A Guide to Reporting Scientific Evaluation in Visualization

Camilla Forsell
C-research, Linköping University,
Norrköping, Sweden

Matthew Cooper
C-research, Linköping University
Norrköping, Sweden

camilla.forsell@liu.se

ABSTRACT
This paper addresses some fundamental issues that need to be considered when reporting an evaluation study in visualization or other areas that develop visual interfaces. Today evaluations are frequently included in publications in these fields, however there exists no uniform standard for reporting formats. Instead there is heterogeneity in the structure and content provided and important information is missing, making it hard to gain insight and trust in the results and conclusions. Consequently there is a need for an easy to access guidance on how to accomplish sound reporting with high quality. This is the only way in which authors can enable an assessment of the validity of a study and enable it to be replicated by others in order to verify it. This paper presents a first effort to introduce guidelines on what constitutes an effective structure and what content to address and how. It also points out common pitfalls and mistakes when reporting and how these can be avoided. The paper could be used as a guide by authors when describing their evaluation studies and it could also be helpful when reviewing publications presenting such work since the same guidelines for content apply.

Categories and Subject Descriptors
H.5.2 [Information Interfaces and Presentation]: User Interfaces – evaluation/methodology, graphical user interfaces, style guides.

General Terms
Documentation, Experimentation, Measurement, Human Factors, Standardization.

Keywords
Reporting, evaluation, visualization, guidelines.

1. INTRODUCTION
Scientific evaluation is now acknowledged to be an important part of all research and development within communities such as human-computer interaction, information visualization, visual analytics and scientific evaluation. In 2000 the International Journal of Human-Computer Studies published a special issue on empirical evaluation [2] and since about this time there has been a continuous call for publications that address evaluation, either as part of development of applications and techniques or as a core research topic in itself. There is also an appreciation of the difficulty of designing and conducting sound scientific evaluations and inexperienced authors are often encouraged to partner with colleagues with greater knowledge to ensure high quality studies. The process of sound scientific evaluation also encompasses sound reporting: for the reader it should be possible to clearly understand every aspect of the study. This is the only way authors can enable an assessment of the reliability, validity and generalizability of a study and enable it to be replicated by others in order to verify it. To accomplish such a description is, however, difficult to master, and yet, to our knowledge, there exists no uniform standard for reporting formats specific to our research community nor any easy to access guidance on appropriate structures and what content to include and how. Authors are referred to textbooks [5] and guidelines [1], which may be too general and/or too comprehensive to be appealing to consult. Also, it is hard to use conference and journal publications from one's own field as guidance since there exists great heterogeneity in the structure and content provided. Consequently, for the inexperienced author it is impossible to know what is appropriate or not. The result is that important information is often missing in publications making it hard to gain insight and trust in the results and conclusions. The contribution of this paper is to present a first effort to introduce guidelines on what constitutes an effective structure and what content to address and how. It also points out common pitfalls and mistakes when reporting and how these can be avoided. The paper could be used as a guide by authors when describing their evaluation studies and it could also be helpful when interpreting or reviewing publications since the same guidelines apply.

2. MOTIVATION AND DELIMITATION
The objective with guidelines for reporting of scientific publications is to ensure that they include all relevant information for the readers to gain insight and trust in the results and conclusions. Based on our experience as authors, reviewers, papers chairs, symposium chairs, and tutorial presenters we argue that the current status of publications reporting on evaluation studies clearly shows a great need for such guidelines. This is particularly true in reporting the method and results, which are essential to communicate what and how was done and how the results relate to this. This opinion is also well supported by recent discussions with researchers at conferences and tutorials when reporting is discussed. Many publications have similar flaws that could easily be prevented if authors were aware of what information to include and how, and how to avoid common omissions and errors. The aim of this paper is a first effort to provide our community with such guidelines. These are based on existing literature such as textbooks in experimental design [5],
and borrow from reporting guidelines in other areas [1, 8] and also on our experience and expertise in the field of evaluation in visualization. The main focus is on reporting quantitative studies (controlled experiments) but the general knowledge applies to all kinds of studies. When reporting on an evaluation study this will always require some introduction and background and an overall description of the study itself along with the specific aims and hypothesis. That can be organized in several ways and depends on the type of paper (core evaluation paper or evaluation as part of a larger work). This is not included in this paper, which covers guidelines only for reporting what is regarded as the method and results section in a publication.

The objective of presenting this work in progress at the AVI 2012 conference is that it would provide an excellent route to discuss the need for guidelines, and to receive important feedback from the research community on what they would like to be incorporated into future improvements. By pursuing this work in the future the quality of published evaluation studies will be improved and thus also the evidence-base of the community.

3. REPORTING

The information provided should be presented in a way that is scientific, unambiguous and useful. Here terminology is highly important but unfortunately, there is no ‘universal code’ to apply. Several words can be used to refer to the same thing. For example, a trial is commonly one of a number of repetitions in an experiment but is also used to refer to an entire experiment. Also, one of a number of all repetitions can be called a task or a case whereas different phases of experiments are called parts, tests, sessions etc. The most important point is not to choose the ‘right’ word, but to be consistent. We have come across many publications where several different words have been used throughout a publication to refer to the same thing, making the description impossible to follow. When you first describe something, that is when you give it an operational definition, it should be crystal clear what you mean and then you should use that term consistently throughout the text. Good operational definitions define and describe variables, measurements and procedures so that they cannot be misunderstood and so that other researchers can replicate them by following the definitions [6, p. 75]. For example if the objective of the evaluation is to assess the usability of a visualization technique, usability has to be clearly defined which could be done using the ISO definition (International Organisation for Standardisation) [7]. Then also its constructs of effectiveness and efficiency (and perhaps satisfaction) need to be defined in an operational form describing exactly how these should be measured, for example by recording response times and accuracy and how this is done.

Throughout the description the author must show awareness of possible biases by making clear statements about their potential impact on the outcome of the study and how this is considered and handled. The content of high quality reporting needs to be presented in sufficient detail to allow the reader a clear understanding of its value and contribution and to replicate a study without having to contact the author. This should be carried out in a way that is easy to follow and avoids repetition when redundancy is unnecessary.

4. THE METHOD SECTION

The method section is should describe all relevant details about what you have done, how you proceeded when doing it and what you used when doing it. Considered the section as a recipe, it should be possible to fully understand all the ingredients, measurements and procedures and to replicate a study by following the description. There are several fundamental issues that need to be covered. These are usually discussed in separate subsections and some typical ones are described below. How to organize these subsections is flexible and depends on what will constitute a logical order of presentation both to enhance readability and to avoid repetition. Naturally, some subsections can be collapsed or perhaps some of their content could be more sensibly placed elsewhere. When a publication consists of several experiments it can be beneficial for good readability to have an overall experimental method section describing what is common for all experiments and then only report what differs in each of them. This is also helpful when space is short.

What should be reported on in each subsection could be presented as checklists. However such an approach is too superficial to be helpful. Therefore, in each subsection some example sentences have been included to exemplify how to describe certain items.

4.1 Participants

This section should describe essential information about the participants that were engaged to take part in the evaluation. It should include all characteristics relevant to the purpose of the study, and discuss how they may effect the outcome. Common demographic characteristics to report are sex, age, education level, and level of experience in terms of expertise or number/range of years. It should also be reported how the participants were selected and from where (volunteers or recruited from a company, a university, students at a specific course etc). Regarding age we advise to state the median age instead of the mean age, since that is more informative, together with the full range of years, for example: “The participants were aged between 22 and 35 years, the median age was 25”. In this section it might also be applicable to state whether any power analysis or other presumption was made to decide on the number of participants. How the participants were assigned to different groups/conditions is, however, best described in the experimental design section. Further, it is also common to state whether the participants had normal or corrected to normal vision (via glasses, contact lenses) and, if relevant for the study, whether they had normal colour vision, stereo vision etc. Here the test used to assess these abilities should be stated, for example: “All participants had normal colour vision according to the Nishiara test for colour vision”. The actual testing procedure is later described in the procedure section. It should also be described whether the participants received any compensation or earned educational credits or similar, since this affects the motivation for taking part for taking part. Finally, journals and conferences in visualization do not require (as is often the case in psychology and medicine) a specific statement that “Informed consent was obtained from all participants” (indicating fully informed participants see for example [3] for more detail), but doing so shows good ethics and is recommended.

4.2 Task

The actual task(s) used to obtain the evaluation data should be thoroughly explained. This includes the nature of the task and how it was executed, why it was chosen, its representativeness for the research question and to what extent it can be generalized beyond the evaluation context. The task is closely related to the main objective and aim with the study. This means that when the reader reaches the method section it has already been described, at least in general terms, in previous sections such as the introduction, the description of the technology or in the introduction to the evaluation. Therefore a separate section called
‘Task’ might not be required to avoid unnecessary redundancies. The description of the specific task performed by the participants and how it was carried out in terms of operating the software (if applicable) and giving response could therefore be included in another section such as the procedure section (exemplified in that section below).

The method section needs to describe all material and equipment used in the evaluation. This can be done in several ways, where different headings and divisions of content may be suitable. It is a matter of convenience and what works best, what’s important is that all relevant information is included and presented logically. We propose the following two common subsections: ‘Stimuli’ for details of stimuli (what is presented to the subject) and materials such as questionnaires, and ‘Apparatus and usage conditions’ for hardware and setup.

4.3 Stimuli (or Material)
This subsection should present more detail about the stimuli for each type of task and trial. More general issues about the visualization technique could have been explained earlier in the paper and then this description need only cover what the actual images presented on the display looked like. For example: “Each stimulus display was comprised of a 30x16 cm image that showed a parallel coordinates visualization with ten axes in parallel. Between each pair of axes a relationship i.e. a pattern in data was depicted. There were six specific stimulus patterns, which were constructed from a synthetic data set containing…” The stimuli should also be illustrated by figures to further enhance understanding of their appearance. If not explained earlier this is also where you describe any data set used in the study. If it is a well known data set, it is enough to cite it. Otherwise the data set should be clearly defined to allow for verification and replication. It is possible to refer to previous publications or a web source to reduce the description but for best practice sufficient detail should still be included so that the description is self-contained. Any additional information of relevant aspects of the circumstances in which the evaluation was conducted is also crucial to mention and motivate. Examples here are customization of a technique such as restricted rotation to avoid confusion for inexperienced participants or whether the stimuli are only implemented for the evaluation (instead of being authentic) to allow full control and clearly defined answers. These are common and valid approaches in evaluation but they all need to be thoroughly explained and motivated with sound arguments. This section should also describe any other materials used such as questionnaires. If these instruments are newly designed their content should be well described, if they are validated report its name and reference.

4.4 Apparatus and usage condition (or Equipment)
This is the place to describe the display system and other equipment used, and also what were the settings of the equipment and usage conditions. This includes the type of computer, response apparatus and other hardware, and also what software was used to present stimuli and record responses. For example: “The images were rendered using OpenGL”. “A program was constructed that showed each stimulus against a black background”. The name of the manufacture’s products may sometimes be needed, e.g., “The computers were equipped with Nvidia TNT graphics cards”. How the participants were seated or standing should be specified and, when relevant, detail specific conditions such as lighting, viewing distance, and visual angle.

4.5 Experimental design
In this subsection you provide a description of the structure of the evaluation, i.e. the experimental design. This includes specification of what design was used, what were the independent variables and how many levels did each of them have, and also their nature in terms of being a within- or between-subject variable. For example: “The study was performed as a three-variable mixed design with one within-subject variable having two levels: task type A (simple) vs. task type B (complex). The two between-subject variables were: type of visualization (2Dm vs. 3Dm vs. 2Da) and order of presentation of task type (task A or task B first)”. The motivation for choosing a certain design could also be emphasized, for example motivating the nature of a certain variable: “Having order of presentation of the different tasks as a between-subjects variable was chosen to avoid any systematic effects of the different orderings and also provide an opportunity to statistically determine whether there are any effects of the different orderings”. This shows the author’s awareness of a potential impact or bias on the outcome and how this was considered/handled. The above description also needs to be supplemented with a specification of the resulting number of conditions, what actions were taken to assign participants to conditions and how tasks/stimuli were presented in each. This includes describing procedures for randomization, blocking and balancing. For example: “Participants were randomly assigned to one of the three types of visualization methods”. It is also appropriate to describe here how many trials each condition consisted of and how many trials this yielded in total per participant. For example: “Each participant viewed a total of 50 (2x25x1) images”. The figures within brackets illustrate the conditions and number of trials in each (2 different task types x 25 trials per type x 1 visualization method). Description of design types and assignment procedures can be obtained from [5, 6].

4.6 Procedure
This part of the method description gives details of how the evaluation was conducted in a practical sense. Many aspects need to be covered and it is important to make the text easy to follow and not lengthy. Still, there should be sufficient information about the different phases to provide the reader with a clear picture of what happened from the moment the participants arrived to the moment they left. Examples of common parts to cover are: introduction, filling in forms (demographic information, informed consent etc.), review of instruction material, training phase (if any) and the actual testing phase which often consists of one or more timed-trial session(s). The different parts should be described in chronological order and for each it should be specified what was taking place, how and with what material. Examples of issues to specify here are: whether any feedback or other assistance was given during training, how the task was presented to the participants on the screen and how it was executed, whether the sequence of trials was self-paced or not, how response was given (e.g., clicking on an interface, using the numerical keyboard or orally), and the number of trials and participation time for different parts and in total. This section obviously needs to cover a description of the task(s) to be performed. To enhance readability and good flow a recommendation is to keep this description as short as possible here and give the thorough description with motivation in a
4.7 Results
This section presents the results from applying the method by presenting mainly the data obtained from the evaluation. The reason is that the results should stand for themselves, i.e. a result could be accurate whereas a conclusion based on them may not be. The interpretation and discussion of the results should therefore be left to the final discussion (and conclusion) section. Yet, in some cases it may be appropriate to include a short discussion of the results in the end of the results section, or in a subsection (named discussion of results) especially if the results lead to follow up experiments.

Regarding any obtained demographic information such as gender, or level of experience, this is often not considered as results and is, in the majority of cases, reported in the section describing the participants.

First it should be stated whether all data from all participants were analysed or not. Sometimes data needs to be omitted due to equipment failure, dropouts, missing values for some trials etc. Such cases, and how they may affect the analysis, must be clarified. The section should show awareness of any potential data analysis biases. It should start by reporting any tests done to check whether data is appropriate for parametric testing or not (e.g. check for normality and variance), describe any treatments of data (e.g. logarithmic transformation to fit data to a normal curve) and the outcome of that treatment [4]. This could be explained in detail but when space is an issue it could be enough to state: “Data was analysed with Analysis of Variance (ANOVA) tests. When data did not adhere to the assumptions for this parametric test, non-parametric tests were used”. What response measures (dependent measures) were used for analysis should be clarified, for example if accuracy was measured and you base your analysis using error rates (rather than raw scores of success or failure) you state: “For each participant the mean error rate was calculated and these values were used as dependent measures in subsequent analysis”. Also, since there are, for example, variants of ANOVA tests for different experimental designs it should always be stated which type was used, and state what decision criterion that was applied: “Group mean values were calculated and a between-subject ANOVA was carried out using a decision criterion of 0.05. Variables were visualization type (2D versus. 3D) and order of presentation”. Thereafter you state the finding to which the test relates, report the test statistic, usually with its degrees of freedom, the probability value and associated descriptive statistic. For example, “There was a significant effect of visualization type $F(1,24) = 5.528, p < 0.05$ but no significant effect of order. The response times for the 2D visualization were significantly faster than for the 3D visualization. The group mean value for the search times with 2Dm was 23.9 seconds with a standard deviation of 1.35 while in the 3D condition it was 37.2 seconds with a standard deviation of 1.61”. The descriptive statistics in running text are informative but they can be replaced with a reference to tables/graphs, especially when space is an issue. For example in the beginning of the section it could state “Descriptive statistics are illustrated in figures 7-9” and that would be sufficient.

To conclude, to write the method section is not an easy task. A lot of information needs to be included and a thorough description of a study requires a considerable amount of space. This can be difficult when faced with a page limit, and there is always a trade-off between including irrelevant information and leaving relevant aspects out. A good approach is to have someone without prior knowledge about your study provide feedback about what is missing and what can be left out. When space is critical it is more important than ever to focus on the details most important to the outcome of the experiment and for replication. However, the aim should always be for the description to be as complete as possible. This is how you make it possible for others to assess and verify the value of your work and allow them to replicate it to see if they will reach the same findings.

5. CONCLUSION AND FUTURE WORK
This paper discusses the importance of sound reporting and hopefully serves as a starting point for a standardized format that can be used to guide authors. The contribution is a guide as to how to structure a sound report of an evaluation study with a number of details on what content is important to address and how to do so. It outlines the basics of how to present a text that allows the reader to fully verify the value of the work and that allows replication. The reader of this paper should come away more capable of knowing what is important to consider and how to avoid common pitfalls and mistakes. They should also be more capable of finding potential problems in an evaluation paper when reviewing such work.

This initial work is presented for discussion among evaluation experts, experienced in reporting and reviewing while calling for feedback. The long-term objective is to reach an established and agreed upon set of guidelines providing enough detail and examples for authors to feel confident when reporting.

6. REFERENCES