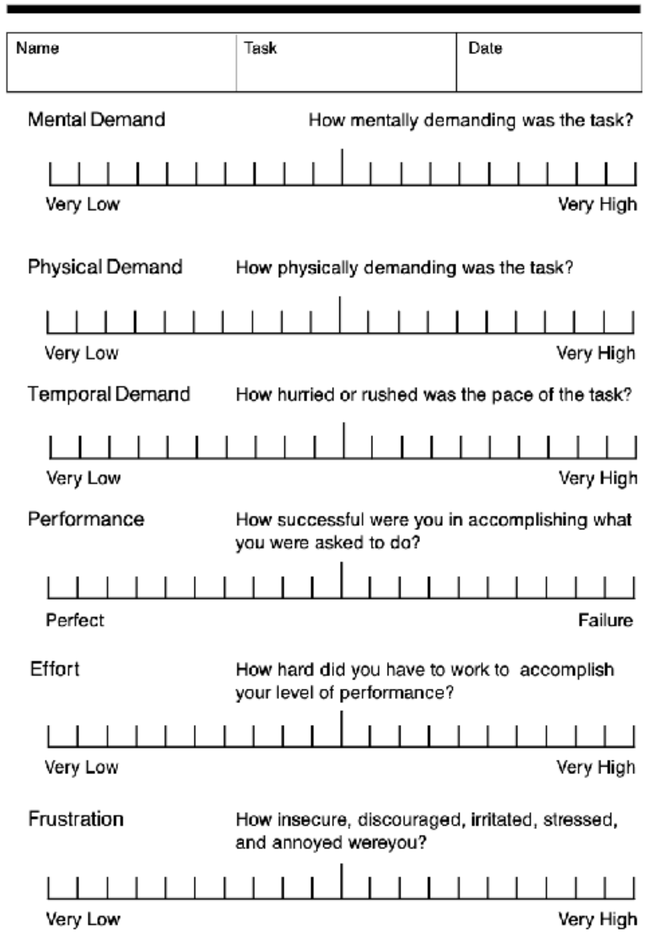
CS122 Class 14: Evaluation

Figure 8.6

NASA Task Load Index

Hart and Staveland's NASA Task Load Index (TLX) method assesses work load on five 7-point scales. Increments of high, medium and low estimates for each point result in 21 gradations on the scales.



Exploration

It was easy for me to explore many different options, ideas, designs, or outcomes without a lot of tedious, repetitive interaction.

Agree |||||||||||||||||||| Disagree

Collaboration

I was able to work together with others easily while doing this activity.

Agree |||||||||||||||||||| Disagree

Engagement

I was very absorbed/engaged in this activity - I enjoyed it and would do it again.

Agree |||||||||||||||||||| Disagree

Effort/Reward Tradeoff

What I was able to produce was worth the effort required to produce it.

Agree |||||||||||||||||||| Disagree

Tool Transparency

While I was doing the activity, the tool/interface/system "disappeared," and I was able to concentrate on the activity.

Agree |||||||||||||||||||| Disagree

Expressiveness

I was able to be very expressive and creative while doing the activity.

Agree |||||||||||||||||||| Disagree

NASA-TLX and CSI, two likert-based evaluation schemes

Class 14 agenda

- Zipcrit
- Next milestone 3: Task analysis & video prototype
- Lecture: Evaluating tools
- Break
- Studio: From design goals to evaluation goals

Milestone 3: Task Analysis & Video Prototype

Milestone 3: Task Analysis & Video Prototype

Due 11:00am Weds 3/25.

In class 3/9, we began to create a *scenario* paper prototype of the most salient task users should be able to complete with your tool. Now it's time to further flesh out your tool by creating a *task analysis* detailing all possible tasks, and by recording a *video prototype* of you interacting with the paper prototype across not one, but the three most important tasks/scenarios.

Personas

Before you decide on what tasks your tool should support, it may be helpful to envision the kinds of people who would be likely to use your tool. One common design strategy is to create *personas*: basically fake people with realistic backstories and realistic goals that will help you concretely imagine users–and thus usage scenarios–for your tool.

For instance, an example persona for our previously mentioned blackout poetry tool could be Maya:

> **Background**: Maya is a 27-year-old based in Los Angeles. With a BS in English, she has a passion for creative writing, though has yet to find success in that as a career (she works a boring spreadsheet desk job to pay the bills). Maya loves to travel over the world for inspiration for her creative practice.

> **Goals:** Maya mainly does blackout poetry to relax as a hobby, but also enjoys being able to reappropriate texts to give them new meaning and new life. She likes sharing her blackout poetry on her IG stories. Because she travels so much, she would like to be able to digitally do black out poetry on her phone on trains, as opposed to at a desk with physical sheets of paper.

Come up with 2 personas for your tool, and try to make them have as different background as possible. For instance, if Maya is a user that is close to your own lived experiences, what about someone who is much older? Someone who is a professional rather than a hobbyist? Try to capture the wide walls of your tool in your personas: your imagined users should span a wide space of usage potential.

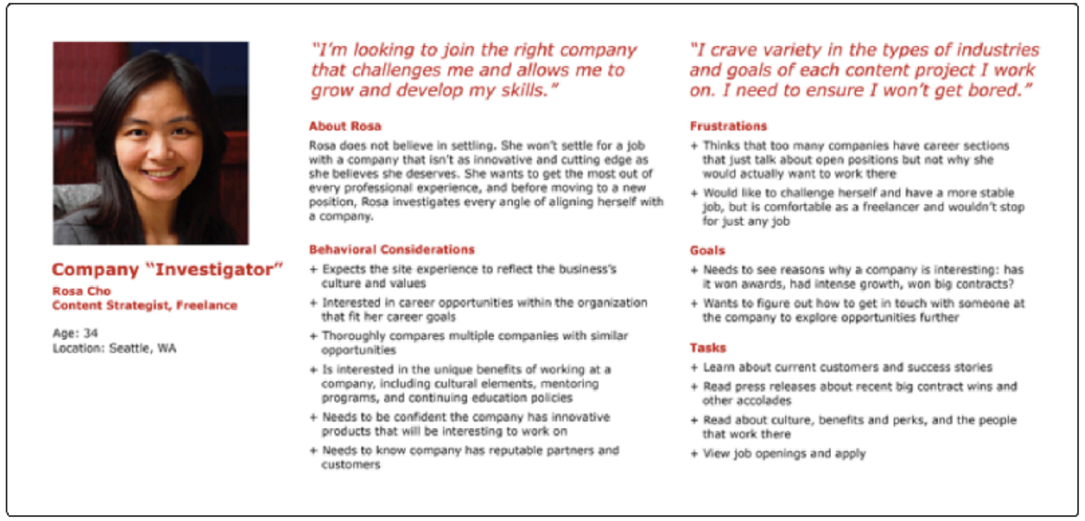
1. Create 2 personas

2. Make a task analysis diagram of the three most frequent goals

3. Make these goals into a paper prototype & take a video

Personas

- Personas are fictional characters you create to represent real users and captures their diverse backgrounds, interests, and goals
  - Cognitive aids to concretely think through interactions - “given their background, what would my persona do?”
  - Again, concreteness is more powerful in design than abstractness



Company “Investigator”

Rosa Cho
Content Strategist, Freelance

Age: 34
Location: Seattle, WA

“I’m looking to join the right company that challenges me and allows me to grow and develop my skills.”

“I crave variety in the types of industries and goals of each content project I work on. I need to ensure I won’t get bored.”

About Rosa

Rosa does not believe in settling. She won’t settle for a job with a company that isn’t as innovative and cutting edge as she believes she deserves. She wants to get the most out of every professional experience, and before moving to a new position, Rosa investigates every angle of aligning herself with a company.

Behavioral Considerations

+ Expects the site experience to reflect the business’s culture and values
+ Interested in career opportunities within the organization that fit her career goals
+ Thoroughly compares multiple companies with similar opportunities
+ Is interested in the unique benefits of working at a company, including cultural elements, mentoring programs, and continuing education policies
+ Needs to be confident the company has innovative products that will be interesting to work on
+ Needs to know company has reputable partners and customers

Frustrations

+ Thinks that too many companies have career sections that just talk about open positions but not why she would actually want to work there
+ Would like to challenge herself and have a more stable job, but is comfortable as a freelancer and wouldn’t stop for just any job

Goals

+ Needs to see reasons why a company is interesting: has it won awards, had intense growth, won big contracts?
+ Wants to figure out how to get in touch with someone at the company to explore opportunities further

Tasks

+ Learn about current customers and success stories
+ Read press releases about recent big contract wins and other accolades
+ Read about culture, benefits and perks, and the people that work there
+ View job openings and apply

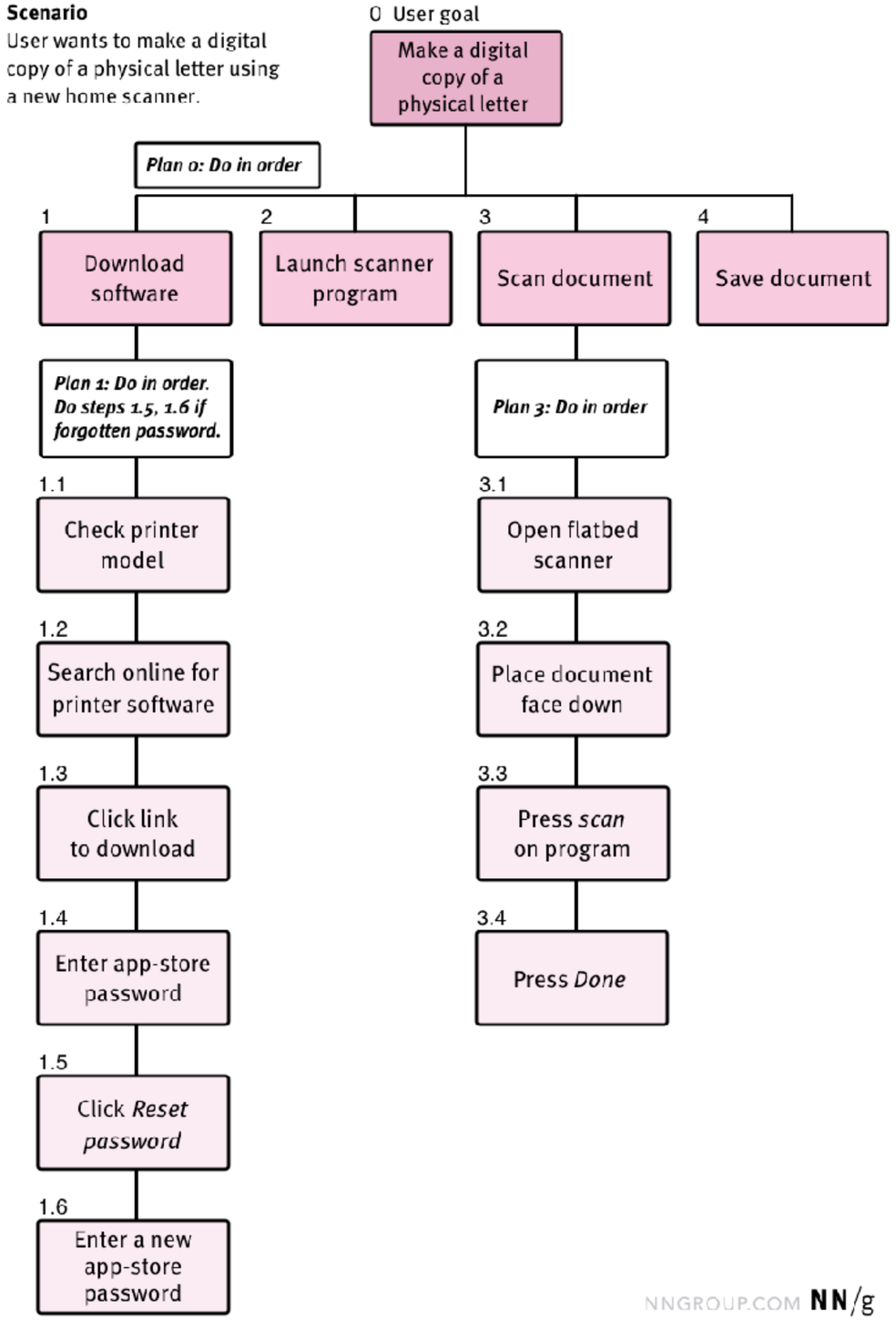
- Personas could hold the “promise of empathy” but may also fall into stereotypes or not consider social and structural power relations—more in your next reading response!

https://www.nngroup.com/articles/persona/

Task analysis

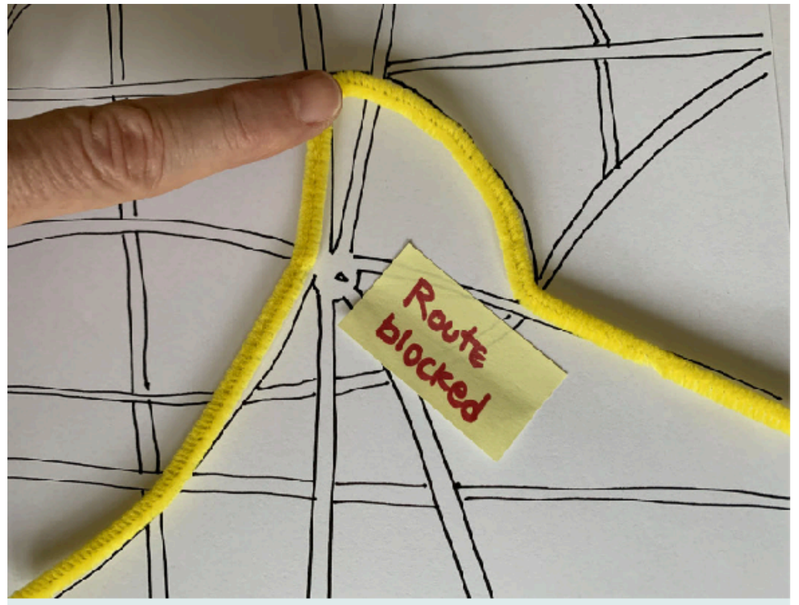
- Concreteness is more powerful in design than abstractness
  - Break down abstract “goals” into specific tasks (which buttons are clicked? What kinds of user input are needed?)
- In class, you identified **one** main scenario. For task analysis/video prototype, you should cover all important tool scenarios (~3)

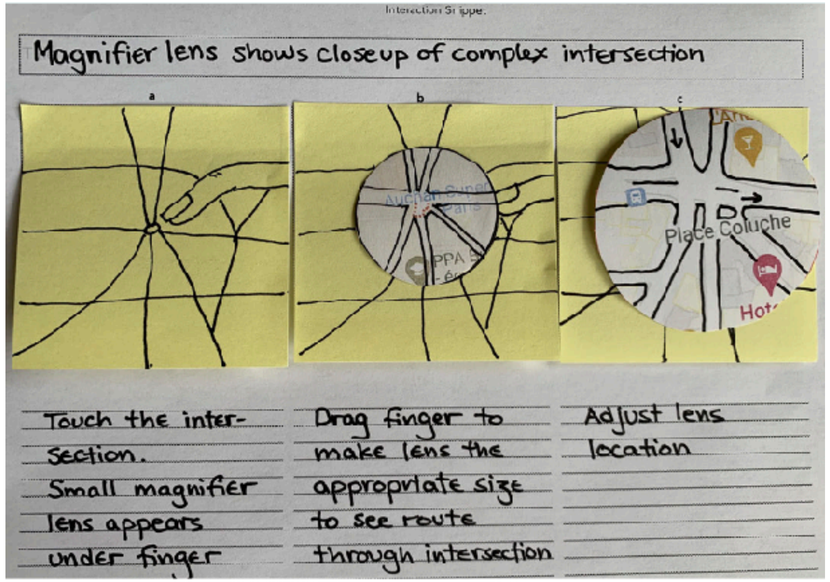
https://www.nngroup.com/articles/task-analysis/



Video prototype

- Translate task analysis to UI & UX
- Create usable design resources
- Have interaction snippets to prepare for complex interactions
- Take a video of a wizard-of-oz interaction with the paper prototype
- Main evaluative goal: investigate **breakdowns** in conceptual models through interactivity



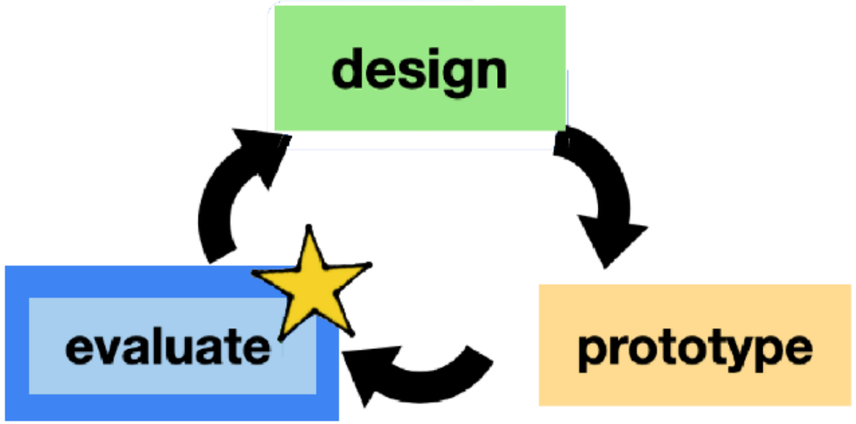


https://www.lri.fr/~mbl/pdf/DOIT-CHI23-Preview-Web.pdf

Evaluating tools

Formative feedback ≠ evaluation

- Even though both may involve going to users and collecting feedback, your paper prototype "evaluation" is for *formative design feedback*: evaluation to *iterate* in the design cycle



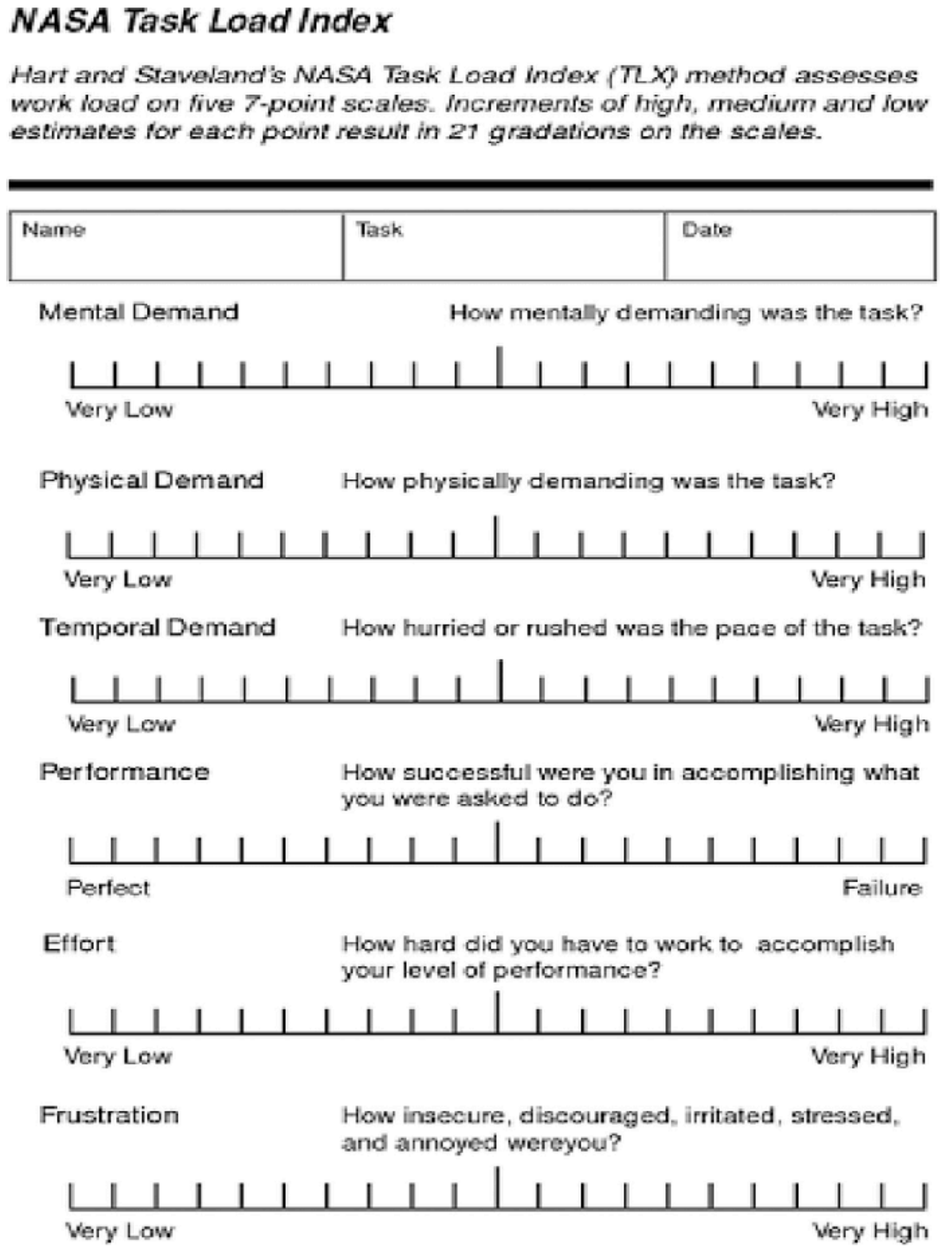
- This lecture will be talking about evaluating tools at the very end (e.g., before publishing a research paper) to "prove" that they're "good"

Why evaluate?

- How do we know if we met our design goals?
- How do we know if our tool is good?
  - Good could mean useful, expressive, helps you do something faster, enables an interaction that isn't enabled before, gives users more power…up to you to choose what "good" is, as long as you have *operationalizable metrics*

Common metrics: NASA-TLX

- NASA-TLX uses self-reported **likert scales** (rating 1-7) to convert qualitative feelings into quantitative numbers (ordinal data)
- Across categories of
  - Mental demand
  - Physical demand
  - Temporal demand
  - Performance
  - Effort
  - Frustration



NASA Task Load Index

Hart and Staveland's NASA Task Load Index (TLX) method assesses work load on five 7-point scales. Increments of high, medium and low estimates for each point result in 21 gradations on the scales.

Name	Task	Date

Mental Demand — How mentally demanding was the task?
Very Low — Very High

Physical Demand — How physically demanding was the task?
Very Low — Very High

Temporal Demand — How hurried or rushed was the pace of the task?
Very Low — Very High

Performance — How successful were you in accomplishing what you were asked to do?
Perfect — Failure

Effort — How hard did you have to work to accomplish your level of performance?
Very Low — Very High

Frustration — How insecure, discouraged, irritated, stressed, and annoyed wereyou?
Very Low — Very High

Common metrics: SUS

- System usability scale (1986), 10 questions

1. I think that I would like to use this system frequently.
2. I found the system unnecessarily complex.
3. I thought the system was easy to use.
4. I think that I would need the support of a technical person to be able to use this system.
5. I found the various functions in this system were well integrated.
6. I thought there was too much inconsistency in this system.
7. I would imagine that most people would learn to use this system very quickly.
8. I found the system very cumbersome to use.
9. I felt very confident using the system.
10. I needed to learn a lot of things before I could get going with this system.

Common metrics: CSI

- Similar to the NASA-TLX, but specifically developed for creativity support tools
- Across categories of
  - Exploration
  - Collaboration
  - Engagement
  - Effort/Reward Tradeoff
  - Tool Transparency
  - Expressiveness

Discuss: What are some shortfallings of self reported likert scale evaluations?

Exploration

It was easy for me to explore many different options, ideas, designs, or outcomes without a lot of tedious, repetitive interaction.

Agree Disagree

Collaboration

I was able to work together with others easily while doing this activity.

Agree Disagree

Engagement

I was very absorbed/engaged in this activity - I enjoyed it and would do it again.

Agree Disagree

Effort/Reward Tradeoff

What I was able to produce was worth the effort required to produce it.

Agree Disagree

Tool Transparency

While I was doing the activity, the tool/interface/system "disappeared," and I was able to concentrate on the activity.

Agree Disagree

Expressiveness

I was able to be very expressive and creative while doing the activity.

Agree Disagree

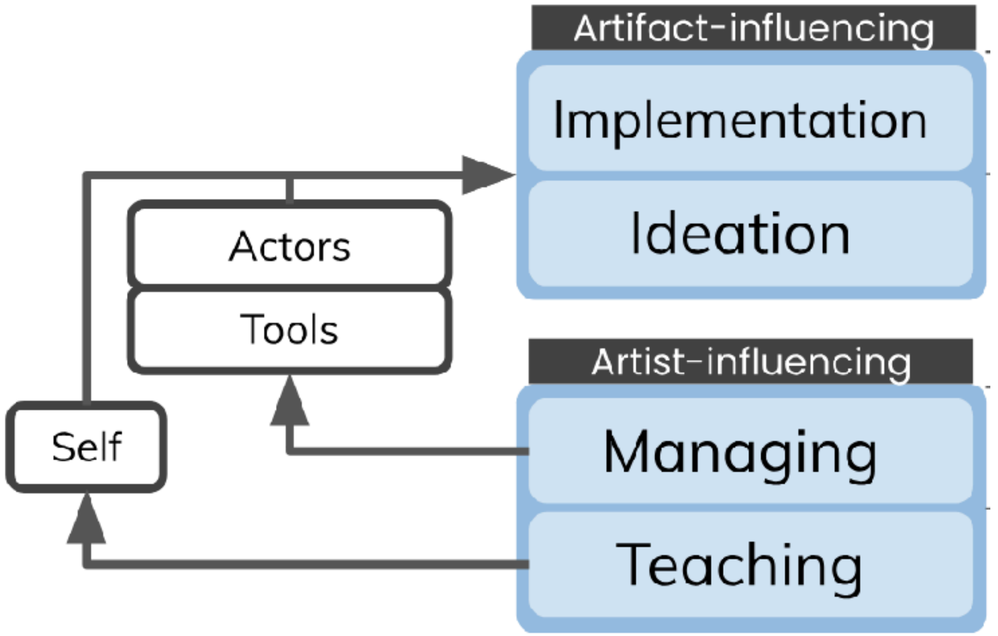
Creativity is hard to measure

- Seriously, there is no research or literature that agrees on how to measure creativity
- Part of this is that creativity is often *domain specific* and *social*
- My personal opinion is *quantitative* studies are less well suited for creative tools (but certainly useful for other kinds of tools, like productivity tools)
- Many CST researchers have moved towards collecting *qualitative* data, like participant thinkaloud quotes, or short post-interviews

Critiques of ecological validity

- Most studies are done in-lab in 1-2 hour sessions, yet we develop relationships with tools *over time*
  - **Ecological validity**: Is the context of the tool use for evaluation actually the context in which users will use it in “real life”?
  - A push for “in the wild” longitudinal studies, observing use over 3+ weeks

Theory: Artist support networks

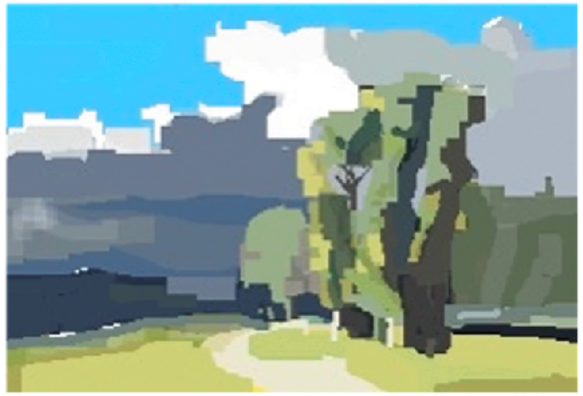


- We can't just think about evaluating tools, but *people* play a big role in how tools are used as well
- Artists teach other other, support each other, and help develop wholistic emotional wellbeing. These measures are often not captured with traditional evaluation metrics, but are maybe a big part of the reasons why we use tools?

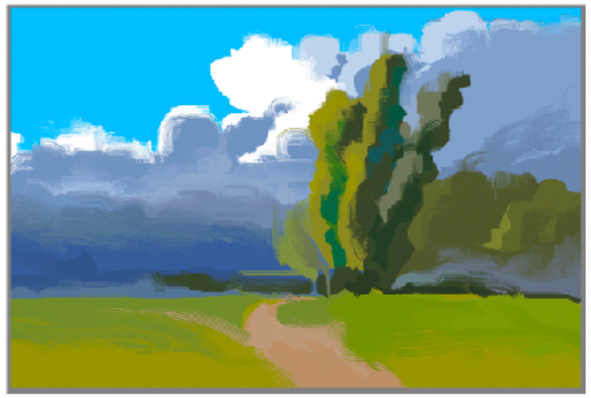
Artist Support Networks: Implications for Future Creativity Support Tools. Chung et al., DIS 2022. https://dl.acm.org/doi/10.1145/3532106.3533505

Artist support networks provide inspiration, creative safety, and validate individual creativity

1B: Trackpad pixel artist
(Week 1)



1B: Trackpad pixel artist
(Week 2)



Other participants in Group 1
also used trackpad for week 2

Example from my current submission to DIS 2026, a longitudinal group based study of ArtKrit

Prof Li’s philosophy

- If creativity is social, why don’t we evaluate our tools in groups to study how group dynamics
  - Tension: qualitative wholistic view vs “generalizable” knowledge — this is called interpretivism vs positivism in sociology
- Instead of treating evaluations as extractive measures where we gather data, why can’t we create meaningful opportunities for artistic engagement? If our target users are artists, why can’t we create opportunities for them to make art?
- Let’s **co-create evaluations with participants**: Like participatory design, but not co-designing the *tool,* but rather the evaluation experience in artist support networks
- Towards further mitigating power dynamics between “researchers” and “artists”

Other evaluation methods

Hypothesis testing

- We can frame our evaluations as hypothesis tests and conduct quantitative experiments of statistical significance for evaluation.
  - Hypothesis: What do you want to believe to be true about your tool?
  - Independent variable: the thing you’re changing
  - Dependent variable: the metrics you’re measuring to see how they are affected by changing the independent variable

Between vs within subjects design

Between subjects

Two participant groups.

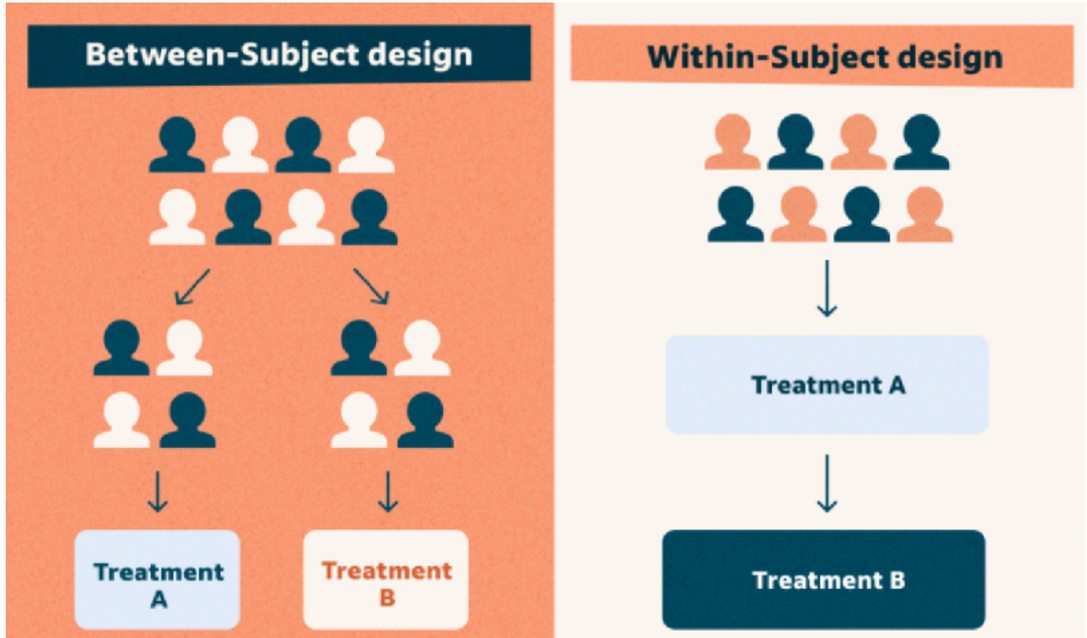
1 group only uses design A.

1 group only uses design B.

Within subjects

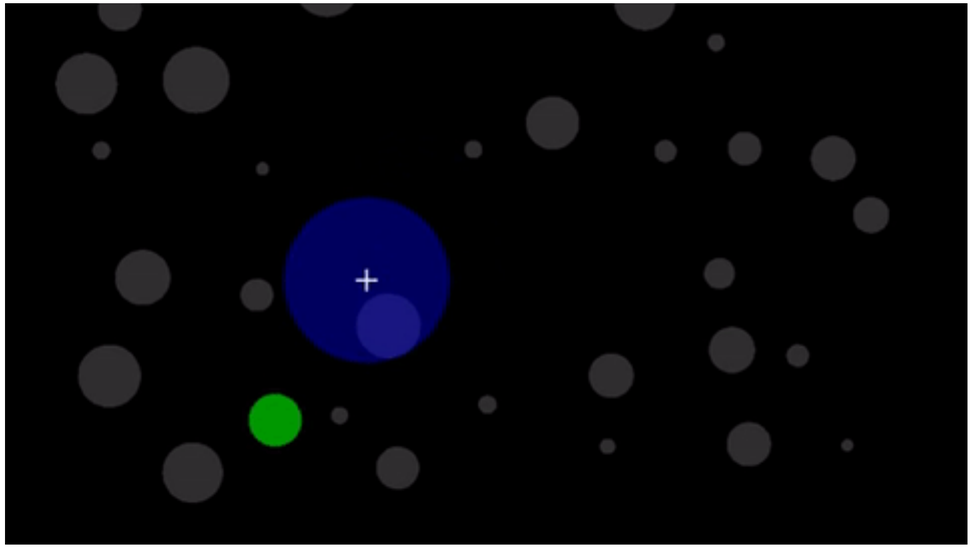
Everyone uses design A and B.

Random ordering (A first or B first) is important to avoid temporal bias!



Example: bubble cursor

- Hypothesis: Users click on targets faster with the bubble cursor
- Independent variable: Cursor type (regular vs bubble)
- Dependent variable: Movement time
- Within subjects study



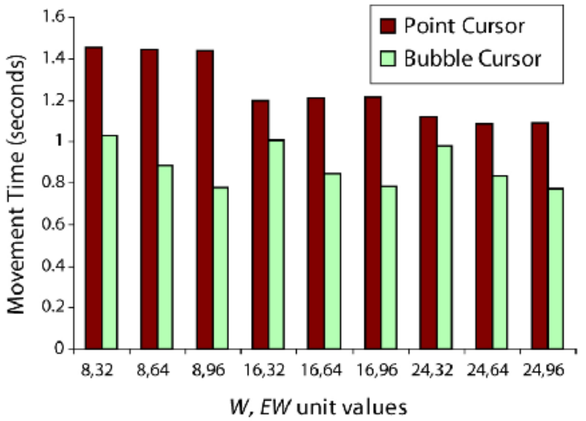


Figure 6. Movement time by *W*, *EW* values for both cursors, averaged over all *A* values.

Grossman et al. The Bubble Cursor. CHI 2005

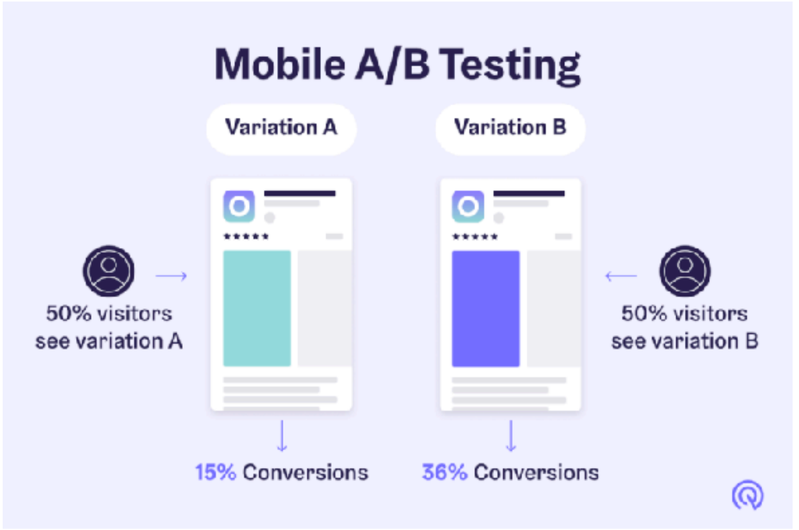
Experiments in the real world

- Research conducted at an institution needs to go through IRB (Institutional review board) approval for ethics
- Requires obtaining the informed consent of participants and identifying potential harms
- Requires detailing study design, variables, randomization, and trials
- Class projects do not need IRB approval :)

Thanks Philip Zimbardo and the Stanford Prison Experiment

How about for design?

- A/B testing: Between subjects testing of one page version or another, usually has dependent variables like click through rate
- For your tool, if you want to do quantitative studies, you could consider comparing to an existing tool as the “control group”

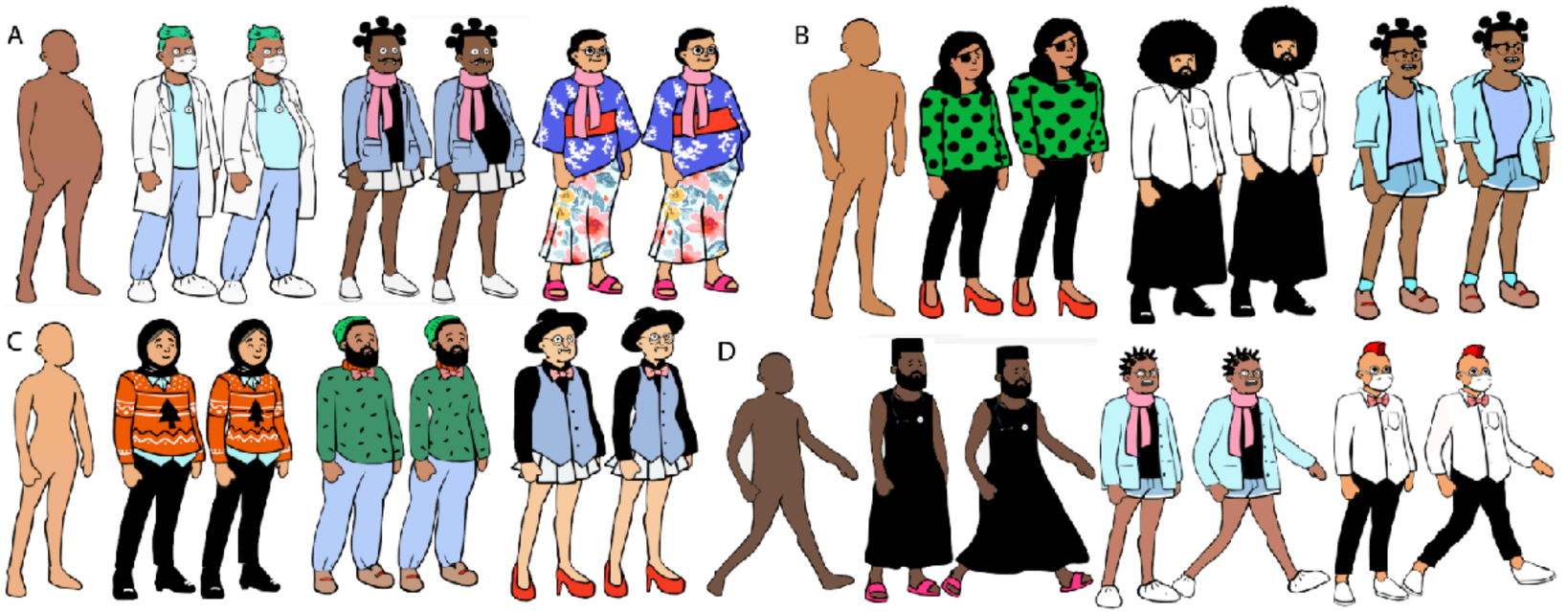


Qualitative studies

- We've already learned about think aloud protocols, semi-structured interviews, and contextual inquiries
- These are all methods of collecting *qualitative* data
- How do we analyze qualitative data (eg text transcripts?)
  - **Thematic analysis**: from your qualitative data, annotate for "codes" or categories. Then look at all codes to create common "themes" that emerge

Existence proof

- Some HCI researchers believe that the tool existing (and showing a range of artifacts the tool can generate) is enough evaluation
- Reviewers can look at the results to make their own judgement calls



Existence proof by generating a wide range of examples with the tool

Your turn

- Activity: **How should you evaluate your tool?**
  - What were your design goals?
  - Matching your design goals, write 2-3 initial hypotheses you have about your tool that can be answered through *observation.*
  - Then write the independent and dependent variables for each hypothesis.
  - Can you pick one metric we've just talked about to match with each design goal? (A/B test it? Likert scales? Post-study interview?)
  - Write this in your design documentation, you'll come back to it the second to last class

- Example: Fading drawing strokes tool
- Design goal: Making drawing not intimidating
- Hypothesis: Using this tool will reduce the pressure of getting started with drawing
- IV: Tool usage; DV: Time it takes to get started drawing.
- Metrics: collect timing information (quant), post-interview asking about feelings getting started (qual)

Class 14 recap

- TODO
  - Fill out the participation form by Friday EOD https://forms.gle/EevGjNs1qVL2uAsp9 and midsemester feedback form https://forms.gle/bWJYu6c7AcKtvDhX9
- Have a great spring break! (Note that your TAs will lead class the following week)
  - Mon: RR and seminar from Harrison & Vika
  - Weds: **Milestone 3:** Task analysis & video prototype