Evaluating Visualizations

PA214

Vít Rusňák

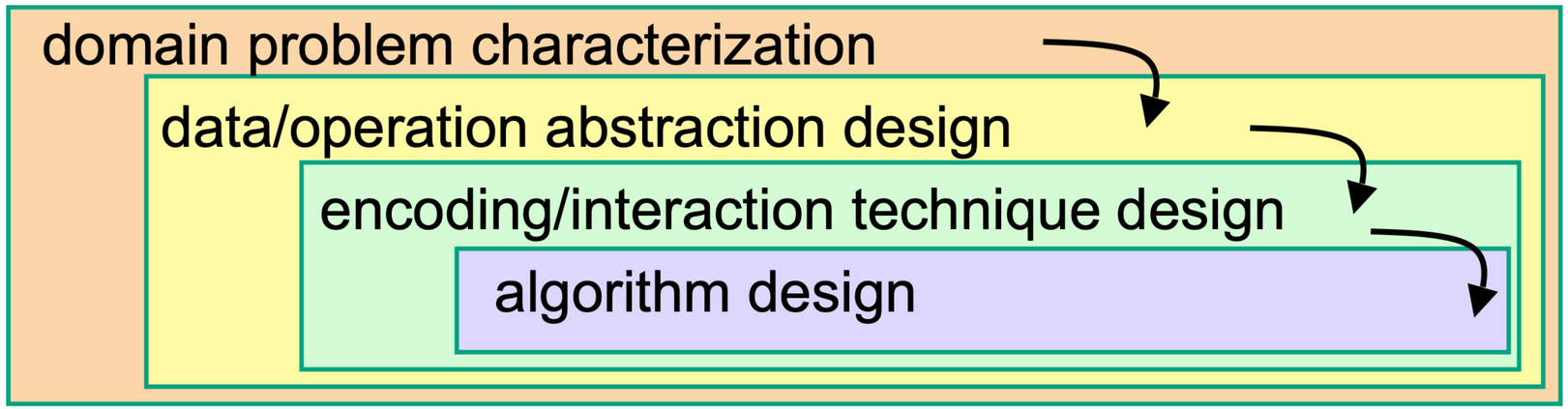
Talk Outline

- Methodologies
- A Bit of Terminology
- Visualization Evaluation Categories
  - Evaluating User Performance
  - Evaluating User Experience
- Doing the Evaluation

Methodologies

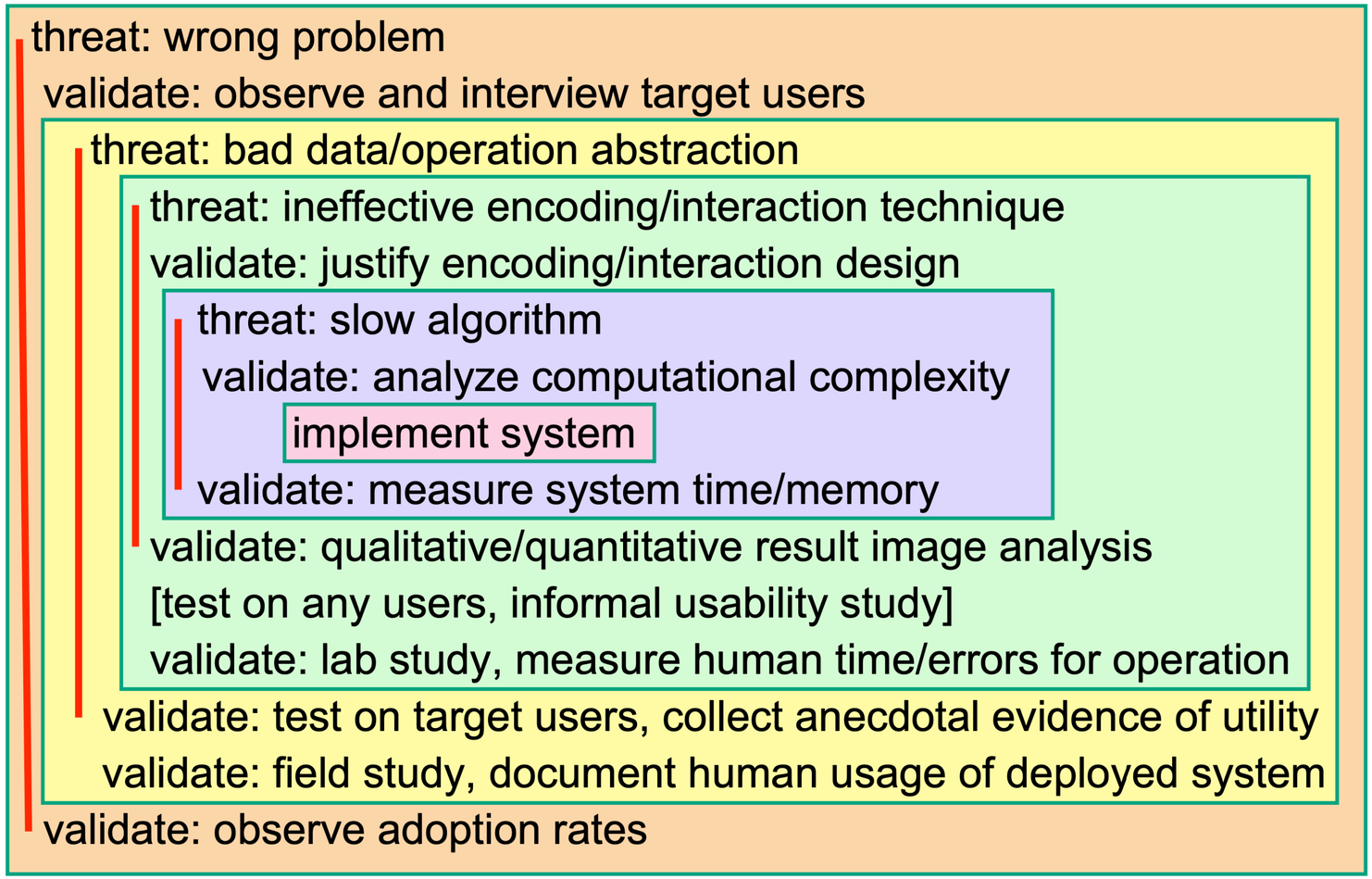
- Systematic approaches
- Should help in the whole process
- Examples:
  - Design Study
  - Nested Model

Nested Model



Source: T. Munzner. A Nested Model for Visualization Design and Validation. IEEE TVCG. 2009

Nested Model



Source: T. Munzner. A Nested Model for Visualization Design and Validation. IEEE TVCG.

Design Study

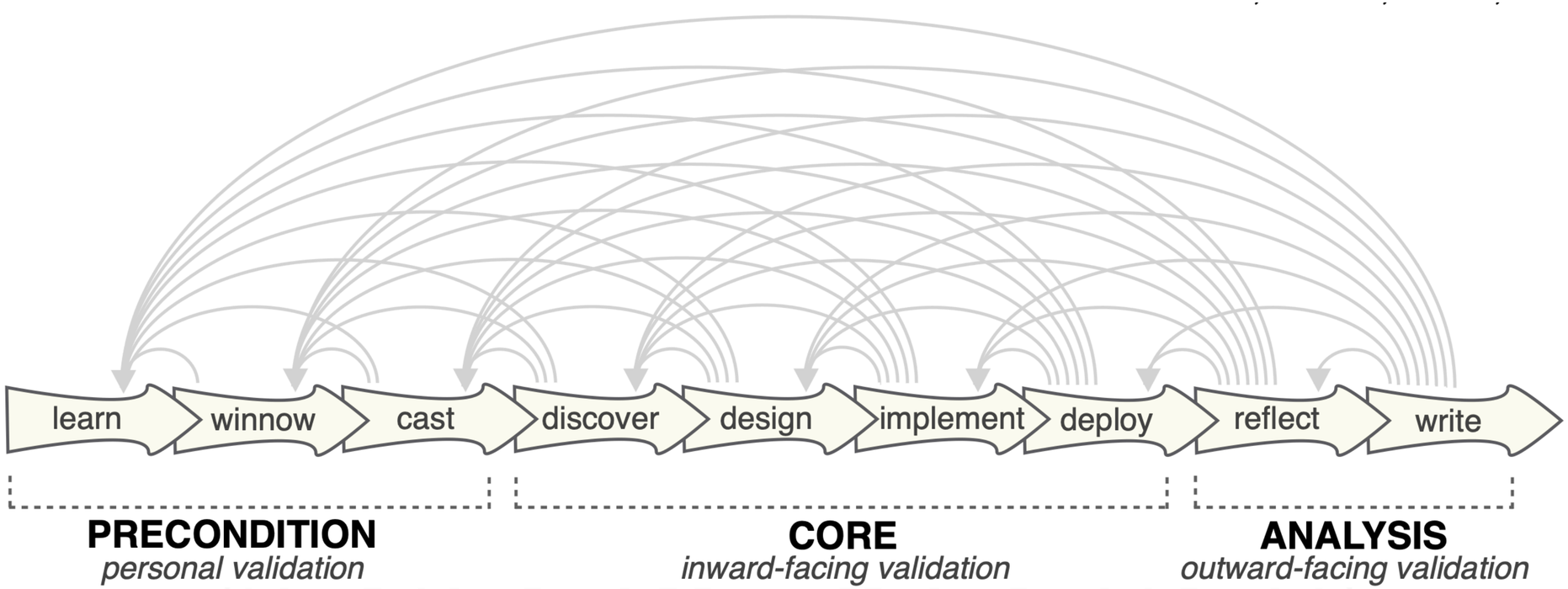


Fig. 2. Nine-stage design study methodology framework classified into three top-level categories. While outlined as a linear process, the overlapping stages and gray arrows imply the iterative dynamics of this process.

Source: M. Sedlmair, M. Meyer and T. Munzner, Design Study Methodology: Reflections from the Trenches and the Stacks. IEEE TVCG. 2012.

Source: http://www.jyukawa.com/images/7050f11-wordle.png

A bit of Terminology

Qualitative vs. Quantitative

TABLE 1.1 Assumed characteristics of research

Qualitative research	Quantitative research
Uses words	Uses numbers
Concerned with meanings	Concerned with behaviour
Induces hypotheses from data	Begins with hypotheses
Case studies	Generalisations

Source: Silverman, David. Doing Qualitative Research: A Practical Handbook. 3rd ed. SAGE, 2010.

Formative vs. Summative

- **Formative** evaluation
  - Typically *qualitative*; takes place at the initial phase of the research project
  - Goals: describe the problem, get better insight, find the same language with target users
- **Summative** evaluation
  - *Qualitative/quantitative*; final phase of the research project
  - Goals: evaluate the results, gain feedback (for the next iteration)

Controlled vs. In-the-Wild

Controlled environment = laboratory conditions

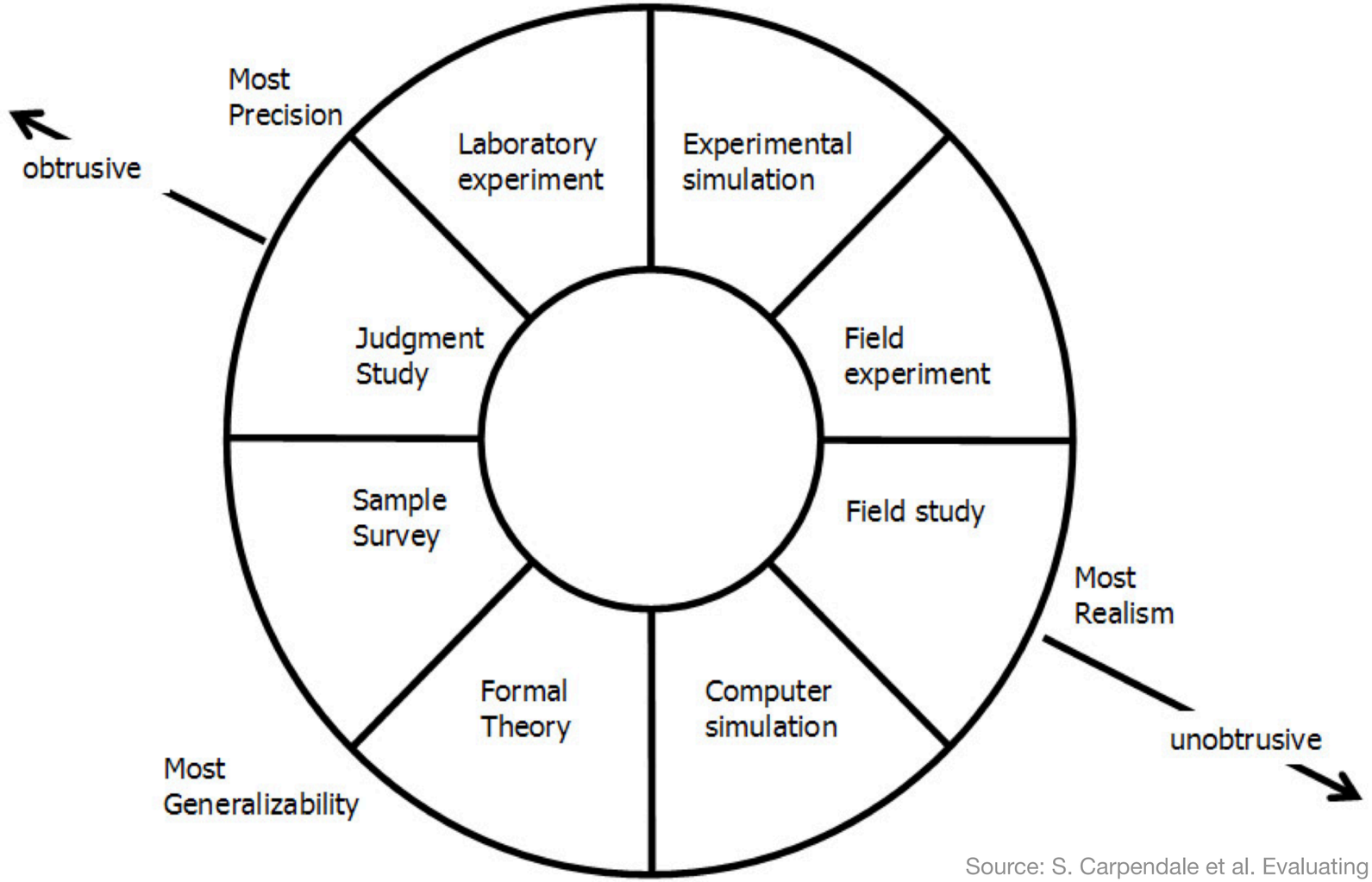
- Pros: elimination of unwanted factors and random variables
- Cons: problem of overfitting, negative influence on the **ecological validity**

In-the-Wild = real-world conditions

- Pros: higher ecological validity and generalizability
- Cons: many uncontrolled conditions

Short term vs. Longitudinal

- **Short term**
  - Usually controlled experiments and performance evaluation
- **Longitudinal**
  - Using methods from grounded evaluation theory
  - Mainly qualitative data (field observations, diaries)



Source: S. Carpendale et al. Evaluating Information Visualizations. 2008.

Visualization Evaluation Categories

Visualization Evaluation Categories

Understanding data analysis processes

- Understanding environments and work practices
- Communication and collaboration
- Visual data analysis and reasoning

Understanding visualization

- Algorithm performance
- Qualitative result inspection
- User experience and performance

Visualization Evaluation Categories

Understanding data analysis processes

- Understanding environments and work practices — formative evaluation
- Communication and collaboration — almost non-existent
- Visual data analysis and reasoning — case studies

Understanding visualization

- Algorithm performance
- Qualitative result inspection
- User experience and performance

Visualization Evaluation Categories

Understanding data analysis processes

- Understanding environments and work practices
- Communication and collaboration
- Visual data analysis and reasoning

Understanding visualization

- Algorithm performance
- Qualitative result inspection
- User experience and performance

Case studies

- “A detailed reporting about a small number of individuals working on their own problems in their normal environment” [1]
  - Case study from domain expert | close collaboration | vis. researcher
  - ~~Usage scenario~~
- Small number of participants (often up to 5)
- New tool/visualization + dataset
- Almost step-by-step description of how the participant use the tool
- Summarized feedback (feature requests, opinion of participants on the tool functions and limits and its applicability in their work)

[1] B. Shneiderman and C. Plaisant. 2006. Strategies for evaluating information visualization tools: multi-dimensional in-depth long-term case studies. BELIV ’06

Visualization Evaluation Categories

Understanding data analysis processes

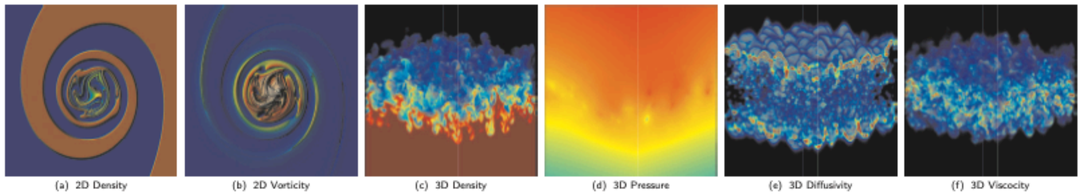
- Understanding environments and work practices
- Communication and collaboration
- Visual data analysis and reasoning

Understanding visualization

- Algorithm performance — benchmarking
- Qualitative result inspection — qualitative inspection, heuristics
- User experience and performance —common for Vis and HCI communities

Algorithm Performance

- Quantitative
- Usually benchmarking and reporting performance of a (novel) algorithm or technique
- Computation time, rendering speed (fps), memory footprint, …
- The importance of test datasets and their availability



(a) 2D Density (b) 2D Vorticity (c) 3D Density (d) 3D Pressure (e) 3D Diffusivity (f) 3D Viscocity

Fig. 1. Visualizations of 2D data (as pseudocolored height fields) and 3D data (volume rendered) used in our experiments.

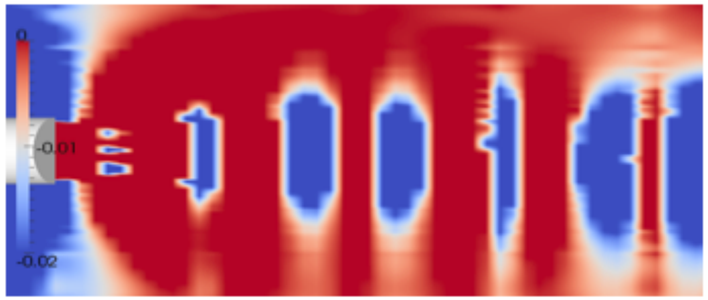
data set								compressed size (MB) and compression time (seconds)									
name	unique (%)	entropy (bits)	range (bits)	min	max	size (MB)	time (sec)	zlib		[RKB2006]		[EFF2000]		[ILS2005]		new scheme	
m2d density	3.89	3.49	21.83	8.7E−01	1.2E+00	19.6	0.71	1.6	0.86	4.3	**0.49**	4.4	0.56	**1.3**	1.08	1.3	0.56
m2d vorticity	99.20	22.25	31.05	-1.4E+02	2.5E+01	19.6	0.71	18.4	2.14	**11.8**	**1.21**	15.5	1.29	12.9	2.22	13.8	1.49
m3d density	7.67	5.16	23.60	1.0E+00	3.0E+00	364.5	12.81	50.4	17.55	100.5	9.06	96.3	**8.48**	35.7	19.03	**35.5**	9.25
m3d pressure	27.29	23.91	31.06	-3.7E+00	2.3E+03	364.5	12.80	229.2	99.76	95.6	9.31	87.9	**8.87**	**40.1**	18.79	40.4	9.96
m3d diffusivity	36.87	23.19	30.02	0.0E+00	6.8E+00	364.5	12.68	297.6	42.90	250.8	19.09	239.3	**15.02**	**198.8**	31.92	203.0	18.47
m3d viscocity	50.07	24.86	28.59	8.6E−15	2.9E+00	364.5	12.62	314.0	46.09	249.4	18.95	246.1	**14.68**	209.2	32.66	**207.5**	19.45
h3d temp	65.70	23.54	31.56	-7.7E+01	1.0E+35	95.4	3.77	75.8	14.56	59.3	4.64	53.0	**4.27**	**44.1**	8.04	44.1	5.06
h3d pressure	81.82	24.13	31.58	-3.4E+03	1.0E+35	95.4	3.78	82.3	12.00	64.3	5.14	52.9	**4.87**	**45.0**	7.78	45.2	5.34
h3d x velocity	84.18	24.18	31.55	-5.3E+01	1.0E+35	95.4	3.89	86.1	11.27	67.4	6.22	63.3	**4.59**	**54.5**	8.86	55.4	5.44
h3d y velocity	84.32	24.18	31.55	-4.6E+01	1.0E+35	95.4	3.83	84.5	11.42	67.1	5.74	62.3	**5.04**	**53.5**	8.64	53.8	5.53
h3d z velocity	86.82	24.24	31.54	-3.2E+00	1.0E+35	95.4	3.87	88.4	10.76	85.6	8.50	76.9	**5.29**	**68.9**	9.83	69.1	6.65
M3d density	40.14	18.84	52.59	1.0E+00	3.0E+00	288.0	11.28	136.8	41.91	160.3	11.63	121.6	**10.94**	-		**105.2**	11.63
M3d pressure	100.00	25.17	63.00	-2.2E+00	2.2E+00	288.0	11.20	272.6	35.18	237.3	**14.91**	225.1	16.59	-		**208.4**	17.20
M3d x velocity	100.00	25.17	63.00	-2.2E+00	2.3E+00	288.0	10.83	275.6	32.30	230.4	**14.73**	215.1	15.91	-		**197.7**	16.84
M3d y velocity	100.00	25.17	63.00	-2.1E+00	2.3E+00	288.0	10.54	275.1	32.19	223.1	**14.27**	215.2	15.16	-		**197.7**	16.65
M3d z velocity	100.00	25.17	63.00	-5.2E+00	9.0E+00	288.0	10.32	275.5	32.62	226.6	**14.74**	213.7	16.05	-		**196.8**	16.14
atom x position	61.10	23.82	31.01	-4.8E−02	4.6E+02	107.7	7.07	84.3	21.18	76.0	7.88	78.8	**7.61**	**67.3**	12.88	68.6	9.07
atom y position	45.90	23.32	26.99	3.7E−02	2.1E+03	107.7	7.08	65.9	30.76	60.4	6.97	56.4	**6.31**	47.0	10.49	**46.9**	7.73
atom z position	61.68	23.84	27.48	9.1E−05	4.6E+02	107.7	7.46	94.6	19.86	82.6	9.00	86.1	**8.25**	**75.7**	13.80	78.2	9.93
atom y velocity	64.65	23.87	30.96	-1.5E−01	1.4E−01	107.7	7.30	95.7	19.88	93.8	10.07	99.1	**9.65**	**84.3**	14.93	87.6	9.92
atom temp	64.91	23.94	27.41	3.0E−03	7.1E+03	107.7	6.69	95.7	19.76	91.6	10.27	95.9	**8.34**	**84.6**	15.02	84.6	10.31
atom energy	3.45	18.57	21.79	-3.6E+00	-2.7E+00	107.7	7.15	77.9	38.59	74.1	7.98	71.8	**7.01**	60.8	12.66	**60.5**	8.30
lucy	61.39	24.38	31.09	-6.1E+02	1.2E+03	160.5	-	137.8	-	99.5	-	90.0	-	**73.6**	-	77.8	-
david$_{1mm}$	25.23	17.08	31.11	-4.4E+03	1.8E+03	322.5	-	144.9	-	155.7	-	163.4	-	**108.6**	-	131.9	-
torso	84.72	18.48	31.08	-2.7E+02	5.8E+02	1.9	-	1.7	-	1.5	-	1.5	-	**1.3**	-	1.3	-
rbl	71.90	20.14	25.99	1.5E+00	3.6E+02	8.4	-	7.1	-	5.8	-	5.6	-	**4.7**	-	4.8	-

Table 1. Compression results for the Miranda (m2d, m3d, M3d) and hurricane (h3d) structured grids, the atom point set, the lucy and david triangle meshes, and the torso and rbl tetrahedral meshes. All data but M3d is represented in single precision. The [ILS2005] scheme operates on single precision only, hence the missing values. For the meshes we report only the compressed size of vertex coordinates; timings are dominated by connectivity coding, and are hence excluded. The range measures (the logarithm of) the number of floating-point values between min and max. Note that the first-order entropy is limited by the number of samples in a data set.

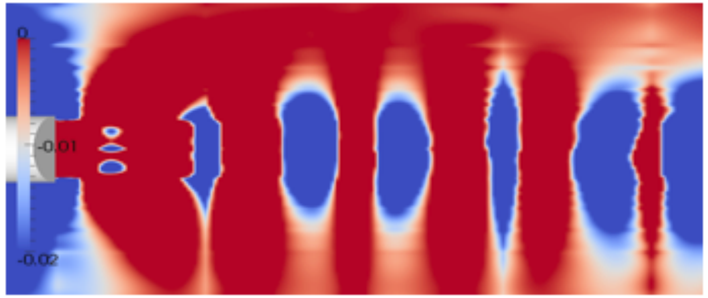
Source: P. Lindstrom and M. Isenburg, "Fast and Efficient Compression of Floating-Point Data," in *IEEE Transactions on Visualization and Computer Graphics*, vol. 12, no. 5, pp. 1245-1250, Sept.-Oct. 2006.

Qualitative Results Inspection

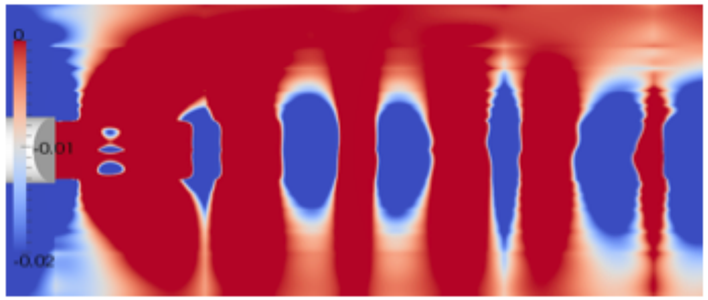
- Three main types:
  - Image quality
  - Visual encoding
  - Walkthrough
- Comparative vs. isolated



(a) Cut-plane with 902,289 triangles, VTK Rendering Time = 0.08 seconds.



(b) Cut-plane with 8,388,608 triangles, VTK Rendering Time = 2.0 seconds.



(c) Pixel-exact cut-plane color map. Rendering time is 0.015 seconds for a 1800×800 image.

Source: B. Nelson, R. M. Kirby and R. Haimes, "GPU-Based Interactive Cut-Surface Extraction From High-Order Finite Element Fields," in *IEEE Transactions on Visualization and Computer Graphics*, vol. 17, no. 12, pp. 1803-1811, Dec. 2011.

Qualitative Discussions and Heuristics

- Performed by visualization researchers
- Do not involve end users or participants
- Based on the objective assessment
- Objective description of the result
- Heuristic evaluation

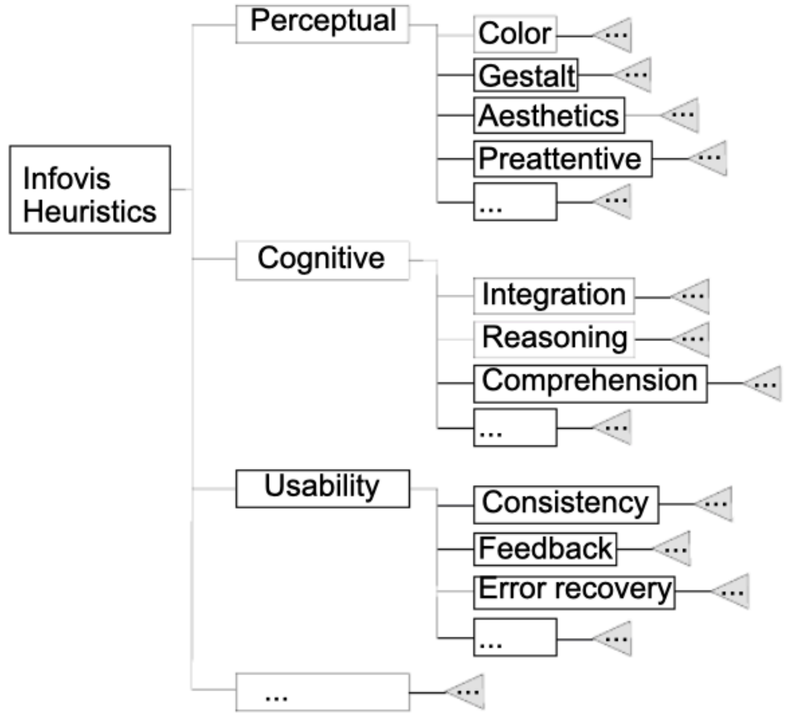
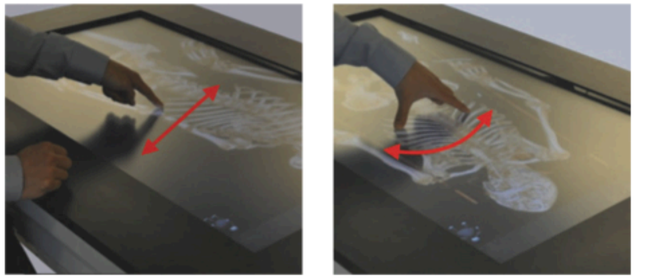


Figure 1: Evaluation Tree.

Source: T. Zuk, et al. Heuristics for information visualization evaluation. BELIV '06. DOI:10.1145/1168149.1168162

User Performance and Experience

- The most common types
- Controlled experiments with ***participants***
- **User performance** (quantitative)
  - time and resource intensive (10+ participants)
  - time and/or errors using new technique
  - comparison against the automatic technique
- **User experience** (qualitative)
  - feedback from experts, reports from demonstrations



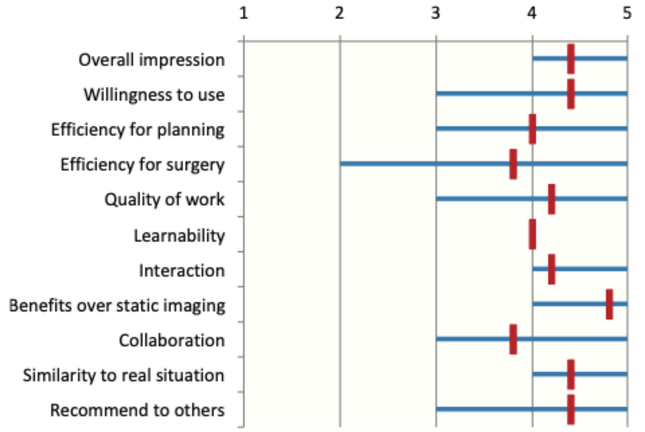


Fig. 12. The quantitative results of the user study questionnaire. Subjective satisfaction regarding use of the table was measured for 11 questions, see section 6. The 5-point rating scale ranges from *Strongly unfavorable* (1) through *Unsure (3)* to *Strongly favorable* (5). Vertical red bars denote the mean value and horizontal blue lines denote the full span of given ratings.

Source: C. Lundstrom, et al., "Multi-Touch Table System for Medical Visualization: Application to Orthopedic Surgery Planning," *IEEE TVCG*, 2011.

User Performance: Terminology

- Independent variable (test conditions)
- Dependent variable (measured behaviors)
- Control variable
- Random variable
- Confounding variable
- Participant
- Within subjects vs. between subjects
- Counterbalancing & Latin Square

Independent vs Dependent Variables

- **Independent** variable (also factor)
  - a circumstance that is manipulated through the design of the experiment
  - independent of participant behavior (i.e., there is nothing a participant can do to influence it)
  - examples: interface, device, visual layout, expertise, age, gender
- **Dependent** variable
  - any measurable aspect of the interaction involving a factor
  - examples: task completion time, error rate, accuracy, throughput
  - make sure you identify all of them

Control Variable

- *A circumstance (not under investigation) that is kept constant to test the effect of an independent variable.*

- More control => the experiment is less generalizable

- Example: measure effect of font color and background color on reader comprehension
  - independent variables: font color, background color
  - dependent variables: comprehension test scores
  - control variables: font size, font family, ambient lighting

Random Variable

- *A circumstance that is allowed to vary randomly.*
- Outcomes => more generalizable results (good) but also more variability in the measures (bad)
- Example: the amount of coffee consumed prior to testing

Confounding Variable

- *A circumstance that varies systematically with an independent variable.*
- Should be controlled or randomized to avoid misleading results
- Example: prior experience of participants

Participants

- People ***participating*** in the experiment (don't use ~~subjects~~)
- How many?
  - Short answer: use the same number as used in similar research
  - Too many: unnecessary work
  - Too few: fail to get statistically significant results => paper reject

Within vs. Between Subjects

Within-subjects design
The same participant tests all conditions corresponding to a variable.

Between-subjects design
Different participants are assigned to different conditions corresponding to a variable.

Source: https://www.nngroup.com/articles/between-within-subjects/

Counterbalancing

- Compensation of (unwanted) learning effect
- The order of tasks or datasets used in the experiment
  - (pseudo)Randomized order — generate one for each participant
  - Latin Square - an $n \times n$ array filled with n different symbols, each occurring exactly once in each row and column (i.e., Sudoku).

Evaluating User Performance

- Gathering evidence, not proving things (mathematicians do)
- **Hypotheses testing**
  - Null hypothesis: “There is no difference between A and B”
- *Parametric tests:* ANOVA, t-test, F-test, …
- *Non-parametric tests:* Chi-square, Mann-Whitney’s U test, Friedman test, …
- https://yatani.jp/teaching/doku.php?id=hcistats:start

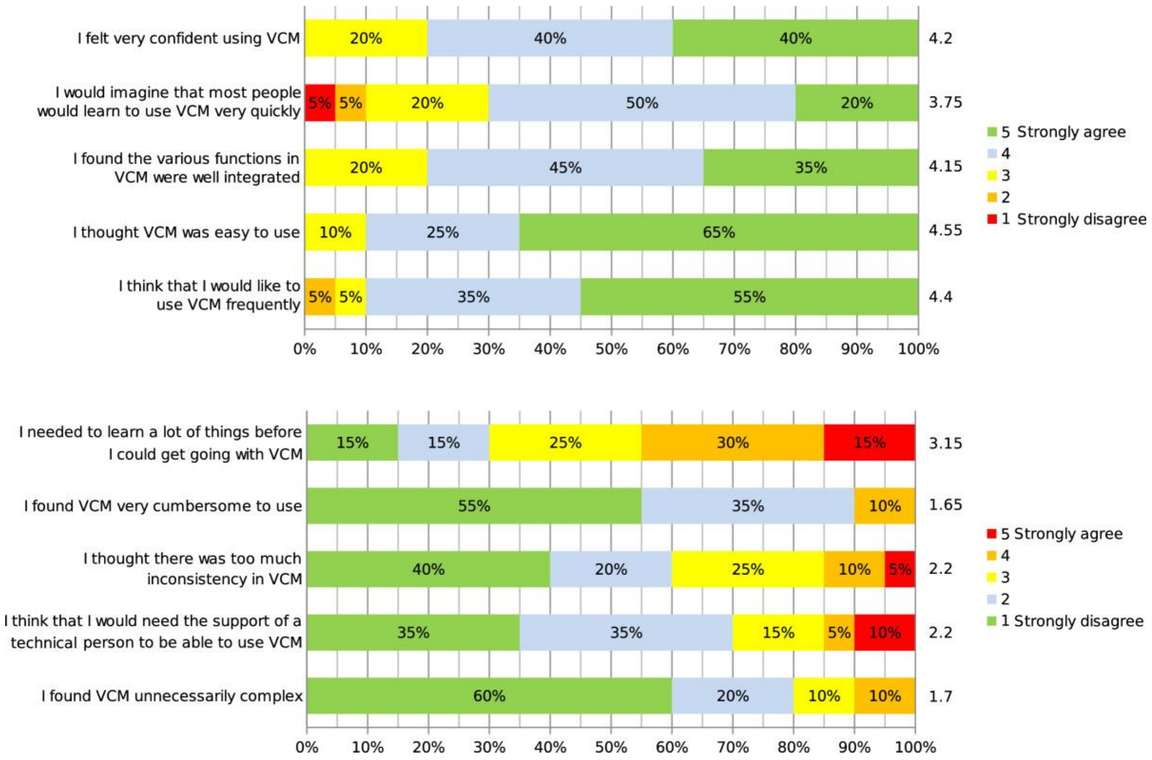
Evaluating User Experience

- Interpretation of (standardized) questionnaire results
- Synthesis of anecdotal experience (direct quotes)
- Grounded evaluation methods — diaries, observations
- (Open) Coding — the process of subdividing and labeling raw data in order to form a theory

Standardized Usability Questionnaires

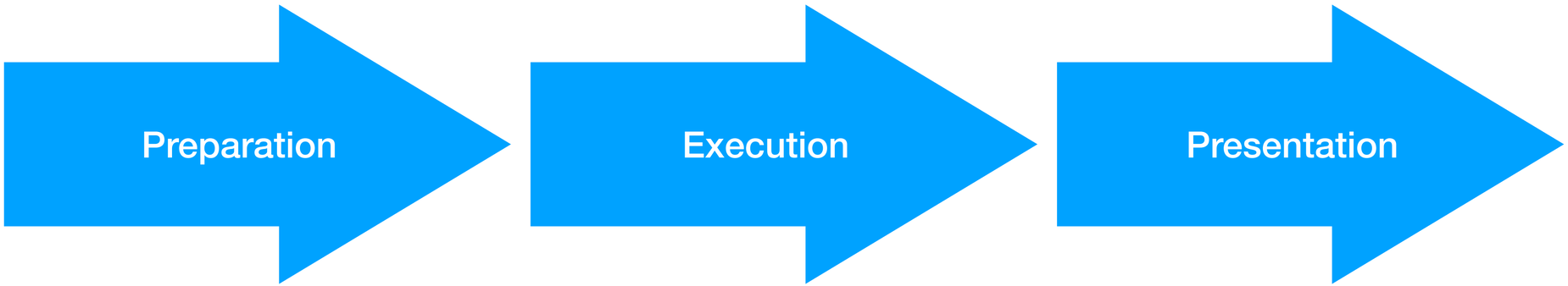
“Questionnaires designed for the assessment of perceived usability, typically with a specific set of questions presented in a specified order using a specified format with specific rules for producing scores based on the answers of respondents.”
[3. Sauro, Lewis, Quantifying the User Experience, 2016]

- There is plenty of them:
  - Post-task: SEQ, NASA-TLX, …
  - Post-study: SUS, SUMI, PSSUQ, …
- Benefits: objectivity, replicability, quantification, economy, communication, scientific generalization



Source: Pereira, S., Hassler, S., Hamek, S. *et al.* Improving access to clinical practice guidelines with an interactive graphical interface using an iconic language. *BMC Med Inform Decis Mak* **14,** 77 (2014). https://doi.org/10.1186/1472-6947-14-77

Doing the Evaluation



Data
Workflow
Documents
Pilot execution

Introduction
Demonstration
Familiarization
Debriefing

Data analysis
Outcomes
Summarization

Preparation

- Set the goal, then choose the method
- Prepare data and related documents, datasets, consent forms, questionnaires (pre-, post-)
- Always do the pilot or dry run => identification of unexpected problems
- Make a checklist — always follow the same steps
- Get the participants

Consent Form

- Who you are
- What you are asking the participants to do
- What kind of data you will be collecting and how it will be used
- What rights the participant has
- If they will be compensated
- The participant must explicitly say "yes" to the consent form

SIMON FRASER UNIVERSITY

INFORMED CONSENT BY SUBJECTS TO PARTICIPATE IN EVALUATION OF AN INTERACTIVE COMPUTER SYSTEM FOR DATA VISUALIZATION

The University and those conducting this project subscribe to the ethical conduct of research and to the protection at all times of the interests, comfort, and safety of subjects. This form and the information it contains are given to you for your own protection and full understanding of the procedures. Your signature on this form will signify that you have received a document which describes the procedures, possible risks, and benefits of this research project, that you have received an adequate opportunity to consider the information in the document, and that you voluntarily agree to participate in the project.

Knowledge of your identity is not required. You will not be required to write your name or any other identifying information on the research questionnaires. An audio recording of your voice and a video recording of the computer screen only will be made during the session. The video and audio recordings of the session will be reviewed only by the Principal Investigator. All research materials will be held confidential by the Principal Investigator and kept in a secure location. These research materials will be destroyed after the completion of the study.

Having been asked by Daryl H. Hepting of the School of Computing Science of Simon Fraser University to participate in a research project study, I have read the procedures specified in the accompanying information sheet. I understand the procedures to be used in this study and the personal risks and benefits to me in taking part. I understand that I may withdraw my participation in this study at any time.

I understand that my decision to participate in this study, and my subsequent involvement in it, will have absolutely no bearing on any other dealings I have with Mr. Hepting. This includes, but is not limited to, the case that I am a student in the CMPT 361 course taught by Mr. Hepting, offered at SFU during the 99-2 semester.

I understand that I may register any complaint I might have about the study with the researcher named above or with Dr. Jim Delgrande, Director, School of Computing Science of Simon Fraser University, Burnaby, BC, V5A 1S6, telephone 604-291-4277.

I may obtain copies of the results of this study, upon its completion, by contacting Mr. Daryl Hepting, in care of the School of Computing Science at Simon Fraser University.

I understand that my supervisor or employer may require me to obtain his or her permission prior to my participation in a study such as this.

I agree to participate by completing: a pre-task questionnaire; a training session on the prototype software system; a task with the prototype software system; and a post-task questionnaire. I understand that these activities will require approximately one hour at a time scheduled with Mr. Hepting. I understand that the experiment will be conducted in Room 9836 in the Applied Science Building of Simon Fraser University.

NAME (please type or print legibly): _____

ADDRESS: _____

_____

SIGNATURE: _____

WITNESS: _____

DATE: _____

A COPY OF THIS SIGNED **CONSENT FORM** AND A **SUBJECT FEEDBACK FORM** WILL BE PROVIDED TO YOU AT YOUR EXPERIMENT SESSION.

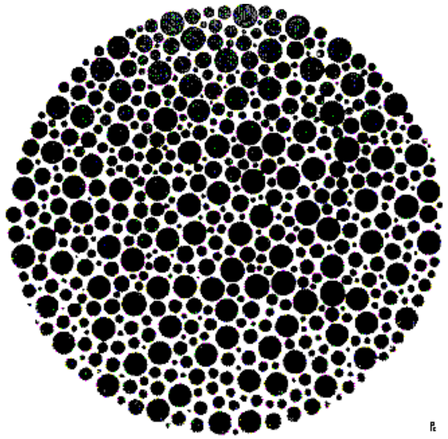
Source: D. Hepting. “A New Paradigm for Exploration in Computer-Aided Visualization”. Dissertation thesis. Simon Fraser University. 1999.

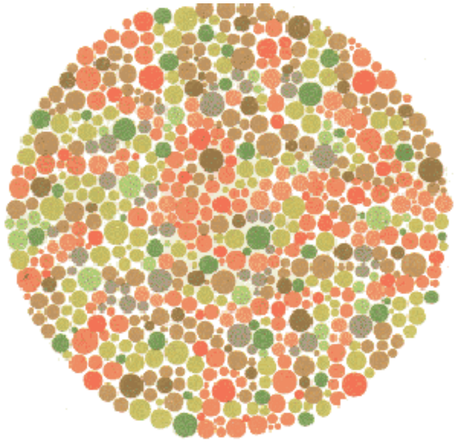
Demographic Questionnaire

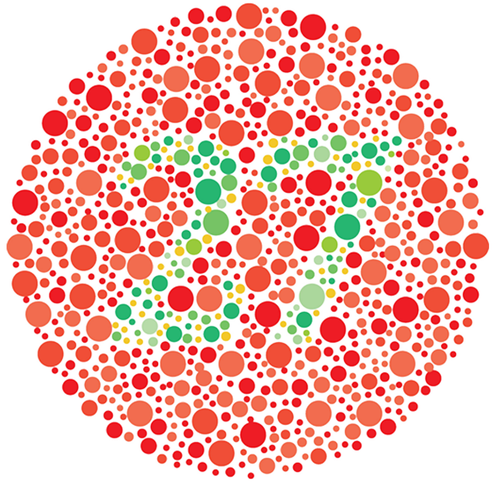
- Age — be careful when doing experiments with <18yo
- Gender — female, male, prefer not to say
- Occupancy, experience and other **relevant** information

Color Perception Test

- Shinobu Ishihara, 1917
- Ishihara plates
- Diagnostic test for color perception deficiencies
- 38 plates (full set)
- Variants with 10, 12 or 24







Execute

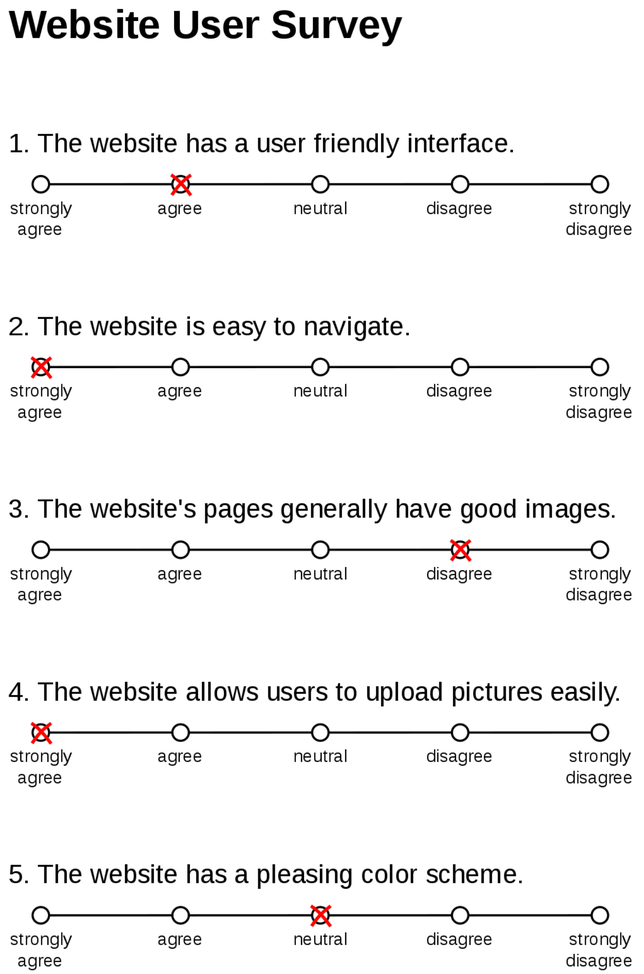
- Follow the checklist
- Do not change experiment design or conditions after starting it
- Get consent first, debrief participants afterwards
- Use different dataset for practice trials and main experiment
- Record: audio/video, mouse traces, make notes

Presenting Results

- Introduce the number of **participants** and their demography
- Describe the **environment**
  - Place where the evaluation happened, screen size and resolution
- Describe the **procedure** (protocol) in details
  - Step-by-step protocol, duration of parts and the whole
- Outline **design** — key for the reproducibility
  - Dependent and independent variables, used datasets, …
- Discuss the **results**
  - use of statistical tests, charts, tables, summarizing ***important outcomes*** *and deriving* ***insights***

Likert Scales

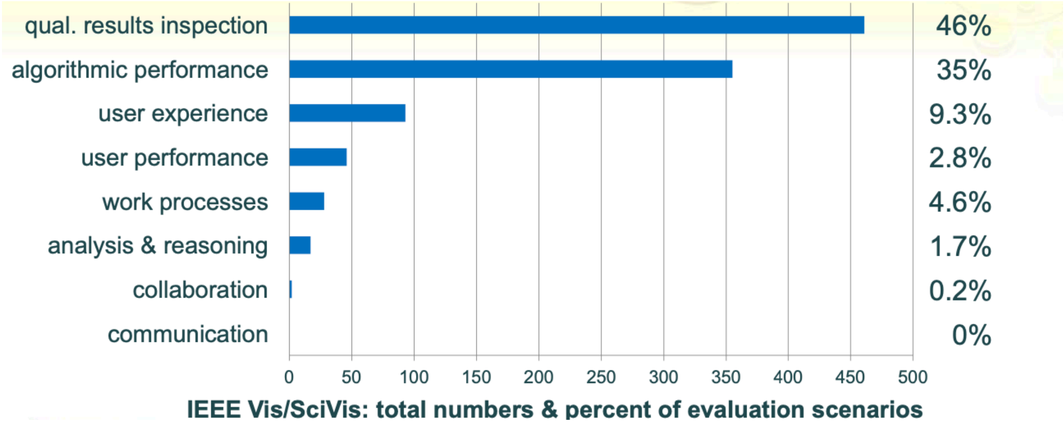
- Statement soliciting level of agreement
- Gradations between responses are (more or less) equal
- Ordinal data => Be careful with averaging (median is often better)
- Even vs. odd number of options



Source: https://en.wikipedia.org/wiki/Likert_scale

Take away…

- In SciVis, InfoVis, VAST, we mostly do:
  - algorithm benchmarking, user performance (quantitative)
  - case studies, qualitative inspection, user experience (qualitative)
  - not find many longitudinal evaluations, diary studies or ethnography methods
- Doing the evaluation right is very tricky
- Use **methodologies** and **best practices** from the field (learn from papers)
- Contribution of **real users** is invaluable but also painful (involve them ASAP)



Source: http://tobias.isenberg.cc/personal/papers/Isenberg_2013_SRP_Slides.pdf

References

1. J. Lazar, J. H. Feng, and H. Hochheiser. 2010. Research Methods in Human-Computer Interaction. Wiley Publishing.

2. I. S. MacKenzie. 2013. Human-Computer Interaction: An Empirical Research Perspective (1st. ed.). Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

3. J. Sauro and J. R. Lewis. 2016. Quantifying the User Experience, Second Edition: Practical Statistics for User Research (2nd. ed.). Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

4. T. Munzner. 2009. A Nested Model for Visualization Design and Validation. IEEE Transactions on Visualization and Computer Graphics 15, 6 (November 2009), 921–928. DOI:https://doi.org/10.1109/TVCG.2009.111

5. M. Sedlmair, M. Meyer and T. Munzner, "Design Study Methodology: Reflections from the Trenches and the Stacks," in *IEEE Transactions on Visualization and Computer Graphics*, vol. 18, no. 12, pp. 2431-2440, Dec. 2012.

6. T. Isenberg, P. Isenberg, J. Chen, M. Sedlmair and T. Möller, "A Systematic Review on the Practice of Evaluating Visualization," in *IEEE Transactions on Visualization and Computer Graphics*, vol. 19, no. 12, pp. 2818-2827, Dec. 2013.