# Experimental Pitfalls

**Helen C. Purchase**  
University of Glasgow

---

When we have an idea, we want to test it to see if it is a good one.

---

Ideas that relate to the human use of computers need to involve participants in the testing.

---

And after all the effort…

No experiment can ever be perfect.

---

<table>
<thead>
<tr>
<th>Conditions</th>
<th>Decisions</th>
</tr>
</thead>
</table>
| Experimental objects | Nature of the participants  
Number of participants |
| Tasks | Pre- and post-experimental activities |
| Location  
Equipment  
Online? |  |
| Experimental timing |  |
| Data collection methods | Data analysis methods |

---

“Experimental Pitfalls”  
or:  
“Five Experimental Failures and a Joke”  
or:  
“Five Things I have Learned (and a Joke)”
Five Things I have Learned

... subject variability
... use of randomisation
... random factors
... piloting
... decision making

Object-oriented class diagrams

Aesthetics:
  - bends, crosses, orthogonality, upward-flow

Eight conditions:
  - b+ b- c+ c- o+ o- f+ f-

Experimental object (program code):
  - System for storing information about keys and the doors they can unlock – two versions (distributed, centralised)

<table>
<thead>
<tr>
<th></th>
<th>Green (distributed)</th>
<th>Blue (centralised)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Few bends (b-)</td>
<td>![Image]</td>
<td>![Image]</td>
</tr>
<tr>
<td>Many crossings (c+)</td>
<td>![Image]</td>
<td>![Image]</td>
</tr>
<tr>
<td>Not much orthogonality (o-)</td>
<td>![Image]</td>
<td>![Image]</td>
</tr>
<tr>
<td>Mostly upward direction (f+)</td>
<td>![Image]</td>
<td>![Image]</td>
</tr>
</tbody>
</table>

Steve Grunden
Unpublished, 1997
The experiment

- Third-year computer science class
- Top quartile of the relevant class in mid-semester test
- N=49, one-to-one
- Between-participants (6 per condition + 1)
- Pre-experiment tutorial
- Data: time to get correct answer

So...

... it is hard to ensure equivalent domain knowledge in a between-participants’ experiment
Screen Layout Principles

<table>
<thead>
<tr>
<th>Good</th>
<th>Bad</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>


90 stimuli
10 for each of nine aesthetic categories

9 conditions: 3 overall average aesthetic, cohesion, economy, regularity, sequence, symmetry, unity

N=21
• RQ: “Does layout aesthetic affect visual effort?”
• Task: count the number of upright triangles

• Within-participants experimental design

• Dependent variables: accuracy, response time, scan path length, scan path duration, number of fixations, fixation duration/gaze time
• Independent variables: aesthetics levels (high, medium, low), layout metrics

• 10 practice tasks
• 90 stimuli => 90 factorial possible sequences

• 10 practice tasks
• 90 stimuli => 90 factorial possible sequences
• Only two used

• 10 practice tasks
• 90 stimuli => 90 factorial possible sequences

So…

... randomisation introduces important variability that can mitigate against unwelcome learning effects

N=21
Subjective Complexity = -0.024 + 0.00905col + 0.094hog
The "Histogram of Oriented Gradients" algorithm measures the number of distinct objects in an image

Subjective Aesthetics = 0.659 – 0.0005926col – 0.001har
The "Harris Corner" algorithm measures the number of corners in an image

RQ: does complexity/aesthetics affect search efficiency?
Task: icon search time

<table>
<thead>
<tr>
<th>Complexity</th>
<th>Target icon</th>
<th>High comp</th>
<th>Medium comp</th>
<th>Low comp</th>
</tr>
</thead>
<tbody>
<tr>
<td>Distractors</td>
<td>high complexity</td>
<td>hh-c</td>
<td>mh-c</td>
<td>lh-c</td>
</tr>
<tr>
<td>medium complexity</td>
<td>hm-c</td>
<td>mm-c</td>
<td>lm-c</td>
<td></td>
</tr>
<tr>
<td>low complexity</td>
<td>hl-c</td>
<td>ml-c</td>
<td>ll-c</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Aesthetics</th>
<th>Target icon</th>
<th>High aesth</th>
<th>Medium aesth</th>
<th>Low aesth</th>
</tr>
</thead>
<tbody>
<tr>
<td>Distractors</td>
<td>high aesthetic</td>
<td>hh-a</td>
<td>mh-a</td>
<td>lh-a</td>
</tr>
<tr>
<td>medium aesthetic</td>
<td>hm-a</td>
<td>mm-a</td>
<td>lm-a</td>
<td></td>
</tr>
<tr>
<td>low aesthetic</td>
<td>hl-a</td>
<td>ml-a</td>
<td>ll-a</td>
<td></td>
</tr>
</tbody>
</table>
Does complexity affect search time?

The complexity of the target icon only has a search time effect (simple is quicker) when the distractors are complex.

Does aesthetics affect search time?

The aesthetics of the target icon only has a search time effect (beautiful is quicker) when the distractors are ugly.

But what about other features of the icons we have not considered: e.g. elegance, or metaphoric association, or balance, or symmetry?

<table>
<thead>
<tr>
<th>Complexity</th>
<th>Target Icon</th>
</tr>
</thead>
<tbody>
<tr>
<td>high comp</td>
<td>hh-c</td>
</tr>
<tr>
<td>medium comp</td>
<td>hm-c</td>
</tr>
<tr>
<td>low comp</td>
<td>hl-c</td>
</tr>
<tr>
<td>high comp</td>
<td>mm-c</td>
</tr>
<tr>
<td>medium comp</td>
<td>lm-c</td>
</tr>
<tr>
<td>low comp</td>
<td>ml-c</td>
</tr>
<tr>
<td></td>
<td>lh-c</td>
</tr>
</tbody>
</table>

Third dimension: symmetry
Control/ Random/ Confounding factors

Conditions: deliberately change

Control: deliberately ensure that they don’t change

Random: deliberately allow to change randomly to ensure generalisability

Confounding: factors that change together with the conditions (even though you don’t want them to)

I. Scott McKenzie, Human-Computer Interaction, 2013

Conditions: effectiveness of two biological diagrams for learning (A and B)

Control: deliberately ensure that they don’t change (first year biology students)

Random: deliberately allow to change to ensure generalisability (age)

Confounding: factors that change together with the conditions (the biology students who study chemistry performed better than those who didn’t)
So...

... we need to think carefully about what we want to control, can control, can’t control, and don’t care about controlling.

Scrolling Behaviour with Single- and Multi-column Layout

Comparing:
• Vertical scrolling: single column (web browsers)
• Horizontal scrolling text: multiple columns of same height (electronic readers)

How do people read?
How do people scroll?

“...Our results suggest that horizontal-scroll layout will be particularly popular on devices such as e-book readers that have slow display refresh and so are not well-suited to continuous scrolling...

...We plan to conduct further studies to see if our findings generalize to other kinds of participants, devices and reading material.”
Experimental process

• Demo of device
• Condition A
  – demo and training
  – read story, answer three simple questions
• Condition B
  – demo and training
  – read story, answer three simple questions
• Data:
  – Logging (eye-tracking)
  – Preference questionnaire

From my notes (verbatim)

• P1: One stylus is not enough
• P2: Problem with vertical scrolling using the stylus directly on the text — text jumps DOWN a little before moving UP
• P3: Definitely a problem with the vertical scrolling
• P5: System hung during horizontal training (totally unresponsive). Reset. System crashed during reading of HR. Reset and put charger in. System crashed again near the end of reading HR. Reset. Crash during the HR questions. Experiment abandoned
• P8: rapid, uncontrollable scrolling
• P9: a problem with sticking buttons

38 students had been recruited in advance:
“I’m afraid that I am going to have to cancel our experimental session next week – the mobile device we use has unexpectedly developed a fault.”

So...

... never (ever, ever) remove the piloting step!

Dynamic graphs

• “Does maintaining the ‘mental map’ help in understanding evolving graphs?”
• Conditions:
  – low mental map
  – medium mental map
  – high mental map
• Three different evolving graphs
  – 14-20 nodes, 15-30 edges, 4 changes/time-slice
• Four different tasks
  – addition/removal of edges, overall structure

How Important Is the “Mental Map”?  
Graph Drawing Symposium, 2005

Eve Hoggan and Carsten Görg

n=20

Graph 1: low mental map
(lots of movement)

Graph 2: medium mental map

Graph 3: high mental map
(minimal movement)

Aggregating over all three graphs and all four questions

Errors: no significant difference

Response time: no significant difference
Individual questions

- Q1: number of new edges
- Q2: node with most changes
- Q3: year of extreme reduction in size
- Q4: year a particular page had degree one

As expected, the questions were of different difficulty.

Individual graphs

- We did not expect the graphs to be of different difficulty.
- "We aimed to keep the size and changes of these graphs as similar as possible, while keeping them distinctive...having made an effort to keep the three evolving graphs comparable (similar size, similar number of changes per time-slice)...."

Characterisation of tasks & objects

- It was easy to identify the difference between the tasks in terms of difficulty – and so justifiable to analyse the data according to task.
- It was impossible to identify any difference between the graphs...because they had been arbitrarily defined.

So...

- ...never make arbitrary decisions (they may come back to haunt you!)

Five Things I have Learned

- ... it is hard to ensure equivalent domain knowledge in a between-participants’ experiment
- ...randomisation introduces important variability that can mitigate against unwelcome learning effects
- ...we need to think carefully about what we want to control, can control, can’t control, and don’t care about controlling
- ...never (ever, ever) remove the piloting step!
- ...never make arbitrary decisions (they may come back to haunt you!)

The identification of groups in social networks drawn as graphs is an important task for social scientists who wish to know how the population divides with respect to relationships or attributes.... In this paper, we report on an experiment.... We find that, despite the use of colour as the pre-attentive visual feature to signify group membership, participants tend to rely on structure as the basis for their visual community identification.
The identification of groups in social networks drawn as graphs is an important task for social scientists who wish to know how the population divides with respect to relationships or attributes. In this paper, we report on an experiment. We find that those algorithms that clearly separate communities with large distances are most effective, while the use of colour to represent community membership is more successful than reliance on structural layout.

In summary...

- Experiments are fun...
- ...but time-consuming, difficult, and can never be perfect
- Every decision counts
- We are all still learning...
- ...and it is often easier to imitate others’ processes than consider whether they are really appropriate