Evaluating Information Visualization via the Interplay of Heuristic Evaluation and Question-Based Scoring

Marti A. Hearst
UC Berkeley
Berkeley, CA 94720
hearst@berkeley.edu

Paul Laskowski
UC Berkeley
Berkeley, CA
paul@ischool.berkeley.edu

Luis Silva
UC Merced
Merced, CA
lsilva4@ucmerced.edu

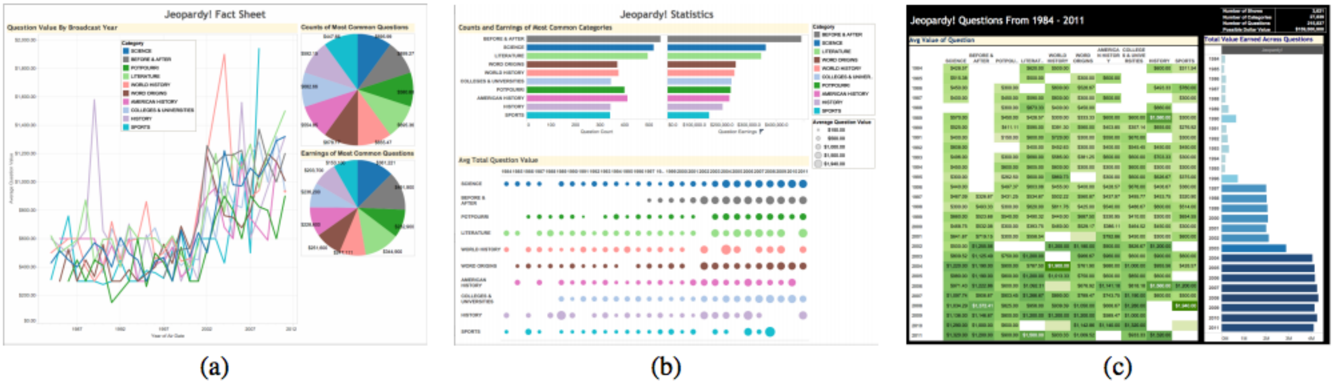

Figure 1. Three visualizations of the same dataset (from the game show *Jeopardy!*), designed to be of (a) low, (b) high, and (c) middling usability.

ABSTRACT

In an instructional setting it can be difficult to accurately assess the quality of information visualizations of several variables. Instead of a standard design critique, an alternative is to ask potential readers of the chart to answer questions about it. A controlled study with 47 participants shows a good correlation between aggregated novice heuristic evaluation scores and results of answering questions about the data, suggesting that the two forms of assessment can be complementary. Using both metrics in parallel can yield further benefits; discrepancies between them may reveal incorrect application of heuristics or other issues.

ACM Classification Keywords

H.5.m. Information Interfaces and Presentation (e.g. HCI): Miscellaneous

Author Keywords

Visualization; Education; Assessment; Heuristic Evaluation; Usability Testing; Question Answering.

http://dx.doi.org/10.1145/2858036.2858280

INTRODUCTION

Information visualization is an exciting area both to teach and practice in today, due to the explosion of interest and tools now readily available for use and view [10]. With the increasing interest in data science on college campuses and beyond, the teaching of visualization is moving beyond its standard home of computer and information science graduate departments. With great popularity comes great responsibility on the part of instructors to ensure that good design practices are instilled and encouraged.

Heuristic evaluation (HE) and other qualitative tools are known to work well in practice, but studies show that experts are better than non-experts at HE in many cases [6, 23, 12]. Furthermore, in the instructional setting, a more objective or utilitarian measure assessing designs can be a valuable additional tool to address students' desire for "proof" that the critique is justified. Offering a contrasting measure that shows which aspects of a design causes users to succeed or fail can be especially convincing to students who are swayed by flashy or popular designs. Here we use posing of questions about the underlying data as a straightforward way to assess visualizations.

Consider the three visualizations shown in Figure 1, designed by the experimenters, all of which portray the same subset of data about the game show *Jeopardy!*. Visualization (a) was intended to be the weakest of the three, with its overlapping lines and the use of pie charts, which are well known to have usability problems [8, 17]. Viz's (b) and (c) each have relative strengths, with Viz (b) allowing for comparison of earnings and frequencies by question category and question value by year. Viz (c) allows for accurate assessment of question value by year, and makes outliers stand out, but at the cost of obscuring the trends over years. Viz (c) also does not allow direct comparison of question count and value by category. For these reasons, Viz (b) should receive higher marks at least than design (a), and be close to if not higher than (c) if evaluating the design according to principles of user performance (using Lam et al.'s taxonomy [19]).

Although students may be instructed, for example, that pie charts are a poor choice in terms of usability, it is our experience that they tend to gravitate towards using them due to their familiarity and the visual appeal of circles. In the study described below, students performing a heuristic evaluation rated Viz (a) higher than (b) on average, despite the fact that (b) was more effective when answering questions.

Our contribution in this work is a new way to assess information visualization designs that brings heuristic evaluation and question answering together in a systematic way. Such an approach seems to be needed; Hollingsed and Novick, in a review of 15 years of usability inspection methods [12], state the need to unify the two as a research goal, and Zhu [34], in stating the lack of a universally accepted framework for defining effective visualizations, calls for a method for establishing a correlation between HE and usability studies.

Instead of a focus on finding discrete usability problems, our approach describes designs in terms of continuous quality scales. We also introduce a reliability statistic based on variance in quality scores. Metric scales are useful in a pedagogical setting to, for example, allow students to see where their designs fall within a set of scores. Assessing designs simultaneously with two distinct evaluation methods has the added benefit of identifying outliers, which could suggest that something about a design merits further scrutiny – perhaps signaling an unusual design that is nonetheless effective.

To illustrate the use of this evaluation instrument, we report the results of a controlled study with 47 students new to HCI and visualization, comparing simple HE to question-based assessments. We apply the evaluation instrument to these designs and show how it can be used to compare results in a classroom setting.

The insights gained by comparing individual designs, as well as the specific statistics we selected to quantify similarity and relative performance of assessment tools, can find application beyond the classroom, in any context in which visualizations are assessed. The desire for objective proof for design guidelines has been observed in the commercial world, in which battles over appearance vs usability have been long documented [1, 16].

RELATED WORK

The value of critique and heuristic evaluation in aggregate is well known in HCI [25, 23, 15, 12]. That said, classic HE is not often used in information visualization courses; instead, instructors typically teach the visual principles from Bertin [2], enhanced by the scales put forward by Cleveland and McGill [5] and Mackinlay [22], and cognitive, narrative, and design principles from other books such as Ware [31], Few [8], and Cairo [3], to impart to students how and under which circumstances to bring forward the underlying meaning of data using different combinations of visual marks.

Numerous publications have noted that different usability assessment techniques uncover different usability problems [15, 14]. Nielsen and Landauer's classic work [24] shows that multiple assessors are needed to find key usability problems, and models the tradeoff between assessing early in the design with HE versus after a system has been built with usability testing. However, today for static visualizations, designing with software tools is almost as fast as sketching on paper, thus changing the balance of these older studies.

Recent work has shown how to better align novice critiques with that of experts; in particular, the use of rubrics has been shown in massive open online courses to bring student ratings within one grade of instructors [18] and novice guideline-based feedback in aggregate can be as acceptable to the receivers as expert feedback [30, 21]. However, positively written and emotional critiques received higher average ratings than negative or neutral critiques [33]. To counter this, Robb et al. [28] found that designers seeking feedback tended to shut out negative *textual* feedback, but responded well to *visual* feedback that showed critics did not understand their designs.

Carpendale [4] provides an overview of empirical evaluation methods applied to visualization, and North [26] elaborates on important aspects of qualitative studies, and design for extraction of insight. Lam et al. [19] examined 850 published research papers in information visualization and taxonomized them (our work assumes their category of User Performance). Isenberg et al. [13] build on this, summarizing the frequency of use of each evaluation method in published research; HE is so infrequently used in research papers that they relegate it to a special "novel methods" category.

Heer et al. [11] and Kosara [17] have shown that core visualization perceptual studies, such as proving that bar charts are more effective than pie charts by having participants answer questions about the underlying data, can be reproduced remotely online from their original lab studies.

STUDY

In this study we compared the results of students assessing visualizations by answering questions about the underlying data to those performing heuristic evaluations.

Method

Participants

47 students from an undergraduate computer science HCI course (taken in person during summer session) chose to participate in a 2x2 between-participants study.

Procedure

Prior to the study, one lecture session, with an active learning component, was devoted to the topic of information visualization. This included a homework assignment to read about visualization design principles [7] and practice plotting a simple data set using Google Charts. Earlier in the course, one prior exercise had been devoted to critique via heuristic evaluation for UIs generally.

The study was conducted during one lab session. Participants were split into two rooms according to their randomly assigned condition. They were not restricted as to time, and spent approximately 20 minutes on average on this activity.[1]

Materials

The motivating scenario was the creation of a static visualization for a general interest news publication, with the goal of answering the most common question types identified by [32]. 6 datasets were selected and curated by the experimenters with the intention of displaying similar amounts of data across conditions. Groups of 3 questions were created for each dataset. 3 visualizations per dataset were created, consisting of static charts intended to range in quality and effectiveness from strong to weak. Topics were: dog breeds, tourism, business financials, city repair, vehicle fuel mileage, and game show statistics.

A web application was written to show combinations of visualizations, datasets, and questions to participants in a randomized, counterbalanced order, and record responses.

Conditions

In Condition Q (for question), participants examined a visualization and answered questions about the underlying data using that visualization. They then rated the visualization using an overall effectiveness rating. They repeated this activity for three different visualizations, all on different datasets.

In Condition H (heuristic), instead of answering questions about the underlying data, participants assigned scores on a 5 point scale from strongly disagree to strongly agree for three different heuristics, based on the reading from the earlier homework assignment [7]. These were stated in positive terms to reduce the chance that participants would accidentally reverse the scales. The heuristics used were:

- This visualization makes important information visually salient.
- This visualization uses visual components appropriately.
- This visualization successfully presents multiple relevant facts into a single visual pattern.

The questions were multiple choice, each having one correct answer out of three.[2] Examples questions are shown below, alongside topic and insight category from Yang et al. [32]:

- Which department spent over budget for the first half of the year and under budget for the second half? (Business, Compare Values)
- Which category of dog shown would be best to recommend to someone who wants a moderately expensive long-lived breed? (DogBreeds, Compare Distribution)
- Which year on average was the most lucrative for players in terms of earnings? (Game Show, Identify Extrema)

The final rating wording was identical in both conditions:

- Please assign this visualization an overall score from 1 (low) to 5 (high) that reflects its overall effectiveness and design quality.

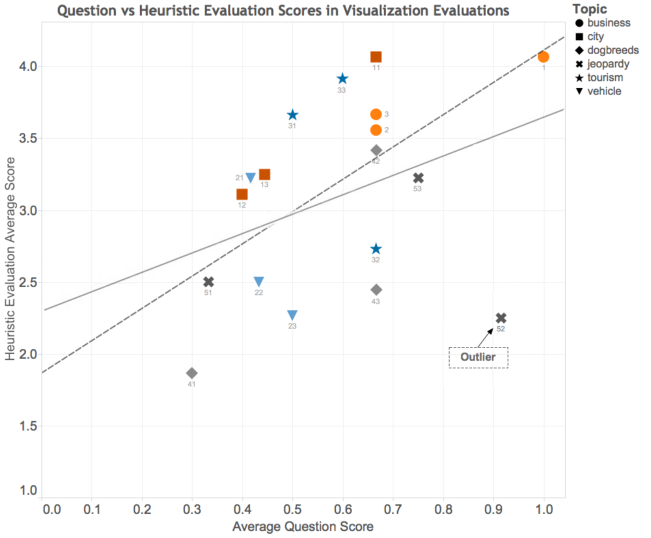

Figure 2. Average correct questions score vs heuristic evaluation rating for each visualization; correlation = 0.38 (solid trend line); with outlier number 52 removed, correlation = .60 (dashed trend line).

Results

Ratings in the heuristic condition were translated to a scale from 1 (strongly disagree) to 5 (strongly agree). The three questions were averaged to generate a composite rating per participant; there were on average 3.8 composite ratings for each design. The mean rating over all designs was 3.05 (sd 0.99). In the question condition, each participant answered 0.59 of questions correctly on average (sd 0.31), and each design was assessed on average by 4.0 participants.

The two metrics share a moderate correlation with each other (Pearson's r = 0.38). That is, if participants tended to agree that the design was strong in the heuristic condition, then they tended to answer correctly in the question condition and vice-versa. This relationship can be seen in Figure 2. There is one notable outlier (according to a Studentized Residual of 3.10); if this design is removed from the dataset, the correlation between the two measures jumps to 0.60, a fairly high degree of correlation (see Figure 2).

This outlier visualization is the one shown in Figure 1(b); it has the next-to-lowest rating based on heuristics, but the second-best score based on the questions. The average score on all three heuristic questions was low for this design, ranging from 2.0 to 2.5. The lowest score was for the question, "This visualization successfully presents multiple relevant facts into a single visual pattern." It could be that participants disliked the combination of disparate elements, though these actually made the design one of the most usable.

In the overall rating task, participants in Condition H assigned a rating more consistent with their scores (correlation = 0.77) than in Condition Q (correlation = 0.34), perhaps because assigning scores is consistent with doing an HE, whereas

[1] Other activities took place during the lab session, but space limitations preclude reporting on them.

[2] Questions used and sample visualizations available in this paper's auxiliary materials.

answering questions invokes a different mode of analysis. It is worth noting that we did not show participants the answers after they attempted the questions; perhaps if we had shown them whether or not they got the questions correct, their final ratings would have aligned more closely with their followup scores. Testing the effects of showing the correct answer is an area for a future work.

Reliability Assessment

The nature of our study means that we have two different measures for each visualization, but no authoritative "ground truth" to compare them to. Because of this, it is not possible to say which procedure is more *accurate*. We can, however, analyze which of them is more *reliable*. In other words, which procedure generates a ranking that is more consistent from one set of participants to another.

In part, reliability depends on the standard deviation of the mean scores for each visualization, which we label σ_m. The more spread out the final scores are, the more confident we can be that we ranked them correctly. Even if a procedure has a high σ_m, however, it will not be very reliable if there is a lot of noise in scoring each design. The higher the average standard error for each final score, σ_v, the more the score will vary from one run of the study to another and the less confidence we can have in it. We therefore consider the ratio, σ_m/σ_v – how spread out the final scores are relative to the noise in each.[3] On this measure, the question condition scores a 1.59, while the heuristic condition scores 1.96, 23% better.

While the question procedure is less reliable than the heuristic procedure, it still performs reasonably well. To get a sense of how big a reliability statistic of 1.59 is, suppose we were to use the question procedure to give each visualization a grade of A, B, C, and so forth. Suppose further that one letter grade corresponds to one standard deviation, a common grading guideline. By approximating the sampling distribution of each final score as a normal curve, we can compute the probability that we are off by a full letter grade from a visualization's true mean value to be only .11.[4]

One reason the question procedure is less reliable is that each question yields a binary result, rather than the more granular scale of the heuristic procedure. This means that there is a lot of variation in the score for a given participant (average standard error 0.23). Adding more questions would reduce this source of error, narrowing the performance gap between the two procedures. Other researchers have noticed the importance of finding the right number of tasks to achieve good coverage for usability tests [20, 27]. Further research is needed to check whether adding more questions ultimately makes the question procedure equally reliable to the heuristics.

DISCUSSION AND CONCLUSIONS

This study has found evidence that having people answer questions about complex visualizations, which is, in effect, a straightforward form of usability testing, can be a useful tool in the teaching arsenal for information visualization assessment. Because it aligns reasonably well with aggregated heuristic evaluation scores by novices, it can be used as an alternative form of assessment, and help make the student designer aware of the problems that might arise in actual use of their design.

The finding that there was less reliability in the questions used in the study partly reflects the binary nature of each question (correct or incorrect). The 5-point scale for each heuristic has the potential to hold more information. One solution is to increase reliability by adding more questions. Since answering questions takes no special training, this can often be accomplished at relatively low cost, via crowdsourcing for example. The measure of reliability we have introduced can help determine when enough questions have been asked to match the characteristics of heuristic evaluation.

Each evaluation method has distinct characteristics: HE is usually applied by experts to anticipate the needs of end users and can be accomplished with fewer participants than usability studies. The question answering method measures whether a visualization successfully imparts needed information. A discrepancy between the two scores may indicate that a heuristic is being misapplied, a specific visualization merits further scrutiny, or some other issue requires attention. An analogy can be drawn to a popular class of techniques in the field of machine learning. Often the output of several algorithms is combined using so-called ensemble techniques. However, it has been shown that these only work well when the signals are diverse in their underlying sources or distributions [9]. We can think of the approach described here as a way for instructors, or for anyone interested in finding alternative ways to assess a design, to combine different assessment approaches, using different "signals," either within one exercise, or across assignments within a course.

This is one step in a vision for a larger framework for teaching and assessing visualizations in which students play different roles at different points in the process, either in a face-to-face course or in a massive online course [10]. Some students could perform an HE and others answer questions about the same design, for a combined and varied analysis. Taking this still further, one set of students would choose datasets and create representative questions, others would design visualizations for the data, and still others would evaluate the designs either with HE or by answering the questions (or with both methods), each rotating in these roles for different datasets. Any of these steps could be performed by people working remotely, and so this procedure can scale to large online courses.

Acknowledgements

The authors thank Colin Gerber for writing the testing software, Valkyrie Savage, Andrew Head, and Björn Hartmann for discussions and supporting the study in their course, the students who volunteered to participate in the study, and the helpful comments of the anonymous CHI reviewers. This research was supported by a Google Social Interactions Grant.

[3] To consistently estimate this parameter, we use the statistic, $g = \sqrt{I/J}f$, where I is the total number of scores, J is the number of designs, and f is Cohen's f [29]. f^2 is sometimes called the signal-to-noise ratio, which we compute as the treatment sum of squares over the error sum of squares.

[4] Information about how to implement this and other aspects of the reliability statistic are included in this paper's auxiliary materials.

REFERENCES

1. Richard I Anderson, Jeremy Ashley, Tobias Herrmann, Justin Miller, Jim Nieters, Shauna Sampson Eves, and Secil Tabli Watson. 2007. Moving UX into a position of corporate influence: whose advice really works?. In *CHI'07 Extended Abstracts on Human Factors in Computing Systems*. ACM, 1905–1908.

2. Jacques Bertin. 1983. *Semiology of graphics: diagrams, networks, maps*. University of Wisconsin press.

3. Alberto Cairo. 2012. *The Functional Art: An introduction to information graphics and visualization*. New Riders.

4. Sheelagh Carpendale. 2008. Evaluating information visualizations. In *Information Visualization*. Springer, 19–45.

5. William S Cleveland and Robert McGill. 1984. Graphical perception: Theory, experimentation, and application to the development of graphical methods. *Journal of the American statistical association* 79, 387 (1984), 531–554.

6. Heather Desurvire, Jim Kondziela, and Michael E Atwood. 1992. What is gained and lost when using methods other than empirical testing. In *Posters and short talks of the 1992 SIGCHI conference on Human factors in computing systems*. ACM, 125–126.

7. Stephen Few. 2006. Visual Communication: Design Principles for Displaying Quantitative Information. *Cognos Innovation Center* (September 2006). http://www.perceptualedge.com/articles/Whitepapers/Visual_Communication.pdf.

8. Stephen Few. 2009. *Now you see it: simple visualization techniques for quantitative analysis*. Analytics Press.

9. Yoav Freund and Robert E Schapire. 1996. Experiments with a new boosting algorithm. In *ICML*, Vol. 96. 148–156.

10. Marti A Hearst, Eytan Adar, Robert Kosara, Tamara Munzner, Jon Schwabish, and Ben Shneiderman. 2015. Vis, The Next Generation: Teaching Across the Researcher-Practitioner Gap. In *IEEE Conference on Information Visualization [Panel]*.

11. Jeffrey Heer and Michael Bostock. 2010. Crowdsourcing graphical perception: using mechanical turk to assess visualization design. In *CHI*. ACM, 203–212.

12. Tasha Hollingsed and David G Novick. 2007. Usability inspection methods after 15 years of research and practice. In *Proceedings of the 25th Annual ACM international conference on Design of communication*. ACM, 249–255.

13. Tobias Isenberg, Petra Isenberg, Jian Chen, Michael Sedlmair, and Torsten Moller. 2013. A systematic review on the practice of evaluating visualization. *Visualization and Computer Graphics, IEEE Transactions on* 19, 12 (2013), 2818–2827.

14. Robin Jeffries and Heather Desurvire. 1992. Usability testing vs. heuristic evaluation: was there a contest? *ACM SIGCHI Bulletin* 24, 4 (1992), 39–41.

15. Robin Jeffries, James R Miller, Cathleen Wharton, and Kathy Uyeda. 1991. User interface evaluation in the real world: a comparison of four techniques. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 119–124.

16. Ron Kohavi, Roger Longbotham, Dan Sommerfield, and Randal M. Henne. 2009. Controlled experiments on the web: survey and practical guide. *Data Min. Knowl. Discov.* 18 (2009), 140–181.

17. Robert Kosara and Caroline Ziemkiewicz. 2010. Do Mechanical Turks dream of square pie charts?. In *Proceedings of the 3rd BELIV'10 Workshop: BEyond time and errors: novel evaLuation methods for Information Visualization*. ACM, 63–70.

18. Chinmay Kulkarni, Koh Pang Wei, Huy Le, Daniel Chia, Kathryn Papadopoulos, Justin Cheng, Daphne Koller, and Scott R Klemmer. 2013. Peer and self assessment in massive online classes. In *ACM Transactions on Computer Human Interaction (TOCHI)*. Vol. 20. ACM.

19. Heidi Lam, Enrico Bertini, Petra Isenberg, Catherine Plaisant, and Sheelagh Carpendale. 2012. Empirical studies in information visualization: Seven scenarios. *Visualization and Computer Graphics, IEEE Transactions on* 18, 9 (2012), 1520–1536.

20. Gitte Lindgaard and Jarinee Chattratichart. 2007. Usability testing: what have we overlooked?. In *Proceedings of the SIGCHI Conference on Human factors in Computing Systems*. ACM, 1415–1424.

21. Kurt Luther, J Tolentino, Wei Wu, Amy Pavel, B Bailey, Maneesh Agrawala, Björn Hartmann, and Steven P Dow. 2015. Structuring, Aggregating, and Evaluating Crowdsourced Design Critique. CSCW.

22. Jock Mackinlay. 1988. Applying a theory of graphical presentation to the graphic design of user interfaces. In *UIST*. ACM, 179–189.

23. Jakob Nielsen. 1992. Finding usability problems through heuristic evaluation. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 373–380.

24. Jakob Nielsen and Thomas K Landauer. 1993. A mathematical model of the finding of usability problems. In *Proceedings of the INTERACT'93 and CHI'93 conference on Human factors in computing systems*. ACM, 206–213.

25. Jakob Nielsen and Rolf Molich. 1990. Heuristic evaluation of user interfaces. In *CHI*. ACM, 249–256.

26. Chris North. 2006. Toward measuring visualization insight. *Computer Graphics and Applications, IEEE* 26, 3 (2006), 6–9.

27. Lucy Nowell, Robert Schulman, and Deborah Hix. 2002. Graphical encoding for information visualization: an empirical study. In *Information Visualization, 2002. INFOVIS 2002. IEEE Symposium on*. IEEE, 43–50.

28. David A Robb, Stefano Padilla, Britta Kalkreuter, and Mike J Chantler. 2015. Crowdsourced Feedback With Imagery Rather Than Text: Would Designers Use It?. In *CHI*. ACM, 1355–1364.

29. James H Steiger. 2004. Beyond the F test: Effect size confidence intervals and tests of close fit in the analysis of variance and contrast analysis. *Psychological methods* 9, 2 (2004), 164.

30. David Tinapple, Loren Olson, and John Sadauskas. 2013. CritViz: Web-based software supporting peer critique in large creative classrooms. *Bulletin of the IEEE Technical Committee on Learning Technology* 15, 1 (2013), 29.

31. Colin Ware. 2012. *Information visualization: perception for design*. Elsevier.

32. Huahai Yang, Yunyao Li, and Michelle X Zhou. 2014. Understand users' comprehension and preferences for composing information visualizations. *ACM Transactions on Computer-Human Interaction (TOCHI)* 21, 1 (2014), 6.

33. Alvin Yuan, Kurt Luther, Markus Krause, Sophie Vennix, Stephen P. Dow, and Björn Hartmann. 2016. Almost an Expert: The Effects of Rubrics and Expertise on Perceived Value of Crowdsourced Design Critiques. CSCW. to appear.

34. Ying Zhu. 2007. Measuring effective data visualization. *Advances in Visual Computing, Lecture Notes in Computer Science (LNCS)* 4842 (2007), 652–661.