CS181DT Class 16: Evaluation

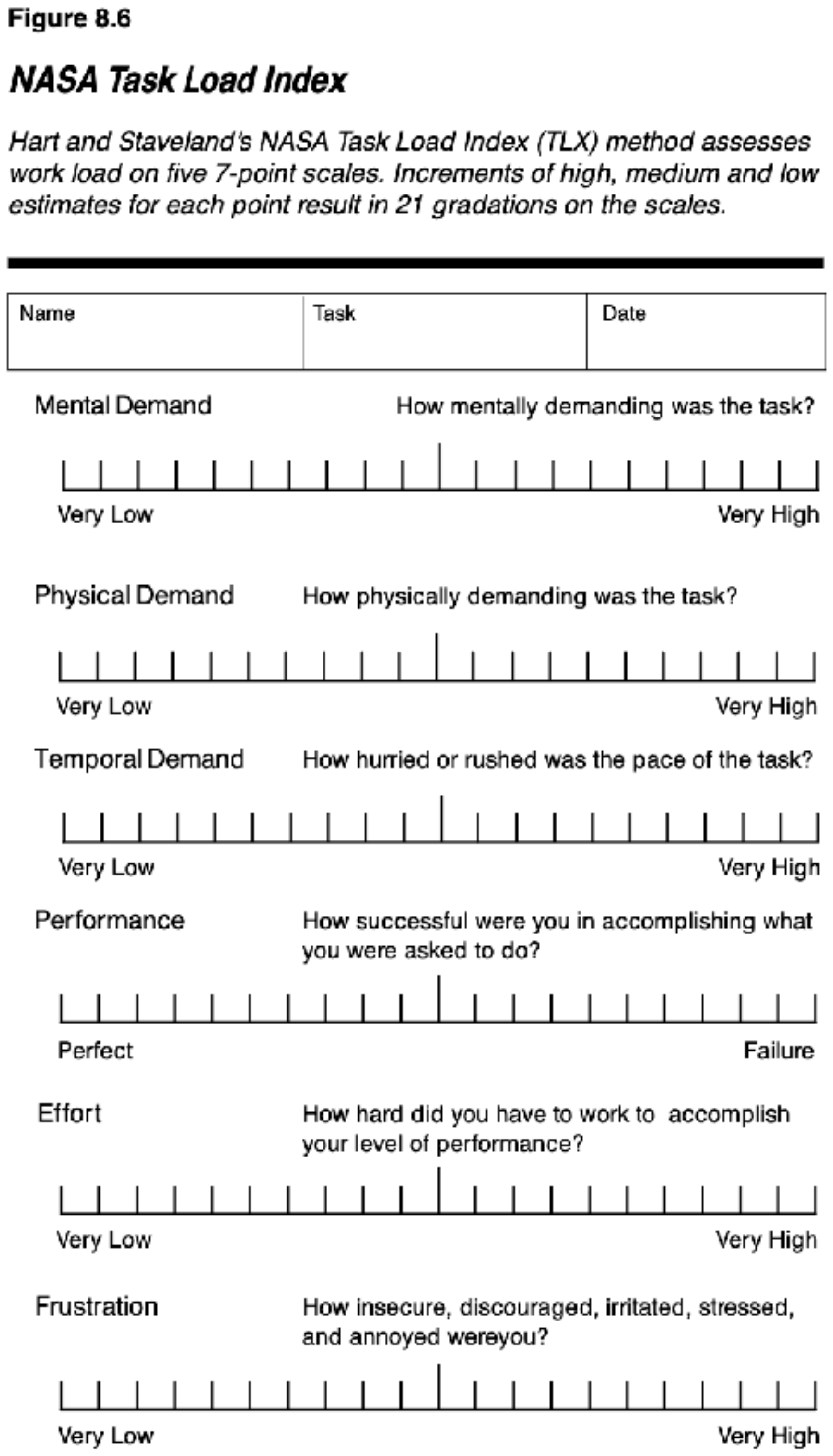


Figure 8.6

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.

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?

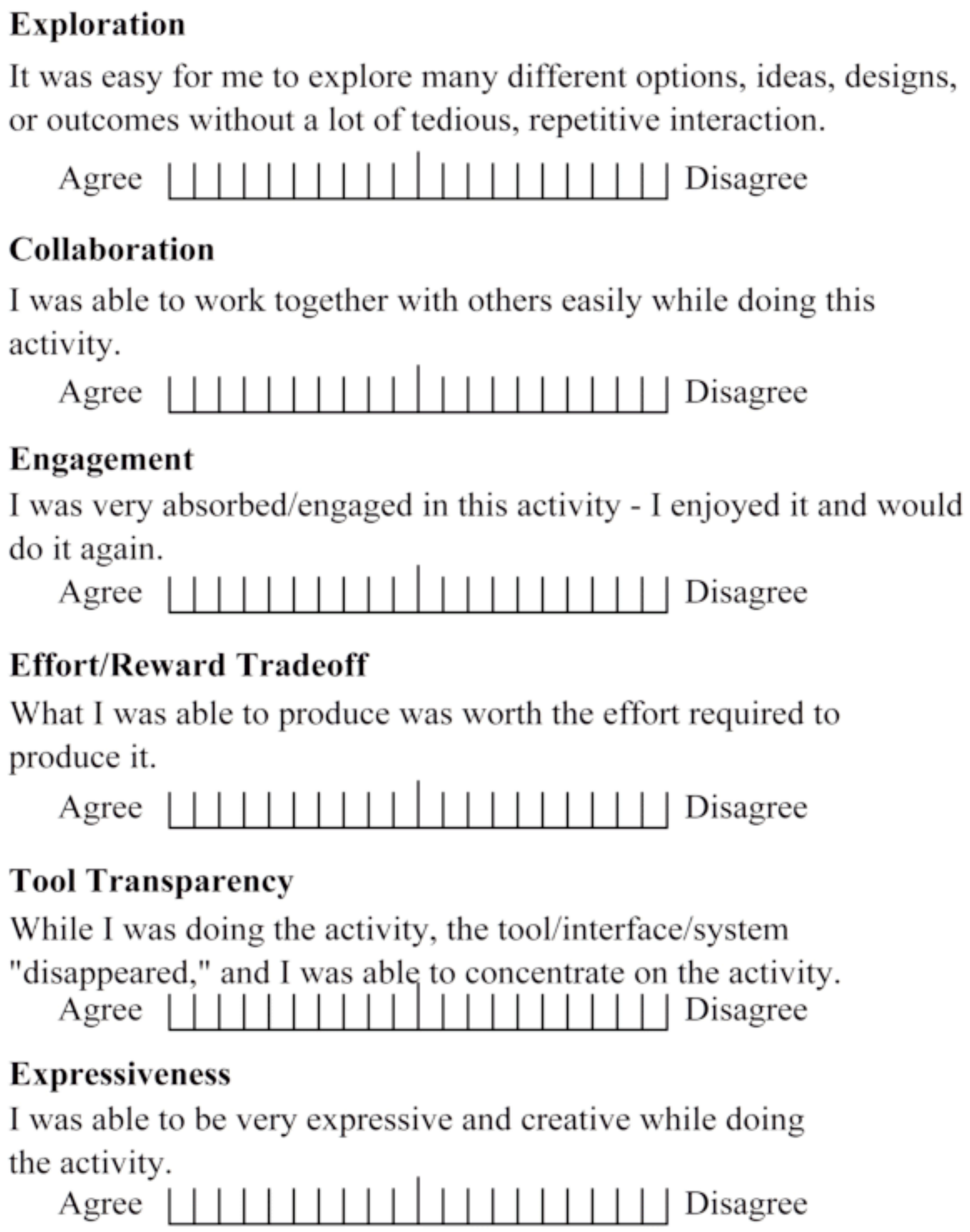
Perfect — Failure

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



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 16 agenda

- Zipcrit
- Studio: Evaluating your paper prototypes
- Break
- Lecture: Evaluating tools
- Milestone 5: Wizard-of-Oz prototype

Evaluating your wireframe paper prototype

Qualitative evaluation strategy: cognitive walkthrough

- A cognitive walkthrough requires a **prototype** and a **goal**
- Ask users to “**think aloud**” to understand what is going on cognitively
  - The user should not be silent. They should ideally always be talking!
    - “So I’m clicking this button because…”
    - “Okay, I’m not sure what to do here. My best guess is that I want to click [X] because I think it would [Y]…”

Your turn: paper prototype

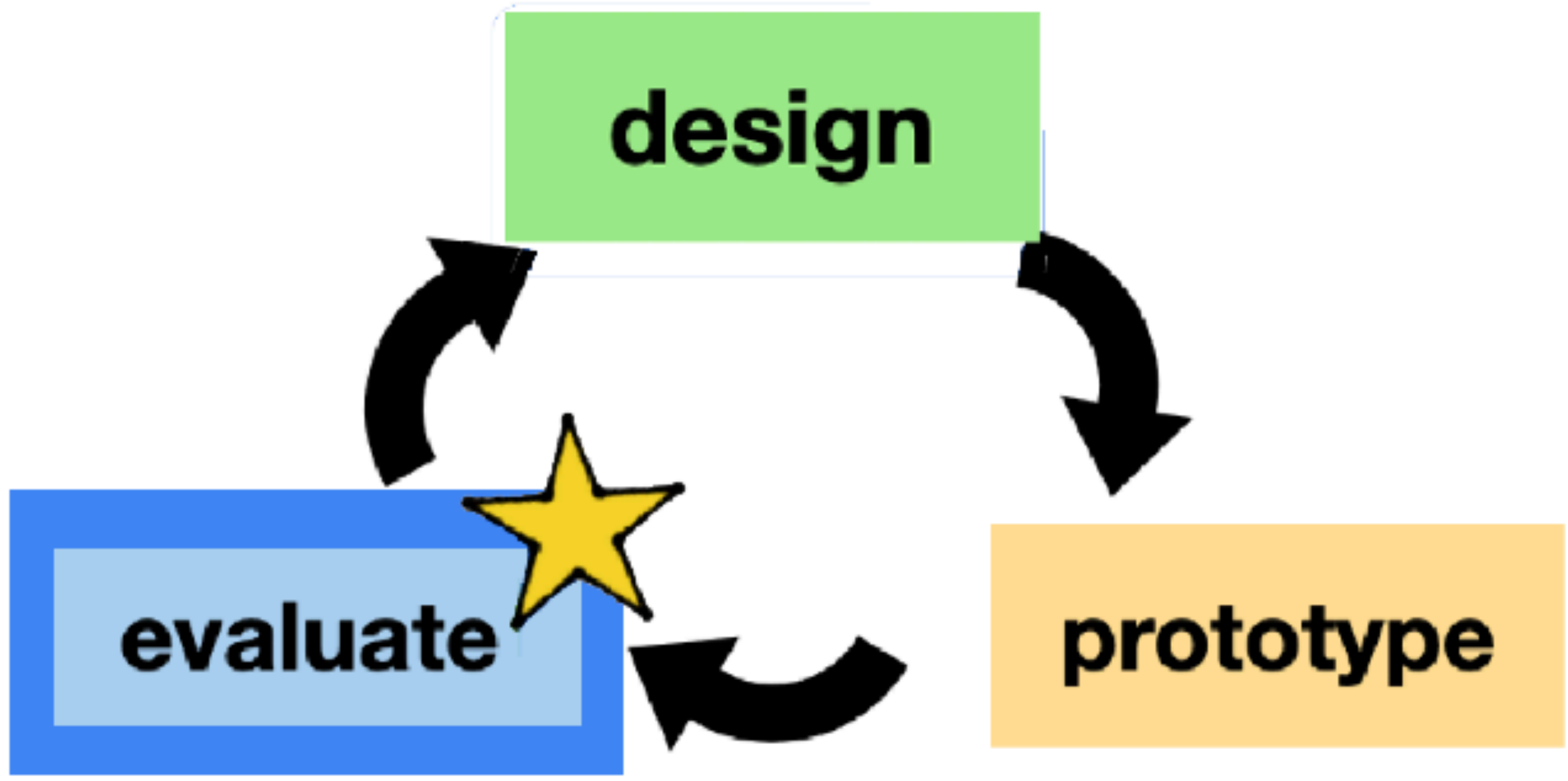
Remember to take photos and write notes in your design documentation!

- Find a group to swap with. Group B uses Group A's prototype first, and then we switch. After both groups are done, find a new pair and continue.
- Roles
  - Group A - WoZ computer: **Computers cannot speak or explain any UI elements** and can only prompt the user with a **goal** and switch out UI elements according to user interaction.
  - Group B - User: use the prototype and think aloud
  - Groups A & B - Observers: take copious notes on the interaction, write your takeaways/analysis of the situation. What is easy for users to do? What do they struggle with? Remind users to think aloud!
  - After the interaction is completed, anyone can give general feedback/thoughts
- Move on at 11:35 (take a break at 11:30)

Evaluating tools

Formative feedback ≠ evaluation

- Even though both may involve going to users and collecting feedback, what we just did was evaluation for *formative design feedback*: evaluation to *iterate* in the design cycle



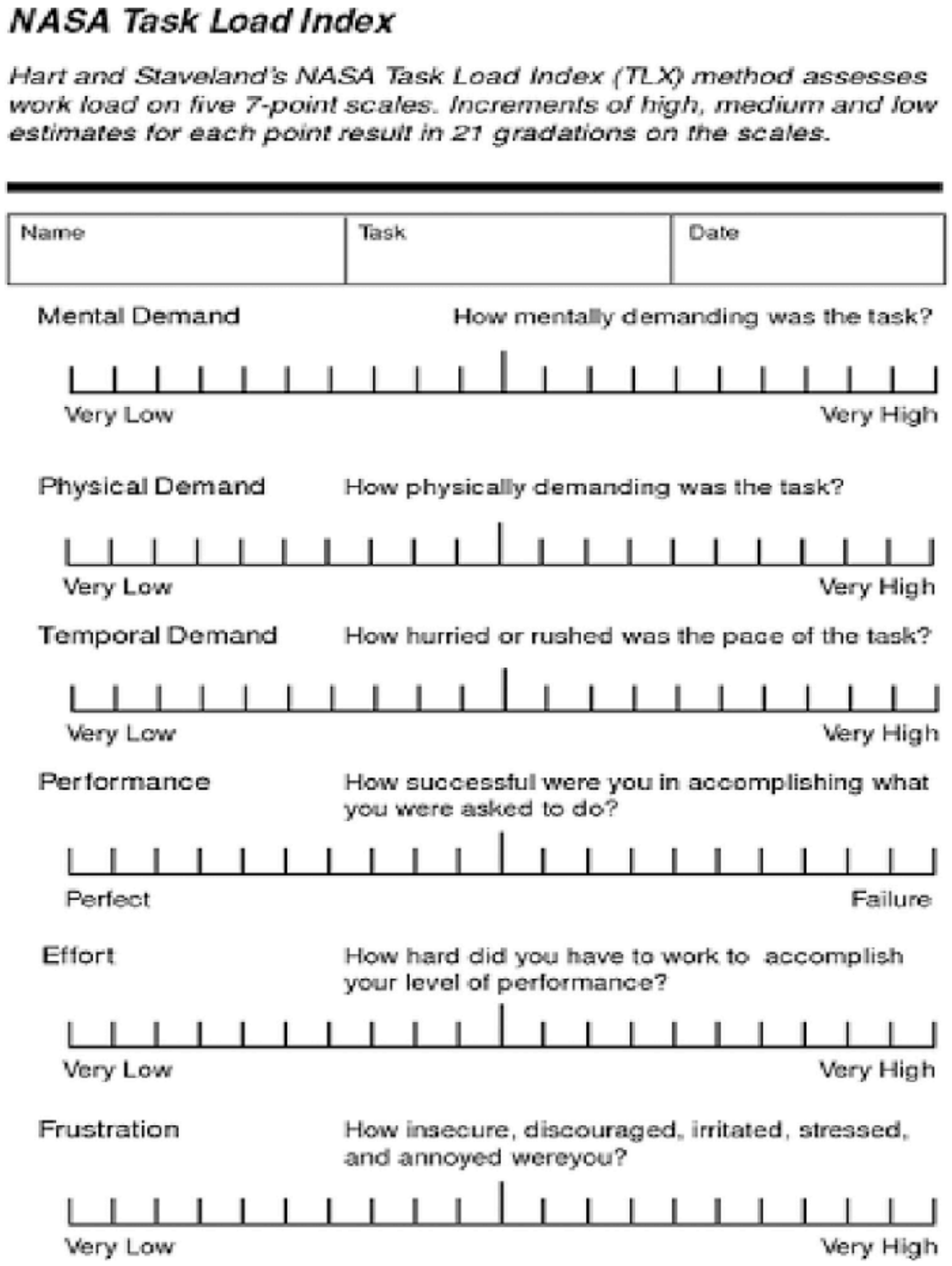
- This lecture will be talking about evaluating tools at the very end (e.g., before publishing a research paper) to "prove" that they're "good"

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

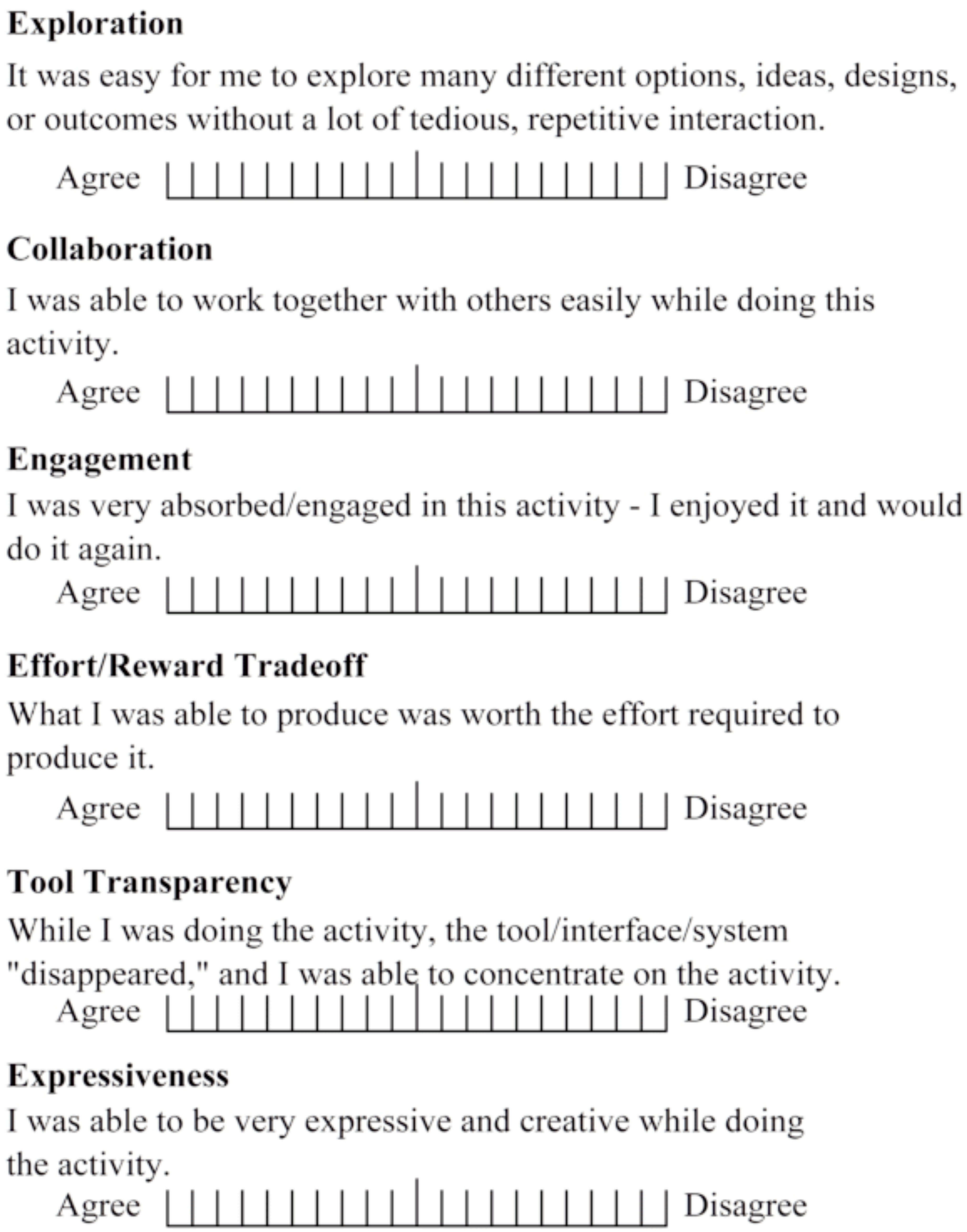
Frustration — How insecure, discouraged, irritated, stressed, and annoyed wereyou?

Very Low — Very High

Common metrics: 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)

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.

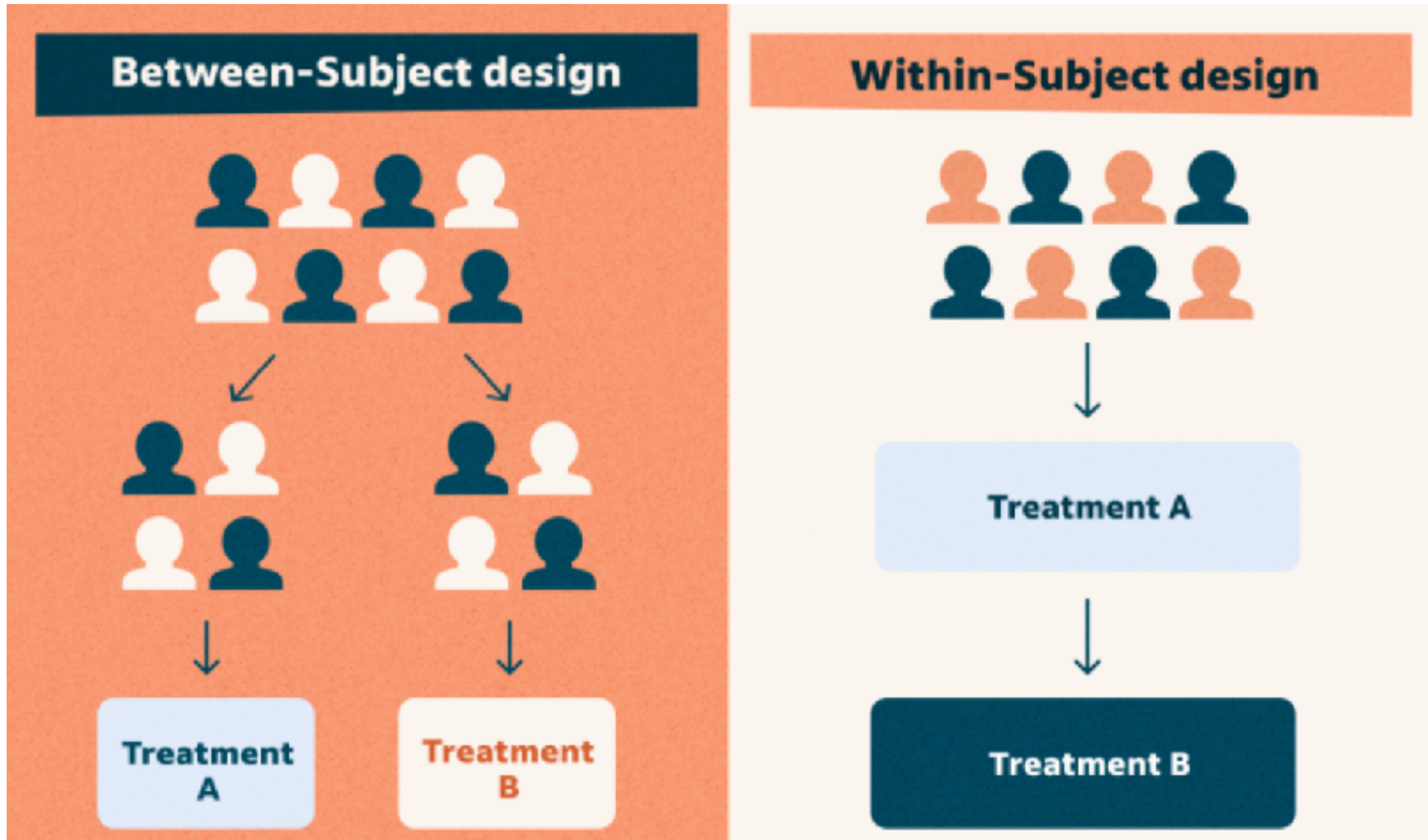
1 group only uses design A.

1 group only uses design B.

Within subjects

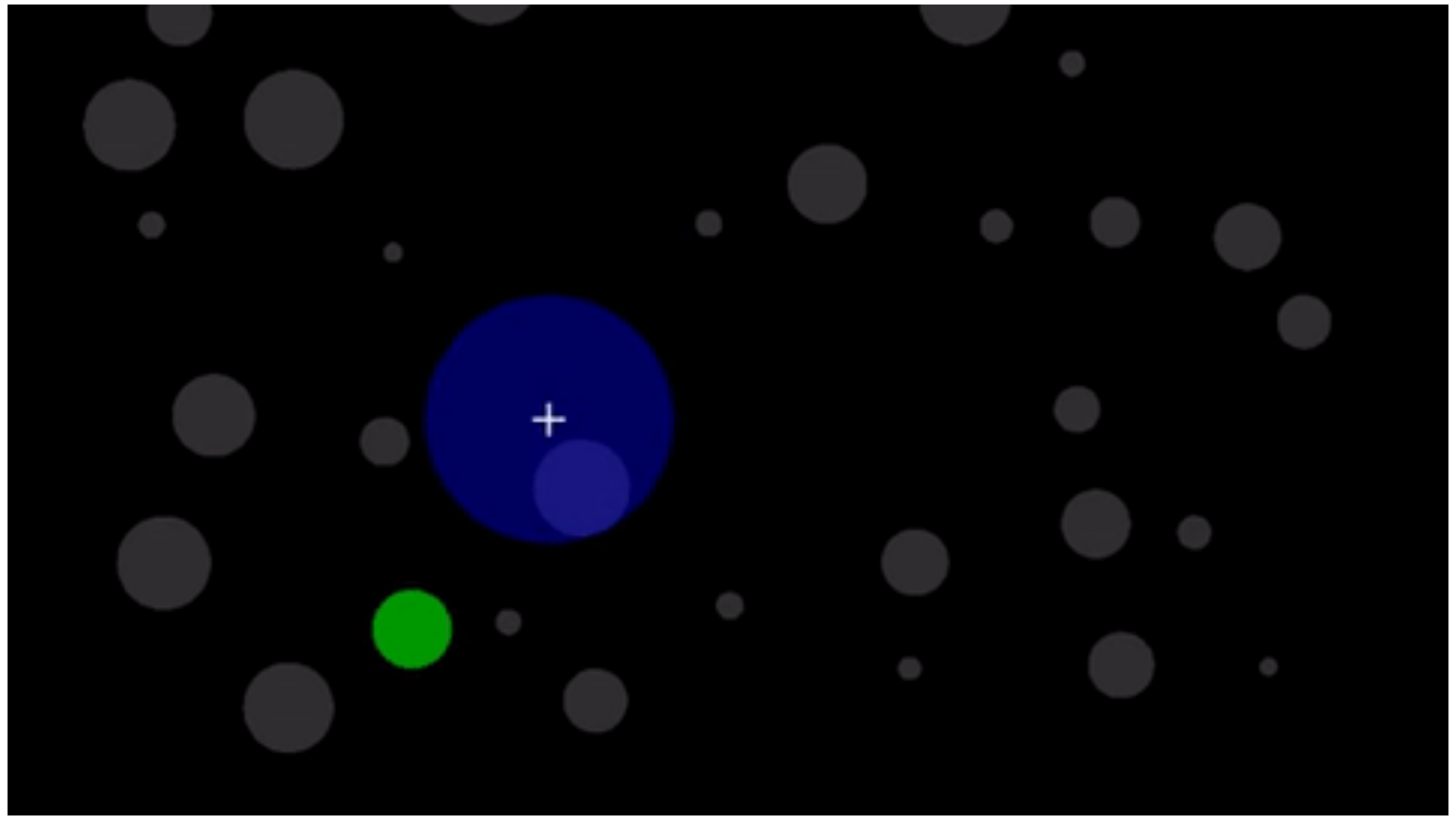
Everyone uses design A and B.

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



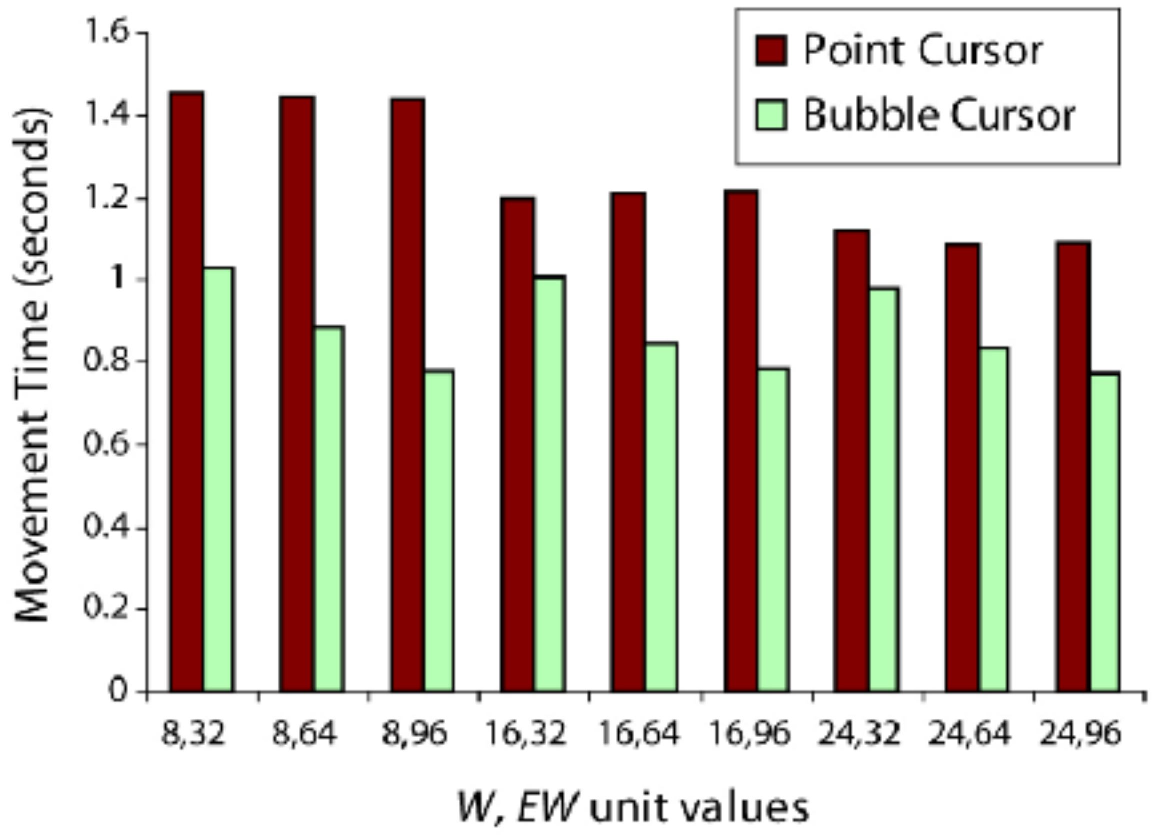


Figure 6. Movement time by *W*, *EW* values for both cursors, averaged over all *A* values.

Grossman et al. The Bubble Cursor. CHI 2005

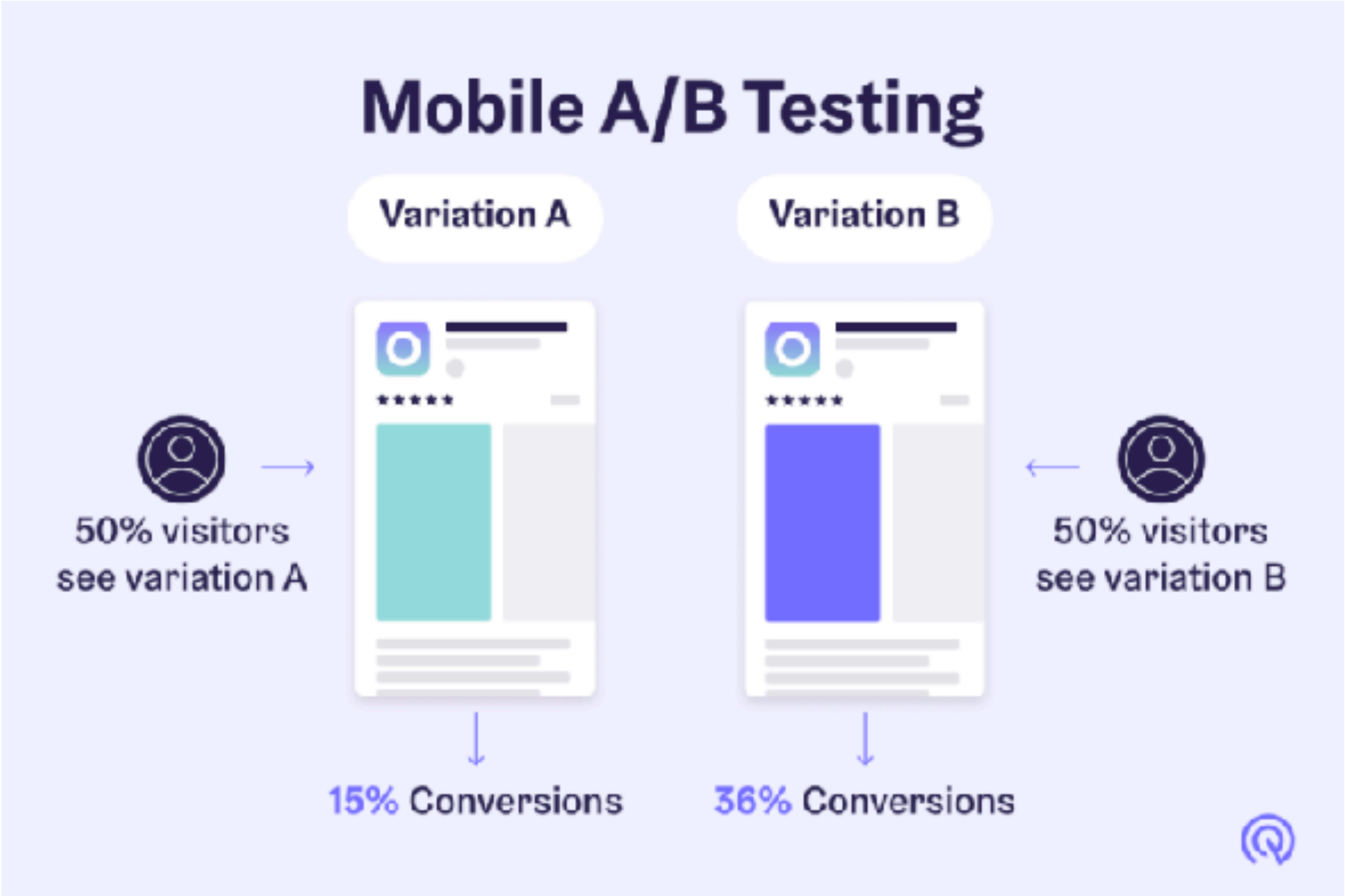
Experiments in the real world

- 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”

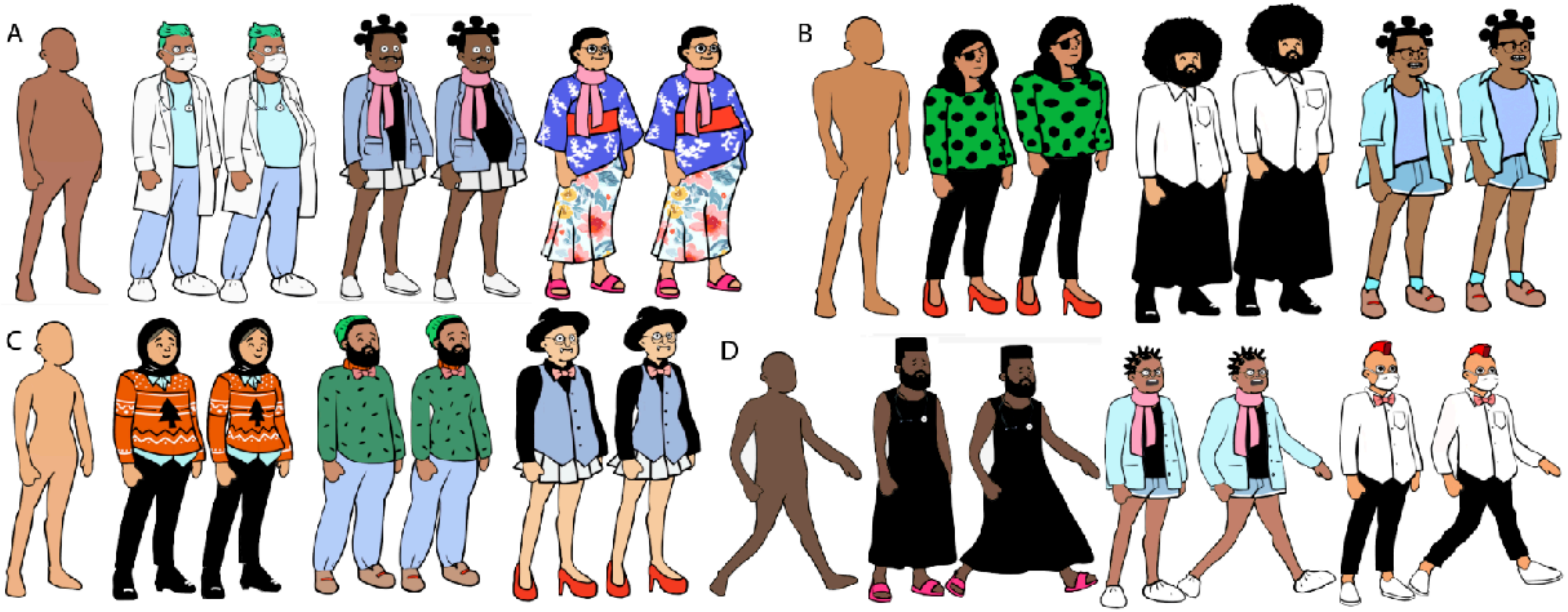


Qualitative studies

- We’ve already learned about think aloud protocols, semi-structured interviews, and contextual inquiries
- These are all methods of collecting *qualitative* data
- Other methods:
    - Longitudinal studies: give the tool to users for weeks+, collect usage data (also quant), conduct post-usage interviews. Benefits: ecological validity (done “in the field” in real contexts of use versus a controlled lab environment)
    - Thematic analysis: from your qualitative data (e.g., interview transcripts), annotate for 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



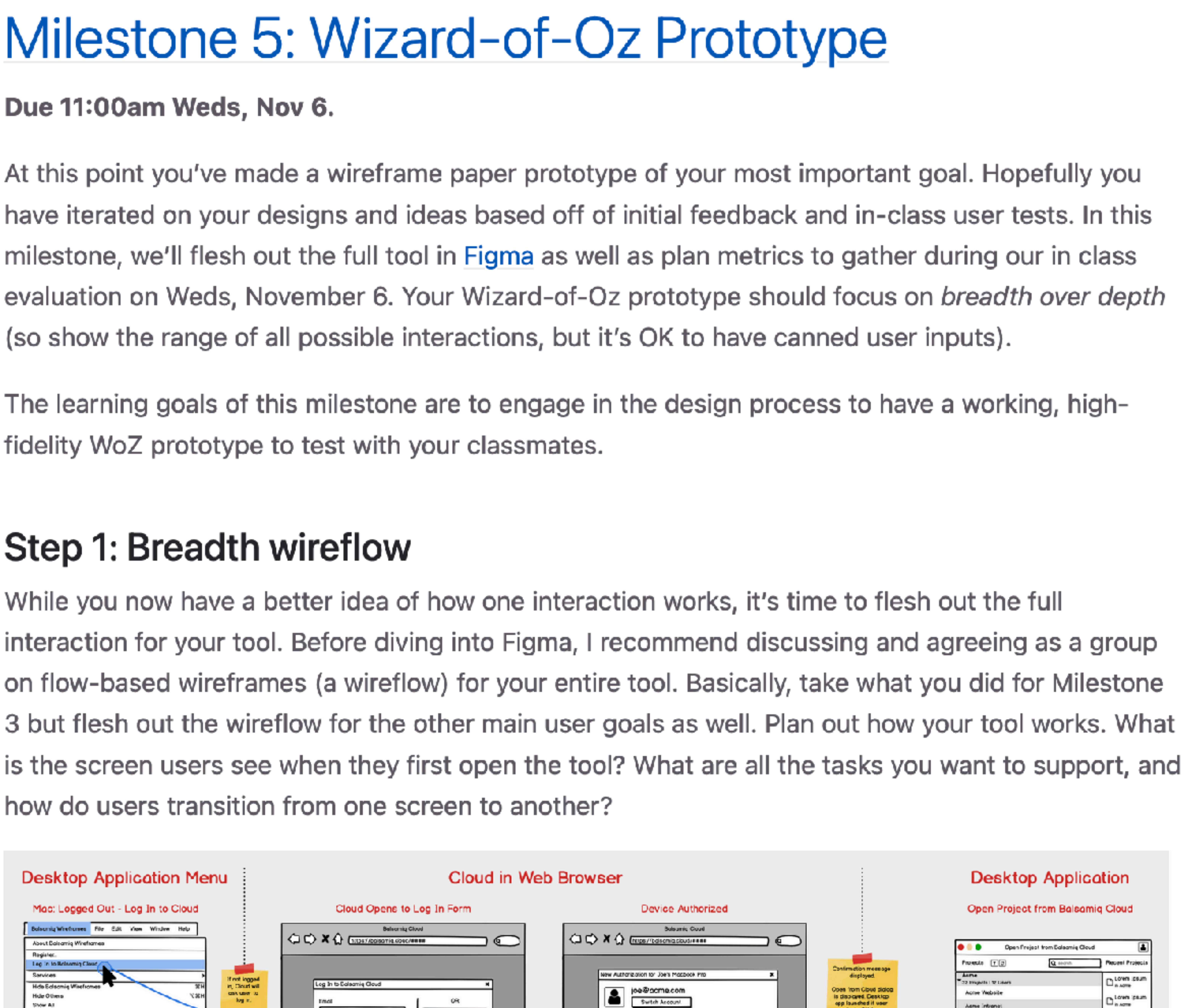
Existence proof by generating a wide range of examples with the tool

Your turn

- Activity: **How should you evaluate your tool?**
  - In groups, first brainstorm and write in your design documentation 2-3 initial hypotheses you have about your tool that can be answered through *observation.* (Go back to your design goals!)
  - Then write the independent and dependent variables for each hypothesis, and potential metrics for *how* you'll get the data. (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
- 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)

Milestone 5: WoZ prototype in Figma



Milestone 5: Wizard-of-Oz Prototype

Due 11:00am Weds, Nov 6.

At this point you've made a wireframe paper prototype of your most important goal. Hopefully you have iterated on your designs and ideas based off of initial feedback and in-class user tests. In this milestone, we'll flesh out the full tool in Figma as well as plan metrics to gather during our in class evaluation on Weds, November 6. Your Wizard-of-Oz prototype should focus on *breadth over depth* (so show the range of all possible interactions, but it's OK to have canned user inputs).

The learning goals of this milestone are to engage in the design process to have a working, high-fidelity WoZ prototype to test with your classmates.

Step 1: Breadth wireflow

While you now have a better idea of how one interaction works, it's time to flesh out the full interaction for your tool. Before diving into Figma, I recommend discussing and agreeing as a group on flow-based wireframes (a wireflow) for your entire tool. Basically, take what you did for Milestone 3 but flesh out the wireflow for the other main user goals as well. Plan out how your tool works. What is the screen users see when they first open the tool? What are all the tasks you want to support, and how do users transition from one screen to another?

Desktop Application Menu
Mac: Logged Out - Log In to Cloud
Cloud in Web Browser
Cloud Opens to Log In Form
Device Authorized
Desktop Application
Open Project from Balsamiq Cloud
If not logged in, Cloud will ask user to log in.
Confirmation message displayed.
Open from Cloud dialog is displayed. Desktop app launched if user
joe@acme.com
Switch Account

We're going from

low to high fidelity!

Class 16 recap

- Wednesday's class is **flipped,** please read the notes on the website before coming to class
  - Most of class time will be a brief lecture + project work time (goal: finish substeps 1 & 2 of Milestone 5)
- TODO
  - Project Intro 11:59pm this Friday
  - Milestone 5: WoZ Figma prototype **next** Weds (1.5 weeks)
  - Please come to OH if you have any trouble!!