Human Factors Research: How to tell what works

Michael Friendly
Psych 6135
http://euclid.psych.yorku.ca/www/psy6135/

Topics

- Why consider human factors in graphic & information design?
- Empirical study of graphs
- Experimental methods
  - Psychophysical methods
  - Eye tracking
  - Computer, web-based experiments
- Visual inference

Control room display

A control room for a nuclear power plant or electrical system for a large city
- How does visual design support important decisions?
- How to warn or know when something fails?

How many things can the operator attend to at one time?
What visual design factors make important events salient?

Traffic control display

A traffic control system for a large metropolitan city (Tokyo)
- How does visual design support important decisions?
- How to warn or know when something fails?
- When is there too much information?

Oops!
An early digital navigation display panel, incorporating visual gauges and charts for a variety of functions—combines separate variables into a single display.

What visual features make it easier or harder to navigate?

Garmin Txi touchscreen device for a small jet

How does the pilot combine the map view with the heads-up view and all the visual dials?

A more modern design integrates a wider variety of displays in a touch-screen device

- Are there too many options or features?
- What features demand the most pilot attention or distract attention from flying?
- How to study the efficacy of alternative visual display designs?

Good dashboards are: dynamic, interactive, customizable

Interactive dashboards use sliders, buttons, pick lists, etc.

From: https://chandoo.org/wp/customer-service-dashboard/
Empirical studies of graphs

- Human factors vs. Psychology
- Experimental methods
- Running graph perception studies
- Some results

Human factors vs. Psychology

- Human factors research often motivated by applied problems in engineering, design, computer science
- Psychological research often motivated by more basic perceptual & cognitive questions.

Psychophysical methods

- Psychophysical methods are used in studies of graphical perception to study the relationship between
  - properties of a stimulus (position, length, area, angle, ...)
  - and a perceived response (how big? which is greater?)
- Magnitude estimation: rated (0-100) size or %
- Matching – adjust size of B so it is as large as A
- Discrimination
  - same/different?
  - which is larger?

Magnitude estimation

How large are transport accidents?
How much bigger than non-transport accidents?

Estimation of length or ratios of length are more accurate than the same judgments of area.

Image:
http://www.cns.nyu.edu/~david/courses/perception/lecturenotes/psychophysics/psychophysics.html
Stevens’ Power Law

- How does perceived magnitude of a sensation relate to stimulus intensity?
- S. S. Stevens (1957) showed that, for many domains
  
  Sensation $\propto$ Intensity$^p$

- These provide ways to assess the accuracy of magnitude estimation for visual encodings
  - length judgments most accurate
- But: graph perception is not always a matter of estimating magnitudes.

![Fig. 5.7 from: Munzner, Visualization Analysis & Design](image)

Magnitude estimation & memory

How does remembered size relate to perceived size? How does it relate to stimulus intensity? Kerst & Howard (1984) propose another power law:

Kerst & Howard, Bulletin of the Psychonomic Society 1984, 22 (6), 517-520

Visual search tasks

Find a target item among the many distractors in this display

![Visual search task](image)

Measure:
- reaction time (RT)
- accuracy (% correct)

Vary:
- # distractors
- # of targets
- complexity (# black squares)
- display format
- ...

Running graph perception experiments

- Paper & pencil tasks
  - little control of experimental presentation features
  - can’t measure RT
- Lab software: run on lab computers
  - e-prime ($$$), https://pstnet.com/
  - matlab ($) & Psychtoolbox: http://psychtoolbox.org/
  - PsychoPy – free, open source, see: http://www.psychopy.org/
- Web-based
  - Survey Monkey
  - Amazon Mechanical Turk
**e-prime**

e-prime is a polished software system for designing psychology experiments
- **E-Studio GUI** → **E-Basic script**; **E-Run** runs experiments

![Diagram](image1.png)

**PsychoPy**

PsychoPy provides a GUI for constructing online experiments
- **Builder interface** → python code that runs the experiment

![Diagram](image2.png)

**Amazon Mechanical Turk**

- Web-based experiments, hosted on Amazon servers
  - **requester jobs**: Human Intelligence Tasks (HITs)
  - **worker pool**: Turkers, get paid for doing tasks ($0.01 – 0.10 per item)
  - each cell of a design can be a separate HIT
  - Amazon provides a markup language for presenting text, movies, images, ... (HTML, javascript)

![Diagram](image3.png)
Early studies: Circles vs. bars

Eells (1929) studied magnitude estimation of proportions of a whole, presented as circles vs. bars.

“What number represents the proportion for each marked segment?”

Conclusions:
1. 1 segment: circles = bars
2. >1 segment: circles better than bars

Graph perception: Elementary perceptual tasks

Cleveland & McGill (1984) proposed that graphical perception could be studied in terms of 10 elementary perceptual tasks involved in most common graphs.

Their study was one of the first modern ones, and set a standard for magnitude estimation tasks.

Cleveland & McGill experiments

Pie charts vs. bar charts – position-angle experiment

Elementary perceptual tasks:
• bar chart: position along a scale
• pie chart: angle? area? arc length?

Experiment:
50 graphs (½ pie, ½ bar), random order
largest marked
“Make a quick visual judgment”
Response on an answer sheet

Measures:
accuracy: \( \log_2(|\text{judged }\% - \text{true }\%| + \frac{1}{8}) \)
bias: judged % - true %

NB: log scale estimates relative error; +1/8 handles zero values

Discussion Q:
• What are the elementary perceptual tasks involved in each of these?
• What is the statistical analysis?

Bar charts tasks – position-length experiment

Experiment:
50 graphs, 10 × 5 types
“What percent is smaller of the larger?”

Measures:
accuracy: \( \log_2(|\text{judged }\% - \text{true }\%| + \frac{1}{8}) \)
bias: judged % - true %
Cleveland & McGill experiments

Cleveland & McGill summarized these experiments in this figure, comparing absolute error in the tasks in these two experiments.

Heer & Bostock: MTurk experiments

- Replicated Cleveland & McGill T1—T5, T6 (angle)
- Added area judgment tasks:
  - T7: Bubble chart
  - T8: Center-aligned rectangles
  - T9: Treemap

Task:
“What % is area A of area B?”

Heer & Bostock: results

Results largely confirmed Cleveland & McGill (1984) with respect to relative order. The area judgment tasks are shown to give even larger errors.

Details:
H & B use a between-S design, n=50 per chart type, 10 charts of each type
10 × 7 = 70 separate HITs (each S responds to 1 chart)
Response: type in a # (% of smaller)
This graph of these results is a great model for data display

Other findings:
For a given graph type, judgments are most accurate when the true difference is extreme.
This within-graph effect is larger than the differences between graphs.

Discussion Q:
What are some problems with this graph? How could it be improved?
Simkin & Hastie: accuracy and RT

Problems with Cleveland & McGill study:

- Assessing accuracy only omits consideration of speed of judgment – should also measure reaction time.
- “Elementary perceptual tasks” give no insight into the cognitive processes used by observers to perform these judgments.

Simkin & Hastie used computer-controlled experiments to measure both accuracy & RT:

- Three types of stimuli x 30 of each = 90 trials
- Discrimination task: Which is larger?
- Judgment task: What % is smaller of the larger?

Analyses: Separate ANOVAs of discrimination RT, judgment RT and errors in each task

Simkin & Hastie (1987), An Information-Processing Analysis of Graph Perception, JASA, 82, 454-465

Simkin & Hastie: processing stages

They propose that tasks using various graph types can be understood in terms of elementary mental processes:

- anchoring: segment a component to serve as a standard for comparison
- scanning: visual sweep across a distance in a graph
- projection: send a ray from one point to another
- superimposition: mentally move elements to a new, overlapping location
- detection: detect difference in size of two components

Hollands & Spence: Discrimination analysis

How do people make decisions about which is larger from different visualizations?

- aligned vs. not aligned
- pie vs. bar chart

Hollands & Spence propose an incremental estimation model to account for speed of processing:

- each stimulus evokes a distribution of a psychological response of “size”
- the response (A>B?) is determined by the separation and overlap between the two distributions
- less overlap → faster response


NB: If these processes are sequential, RTs should reflect additive components. AFAIK, this idea has not been tested or explored.
**Hollands & Spence: results**

Experiment 1
- RT $\downarrow$ as $\Delta p \uparrow$
- RT < for aligned vs. non-aligned
- RT < for divided bars vs. pies

Similar pattern for errors

**Are pies ever better?**

What percent is the indicated region of the total?

Other experiments by Holland & Spence (1992, 1998) show that this judgment task is hardest for separated bars and easiest for pie charts


**More pie studies: Skau & Kosera**

Infographics use many variations of the basic pie chart to show part-whole relations
What properties do people use in making judgments: Angle, area, arc length?

Little difference among these, except for plots showing only angle

Task analysis

Another cognitive approach is to study the visual/mental steps a viewer takes in trying to answer a question based on a graph. Sometimes uses "protocol analysis".

Q: What is the relationship between the length of the eruption and the time between eruptions for Old Faithful?

Mental steps:
1. Understand the Q: identify "length of eruption" & "time between eruptions" as things to search for in the graph.
2. Look at axis labels: See Y: "Eruption length"; X: "Time between Eruptions"
3. Scan data: See "Y increases as X increases"
4. Answer the Q: As the time between eruptions increases, the length of the eruption seems to increase.
5. Notice: Hmm, something weird here!


Eye-tracking studies

Where do people look when viewing graphic displays?

Eye-tracking hardware allows recording of gaze fixation points over time

Eye-tracking software allows some visual analysis

Viewing a bar chart

This 2016 study by Trent Fawcett was a basic test of this methodology.
Uses R and the *saccades* package. There is also: *gazepath*, *emov* and more...

Eye-tracking hardware allows recording of gaze fixation points over time

Eye-tracking software allows some visual analysis

Viewing web pages

This methodology is now well developed, particularly for viewing web pages

This illustration uses heatmap colors to show density of gaze locations

Viewing web pages

This illustration shows a path of eye-movement locations in viewing the same web page.

Visual inference

• To what extent can visual display of real data, against a background of random data, be used as a substitute for standard statistical inference?

One of these plots doesn’t belong. Which one??

Six choropleth maps of cancer deaths in Texas, darker colors = more deaths. Can you spot which of the six plots is made from a real dataset and not simulated under the null hypothesis of spatial independence?

Visual inference: Lineup protocol

Buja et al. (2009) propose an analogy between standard statistical inference and visual inference based on human observers.

Lineup protocol:
• Generate $n-1$ decoys (null data sets).
• Make plots of the decoys, and randomly position a plot of the true data.
• Show to an impartial observer. Can they spot the real data?

With $n=19$ decoys, a correct decision by chance would have $p=1/20 = 0.05$.

**Visual inference: Lineup protocol**

This graphic table makes the comparison more direct:

<table>
<thead>
<tr>
<th>Conventional inference</th>
<th>Lineup protocol</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hypothesis</td>
<td>$H_0: \beta = 0$ vs $H_1: \beta &gt; 0$</td>
</tr>
<tr>
<td>Test statistic</td>
<td>$T(y) = \frac{\hat{\beta}}{SE(\hat{\beta})}$</td>
</tr>
<tr>
<td>Sampling Distribution</td>
<td>$f_{\chi^2(1)}$</td>
</tr>
<tr>
<td>Reject $H_0$ if</td>
<td>Actual $T$ is extreme</td>
</tr>
</tbody>
</table>


---

**Frequency distribution of tips at restaurants.**

Which one is the real data? What features lead you to this conclusion?

Panel 11:
- Skewed
- Multiple peaks
- Outliers

---

**Visual t-test**

For each data set, the observations are shown as points and the group means as crosses.

The accused (real data) is hidden amongst eight innocents.

Can you spot him?

Panel 3: a larger difference among the group means

---

**Visual tests for linear models**

This idea can be extended to visual inference for a wide range of hypotheses in linear models.

Main idea: Numerical test $\rightarrow$ Visual test

<table>
<thead>
<tr>
<th>Case null hypothesis</th>
<th>Statistic</th>
<th>Test statistic</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$H_0: \beta_k = 0$</td>
<td>Scatterplot</td>
<td>Scatterplot with least square line overlaid. For null plots we simulate data from fitted null model.</td>
<td></td>
</tr>
<tr>
<td>$H_0: \beta_k = 0$</td>
<td>Residual plots</td>
<td>Residual vs $X_k$ plots. For null plots we simulate data from normal with mean 0 variance $\sigma^2$.</td>
<td></td>
</tr>
<tr>
<td>$H_0: \beta_k = 0$ (for binary $X_k$)</td>
<td>Boxplot</td>
<td>Boxplot of residuals grouped by category of $X_k$. For null plots we simulate data from normal with mean 0 variance $\sigma^2$.</td>
<td></td>
</tr>
<tr>
<td>$H_0: \beta_k = 0$ (interactive of continuous and binary $X_k$)</td>
<td>Scatterplot</td>
<td>Scatterplot with least square lines of each category overlaid. For null plots we simulate data from fitted null model.</td>
<td></td>
</tr>
</tbody>
</table>

Some statistical tests have greater power to detect a non-null effect. What can be said about the power of different graphical methods for visual inference?

Graphs of wind direction and arrival delays for incoming flights to Phoenix airport.

1: polar scatterplot, delay=radius, wind=angle
2: boxplots of delay grouped by angle
Which is easier?
Which is the real data?

What can we learn from this approach?
- To what extent can visual inference substitute for numerical statistical inference?
- How to study this more?
  - How many observers to declare some effect “significant”
  - Can we use this paradigm to study observer differences?
  - Can we use this to study the effectiveness of different graph types or forms?

It is relatively easy to conduct online studies of graph perception.
- For validity & comprehensibility of results, it is often crucial to examine possible demographic variables that relate to the outcome.

What was the effect of various subject variables on outcome in a lineup task?

Various demographic variables are ordered here by the effect size for differences among various groups.

When several visual displays are compared, important to consider interactions – some people may do better with some displays.

Why is human factor research in graphics useful & important?
- How can it make a difference?
- What methods are available to study this?
  - What is the task?
  - How to measure “performance”?
- What have we learned?
- How to go forward?