EuroRV3: EuroVis Workshop on Reproducibility, Verification, and Validation in Visualization (2017)
K. Lawonn, N. Smit, and D. Cunningham (Editors)

Guidelines and Recommendations for the Evaluation of New Visualization Techniques by Means of Experimental Studies

M. Luz[1], K. Lawonn[2] and C. Hansen[1]

[1]Insitute for Simulation and Graphics, Otto-von-Guericke University Magdeburg, Germany
[2]Insitute for Computational Visualistics, University Koblenz-Landau, Germany

Abstract

This paper addresses important issues in the evaluation of new visualization techniques. It describes the principle of quantitative research in general and presents the idea of experimental studies. The goal of experimental studies is to provide the base for guidelines, which allow testing of hypotheses that newly-developed visualization solutions are better than older ones. Moreover, the paper provides guidelines for successful planning of experimental studies in terms of independent and dependent variables, participants, tasks, data collection and statistical evaluation of collected data. It describes how the results should be interpreted and reported in publications. Finally, the paper points out useful literature and thus contributes to a better understanding of how to evaluate new visualization techniques.

Categories and Subject Descriptors (according to ACM CCS): H.5.1 [INFORMATION INTERFACES AND PRESENTATION]: Multimedia Information Systems—Evaluation/methodology

1. Introduction

Many computer scientists develop promising visualization solutions. However, they are often confronted with the problem of how to evaluate them. There are several approaches which deal with the evaluation issues and emphasize different aspects [KHI*03, Pla04, TM05, LBI*12, IIC*13, Sta14, EY12]. While many research papers focus on the evaluation of technical properties such as algorithm performance, Lam et al. and Isenberg et al. [LBI*12, IIC*13] show that the number of publications evaluating visualization techniques with a more user-focused approach has increased significantly in recent years, reflecting heightened interest in user performance and subjective feedback. However, conduction of studies involving human participants is often not a part of computer scientists' education, and thus might be challenging for them. A research question which computer scientists with focus on visualization may have is whether or not the newly developed solution is in fact beneficial for potential users compared to established visualization techniques. This research question could be investigated by means of experimental studies.

The goal of this work is to give a short overview about the approach of experimental studies, which should help researchers who are not familiar with conducting controlled experimental studies. Links to important textbooks and relevant scientific publications are provided. This paper explains the idea of experimental studies, outlines the development of study design, and discusses the possible consequences of certain decisions in terms of study design and publication of the results. More detailed consideration of the mentioned aspects is given by [FH03, Pur12, FC14].

2. Research Methods

An evaluation with potential users can be performed by means of qualitative or quantitative research methods. Which type of evaluation is performed depends on the purpose of the study. With qualitative methods it is possible to retrieve qualitative information such as visualization requirements in general, advantages and disadvantages of a certain visualization technique, required changes and improvements. This information can be obtained through such means as interviewing or observation. In contrast, quantitative research methods provide user performance data that allow a comparison of different visualization techniques. The usual quantitative method to obtain these data is an experimental study, which the following will focus on.

3. The Principle of Quantitative Research Methods

The goal of quantitative research methods is to test hypotheses on the basis of collected data. In the context of visualization research, the hypothesis is usually that the new visualization technique is better than the old one in terms of selected aspects. Otherwise, the development of new visualization techniques would not make sense. The question is how to test this hypothesis. How much better should the new visualization be in order to be considered preferable? What role does coincidence play in data collection? When is a hypothesis

true and when is it false? The common method to prove the hypothesis involves the statistical evaluation of collected data. It provides clear guidelines, in the form of probability thresholds on which the decision to accept or reject the hypothesis can be made.

Statistical testing of a given hypothesis is based on the idea that conclusions can be drawn from sampled data. However, factors such as random chance and other confounding variables, which cannot fully be overviewed and controlled, may have impacts on the data and in some cases do not reflect the true situation. Thus, the result of a statistical test provides the probability of an error (p-value) that the hypothesis is actually false, despite the result indicating otherwise (type I error).

Usually, the probability threshold is 5% and is called the alpha level. If the p-value is lower than the alpha level, we can accept our hypothesis, at least temporarily, and assume a systematic impact of tested visualization techniques on our results. That means we tolerate that the the accepted hypothesis might be false with a probability of at most 5%. If the p-value is higher than the alpha level, we reject our hypothesis, and assume that the differences between the studied visualization techniques are coincidental.

However, as stated before, the data and hence the p-value depend on many factors which can be manipulated to some degree. This is the reason that even the statistical evaluation may be a matter of discussion.

4. The Idea of Experimental Studies

An experimental study is a method to collect the data we need to make conclusions and test our hypothesis. To that end, we invite participants to our laboratory and let them perform a task. Then we use their performance data for statistical testing. The goal is to identify performance differences between studied visualization techniques, which represent a "signal". These performance differences represent the fact that we are going to determine on the basis of our collected data. But many factors may influence the participant's performance, for example participant characteristics (sex, age, experience, skills) or environmental factors (time of the day, light conditions, reputation of the investigator). That may cause some kind of "noise" (variance) in our data which can obscure our performance differences. To make it easier to find the performance differences, one could (or even should) reduce the data variance by controlling unwanted influences. Therefore, the experiments are called "controlled". For example, we can use same study material, procedures, and apparatus for every participant. This approach is one of the accepted methods which make it easier to reach the decision threshold. However, not only does the data variance vary, but also the performance differences, depending on the studied visualizations. The larger the performance differences, the easier they can be identified, even if the data variance is severe. The relation of the performance difference to data variance is called effect size.

5. Planning of Experimental Studies

Before conducting the experimental study, some considerations about study design should be taken into account, which are discussed in the following. More detailed information and practical guidelines can be extracted from [Pur12]. A good use case of how to evaluate medical visualizations in the example of 3D aneurysm surfaces can be found in [GSB*16].

5.1. Research Question

The very first step is to define the research question or questions based on theoretical considerations. This step is crucial, because all following decisions draw upon the research question, even the decision to conduct an experimental study or to use a different research method. It is important to have an exact idea of what one wants to find out. This would make the research more precise and efficient, and it would help avoid discrepancies in the research publication. In general, experimental studies are appropriate for the research question of whether visualization A is better than B in terms of X, Y and Z. In this case, the visualizations A and B represent so-called independent variables. X, Y and Z represent dependent variables. The research question can be further specified, e.g. for whom and in what context is the visualization A better than B. These two aspects reflect the target group and the task for which the new visualization was developed. In the following, we discuss how these aspects should be specified based on the research question.

5.2. Independent Variables

Independent variables are variables whose impacts have to be investigated. The main independent variable in the context of visualization is of course the visualization technique. Therefore, the impact of newly-developed visualization techniques should be compared with other techniques. For example, the newly-developed illustrative visualization approach was compared with standard Phong shading [RHD*06,LLPH15,LLH17], and an advanced medical visualization approach was compared with a standard 3D medical visualization [HZR*13, HZS*14]. It is possible to investigate several independent variables, for example different aspects of one visualization technique and its combinations such as illumination and perspective [WB08].

5.3. Dependent Variables

As described above, the goal of the experimental study is to test whether one visualization technique is better than another. But better in terms of what? In case of visualization, it could be such aspects as depth and shape impression, intuitiveness, preference, and/or usability. The next question which arises is: how could these aspects be measured/operationalized? Usually, the dependent variables can be operationalized with task performance accuracy and task performance duration as objective measures [GSB*16]. Moreover, it is informative to measure participants' subjective opinion such as acceptance or preference for one visualization or another, or subjective judgment of investigated aspects, since they may differ from objective performance.

5.4. Participants

After the specification of the research questions and independent and dependent variables, the researcher should define the study participants and the task they have to perform.

The questions that the researchers often raise in terms of participants are who should the participants be, and how many participants are needed?

First and very important, the participants should be "naive". That means that they must not know what results you are expecting as a researcher, i.e. what hypothesis you have. That means that enrolling colleagues as participants may be not the best idea. On the other hand, you should ensure that this information is at no time before or during the data collection passed to participants, either explicitly or implicitly (experimenter bias). If it is, the influence on your results can help to confirm your hypothesis, which may make you happy, but such a situation does not reflect the facts and may not be replicable. Second, the sample should be representative, which means the participant should have the important characteristics which the majority of the target group of visualization have, for example special knowledge or skills. That means, on the other hand, that people with exceptional characteristics, such as color blindness if colors provide an important cue in your visualization technique, should be excluded from participation in the experiment.

The issue of number of participants is a tricky one. On the one hand, by increasing the number of participants, you make the influence of a coincidence smaller, which makes it easier to reach the alpha level and confirm the hypothesis. But on the other hand, with increased number of participants the probability of detecting smaller effect size increases as well. This may mean that with a large sample the alpha level could be reached, but the difference between the studied visualizations is not practically relevant. That's why, when interpreting the data, an experienced researcher considers not only the exceedance of an alpha level, but also the number of participants and the effect size. Thus, for the determination of a sample size, it is recommended to first consider what effect size you want to detect and based on this consideration calculate the required sample size. To do this is possible using tables [Coh88] or special software as G*Power. Moreover, it is described in statistical books [MMC17]. The issue regarding appropriate number of participants in visualization evaluation is further discussed in [IIC*13].

5.5. Task

The goal of the visualization techniques is to make data more readily interpretable, reduce error of data interpretation, emphasize/make information visible, and show relationship between the data. The task should be chosen in a way that the participants' task performance allows conclusions to be made about these investigated aspects, for example such tasks as depth judgment, orientation matching or surface categorization. A good overview about the possible tasks for comparing different medical visualizations is provided by [PBC*16].

5.6. Data Collection

Before conducting the experimental study, researchers should plan how the data should be collected, treated and evaluated. This will help to avoid mistakes which could render the data useless.

Data collection is the most important step where the data variance can be reduced. In section 4, we explained that this could be achieved by using same study material, same procedures, and same apparatus for every participant. By standardizing these conditions their influence is kept constant, which reduces the data variance. However, another important aspect should be considered, namely the possible confounding variables. Confounding variables are correlated to independent variables and may offer alternative explanations for the performance differences and therefore limit the attribution of these performance differences to independent variables, in our case to visualization techniques. Confounding variables can also be inversely correlated to independent variables, which will make the performance differences less identifiable. These confounding variables could be learning and sequence effects, or effects of fatigue. To minimize this bias, the sequence of studied independent variables, e.g. visualization techniques, should be balanced across the participants. Moreover, it is worth the consideration to randomize used stimuli.

5.7. Data Treatment and Statistical Evaluation

Before conducting the study, the researcher should consider how the collected data should be treated and how the postulated hypothesis can be tested by means of statistical tests.

Data treatment can be run- or participant-related and depends on the study purpose [LSM16]. Run-related data treatment means that every data set is connected to a certain run, for example a certain stimulus. Run-related data treatment is indicated when technical characteristics of new methods or techniques are studied by means of standardized procedure to make the performance independent from the human participant e.g., to measure the performance or the preprocessing time for an algorithm. Participant-related data treatment means that every data set is connected to a certain participant and is indicated when the impact of phenomena on human participants is investigated, for example the impact of visualization techniques on depth impression. Therefore, during data processing, values which come from various task runs should be averaged for each participant and each condition for statistical evaluation. Otherwise, by treating data run-related, the sample size would be artificially enlarged, which definitely would reduce data variance and contribute to getting statistically significant results, but would not represent the facts and therefore would be difficult to replicate.

In terms of choosing of appropriate statistical test see statistical books [JMR11, FJF12, Fie13, MMC17, FH03, Hin04] because this is a complicated issue. Some of these books even provide decision trees which facilitate the choice.

6. Publishing the Research

After the study is planned, the researcher can conduct the study according to developed study design. Usually, researchers want to publish their research. The publication follows the study design. However, additional aspects should be included in the publication, which are explained in following.

6.1. Reporting Results

It is important to report all statistical parameters and not simply report whether the test revealed significant results. This helps the

readers understand how the data was treated and how statistical analyses were performed, and therefore to identify possible biases. Moreover, the effect size should be reported as well. Effect size is the only parameter that allows comparison of the results between different studies. The reported statistical parameters permit well-justified statements about the investigated effects and therefore increase the chances that the study will be included in reviews and meta-analyses.

It is very important for scientific progress to report non-significant results and publish studies which reveal these results, despite having only small potential to be cited. These studies ensure that scientists don't spend resources on topics which don't have potential and redirect their effort to a development of alternative solutions. Unfortunately, many scientists and, more importantly, peer reviewers are often not aware of this relationship, and/or are just interested in increasing the amount of citations and the impact factor, which leads to a so-called publication bias. That means that only predominantly significant results end up published. But in some cases they might be only coincidental, and a much greater number of studies with non-significant results exist in archives.

6.2. Interpretation of the Results

The results of each study are usually limited to the conditions under which the data was collected: they are valid only for the used sample, for the particular task the participants had to perform, and for other studied aspects. One study cannot provide the answers to all research questions, but only to their limited number. Moreover, the collected data may to some degree not reflect the facts (type I error). To draw definite conclusions about the investigated effects and the generalization of the results, several studies with the same research focus should be considered, which may be summarized in a review or even meta-analysis. In this light, the replication of studies will provide an important benefit.

7. Summary

This work provided a short overview of main important aspects that should be considered when conducting experimental studies to evaluate visualizations. It is such a broad topic that whole books are dedicated to it, and it is impossible to address all issues related to it in a small paper. Thus, while conducting experimental studies, it is reasonable to ask for help from experts, for example psychologists for study design and data interpretation as well as statisticians for statistical evaluation. Nevertheless, with this paper we hope to contribute to a better understanding of how to evaluate new visualization techniques.

Acknowledgement

This work was funded by the German Research Foundation (DFG) within the projects HA 7819/1-1 and LA 3855/1-1.

References

[Coh88] COHEN J.: *Power Analysis for the behavioral Science*. Academic Press, 1988. 3

[EY12] ELMQVIST N., YI J. S.: Patterns for visualization evaluation. In *Proceedings of the 2012 BELIV Workshop: Beyond Time and Errors - Novel Evaluation Methods for Visualization* (2012), BELIV '12, ACM, pp. 12:1–12:8. 1

[FC14] FORSELL C., COOPER M.: An introduction and guide to evaluation of visualization techniques through user studies. In *Handbook of human centric visualization*. Springer, 2014, pp. 285–313. 1

[FH03] FIELD A., HOLE G.: *How to Design and Report Experiments*. Sage, 2003. 1, 3

[Fie13] FIELD A.: *Discovering statistics using IBM SPSS statistics*. Sage, 2013. 3

[FJF12] FIELD A., J. M., FIELD Z.: *Discovering Statistics Using R*. Sage, 2012. 3

[GSB*16] GLASSER S., SAALFELD P., BERG P., MERTEN N., PREIM B.: How to Evaluate Medical Visualizations on the Example of 3D Aneurysm Surfaces. In *Eurographics Workshop on Visual Computing for Biology and Medicine (VCBM)* (2016), pp. 153–162. 2

[Hin04] HINTON P.: *Statistics Explained: A Guide for Social Science Students, 2nd Edition*. Taylor & Francis, 2004. 3

[HZR*13] HANSEN C., ZIDOWITZ S., RITTER F., LANGE C., OLDHAFER K., HAHN H. K.: Risk maps for liver surgery. *International Journal of Computer Assisted Radiology and Surgery 8*, 3 (2013), 419–428. 2

[HZS*14] HANSEN C., ZIDOWITZ S., STRAVROU G., OLDHAFER K. J., HAHN H. K.: Impact of model-based risk analysis for liver surgery planning. *International Journal of Computer Assisted Radiology and Surgery 9*, 2 (2014), 473–480. 2

[IIC*13] ISENBERG T., ISENBERG P., CHEN J., SEDLMAIR M., MOLLER T.: A systematic review on the practice of evaluating visualization. *IEEE Trans Vis Comput Graph 19*, 12 (2013), 2818–2827. 1, 3

[JMR11] JUDD C. M., MCCLELLAND G. H., RYAN C. S.: *Data analysis: A model comparison approach*. Routledge, 2011. 3

[KHI*03] KOSARA R., HEALEY C. G., INTERRANTE V., LAIDLAW D. H., WARE C.: Thoughts on User Studies: Why, How, and When. *Computer Graphics and Applications 23*, 4 (2003), 20–25. 1

[LBI*12] LAM H., BERTINI E., ISENBERG P., PLAISANT C., CARPENDALE S.: Empirical studies in information visualization: Seven scenarios. *IEEE Transactions on Visualization and Computer Graphics 18*, 9 (2012), 1520–1536. 1

[LLH17] LAWONN K., LUZ M., HANSEN C.: Improving spatial perception of vascular models using supporting anchors and illustrative visualization. *Computers & Graphics 63* (2017), 37 – 49. 2

[LLPH15] LAWONN K., LUZ M., PREIM B., HANSEN C.: Illustrative Visualization of Vascular Models for Static 2D Representations. In *International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI)* (2015), pp. 399–406. 2

[LSM16] LUZ M., STRAUSS G., MANZEY D.: Impact of image-guided surgery on surgeons' performance: a literature review. *International Journal of Human Factors and Ergonomics 4*, 3-4 (2016), 229–263. 3

[MMC17] MOORE D. S., MCCABE G. P., CRAIG B. A.: *Introduction to the Practice of Statistics*. Freeman and Company, 2017. 3

[PBC*16] PREIM B., BAER A., CUNNINGHAM D., ISENBERG T., ROPINSKI T.: A survey of perceptually motivated 3d visualization of medical image data. *Computer Graphics Forum 35*, 3 (2016), 501–525. 3

[Pla04] PLAISANT C.: The challenge of information visualization evaluation. In *Proceedings of the Working Conference on Advanced Visual Interfaces* (2004), AVI '04, ACM, pp. 109–116. 1

[Pur12] PURCHASE H. C.: *Experimental Human-Computer Interaction: A Practical Guide with Visual Examples*, 1st ed. Cambridge University Press, 2012. 1, 2

[RHD*06] RITTER F., HANSEN C., DICKEN V., KONRAD-VERSE O., PREIM B., PEITGEN H.-O.: Real-time illustration of vascular structures. *IEEE Trans. Vis. Comput. Graph. 12*, 5 (2006), 877–884. 2

[Sta14] STASKO J.: Value-driven evaluation of visualizations. In *Proceedings of the Fifth Workshop on Beyond Time and Errors: Novel Evaluation Methods for Visualization* (2014), ACM, pp. 46–53. 1

[TM05] TORY M., MOLLER T.: Evaluating visualizations: do expert reviews work? *IEEE Comput Graph Appl 25*, 5 (2005), 8–11. 1

[WB08] WEIGLE C., BANKS D.: A comparison of the perceptual benefits of linear perspective and physically-based illumination for display of dense 3d streamtubes. *IEEE Transactions on Visualization and Computer Graphics 14*, 6 (2008), 1723–1730. 2