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?
Navigation

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?

Dashboards: Financial trading

Dashboards combine visual information for decisions on a single screen.

Good dashboards are: dynamic, interactive, customizable.

Dashboards: Customer service

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

Goal for Information Desk is 89.1
NEED TO BE MORE HELPFUL!

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

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

Human factors vs. Psychology

- Human factors research often motivated by applied problems in engineering, design, computer science
  - A/B testing (booking.com, Netflix, ...) for features of user interaction
  - navigation controls: pilot testing, flight simulators
- Psychological research often motivated by more basic perceptual & cognitive questions.
  - accuracy of judgments of graphs
  - judgments of trend, correlation, etc.

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
**Pies vs. dot charts**

Which sector is the largest? Which is the smallest?

Judgement of position along a scale is more accurate than that of area or angle.

**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
    - area: less so
  - But: graph perception is not always a matter of estimating magnitudes.

**Stevens’ Power Law: log-log form**

The exponential form, $S \sim \text{Intensity}^p$ is more easily understood when both are plotted on log scales, where it is linear.

$$\log(S) = p \log(I)$$

$p$ is the multiplier effect of a multiple of Intensity.

**Power Law: Origin story**

In 1953, S. S. Stevens carried out experiments on magnitude estimation of loudness of sounds – measured in db (a log scale).

This graph shows results for 7 individual subjects, offset to show the separate data.

The idea of an average slope, $p$, arose later, as the effect for an average observer.


From: https://santhoshsoundar.blog/power-law/
Power Law: Brightness

What about other stimulus domains?

Luminance of a light source could be measured (on a log scale)

Averaging over observers, log of the magnitude estimates was again linear with log intensity.

Stevens (1975) asserted that he had discovered a fundamental law of the relation between physical properties and perception.

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:

\[ \text{Sensation} \propto \text{Intensity}^p \]
\[ \text{Memory} \propto \text{Sensation}^{p'} \] 

\[ \text{therefore} \]

\[ \text{Memory} \propto \text{Intensity}^{pp'} \]

If \( p \approx p' \), the exponent for memory \( = p^2 \)

\[ \rightarrow \text{remembered area less accurate than area itself} \]

Visual search tasks

Find a target item among the many distractors in this display

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

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

This paradigm is often used in evaluating complex visual displays

Visual search experiment

Visual search for a diagnostic signal (radiology) can vary with the # of targets and whether these are shown stationary or moving.

The response measure here is % accuracy in detecting a given target.

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

Results:
Better for 1 target vs. 2, whether moving or not
A 2nd search is less accurate when the target is moving

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 / Qualtrix
  - 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

Create experiment structure by drag-and-drop in E-Studio
Stimuli: text, audio, images, video, ...
Response devices: keyboard, mouse, external devices, ...

E-Prime GUI

Block of trials

Experiment structure
PsychoPy provides a GUI for constructing online experiments

• Builder interface → python code that runs the experiment

https://www.psychopy.org/

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)

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 segment: circles ≈ bars
• >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.
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 (¼ pie, ½ bar), random order
largest marked
“What percent is each of the others?”
“Make a quick visual judgment”
Response on an answer sheet

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

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

Bar charts tasks – position-length experiment

Experiment:
• 50 graphs, 10 × 5 types
• “What percent is smaller \( \ast \) of the larger \( \ast \)”

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

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

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?

Effect of truncation in bar charts

An Mturk experiment to assess the effect of Y-axis truncation on relative judgements.

How do bookings to (A) compare to bookings to (B)?

Design: Within-S, each subject saw 20 control bar graphs and 20 truncated bar graphs.

Study 1: No warning (n=24)
Study 2: Subjects warned that some graphs might be misleading (n=109)

This graph form ("raincloud plot") combines density estimates, data dots & CIs.
Study 2 tests an interpretation based on task mental set / instructions.
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?

Analysis: 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

Analysis of the three graph types in terms of proposed elementary mental processes

- **bar chart** (position)
- **divided bar** (length)
- **pie chart** (angle)

Figure 7. Proposed Sequence of Elementary Mental Processes to Explain Performance in the Comparison-Judgment Task for Problem (top panel), Length (middle panel), and Angle (bottom panel).

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

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

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

Hollands & Spence: results

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

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?

Variations of these designed to test accuracy of part-whole judgments

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.
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!

Illustration from VanderPlas (2015), Perception in statistical graphics, PhD thesis, Iowa State U

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...

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

Fawcett, T. "The Eyes Have It: Eye Tracking Data Visualizations of Viewing Patterns of Statistical Graphics" (2016).
https://digitalcommons.usu.edu/gradreports/787
Viewing web pages

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

Accessibility of data visualization

- Graph design
  - Design for color deficiencies
  - Favor direct labels over legends
- Assistive technology
  - Data sonification: map data to sounds
  - Screen readers: turn text into speech
- Text, titles, captions
  - Web: use alt-text `<img alt="What to understand here">`
  - PPT: embed descriptive text in images
  - Papers: use better titles and captions

Accessibility of graph captions

Lundgard & Satyanarayan (2022) studied how graph captions affect understanding, particularly for disabled (blind) viewers.

How to design figure captions / alt-text to better communicate what is to be shown?
**Accessibility of graph captions**

They present a model of 4 levels of description and their semantic content.

<table>
<thead>
<tr>
<th>#</th>
<th>LEVEL KEYWORDS</th>
<th>SEMANTIC CONTENT</th>
<th>COMPUTATIONAL CONSIDERATIONS</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Conceptual and domain-specific</td>
<td>Descriptive statistics, trendlines, axes, labels, colors</td>
<td>Focus on the visualization of data and its perception</td>
</tr>
<tr>
<td>2</td>
<td>Perceptual and cognitive</td>
<td>Descriptive statistics, trendlines, axes, labels, colors</td>
<td>Focus on the perception of data</td>
</tr>
<tr>
<td>3</td>
<td>Statistical and relational</td>
<td>Domain-specific insights, current events, social and political context, explanations</td>
<td>Focus on the contextual understanding of data</td>
</tr>
<tr>
<td>4</td>
<td>Domain-specific knowledge and domain-specific expertise (perceiver-dependent)</td>
<td>Complex trends, pattern synthesis, exceptions, commonplace concepts</td>
<td>Focus on the dependency of perception on the perceiver</td>
</tr>
</tbody>
</table>

**Study design**

- **Stimuli:**
  - Chart types (bar, line, scatter)
  - Topics (academic, business, journalism)
  - Difficulty (easy, medium, hard)

- **Subjects:** 90 sighted, 30 blind (proficient with a screen reader).

- **Task:** rank the usefulness of 4 descriptions (Levels 1-4) for understanding.

*Suppose that you are reading an academic paper about how life expectancy differs for people of different genders from countries with different levels of income. You encounter the following visualization. [Table 3.1] Which content do you think would be most useful to include in a textual description of this visualization?*

Read the paper to see the scope and content of this type of research.

**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 \( p=1/20 = 0.05 \)

*From: Buja et al. (2009), Statistical inference for exploratory data analysis and model diagnostics. Phil. Trans. R. Soc. A, 367, 4361–4383.*
Visual inference: Lineup protocol

This graphic table makes the comparison more direct:

<table>
<thead>
<tr>
<th>Hypothesis</th>
<th>Conventional inference</th>
<th>Lineup protocol</th>
</tr>
</thead>
<tbody>
<tr>
<td>$H_0: \beta = 0$ vs $H_1: \beta &gt; 0$</td>
<td>$T(y) = \frac{y - m}{s}$</td>
<td>$T(y) = \frac{y - m}{s}$</td>
</tr>
<tr>
<td>Test statistic</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sampling Distribution</td>
<td>$f_T(x; \nu)$</td>
<td>$f_T(y; \nu)$</td>
</tr>
<tr>
<td>Reject $H_0$ if</td>
<td>Actual $T$ is extreme</td>
<td>Actual plot is identifiable</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>$H_0: \beta_k = 0$ (for binary $X_k$)</td>
<td>Residual plot</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>$H_0: \beta_k = 0$ (interaction of continuous and 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>
</tbody>
</table>

Some statistical tests have greater \textit{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?

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

It is relatively easy to conduct online studies of graph perception.

\begin{itemize}
  \item For validity & comprehensibility of results, it is often crucial to examine possible demographic variables that relate to the outcome
\end{itemize}

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.

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