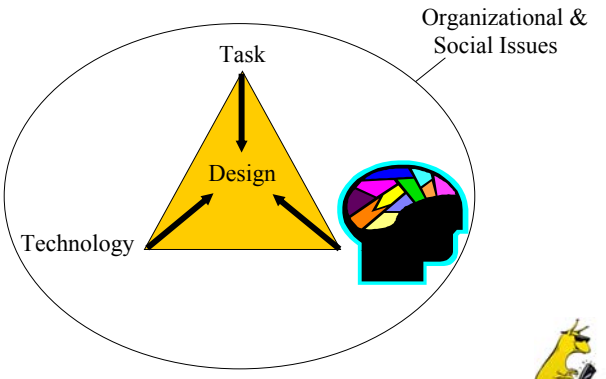
HCI Application: Data Analysis



Content Analysis

- We ran interviews – what to do with the data?
- Content analysis: a technique for making inferences by identifying special characteristics of narratives (written or oral)
- Information is condensed (classified) and made systematically comparable by applying a coding scheme
- What gets coded? Field notes from participant observation, letters, novels, transcripts of recorded communications (T.V shows, interviews, etc.).
- Need to decide:
  - what gets counted (words, pictures?)
  - what levels of analysis (categories, amounts?)
  - what coding frames (every 10th page, every other sentence?)

Steps

1. fully describe the phenomenon to be studied (e.g. perception of Microsoft software)
2. select the media that will be used for data
3. derive coding strategies from theory
   - judges tone/valance from the perspective of the key representative/candidate/character.

   1 = Appears to contribute to positive impression of the representative

   2 = Neutral, mix of positive and negative elements

   3 = Appears to contribute to negative impression of the representative
4. decide on a sampling strategy → you can't count it all (see coding frames before)
5. train the coders/raters (reliability is important)
6. analyze the data (%'s, compare means and variances?)

Coding granularity

- Items: an entire book, a letter, speech, diary, newspaper, or an in-depth interview
- Words: smallest unit, least judgment, results in distribution frequency
- Sentences: more judgment but more contextual
- Paragraphs: very contextual but paragraphs are hard to define in non-written narratives (e.g. interview)
- Characters: the number of times specific persons are mentioned
- Semantics: meanings of overall sentences or paragraphs, requires a lot of judgment

Coding granularity

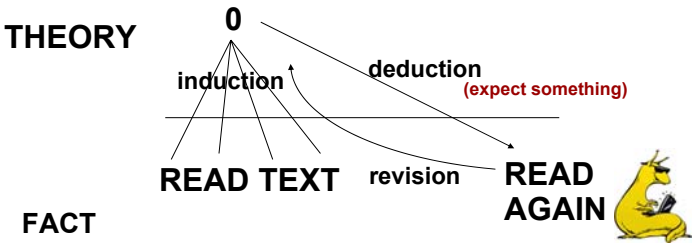
- Concepts: involve words grouped together into conceptual clusters (ideas)
  - crime, delinquency, money laundering, fraud = the conceptual idea of deviance
- Themes: broader than concept
  - must further specify the unit – theme of each sentence, each paragraph, the whole book

Coding approaches

- Common classes
  - used by virtually anyone in society, e.g. age, gender, mother, father, etc
  - essential in assessing whether certain demographic characteristics are related to patterns that arise from other coding
- Special classes
  - colloquial categories
  - includes jargon of various professions, e.g. petty larceny vs. that other category
- Theoretical classes
  - those that emerge in the course of analyzing the data
  - category labels generally borrowed from special classes their substance is grounded in the data
  - not immediately knowable until observers spend considerable time with the content

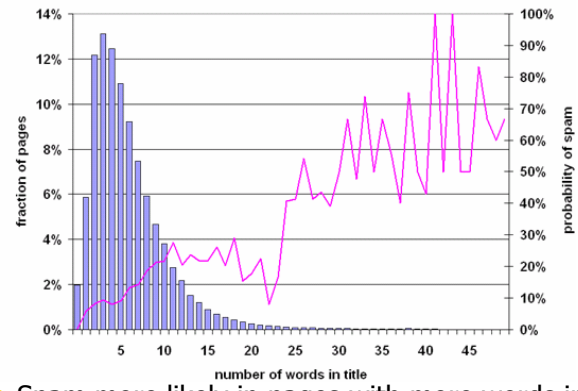
Trade-offs

- In vivo codes vs. conceptual constructs
  - Actual words vs. terms constructed by professionals (obsessive workaholics)
- Established vs. your own codes
  - Individuality in the data vs. being accused of circular reasoning
  - Avoiding accusation: divide the data set in half, develop the code on one half, apply it to the other half



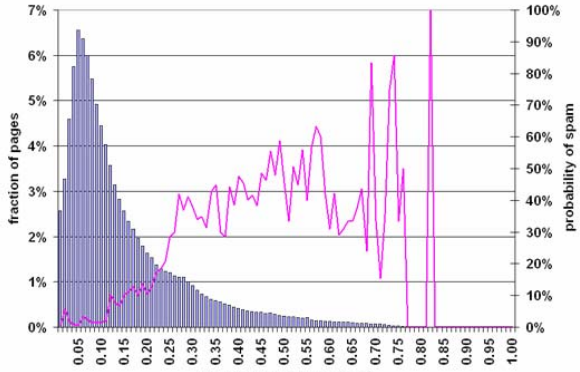
Content Analysis: spam

Distribution of Word-counts in `<title>`



- Spam more likely in pages with more words in title

Distribution of Visible-content



$$\text{Visible Content} = \frac{\text{size (in bytes) of visible words}}{\text{size (in bytes) of the page}}$$

Spam Content Analysis

- size of the page
- static rank
- link depth
- number of dots/dashes/ digits in hostname
- hostname length
- hostname domain
- number of words in the page
- number of words in the title
- fraction of anchor text
- average length of the words
- fraction of visible content
- fraction of top 100, 200, 500, 1000 words in the text
- fraction of text in top 100, 200, 500, 1000 words
- occurrence of strange words
- occurrence of the phrase "Privacy Policy"
- occurrence of the phrase "Privacy Statement"
- occurrence of the phrase "Customer Service"
- occurrence of the word "Disclaimer"
- occurrence of the word "Fax"
- occurrence of the word "Phone"
- occurrence of the word "Copyright"

Exercise

- Evolutionary theory says women will offer (and men will seek) youth, looks, sex appeal while men will offer (and women will seek) age, status, security. Code this set of personal ads and then correlate presence of absence of these items with gender. What other themes emerge?

Research Questions

- Why do we conduct empirical research?
- Simply…
  - To answer (or raise!) questions about a new or existing UI design or interaction technique!
- Questions include…
  - Is it viable?
  - Is it as good as or better than current practice?
  - Which of several design alternatives is best?
  - What are its performance limits and capabilities?
  - What are its strengths and weaknesses?
  - Does it work well for novices, for experts?
  - How much practice is required to become proficient?

Testable Research Questions

- Preceding questions, while unquestionably relevant, are not testable
- Try to re-cast as testable questions (…even though the new question may appear less important)

Scenario…

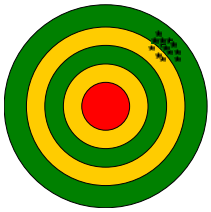
You have invented a new text entry technique for mobile phones. In your view, it's pretty good. In fact, you think it's better than the most widely used current technique, multi-tap. You decide to undertake some empirical research to evaluate your invention and to compare it with multi-tap? What are your research questions?

Research Questions

- Weak question…
  - *Is the new technique better than multi-tap?*
- Better…
  - *Is the new technique faster than multi-tap?*
- Better still…
  - *Is the new technique faster than multi-tap within one hour of use?*
- Even better…
  - *If error rates are kept under 2%, is the new technique faster than multi-tap within one hour of use?*

Reliability and Validity

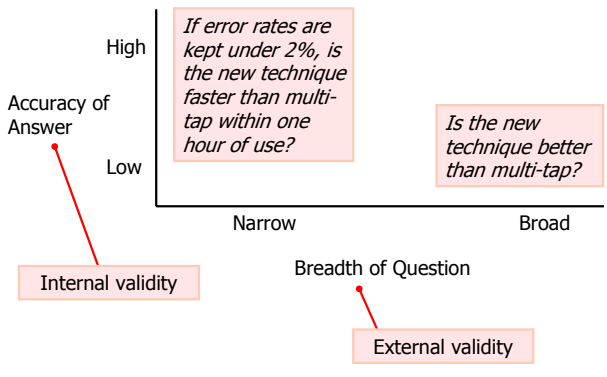
- If the goal was to hit the "Bullseye" with each dart…
  - Then left = consistent but unreliable, right = inconsistent and inaccurate
- Think of reliability ~ consistency, validity ~ accuracy
- Reliability = reproducibility factor (consistency of a measure).
- Reliability is a necessary but not sufficient condition for validity
- Validity = whether you measure what you think you measure



Reliability

- Interrater Reliability (consistency between raters) :
  - Independent observers rate the same sample should produce more or less the same results
- Test-Retest Reliability (consistency over time)
  - A reliable measure should give the same reading at different points in time (for a stable variable).
- Internal Consistency Reliability
  - = consistency among the items that measure the same thing.
  - Relevant when several measurements are made to obtain a score for each participant.
  - A measure that internally consistently measures 1 construct with several independent variables
  - Measured using Cronbach's Alpha

A Tradeoff



Validities

- Internal validity: the extent to which the effects observed are due to the test conditions
  - Differences in the means are due to inherent properties of the test conditions
  - Variances are due to participant differences
  - Other potential sources of variance are controlled
  - Note: Uncontrolled sources of variance are bad news and compromise internal validity
- External validity: the extent to which results are generalizable to other people and other situations
  - Re people, the participants are representative of the broader intended population of users
  - Re situations, test environment and experimental procedures are representative of real world situations where the UI/technique will be used

Test Environment Example

- Scenario…
  - You wish to compare two input devices for remote pointing (e.g., at a projection screen)
- External validity is improved if the test environment mimics expected usage
- Test environment should probably…
  - Use a projection screen (not a CRT)
  - Position participants at a significant distance from screen (rather than close up)
  - Have participants stand (rather than sit)
  - Include an audience!
- But… is internal validity compromised?

Experimental Procedure Example

- Scenario…
  - You wish to compare two text entry techniques for mobile devices
- External validity is improved if the experimental procedure mimics expected usage
- Test procedure should probably require participants to…
  - Enter representative samples of text (e.g., phrases containing letters, numbers, punctuation, etc.)
  - Edit and correct mistakes as they would normally
- But… is internal validity compromised?

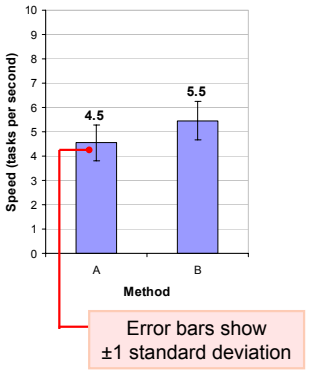
The Tradeoff

- Tension between internal & external validity
- The more the test environment and experimental procedures mimic real-world situations, the more the experiment is susceptible to uncontrolled sources of variation, e.g. pondering, distractions
- Internal and external validity are increased by…
  - Posing multiple narrow (testable) questions that cover the range of outcomes influencing the broader (untestable) questions
  - E.g., a technique that is faster, is more accurate, takes fewer steps, is easy to learn, and is easy to remember, is generally better
- There is usually a positive correlation between the testable and untestable questions

Answering Empirical Questions

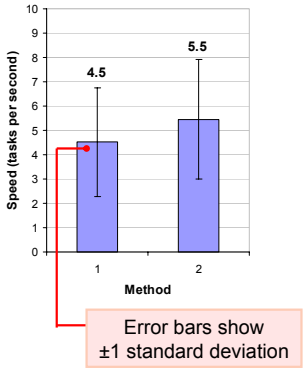
- If you asked participants which one they preferred, they will answer.
- We want to know if the measured performance on a dependent variable is different between test conditions, so…
  - We conduct a user study and measure the performance on each test condition over a group of participants
- Three questions:
  1. Is there a difference? Obvious – some difference is likely
  2. Is the difference large or small? Descriptive statistics can help
  3. Is the difference significant or is it due to chance? Inferential statistics can help (ANOVA)

Two groups data – case 1



Participant	Method	
	A	B
1	5.3	5.7
2	3.6	4.6
3	5.2	5.1
4	3.3	4.5
5	4.6	6.0
6	4.1	7.0
7	4.0	6.0
8	5.0	4.6
9	5.2	5.5
10	5.1	5.6
Mean	4.5	5.5
SD	0.73	0.78

Two groups data – case 2



Participant	Method	
	A	B
1	2.4	6.9
2	2.7	7.2
3	3.4	2.6
4	6.1	1.8
5	6.4	7.8
6	5.4	9.2
7	7.9	4.4
8	1.2	6.6
9	3.0	4.8
10	6.6	3.1
Mean	4.5	5.5
SD	2.23	2.45

Experimental Design

- Treatment: no coffee vs. coffee in the morning
- Factor: independent variable: # cups
- Construct: variable at an abstract level
- Levels: intensity of factors: 0,1,2…5 cups
- Response: dependent variable: alertness
- Covariate: control variable: body mass
- Trial: one simulation execution at one combination of input levels
- Replication: multiple trials at given combination
- Randomization: running the trials in an experiment in random order
- Blocking: dealing with nuisance factors

Scales of Measurement

- Major task in measurement: systematically apply numbers to variables.
- Nominal (naming/category scale)
  - Differences between categories – qualitative.
  - represent categories where there is no basis for ordering the categories, e.g. male vs. female, ford vs. toyota.
- Ordinal (order):
  - involve categories that can be ordered along a pre-established dimension.
  - no way of knowing how different the categories are from one another, e.g. white, green, blue, brown belts.
- Ratio (numbers) :
  - Distance between adjacent numbers are equal.
  - Most ratio scales are counts of things (e.g. temperature).
  - There is reference to zero point.

Scales of Measurement

- Interval :
  - similar to standard numbering scales except that they do not have a true zero (distance between successive numbers is equal), e.g.: IQ (there is no 0).
- Why do we need to make the distinction?
  - It affects the statistical procedures that will be used in describing and analyzing data.
- Effective range of the scale
  - Every measure has an effective range for the population under study.
- Attenuation effect: if effective range is inadequate (distorts data & threatens the validity of the study).
  - Ceiling effect – restricted higher range
  - Floor effect - restricted lower range

Scales of Measurement

	Levels of Measurement			
	Nominal	**Ordinal**	**Interval**	**Ratio**
Examples	Gender, name of places	Socioeconomic class ranks	Scores, personality & attitude scales	Weight ,height, length, time, # of responses
Properties	Identity	Identity Magnitude	Identity Magnitude Equal intervals	Identity Magnitude Equal intervals True zero point
Mathematical operations	-	Rank order	+, -	+, -, x, ÷
Type of data	Nominal	Ordered	Score	Score
Typical statistics used	Chi-square	Mann Whitney *U* - Test	t-test, ANOVA	t-test, ANOVA

Parametric and Nonparametric Tests

- Parametric tests estimate at least one parameter (in t-test it is population mean)
  - Usually for normal distributions and when the dependent variable is interval/ratio
  - Less likely to have type II error
  - Prone to violation to normality of data
- Nonparametric tests do not test hypothesis about specific population parameters
  - Distribution-free tests
  - Although appropriate for all levels of measurement most frequently applied for nominal or ordinal measures
  - Easier to compute and have less restrictive assumptions

Strategy of Experimentation

- "Best-guess" experiments
  - Used a lot
  - More successful than you might suspect, but there are disadvantages…
- One-factor-at-a-time (OFAT) experiments
  - Sometimes associated with the "scientific" or "engineering" method
  - Devastated by interaction, also very inefficient
- Statistically designed experiments
  - Based on Fisher's factorial concept
  - Full factorial, fractional factorial, latin square, etc

Full Factorial Design

- All possible combinations of factor levels are tested
- Start w. two-level design: experiments which include all decision variables at only two levels (usually coded as - and +)
- With this you get the main effects and interactions between pairs and among all 3 variables
- Example: the time to get there in ms (y) from all combinations of three decision variables:
  - T = target distance at 60 pixels, 180 pixels
  - C = CD gain at 20%, 40%
  - K = input device A (mouse) or B (joystick)

Full factorial design with 2 levels per factor

Trial	*T*	*C*	*K*	*y*
1	60 -	20 -	A -	y_{---}
2	180 +	20 -	A -	y_{+--}
3	60 -	40 +	A -	y_{-+-}
4	180 +	40 +	A -	y_{++-}
5	60 -	20 -	B +	y_{--+}
6	180 +	20 -	B +	y_{+-+}
7	60 -	40 +	B +	y_{-++}
8	180 +	40 +	B +	y_{+++}

Fractional Factorial Design

- Full factorial design is time and resource intensive (think if each has more than 2 levels)
- Fractional factorial experiment, meaning simply an experiment involving a subset of the experimental conditions.
- Latin square: get main effects but no interactions (so don't do it if you suspect interaction)
  - Condition: every factors must have the same number of levels
  - Every level of every factor appears with every level of every other factor exactly once
  - So for the full factorial example, how many trials are needed?

ANOVA (Analysis of Variance)

- While t-test is for comparing 2 means, ANOVA is for >2
- Why not do multiple t-tests? If you want to test H0: m1 = m2 = m3
- Why not test:

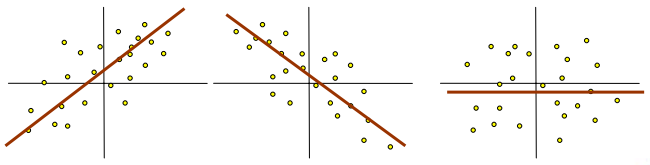
m1 = m2
m1 = m3
m2 = m3 } For each test 95% probability to correctly fail to reject (accept?) null, when null is really true

0.95^3 = probability of correctly failing to reject all 3 = 0.86

- ANOVA: calculate ratios of different portions of variance of total dataset to determine if group means differ significantly from each other (Excel – Data analysis – ANOVA single f.)
- Calculate 'F' ratio, named after R.A. Fisher
- Same rule as t-test, observe p to see significance

Correlation

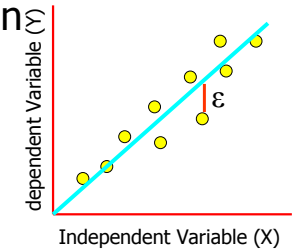
- A total of 4000 cans are opened in Texas every minute. 10 babies are conceived in Texas every minute. Therefore, each time you open a can in Texas, you stand a 1 in 400 chance of becoming pregnant.
- R = correlation coefficient (under Data analysis on Excel, check p to see if the correlation is significant)



Positive correlation | Negative correlation | No correlation

Simple Linear Regression

- linear relationship between a predictor variable, plotted on the x-axis, and a response variable, plotted on the y-axis
- R^2 = Coefficient of Determination (to judge the adequacy of the regression model)
- Remember R = correlation coefficient
- Regression on Excel:
  - Scatter plot
  - Format Trendline
  - Type: linear
  - Option: Display R-squared and equation on chart



$$Y_i = \beta_o + \beta_1 X_i + \varepsilon_i$$

intercept slope residuals

Putting it all together

- Scenario…

Researcher R has an interest in the application of eye tracking technology to the problem of text entry. After studying the existing body of research and commercial implementations, R develops some ideas on how to improve the interaction. R initiates a program of empirical inquiry to explore the performance limits and capabilities of various feedback modalities for keys in on-screen keyboards used with eye typing.

Experiment Design

- 4 x 4 repeated measures design
- Control variables (viz. factors)…
  - Feedback modality (A0, CV, SV, VO)
  - Block (1, 2, 3, 4)
- Dependent variables (viz. measures)
  - Speed (in "words per minutes")
  - Accuracy (in "percentage of characters in error")
  - Key activity (in "keystrokes per character")
  - Eye activity (in "read presented text events per phrase")
  - Etc. (other "events" of interest)
  - Also… responses to "broad" questions
- Order of conditions
  - Feedback modality order differed for each participant

Procedure

- General objectives of experiment explained
- Eye tracking apparatus calibrated
- Practice trials, then
- Data collection begins
- Phrases of text presented by experimental software
- Participants instructed to enter phrases "as quickly and accurately as possible"
- Five phrases entered per block
- Total number of phrases entered in experiment…
  - 13 x 4 x 4 x 5 = 1040

Anova Data Table

Factors and levels

Speed																	
	A	A	A	A	C	C	C	C	S	S	S	S	V	V	V	V	
Participant	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	Mean
1	6.17	7.19	7.04	7.09	6.76	7.40	7.54	7.94	6.44	6.17	7.84	6.81	5.20	6.29	7.39	7.63	6.93
2	6.71	7.25	7.05	7.15	7.73	7.57	8.04	7.26	7.00	6.75	7.68	7.46	7.50	7.07	7.32	7.06	7.29
3	6.80	6.65	7.62	7.98	6.61	7.18	7.34	8.19	6.65	7.53	7.09	7.90	5.73	7.24	6.94	7.13	7.16
5	6.30	6.31	7.59	7.38	6.85	7.64	7.58	7.88	7.07	6.43	7.26	7.65	6.75	6.59	6.97	7.72	7.12
7	6.68	6.89	7.32	7.51	7.00	7.81	7.64	7.24	6.80	7.35	7.4		6	6.72	7.57	7.20	7.11
8	6.08	6.55	6.83	5.92	7.44	6.93	7.56	6.41	7.38	7.07	7.0		2	7.93	7.45	7.16	6.98
9	7.62	7.01	6.60	7.07	6.91	6.81	6.91	7.73	6.50	7.57	7.59	7.80	6.62	7.06	7.16	7.41	7.15
10	5.88	5.71	7.33	7.11	6.66	7.97	7.64	8.15	6.35	7.21	6.56	7.33	5.00	6.97	6.54	6.36	6.80
12	6.89	7.61	7.42	7.88	7.79	8.28	8.20	8.39	6.62	6.87	7.99	8.23	9.57	8.17	7.91	7.09	7.81
13	6.85	6.57	8.14	6.00	5.92	7.89	7.49	6.98	6.05	7.45	5.34	7.46	7.21	6.81	6.80	8.24	6.95
14	5.37	5.56	6.04	6.86	6.20	6.82	7.71	7.76	5.85	6.37	6.74	6.69	5.98	6.43	6.38	5.87	6.41
15	5.51	6.12	6.32	7.00	6.16	6.49	7.21	7.19	5.65	6.52	6.49	7.10	5.31	6.88	6.36	6.93	6.45
16	5.88	7.18	5.95	6.00	4.85	6.98	7.37	6.98	6.88	6.21	4.96	5.34	6.72	7.14	4.96	6.80	6.26
																	6.96

Outlier

Each cell is the mean for five phrases of input

Anova Table

ANOVA Table for Entry Speed (wpm)

	DF	Sum of Squares	Mean Square	F-Value	P-Value	Lambda	Power
Subject	12	32.319	2.693				
Feedback Mode	3	8.210	2.737	8.772	.0002	26.317	.994
Feedback Mode * Subject	36	11.231	.312				
Block	3	13.310	4.437	10.923	<.0001	32.768	.999
Block * Subject	36	14.623	.406				
Feedback Mode * Block	9	1.772	.197	.633	.7669	5.694	.294
Feedback Mode * Block * Subject	108	33.606	.311				

Verbal statement and discussion of findings will include…

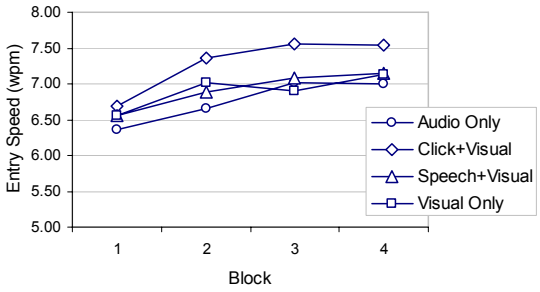
- Main effect for Feedback mode *significant*: $F_{3,36} = 8.77$, $p < .0005$
- Main effect for Block *significant*: $F_{3,36} = 10.92$, $p < .0001$
- Feedback mode by block interaction *not significant*: $F_{9,108} = 0.767$, ns

Summary Table for Speed

Speed (wpm)					
	Feedback Mode				
Block	Audio Only	Click+Visual	Speech+Visual	Visual Only	mean
1	6.36	6.68	6.56	6.55	6.54
2	6.66	7.37	6.88	7.02	6.98
3	7.02	7.56	7.09	6.90	7.14
4	7.00	7.55	7.14	7.12	7.20
mean	6.76	7.29	6.92	6.90	6.97

5.7% faster on 4th block

Summary Chart



The Broad Questions

- Participants were asked to rank the feedback mode based on personal preference
- Six of 13 participants gave a 1st place ranking to the fastest feedback modality
  - Not a strong result
  - Probably the differences just weren't large enough for participants to really tell the difference in overall performance.
- Notably, ten of 13 participants gave a 1st or 2nd place ranking to the fastest feedback modality
  - Thus, there is a modest trend that better performance yields a better preference rating (but empirical research is the key!)