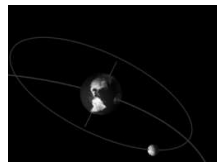


Assessing Visualization Tools with Quantitative Methods

Shaaron Ainsworth
School of Psychology & Learning Sciences
Research Institute
University of Nottingham

Imaginary Scenario

- ◆ Understanding Earth's rotation around the sun.
 - Children have many misconceptions about earth's motion in space, relation to the sun, causes of seasons, etc. So they interact with a 3d animation which shows cycles of days, months, years, tilt of the earth and its rotation. They zoom and pan to see the relative sizes of the earth, moon and sun. Later, they write essays on the causes of day and night, and the patterns of low and high tides.



Evaluation – my goal is to discover whether this animation enhances these children's understanding

Four Key Questions to Answer

- ◆ What are appropriate forms of measurement?
- ◆ What is an appropriate design?
- ◆ What is an appropriate form of comparison?
- ◆ What is an appropriate context

One important distinction

- ◆ Twisting Salomon (1991): Effects with or effects of Visualisation
 - Effects with technology: amplify someone's performance by using technology (e.g. by performing complex calculations).
 - effects of technology: amplifying someone's skills and knowledge as a result of using technology to bring about lasting changes - cognitive residue
- ◆ If scientific visualization is concerned with helping us understand scientific information then although it is partly about the former, it is mostly about the latter.

Four Key Questions to Answer

- ◆ What are appropriate forms of measurement?
- ◆ What is an appropriate design?
- ◆ What is an appropriate form of comparison?
- ◆ What is an appropriate context

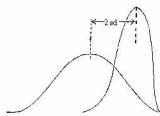
Common Measures (Dependent Variables)

- ◆ Knowledge (gains)
 - Post-test performance
 - Post-test – Pre-test
 - (Post-test – Pre-test)/Pre-test: to account for high performers
- ◆ Efficiency
 - Does it reduce time spent in learning and performance
- ◆ How the system is used in practice (and by whom)
 - What features in a SciVis are used
- ◆ User's attitudes
 - Beware happy sheets
- ◆ Cost savings
- ◆ Teachbacks
 - How well can people now teach what they have learnt

Knowledge Gains: Effect Size

(Gain in Experimental – Gain in Control)/ St Dev in Control

Comparison	Effect size
Classroom teaching v Expert Tutoring	2 sd
Classroom teaching v Non Expert Tutoring	0.4 sd
With Visualizations V Without	??

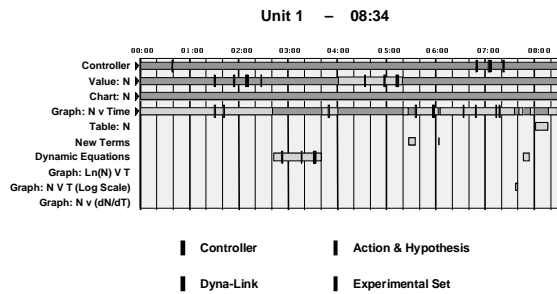


A 2 sigma effects means that the average tutored student performed as well as the top 2 percent of those receiving classroom instruction

Interaction Data

- ◆ Time on task
- ◆ Progression through curriculum or environment
- ◆ Use of system features, for example:
 - which representations
 - in which order
 - Interacted with in what way (e.g. zooming, action, for which dimensions of information)

An Example of Interaction Data: DEMIST Traces (Van Labeke & Ainsworth, 2002)



Process Data

- ◆ Protocols
- ◆ Dialogue turns
- ◆ Gesture and Non-verbal behaviour
- ◆ Eye movement data
- ◆ Poor men's eye tracker
- ◆ Brain Imaging ...

DV Summary

- ◆ Rarely the case that a single DV will be sufficient
- ◆ Could look for more innovative outcome measures in education situations
- ◆ Beware the Law of Gross Measures
 - Subtle questions require subtle DVs which may be impossible in many situations
- ◆ Interaction data often got for free and it's a crime not to look at it! But it may not mean what you think it does....
- ◆ Process data hard work but often worth it.
- ◆ Capturing interaction data rarely changes learners' experiences, but capturing process data often does.
- ◆ The Future: more integration of quantitative data with process data (such as video) as e-science and e-socialscience provides us with the tools...

Four Key Questions to Answer

- ◆ What are appropriate forms of measurement?
- ◆ What is an appropriate design?
- ◆ What is an appropriate form of comparison?
- ◆ What is an appropriate context?

Experimental Methods

- ◆ State a causal hypothesis
- ◆ Manipulate independent variable
- ◆ Assign subjects randomly to groups
- ◆ Use systematic procedures to test hypothesised causal relationships
- ◆ Use specific controls to ensure validity

Quasi-Experimental Methods

- ◆ State a causal hypothesis
- ◆ Include at least 2 levels of the independent variable
 - we may not be able to manipulate it
- ◆ Cannot assign subjects randomly to groups
- ◆ Use specific procedures for testing hypotheses
- ◆ Use some controls to ensure validity

Prototypical experimental designs

- ◆ (intervention) post-test
- ◆ Pre – (intervention) - post-test
- ◆ Pre – (intervention) - post-test – delayed test
- ◆ Interrupted time-series
- ◆ Cross-over
- ◆ Illustrated with a simple comparison
 - Good old fashioned tables
 - Sexy new SciVis

	1	2	3	4
1	11	7	3	56
2	23	43	13	4
3	34	2	12	54
4	2	1	77	32
5	1	99	5	5



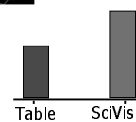
Post-test

	1	2	3	4
1	11	7	3	56
2	23	43	13	4
3	34	2	12	54
4	2	1	77	32
5	1	99	5	5

⇒ Post-test



⇒ Post-test



Post-test

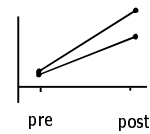
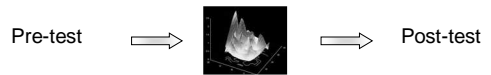
◆ Advantages

- Quick

◆ Disadvantages

- A lot!
- Need random allocation to conditions
- Can't account for influence of prior knowledge on performance or system use

Pre-test to Post-test



Pre-test to Post-test

◆ Advantages

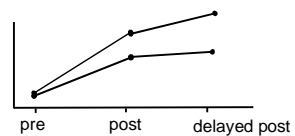
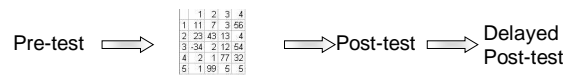
- Better than just post-test as can help explain why some people understand more than others
- Can show whether prior knowledge is related to how system is used
- If marked prior to study can be used to allocate subjects to groups such that each group has a similar distribution of scores

◆ Disadvantages

- No long term results
- Can not tell **when** improvement occurred if long term intervention



Pre-test to Post-test to Delayed Post-test



Pre-test to Post-test to Delayed Post-test

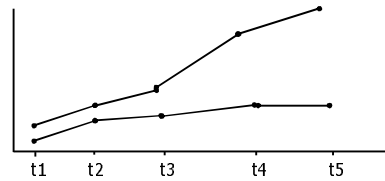
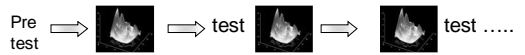
◆ Advantages

- Does improvement maintain?
- Some results may only manifest sometime after intervention
- Different interventions may have different results at post-test and delayed post-test (e.g. individual and collaborative learning)

◆ Disadvantages

- Practical
- Often find an across the board gentle drop off

Interrupted Time-Series Design



Interrupted Time-Series Design

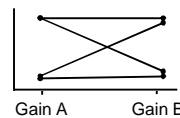
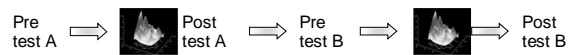
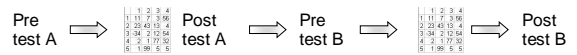
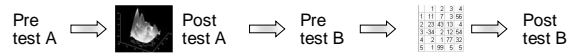
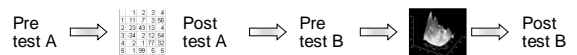
◆ Advantages

- Time scale of improvement
- Ceiling effects

◆ Disadvantages

- Time-consuming
- Effects of repeated testing

Full Cross-over



Full Cross-over

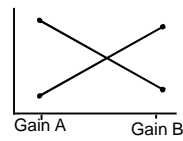
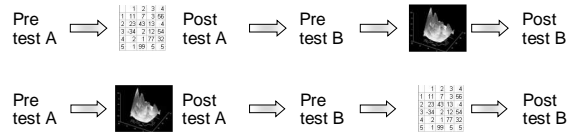
◆ Advantages

- Controls for the (often huge) differences between subjects
 - ◆ Each subject is their own control
- May reveal order effects

◆ Disadvantages

- Four groups of subjects rather than two!
- Statistically complex – predicting at least a 3 way interaction

Partial Cross-over

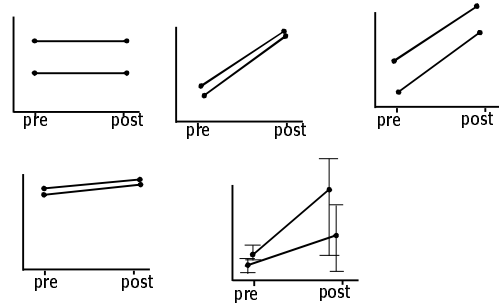


Partial Cross-over

◆ Same as full cross over but

- Advantages
 - ◆ less complex and subject hungry
- Disadvantages
 - ◆ less revealing of order effects

Some Common Problems/Results



Four Key Questions to Answer

- ◆ What are appropriate forms of measurement?
- ◆ What is an appropriate design?
- ◆ What is an appropriate form of comparison?
 - SciVis alone
 - SciVis v non-interventional control
 - SciVis v alternative representation
 - SciVis_(a) v SciVis_(b) (within system)
- ◆ What is an appropriate context

SciVis alone

- ◆ Uses
 - Does something about the user or the system predict outcomes?
 - E.g. Do users with high or low prior knowledge benefit more?
 - E.g. Does focusing on dynamic or static images lead to better performance?
- ◆ Disadvantages
 - No comparative data – is this a good way of representing information?
 - Identifying key variables to measure

SciVis v Non-interventional Control

- ◆ Uses
 - Is this a better way of visualizing something than not visualizing it at all?
 - Rules out improvement due to repeated testing
- ◆ Disadvantages
 - Often a no-brainer!
 - Does not answer what features of the system lead to understanding
 - Ethical ?

SciVis v Alternative Representations

- ◆ Uses
 - Proof of concept
 - Real world validity
 - Addresses which visualizations are useful (for which users, in which domains, performing what tasks and expecting what outcomes)...
- ◆ Disadvantages
 - Sometimes the alternative representations and SciVis differ in many ways...
 - It always a four way interaction...

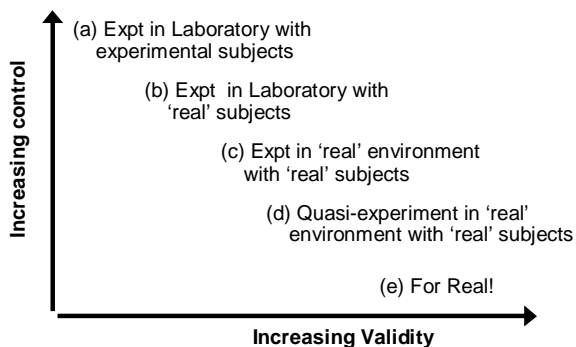
SciVis_(a) v SciVis_(b) (within system)

- ◆ For example
 - Only certain visualizations available
 - Only certain actions on visualizations
- ◆ Uses
 - Much tauter design, e.g. nullifies Hawthorne effect
 - Identifies what aspects of the visualization aided understanding
 - Aptitude by treatment interactions
- ◆ Disadvantages
 - Identifying key features to vary – could be very time consuming!

Questions to answer

- ◆ What are appropriate forms of measurement?
- ◆ What is an appropriate design?
- ◆ What is an appropriate form of comparison?
- ◆ What is an appropriate context?

Context



Choosing a context

- ◆ There is no “perfect” context! Real is not necessarily better for experiments...
- ◆ I try to avoid (a) but can't always
- ◆ Pick depending on access and nature of question
 - E.g. beware expts which need effort in artificial situations
 - Why should subjects who have no need to learn something apart from payment or course credit, work hard at learning?
 - Remember the Law of Gross Measures, time data often impossible in real contexts

Beware of...

- ◆ Evaluating on an inappropriate population
 - Develop a system aimed for one age of learners but test it on another
- ◆ Two many or two few subjects
 - Normally see too few (try to keep a minimum of 12 per cell) but this will change depending on variability
 - Too many also a problem – want to find differences that are educationally as well as statistically significant
- ◆ Inappropriate control
 - Credit assignment problem

Beware of... Inappropriate Generalisations

Learner Characteristics

- ◆ Ability levels
- ◆ Prior knowledge
- ◆ Developmental levels
- ◆ Gender
- ◆ Attitudes
- ◆ Motivation

Task Characteristics

- ◆ Procedural v conceptual learning
- ◆ Collaborative v Individual
- ◆ Time on task
- ◆ Timescale of intervention
- ◆ Frequency of use
 - e.g. 10 minutes a day v 1 hour a week

Conclusions

4 main things to ask yourself

- What are appropriate forms of measurement?
- What is an appropriate design?
- What is an appropriate form of comparison?
- What is an appropriate context
- ◆ No bun fights – accept all methods as valid and interesting if well performed and targeted at appropriate questions

Types of Questions...

◆ Good for quantitative methods...

- Is my visualization more effective than an alternative..
- How effective is my visualization
- Are my results robust and replicable

◆ But less good for quantitative methods...

- Who benefits from learning with visualizations
- How do people learn with visualization
- How does learning with visualizations change over time?
- How does the wider context influence learning with visualizations?