did not find an overall time effect between the two interfaces, the researchers found a task effect and thus divided the tasks into *simple* and *complex* tasks based on task-completion time. The study concluded that the DateLens trials were faster in *complex* tasks, while the Pocket PC trials were faster in *simple* tasks.

On closer inspection, we realized that while their DateLens interface provided a day, week, month, and year view, it also provided a three-month and a six-month view, with the three-month view being the default in the study. On the other hand, the Pocket PC interface did not seem to have a corresponding three-month overview. Since three of their six *complex* tasks (tasks 5, 10, 11) required scheduling and counting events within three-month periods, we could not determine if the benefits of the DateLens interface in these tasks came from providing a three-month overview, or from providing multiple levels of details in the same view. Again, our need to understand performance contribution from individual factors forced us to exclude these results from our analysis.

2.2 Information content

When the information displayed on the different interfaces is different, the interfaces may be used for different purposes, making direct comparisons difficult. One example is Hornbæk *et al.*'s study on online document reading [14, 15].

Their study looked at two interfaces that provided multiple levels of data detail simultaneously. The overview in their Overview+Detail interface showed document header and subheaders and acted as a table of contents. Their Fisheye interface showed context based on a degree-of-interest algorithm, thus the content was dynamic based on the focal point of the document. Not surprisingly, the participants used the two interfaces differently. Reading patterns indicated that when using the Fisheye interface, participants spent more time in the initial orientation mode, but less time in the linear read-through mode, suggesting that the Fisheye interface shortened navigation time by supporting an overview-oriented reading style. In contrast, reading patterns in the Overview+Detail interface was found to be less predictable and "shaped by situation-dependent inspiration and associations", and "the overview pane grabs subjects' attention, and thereby leads them to explorations that strictly speaking are unnecessary" (p.144), probably because display was similar to a table of contents. Study results reflected the differing information content displayed in these overview/context components. Compared to the Fisheye interface, participants who used the Overview+Detail interface produced better results in the essay tasks at the expense of time, and the study failed to find differences between the two interfaces for the question-answering tasks. While the different information content is arguably part of the interface design, we could not incorporate results from this study in our analysis as we could not separate out visual spatial effects from those of displaying different kinds of information in the overview/context.

2.3 Levels of display

The total amount of information and levels of detail encoded by the interface is also important, as extra information or levels may be detrimental to performance. One example is Plumlee and Ware's study on visual memory in zooming and multiple windows [24]. Their Zooming interface had a continuous-zoom mechanism that shows intermediate levels of detail that did not seem to be present in the Multiple Windows interface, which seemed to have only two levels based on the authors' descriptions. Their study task required the participants to match a complex clusters of 3D objects. To do so, the participants needed to first locate clusters at the low-zoom level, and match cluster components at the high-zoom level. Intermediate-zoom levels did not seem to carry information required by the tasks as the clusters themselves were not visible given the textured backgrounds.

Plumlee and Ware stated that the participants needed 1.5 seconds to go through a magnification change of at least 30 times between the lowest and highest zoom levels. During this time, the participants needed to keep track of the components in various objects in the task in their short-term memory. We wondered if having the extra levels of details in their Zooming interface unnecessarily degraded the participants' visual memories and made the interface less usable. This extra cognitive load may explain the relatively small number of items the participants could handle before the opponent Multiple Windows interface became more appropriate for the task, in contrast to the results of a 2005 study on graph visualization by Saraiya *et al.* [26]. Saraiya *et al.*'s Single-Attribute zooming interface supported better performance than their Multiple-Attribute detailed interface even when the task involved a 50-node graph, each node with 10 time points. Due to the differing levels of data displayed in the two study interfaces, we excluded Plumlee and Ware's 2006 study from our analysis to understand the conditions in which simultaneous display of multiple data levels is beneficial.

Similarly, Baudisch *et al.* studied static visual path-finding tasks and dynamic obstacle-avoidance task using three interfaces each providing multiple levels of details [2]. Their z+p, or zoom and pan, interface and their o+d, or overview plus detail, interface seemed to support more levels of detail than their f+c, or focus plus context, interface, which had two levels only. Their f+c trials were faster than the z+p and the o+d trials for the static visual path finding tasks, and were more accurate in the dynamic obstacle-avoidance task. While the special hardware setup in their f+c interface undoubtedly contributed to the superior performance of their participants when using the interface, we wondered if the extra resolutions may have distracted the participants in the other two interface trials, even though we did include this study in our analysis of simultaneous displays of multiple levels of detail as we believed the difference in the number of display level was small.

2.4 Levels in data

Since researchers have argued that the interface should only display a different data resolution if it is meaningful to the task at hand (e.g., [9, 18]), the number of organization levels in the data displayed is an important consideration in studying interfaces that show multiple data levels. For example, in Hornbæk *et al.*'s study on map navigation, there were surprisingly large differences in usability and navigation patterns between the two study maps, despite being similar in terms of the number of objects, area occupied by the geographical state object, and information density [13]. The maps differ by the number of levels of organization: the Washington map had three levels of county, city, and landmark, while the Montana map was single-leveled. Perhaps for this reason, the study failed to find differences in participant performance when using the two study interfaces with the Montana map, but their participants were faster in the navigation task and more accurate in the memory tasks using just the zoomable interface without an overview with the Washington map. We took advantage of this unintended data-level difference to study how interfaces with multiple levels of display data support single-leveled data. These fortuitous opportunities for re-analysis were rare.

2.5 Interaction complexity

In some cases, interaction style may be a factor in the study, and in others, having different interaction complexities may not be avoidable in the different study interfaces. Nonetheless, interaction complexity differences make comparison difficult, as seen in Hornbæk and Frokjær's study on fisheye menus [16].

Hornbæk and Frokjær's intention was to study the visual design and use of fisheye menus. They had four interfaces: a traditional cascading menu, the Fisheye menu as described by Bederson [4], the Overview menu, and the Multi-focus menu. The Overview menu and the Multi-focus menu are both based on the Fisheye menu, and all three implement the focus-lock interaction to aid menu-item selection. The Overview menu did not implement distortion, and showed a portion of the menu items based on mouse position along the menu, showing the field-of-view in the overview. The Multi-focus menu shows important menu items in readable fonts, and did not have an index of letters as in the fisheye or the overview menu.

Surprisingly, their cascading-menu interface outperformed all other interfaces. The author suggested that one possible reason is due to the relatively simple navigation in the cascading-menu interface: their participants encountered obvious and severe difficulties in using the focus-lock mode in the other interfaces. While the authors successfully identify a usability problem in the fisheye menu, we could not conclude if the visual designs of the other three interfaces, which showed menu data at multiple levels simultaneously, were truly inferior to the cascading-menu interface, which showed data one level at a time.

3. CAPTURE USAGE PATTERNS

In most studies, the main measurements are performance time, accuracy, and subjective preferences. While these measurements provide valuable information about overall interface effectiveness, efficiency, and user acceptance, they may not be sensitive enough to illuminate the factors involved in interface use and to tease out design tradeoffs, especially when the study failed to find differences between the interfaces based on overall results.

In our analysis to extract design guidelines for interfaces that display multiple levels of detail, most reported experimenter observations on participant strategy and comments to interpret performance results. However, only five of the 19 studies reported usage patterns, constructed either based on non-intrusively collected interactivity recordings such as eye-tracking records [21, 16] or navigation action logs [13, 15, 17].

We found these five studies were most useful in our analysis to extract design guidelines for interfaces that display multiple levels of data organization. For example, Hornbæk *et al.*'s study on online document reading used progression maps to investigate reading patterns [15]. Progression maps showed the part of the document that was visible to the participants during the reading process. The authors interpreted their performance time and interface effectiveness results using reading patterns derived from the progression maps, and provided a richer understanding of how the study interfaces were used. For example, their reading pattern explains the longer performance time in the question-answering task trials using the Overview+Detail interface: "further explorations were often initiated by clicking on the overview pane", and "further exploration [of the displayed documents] happen[ed] because of the visual appearance of the overview and because of the navigation possibility afforded by the ability to click the overview pane". They therefore concluded that "the overview pane grabs subjects' attention, and thereby leads them to explorations that strictly speaking are unnecessary" (p. 144).

In another of their studies, Hornbæk and Hertzum looked at fisheye menu use [16]. Despite not finding performance differences between their Fisheye menu interface and two other visual variants, the Overview and the Multi-focus interfaces, eye-tracking results showed interesting insights into how the interfaces were used: their participants used the context regions more frequently in the Multi-focus interface trials, possibly due to the readable information included in the context regions. The researchers were therefore able to conclude, based on usage pattern, that designs should make "the context region of the [fisheye menu] interfaces more informative by including more readable or otherwise useful information" (p. 28, [16]).

4. ISOLATE INTERFACE FACTORS

Information visualization systems are complex interfaces that typically involve visual encoding and interaction, and for some implementations, view coordination and image transformation. As discussed in Section 2, identifying such factors to ensure comparable test interfaces are probably sufficient when the study aims to evaluate overall system effectiveness. However, studying overall effects may obscure contributions from each factor, a difficulty we encountered during our analysis to draw design guidelines based on these factors.

That was the case when we looked at Gutwin and Skopik's study on 2D-steering, where at least three factors were at play [10]. Their study looked at five overview+detail and two fisheye interfaces. In addition to the different spatial arrangements of the different levels of details in their interfaces, there were also different effective steering path widths and lengths and different interaction styles.

Looking at steering paths of the five interfaces, only one of the overview+detail interfaces, the Panning view, had an increased travel length at higher magnifications. All other interfaces had constant *control/display* ratios for all path magnification levels. One of the overview+detail interfaces, the Radar view, had effectively identical steering paths over all magnifications, as the participants interacted with the miniature constant-sized overview instead of the magnified detailed view. Even though the *control/display* ratios for the Radar view, as for the fisheye interfaces, was constant over all magnifications, the value was 1:6 for the Radar-view interface, but 1:1 for the fisheye interfaces. Not surprisingly, the radar-view trials had similar performance times for tested magnification levels, and were consistently slower and less accurate than their fisheye counterparts.

The interfaces also had different interaction styles. The Panning-view interfaces required two mouse actions on two displays, and the fisheye interfaces only required mouse move on a single display, but with the complications of view distortion. Since the aim of our analysis was to identify factors and quantify their contributions to performance results, we could not fully interpret their study results. The poor performance of the Panning-view interface trials may be due to the view-coordination costs incurred in its separate spatial arrangements of the different levels of data, the mouse-action coordination costs incurred by the complex interaction, or simply due to the increased increased travel lengths at higher magnifications.

Another type of difficulty we encountered in our analysis was to tease out the usability factors involved in the focus+context techniques. While showing all data as a single view in context may provide benefits, the techniques often require more complex interactions, and image distortion has been shown to incur disorientation [6] and visual memory costs [19]. Ideally, we would like to be able to study each of these factors in isolation. However, we were only marginally successful in teasing out the effects of distortion, as our study set had focus+context interfaces that implemented different types and degrees of distortion. For example, Baudisch *et al.*'s 2002 study implemented their focus+context interface with a hardware approach, using different pixel density in their displays to recreate the two regions, thus avoiding the need for distortion in their interface [2]. Their study found performance benefits in all their tasks using their focus+context display. In contrast, studies that implemented drastic and elastic distortion techniques reported null or mixed results, along with observed usability problems, for example the Rubber Sheet Navigation [29] in Nekrasovski *et al.*'s study [20], the Hyperbolic Tree browser in Plaisant *et al.* and Pirolli *et al.*'s studies [23, 21], and the fisheye projections in Schafer's study [30]. Despite this insight, our distortion classification is still rough, both in terms of classifying distortion types and performance effects.

5. REPORT MORE STUDY DETAILS

One of the frustrations we had while analyzing our study set stemmed from the lack of details in study reporting. Indeed, Chen and Yu encountered similar problems in their meta-analysis [7]. Since their meta-analysis synthesized significance levels and effect sizes, they had to exclude many more studies than in our less quantitative analysis. Based on their experience, Chen and Yu recommended four standardizations in empirical studies: testing information, task taxonomy (for visual information retrieval, data exploration, and data analysis tasks), cognitive ability tests, and levels of details in reporting statistical results. They also asked for better clarity in descriptions of visual-spatial properties of information visualization systems, and more focus on task-feature binding in studies. The researchers concluded that "it is crucial to conduct empirical studies concerning information visualization systematically within a comparable reference framework" (p. 864).

In addition to supporting Chen and Yu's recommendations [7], we have two further recommendations. We advocate providing full task instructions. We also advocate documenting the interface interactions with video, or even making the interface prototype software and trial experiments available for download. Allowing others to see or experience the exact instructions and interface behavior seen by study participants would help reproducability and help clarify study procedures for later meta-analysis.

Although nine of the 19 study papers we looked at in our analysis provided detailed descriptions of the study tasks, only five provided actual task instructions. Since interface use can be severely affected by task nature, it is difficult to analyze study results when the publications did not provide the written task instructions given to participants before the trials, and any verbal hints given during the trials. For example, in our analysis, we needed to ascertain the factors that lead to successful use of simultaneous display of multiple levels of data detail. One possibility is when the task instruction provides clues that spans multiple data levels. Since we attempted to reinterpret the results based on different criteria, we encountered difficulties when the study did not provide enough task instructions for us to judge if the the task provided multiple-level clues, for example in Plaisant's SpaceTree study [23].

Even providing detailed task instructions may still be inadequate in some cases. For example, in Pirolli's preliminary task analysis study, their tasks were measured for information scent [21]. Even though the authors did provide a list of tasks, they did not cross-match the list with the information scent scores, making it difficult for us to later associate task nature, information scent score, and study results. We therefore assumed the instructions of high information scent tasks provided useful clues at multiple levels of the tree.

For studies where interaction plays a pivotal role in study results, text descriptions of the interaction, no matter how detailed and carefully constructed, seems inadequate. One example is Hornbæk and Hertzum's 2007 study on the use of fisheye menu, where the focus-lock interaction is one of the major usability problems found in the three fisheye-menu interfaces [16]. Despite the authors' well-constructed descriptions, we did not fully understand the interaction until we tried the online fisheye menu prototype kindly provided by Bederson[2].

We understand the strict page limits for research papers in many venues has required authors to make draconian choices in the amount of detail reported. Even without the page limits, such choices should be guided by the study goals and paper emphasis to ensure readability, as it is impossible to predict how study results may be used in future analysis. We therefore recommend that researchers provide study details as electronic supplementary materials in publication venues that support archival availability of such materials, or as

[2] `http://www.cs.umd.edu/hcil/fisheyemenu/`

information posted on laboratory websites.

6. DISCUSSION AND CONCLUSION

Based on our experience in extracting design guidelines and in carrying out user studies to evaluate visualization interfaces, we present a list of methodology modifications to enhance the utility of quantitative empirical studies: (1) use comparable study interfaces; (2) capture usage patterns in addition to overall performance measures; (3) isolate interface-use factors, and (4) report more study details.

While we understand the area of interface evaluation is an active area of research with many substantial challenges, we believe we can improve our evaluations today by adopting modifications suggested in this paper, for example, record and report detailed observations in study to better capture usage patterns (Section 3) and report more study details to allow re-use of study results (Section 5).

Our two other suggestions, (1) use comparable study interfaces (Section 2) and (4) isolate interface-use factors (Section 4), can be difficult to implement. One challenge is to identify study elements prior to the study to ensure comparability. For example, in Hornbæk's *et al.*'s study on map navigation, the researchers did try to use comparable maps [14, 15]. Differences between the two study maps were only apparent after the study.

Another difficulty in adhering to these suggestions may be due to a conflict of evaluation goals: the goals of the original designs were to compare between systems at the overall-performance level, while our goal was to extract the effects of interface-use factors in systems. It is therefore difficult to modify original study designs without changing these goals, since the systems themselves are complex and are frequently incomparable at the interface-factor level.

In both cases, we believe follow-up studies are needed. Follow-up studies, either performed by the original researchers or by third parties, can take advantage of the knowledge gained in original studies or system-level studies, such as correcting mistakes made in original studies as in using different levels in data (Section 2.4) or different levels in interfaces (Section 2.3). System-level evaluations can be used as a vehicle to identify factors, perhaps by detailed observations of how participants interact with the systems. These factors can then be studied in more detail and in isolation in subsequent studies. For example, Baudisch *et al.*'s Fishnet interface study identified at least two factors, the visual components' spatial arrangement and the row-column association with the highlighted popouts [3], which can be studied in isolation with appropriate study designs.

Our paper is similar in spirit to Ellis and Dix's BELIV'06 paper [8]: both papers report existing problems in information visualization; both provide a list of recommendations to enhance the replicability of evaluations as scientific studies and to enhance the applicability of their results for future designs. In our case, our focus is on re-using quantitative studies of information visualization for meta-analysis where a lot of our suggestions are geared towards standardizing study designs based on common units such as the interface-use factors we listed in Sections 2 and 4.

Since our framework emerged from our experience in extracting design guidelines for interfaces that display multiple levels of data details, it is limited by the materials we used. For example, we only have access to published results. In many cases, unpublished studies, such as those with null results, are also informative and valuable in the larger context of furthering knowledge. Also, since the factors we considered are those reported by our surveyed papers, we did not look at important interface-use factors such as user characteristics (e.g., spatial ability, domain expertise) and environment parameters (e.g., workplace settings). We thus welcome the more widespread use of a broader set of methods of interface evaluation approaches (e.g., [31]) to identify interface-use factors.

We believe the process in designing and reporting studies at a more granular level than those found in typical head-to-head system comparison studies will help us better understand interface use and therefore improve their evaluations. Focusing on how interfaces are used, instead of simply recording participant's performance metrics, will help us standardize study designs and reporting at a meaningful level to obtain generalizable study results that can be directly applicable in design. We believe such an exercise, in addition to enabling and therefore encouraging the much needed meta-analysis in our field, is also one of the first steps in constructing a task-encoding taxonomy at a meaningful level and identifying design tradeoffs, which will eventually lead to building a set of basic design guidelines for visual and interaction design elements in systems.

7. REFERENCES

[1] L. Barkhuus and J. Rode. From Mice to Men: 24 years of Evaluation in CHI. In *Alt.Chi*, 2007.

[2] P. Baudisch, N. Good, V. Bellotti, and P. Schraedley. Keeping Things in Context: A Comparative Evaluation of Focus Plus Context Screens, Overviews, and Zooming. In *Proc. ACM SIGCHI Conf. on Human Factors in Computing Systems (CHI'02)*, pages 259–266, 2002.

[3] P. Baudisch, B. Lee, and L. Hanna. Fishnet, A Fisheye Web Browser with Search Term Popouts: A Comparative Evaluation with Overview and Linear View. In *Proc. ACM Advanced Visual Interface (AVI'04)*, pages 133–140, 2004.

[4] B. Bederson. Fisheye Menus. In *Proc. ACM SIGCHI Symposium on User interface software and technology (UIST'00)*, pages 217–226, 2000.

[5] B. Bederson, A. Clamage, M. P. Czerwinski, and G. G. Robertson. DateLens: A Fisheye Calendar Interface for PDAs. *ACM Trans. on Computer-Human Interaction (ToCHI)*, 11(1):90–119, Mar. 2004.

[6] M. S. T. Carpendale, D. J. Cowperthwaite, and F. D. Fracchia. Making Distortions Comprehensible. In *Proc. IEEE Symposium on Visual Languages*, pages 36–45, 1997.

[7] C. Chen and Y. Yu. Empirical studies of information visualization: A meta-analysis. *International Journal of Human-Computer Studies*, 53:851–866, 2000.

[8] G. Ellis and A. Dix. An Explorative Analysis of User Evaluation Studies in Information Visualization. In *Proc. AVI workshop on BEyond time and errors: novel evaLuation methods for Information Visualization (BELIV'06)*, pages 1–7, 2006.

[9] G. W. Furnas. A fisheye follow-up: Further reflection on focus + context. In *Proc. ACM SIGCHI Conf. on Human Factors in Computing Systems (CHI'06)*, pages 999–1008, 2006.

[10] C. Gutwin and A. Skopik. Fisheye views are good for large steering tasks. In *Proc. ACM SIGCHI Conf. on Human Factors in Computing Systems (CHI'03)*, pages 201–208, 2003.

[11] M. Guzdial, P. Santos, A. Badre, S. Hudson, and M. Gray. Analyzing and visualizing log files: A computational science of usability. Technical Report GIT-GVU-94-08, Georgia Institute of Technology, 1994.

[12] M. Guzdial, C. Walton, M. Konemann, and E. Soloway. Characterizing process change using log file data. Technical Report GIT-GVU-93-44, Georgia Institute of Technology, 1993.

[13] K. Hornbæk, B. Bederson, and C. Plaisant. Navigation patterns and usability of zoomable user interfaces with and without an overview. *ACM Trans. on Computer-Human Interaction (ToCHI)*, 9(4):362–389, 2002.

[14] K. Hornbæk and E. Frokjaer. Reading of electronic documents: the usability of linear, fisheye and overview+detail interfaces. In *Proc. ACM SIGCHI Conf. on Human Factors in Computing Systems (CHI'01)*, pages 293–300, 2001.

[15] K. Hornbæk, E. Frokjaer, and C. Plaisant. Reading patterns and usability in visualization of electronic documents. *ACM Trans. on Computer-Human Interaction (ToCHI)*, 10(2):119–149, 2003.

[16] K. Hornbæk and M. Hertzum. Untangling the usability of fisheye menus. *ACM Trans. on Computer-Human Interaction (ToCHI)*, 14(2), 2007.

[17] M. R. Jakobsen and K. Hornbæk. Evaluating a fisheye view of source code. In *Proc. ACM SIGCHI Conf. on Human Factors in Computing Systems (CHI'06)*, pages 377–386, 2006.

[18] H. Lam and T. Munzner. A study-based guide to multiple visual information resolution interface designs. Technical Report TR-2007-21, University of British Columbia, 2007.

[19] H. Lam, R. A. Rensink, and T. Munzner. Effects of 2D geometric transformations on visual memory. In *Proc. Symposium on Applied Perception in Graphics and Visualization (APGV'06)*, pages 119–126, 2006.

[20] D. Nekrasovski, D. Bodnar, J. McGrenere, T. Munzner, and F. Guimbretière. An evaluation of pan and zoom and rubber sheet navigation. In *Proc. ACM SIGCHI Conf. on Human Factors in Computing Systems (CHI'06)*, pages 11–20, 2006.

[21] P. Pirolli, S. K. Card, and M. M. van der Wege. The effects of information scent on visual search in the hyperbolic tree browswer. *ACM Trans. on Computer-Human Interaction (ToCHI)*, 10(1):20–53, Mar. 2003.

[22] C. Plaisant. The challenge of information visualization evaluation. In *Proc. ACM Advanced Visual Interface (AVI'04)*, pages 109–116, 2004.

[23] C. Plaisant, J. Grosjean, and B. Bederson. SpaceTree: Supporting exploration in large node link tree, design evolution and empirical evaluation. In *Proc. IEEE Symposium on Information Visualization (InfoVis'02)*, pages 57–64, 2002.

[24] M. Plumlee and C. Ware. Zooming versus multiple window interfaces: Cognitive costs of visual comparisons. 13(2):179–209, 2006.

[25] D. Reilly and K. Inkpen. White rooms and morphing don't mix: setting and the evaluation of visualization techniques. In *Proc. ACM SIGCHI Conf. on Human Factors in Computing Systems (CHI'07)*, pages 111–120, 2007.

[26] P. Saraiya, P. Lee, and C. North. Visualization of graphs with associated timeseries data. In *Proc. IEEE Symposium on Information Visualization (InfoVis'05)*, pages 225–232, 2005.

[27] P. Saraiya, C. North, and K. Duca. An evaluation of microarray visualization tools for biological insight. In *Proc. IEEE Symposium on Information Visualization (InfoVis'04)*, pages 1–8, 2004.

[28] P. Saraiya, C. North, V. Lam, and K. Duca. An insight-based longitudinal study of visual analytics. *IEEE Trans. on Visualization and Computer Graphics (TVCG)*, 12(6):1511–1522, 2006.

[29] M. Sarkar, S. Snibbee, O. Tversky, and S. Reiss. Stretching the rubber sheet: A metaphor for viewing large layouts on small screens. In *Proc. ACM SIGCHI Symposium on User interface software and technology (UIST'89)*, pages 81–91, 2003.

[30] W. Schafer and D. A. Bowman. A comparison of traditional and fisheye radar view techniques for spatial collaboration. In *Proc. ACM Conf. on Graphical Interface (GI'03)*, pages 39–46, 2003.

[31] B. Shneiderman and C. Plaisant. Strategies for evaluating information visualization tools: multi-dimensional in-depth long-term case studies. In *Proc. ACM AVI Workshop on BEyond time and errors: novel evaluation methods for information visualization*, pages 1–7, 2006.