



# Experiment Design



THAI-RS Workshop 5  
June 8, 2022

Sangho Suh  
Design Lab @ UCSD  
@sangho\_suh

---

# Sangho Suh

- Postdoc @ UCSD Design Lab
- PhD @ University of Waterloo
- Human-Computer/AI Interaction,  
Educational Technology,  
Creativity Support

<https://sanghosuh.github.io/>



---

## Previously...



1. How to ask research question



2. How to formulate hypothesis



3. How to collect qualitative & quantitative data



## This session ...



4. How to design experiment



## Experience with Experiment?





## Experience with Experiment?





## Goals of Experiment (WHY)





## Goals of Experiment (WHY)



- Understand real world
  - How users use AI technology
  - Can we help a potential user group with AI?
- Compare things
  - Best/better/worse





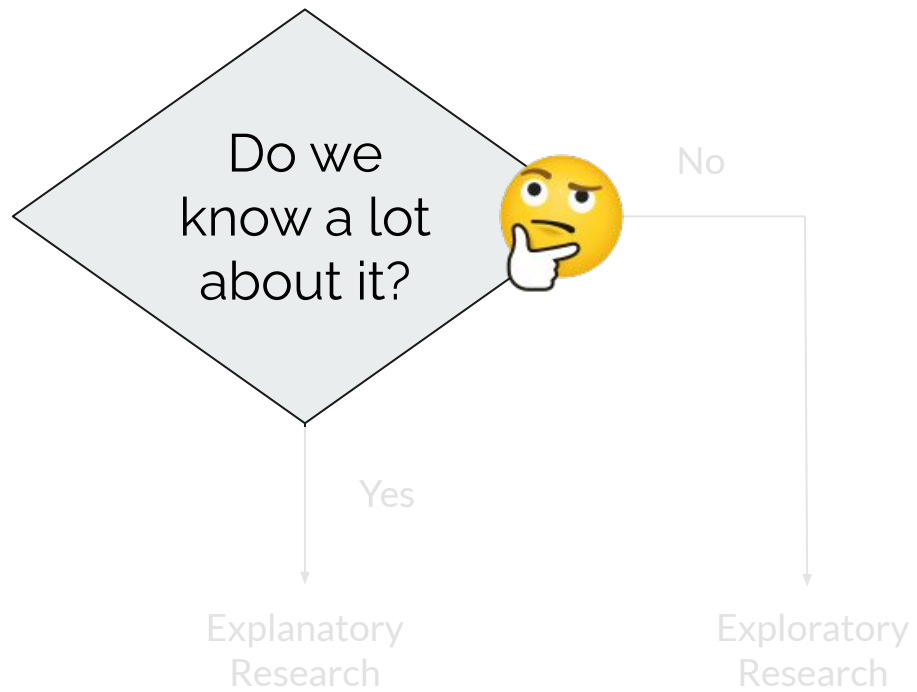
## Goals of Experiment (WHY)

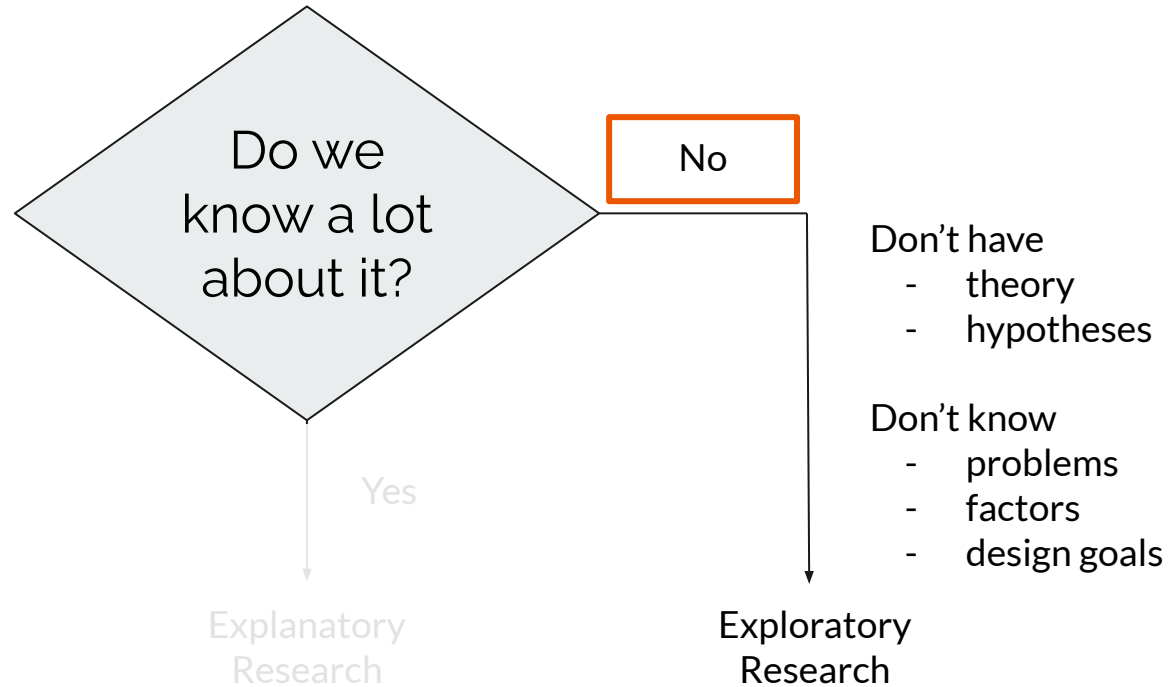


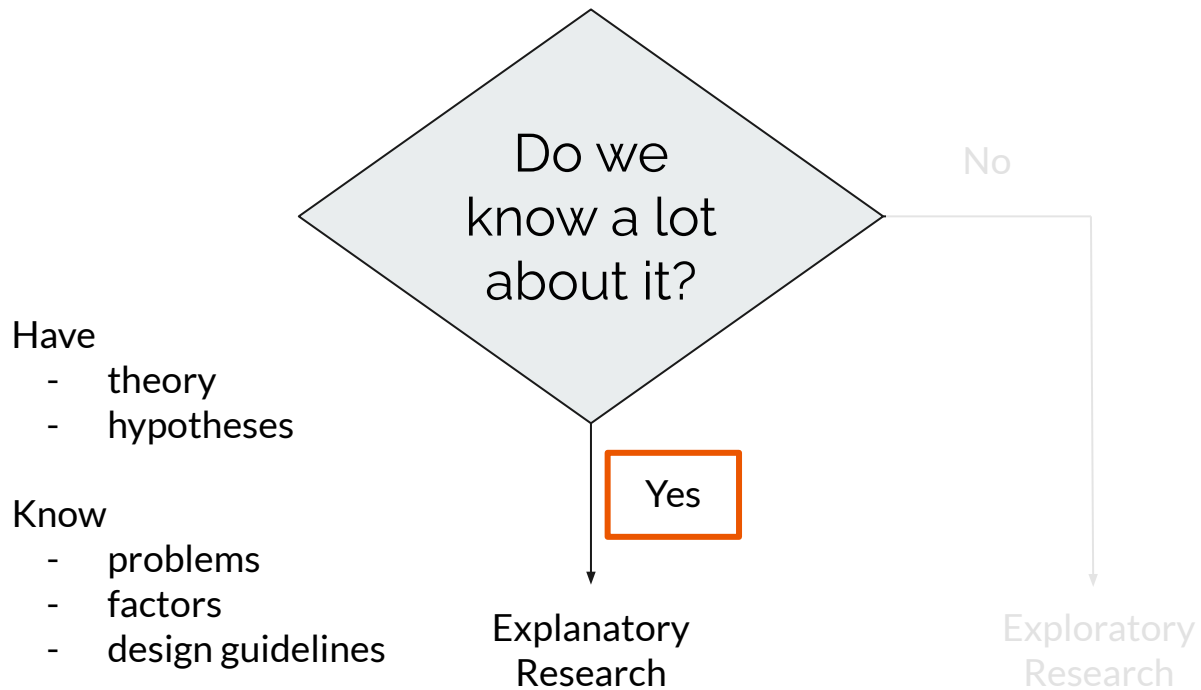
- Engineer toward a target
  - Essential features in AI-driven system
  - Is design X good enough for building trust in AI
- Check conformance to a standard
  - Human-AI Interaction guideline

# Where Do We Start?











## Do We Know?

How do nurses in nursing homes use a conversational agent?

Do conversational agents that inform nurses about their schedules help reduce the task time?

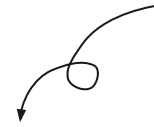


## Do We Know?

How do nurses in nursing homes use a conversational agent?

We **DON'T** know

Independent Variable




Do **conversational agents that inform nurses about their schedules** help **reduce the task time**?




Dependent Variable

We **HAVE** hypotheses




“Navigation applications are becoming ubiquitous in our daily navigation experiences. With the intention to circumnavigate congested roads, their route guidance always follows the basic assumption that drivers always want the fastest route. However, it is unclear how their recommendations are followed and what factors affect their adoption...”





“Navigation applications are becoming ubiquitous in our daily navigation experiences. With the intention to circumnavigate congested roads, their route guidance always follows the basic assumption that drivers always want the fastest route. However, it is *unclear* how their recommendations are followed and what factors affect their adoption.”



“Navigation applications are becoming ubiquitous in our daily navigation experiences. With the intention to circumnavigate congested roads, their route guidance always follows the basic assumption that drivers always want the fastest route. However, it is unclear how their recommendations are followed and what factors affect their adoption. **We present the results of a semi-structured qualitative study with 17 drivers, mostly from the Philippines and Japan. ...**”



## Qualitative Approach

- Motivated by questions that are broad and non-leading
  - How do people follow recommendations from navigation apps?



## Qualitative Approach


- Motivated by questions that are broad and non-leading
  - How do people follow recommendations from navigation apps? (O)
  - Would people follow recommendations from navigation apps if the apps rewarded people with badges? (X)



## Qualitative Approach (hypothesis-generating)


- Motivated by questions that are broad and non-leading
  - How do people follow recommendations from navigation apps? (O)
  - Would people follow recommendations from navigation apps if the apps rewarded people with coupons? (X)
- **Process**
  - Look for patterns
  - Build (theory) from ground up

a.k.a **grounded theory research** :  
building theories and hypotheses from data




“Collaboration is built on trust, and establishing trust with a creative Artificial Intelligence is difficult when the decision process or internal state driving its behaviour isn’t exposed. When human musicians improvise together, a number of extra-musical cues are used to augment musical communication and expose mental or emotional states which affect musical decisions and the effectiveness of the collaboration.”

We KNOW

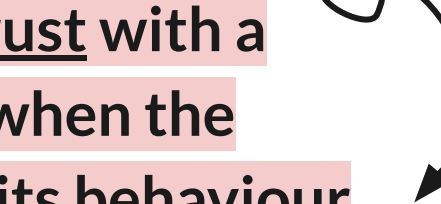


“Collaboration is built on trust, and establishing trust with a creative Artificial Intelligence is difficult when the decision process or internal state driving its behaviour isn’t exposed. When human musicians improvise together, a number of extra-musical cues are used to augment musical communication and expose mental or emotional states which affect musical decisions and the effectiveness of the collaboration.”

We KNOW




“Collaboration is built on trust, and establishing trust with a creative Artificial Intelligence is difficult when the decision process or internal state driving its behaviour isn’t exposed. When human musicians improvise together, a number of extra-musical cues are used to augment musical communication and expose mental or emotional states which affect musical decisions and the effectiveness of the collaboration.”






Because we **KNOW** relevant factors and variables, we can generate hypotheses, e.g.,




“Collaboration is built on trust, and establishing trust with a creative Artificial Intelligence is difficult when the decision process or internal state driving its behaviour isn't exposed. When human musicians improvise together, a number of extra-musical cues are used to augment musical communication and expose mental or emotional states which affect musical decisions and the effectiveness of the collaboration.”




?


Because we **KNOW** relevant factors and variables, we can generate hypotheses, e.g.,



“Collaboration is built on trust, and establishing trust with a creative Artificial Intelligence is difficult when the decision process or internal state driving its behaviour isn't exposed. When human musicians improvise together, a number of extra-musical cues are used to augment musical communication and expose mental or emotional states which affect musical decisions and the effectiveness of the collaboration.”



*Then if we have AI musician that can expose its internal state with musical cues, perhaps human musicians can trust AI musician and collaborate with it?*



“Collaboration is built on trust, and establishing trust with a creative Artificial Intelligence is difficult when the decision process or internal state driving its behaviour isn’t exposed. When human musicians improvise together, a number of extra-musical cues are used to augment musical communication and expose mental or emotional states which affect musical decisions and the effectiveness of the

collaboration. **We developed a collaborative improvising AI drummer that communicates its confidence through an emoticon-based visualisation... extra-musical communication with real and false values were tested by experienced improvising musicians.**”

In a Silent Way Communication Between AI and Improvising Musicians Beyond Sound. CHI'19.



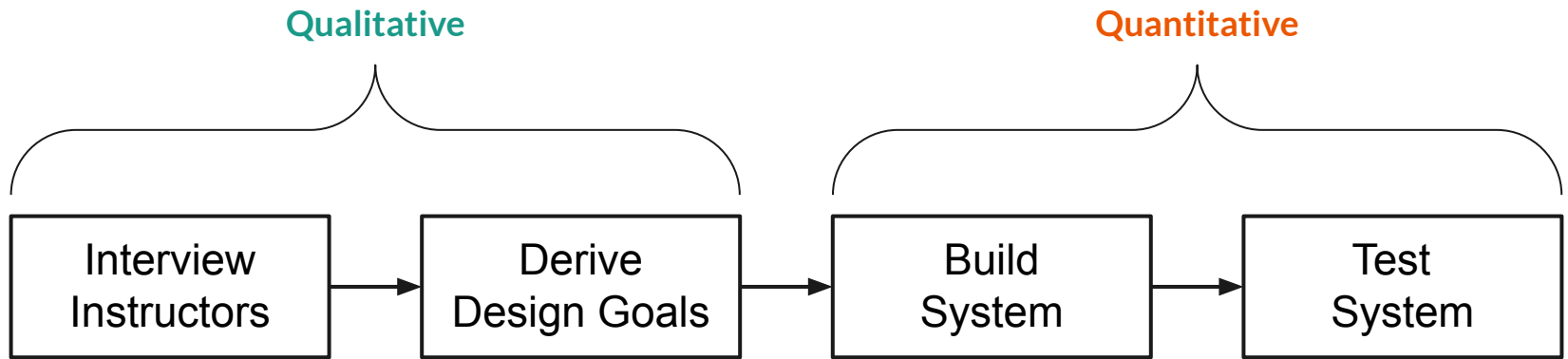
## Quantitative Approach (hypothesis-testing)

- Hypothesis driven or model driven
  - test hypothesis / theory
  - statistics
- Goal
  - be able to say it is unlikely to see effect by chance ( $p \leq 0.05$ )

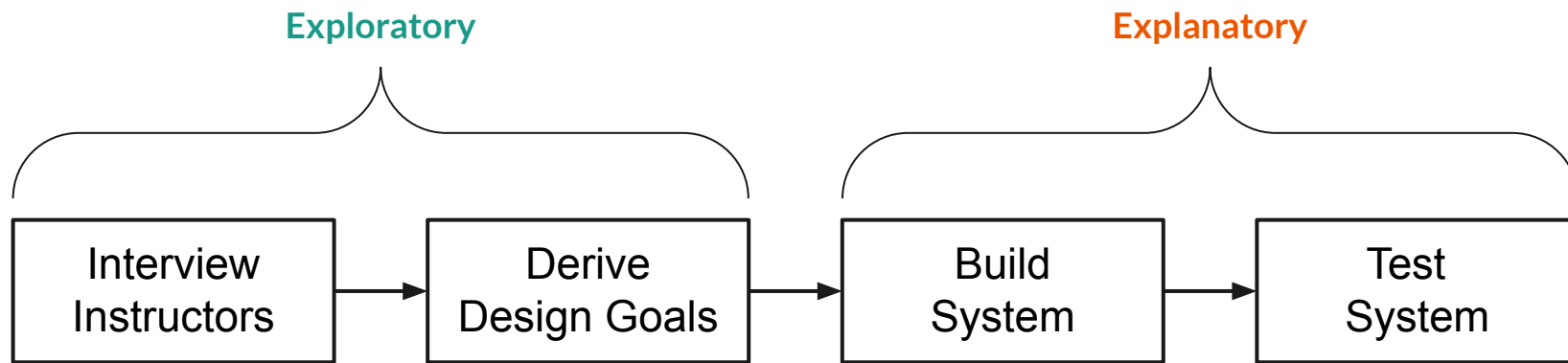


## Mixed Methods Approach

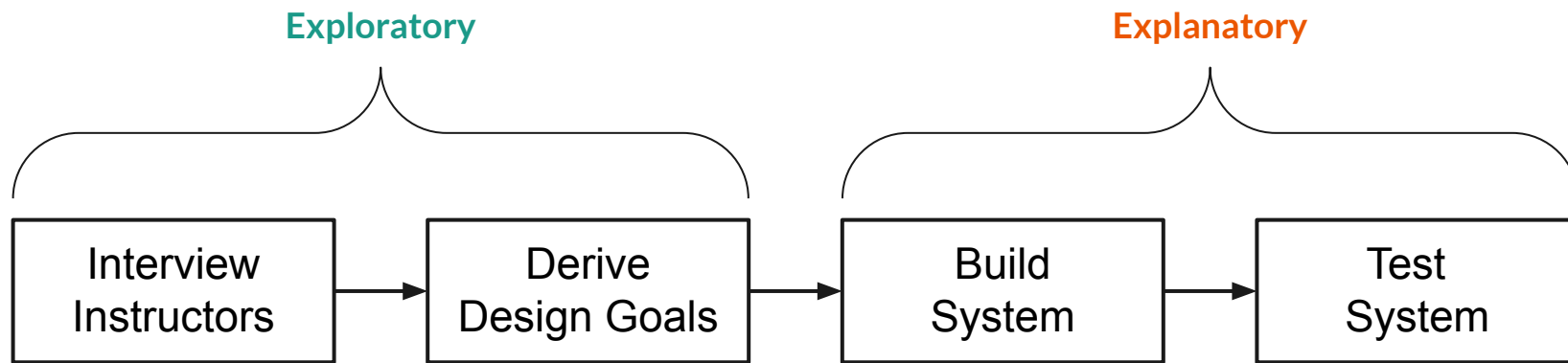
- Qualitative + Quantitative
  - Can do sequentially
    - Typically starts broad using either qualitative or quantitative data
    - Then focuses on another methodology
  - Can do concurrently
    - Use multiple types of data simultaneously to develop a more complete picture



# Why exploratory study?

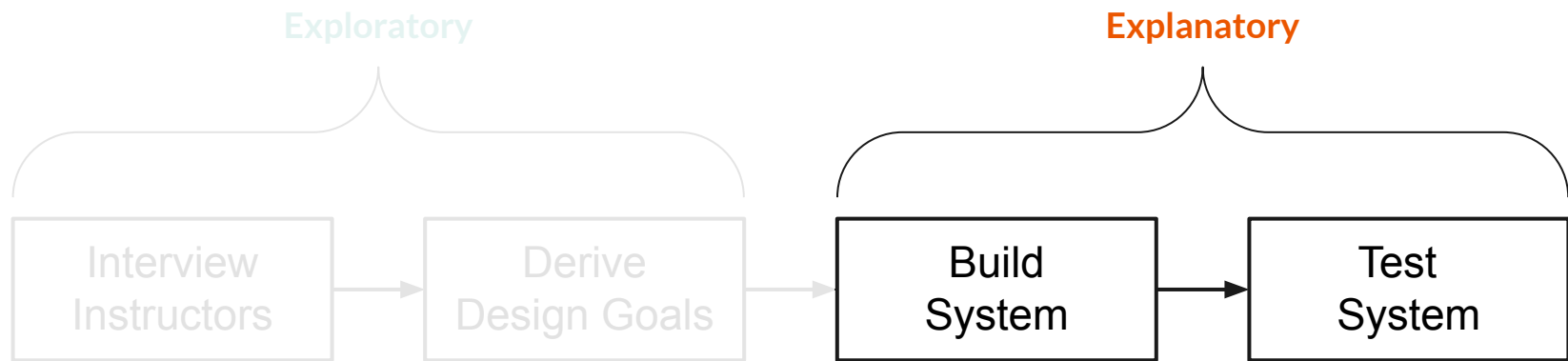


# Is exploratory step necessary before building?





# Can we skip exploratory step?





## Mixed Methods Approach

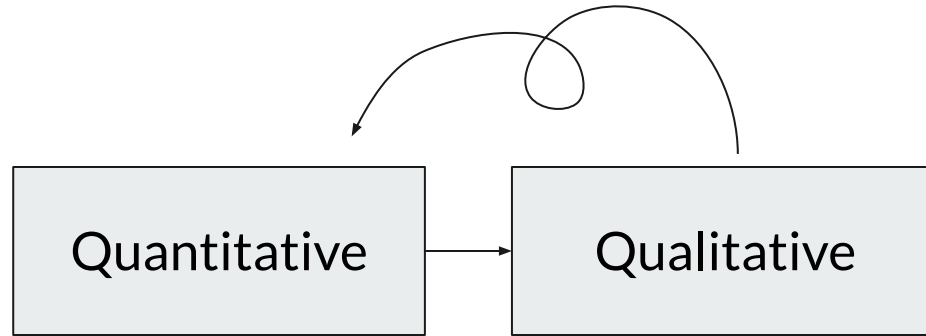
- Goal is to triangulate data
  - Provide a complete picture

**Data triangulation** is the use of a variety of data sources in a study.

**Findings** can be corroborated and any weaknesses in the data can be compensated for by the strengths of other data, thereby **increasing the validity and reliability of the results.**

# For instance

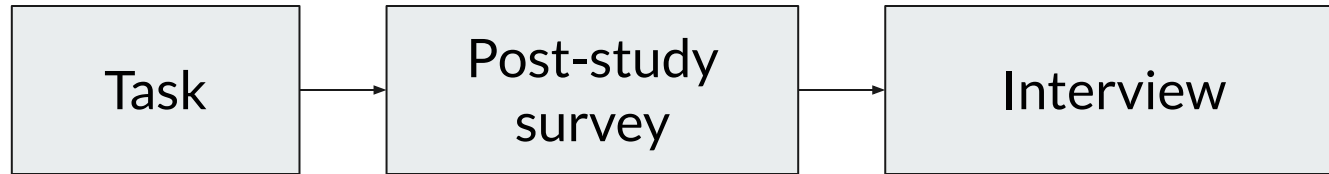
Help explain "WHY"

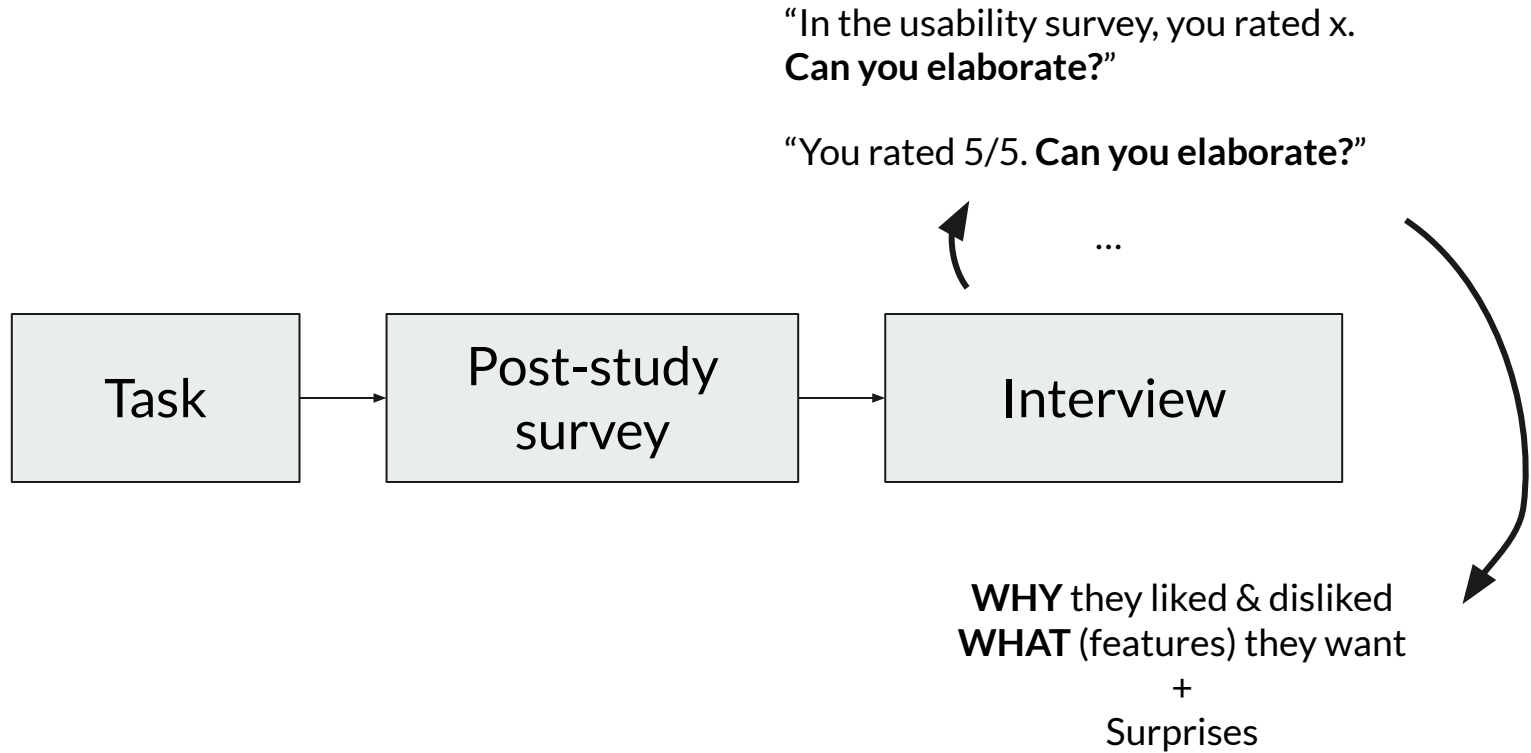


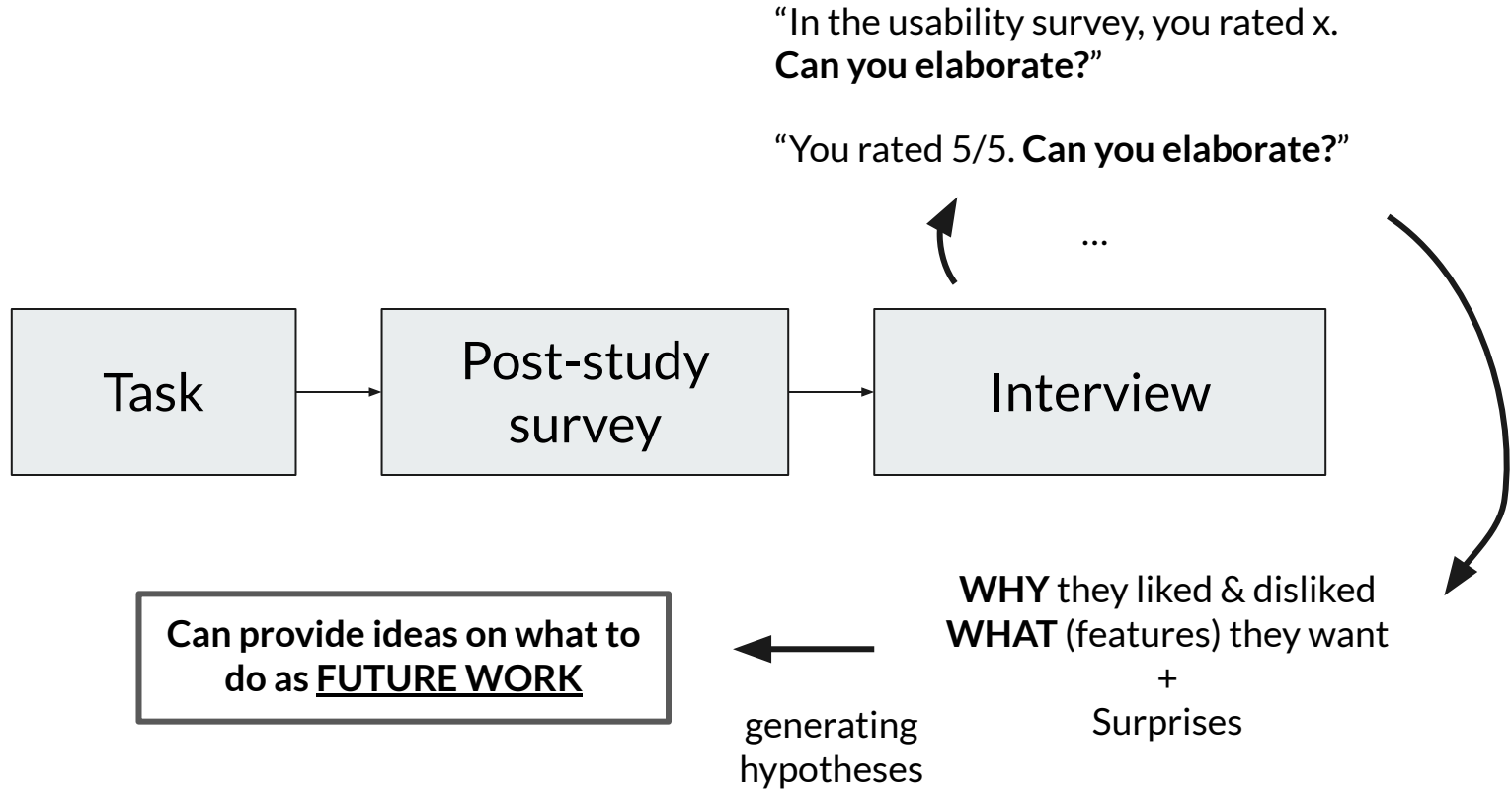
“In the usability survey, you rated x.  
**Can you elaborate?”**

“You rated 5/5. **Can you elaborate?”**

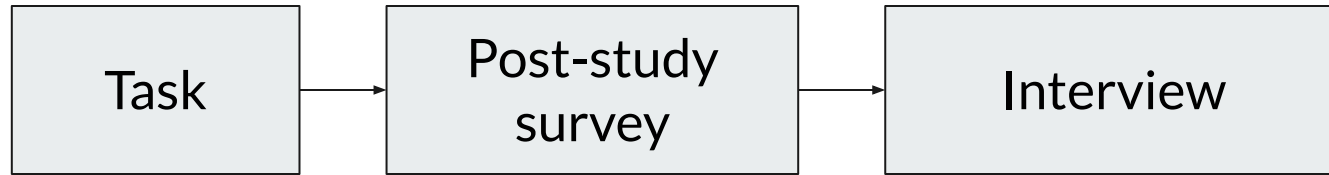
...



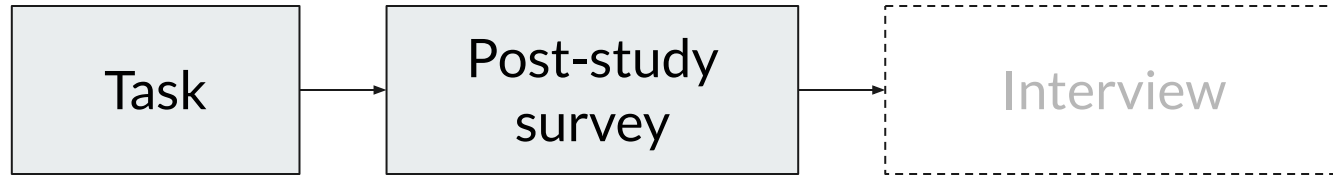




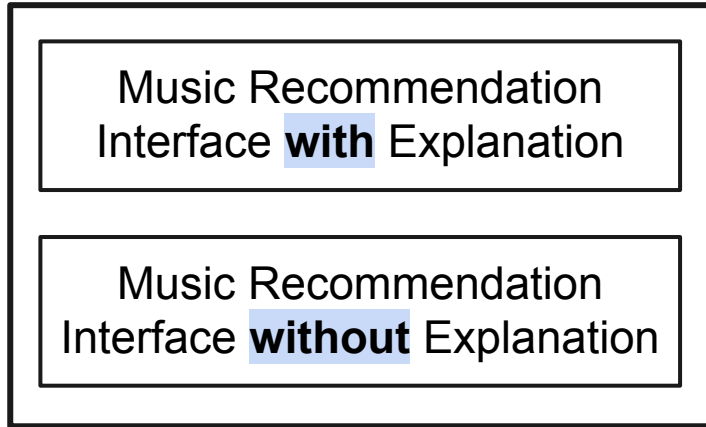
# Is interview necessary for mixed methods?



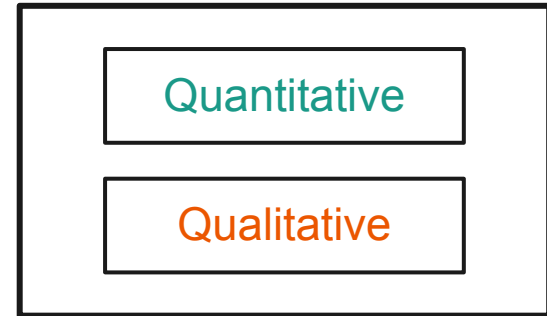
**No, you can collect from survey as well**



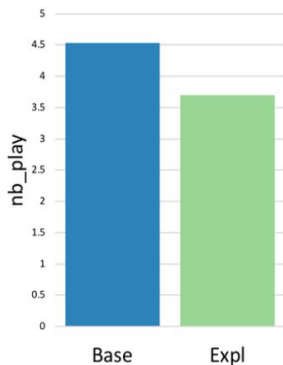




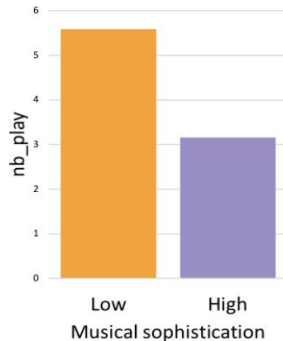
Task



Post-Study  
Survey



(a) The effect of each interface on *nb\_play* showing that participants played more songs in the baseline than in the explanation interface.



(b) The effect of MS on *nb\_play* showing that participants with a low MS played more songs than those with a high MS.

Figure 3: The main effect results of a) interfaces and b) MS on *nb\_play*.

Three participants with the lowest NFC reported that the why component is the most useful part. Two of them mentioned the explanations in general “Where it showed you why it was chosen for you” (P50, P57) “I liked the “WHY”” and one user explicitly mentioned the scatterplot “The dot chart.” (P56). In addition to these three participants in the first NFC quartile, five other participants in the second NFC quartile reported that explanations were the most useful. Two participants reported that they liked the explanations because it explains why the songs are recommended: “it explained why a song was being recommended” (P6) and “The bars showing why it matched search” (P11). The three other users reported that they liked the Why component in general: “why?” (P35), “the chart” (P33) and “It was really cool seeing the Why? bar graph and scatter plot, cause it kinda lets you compare songs you enjoy.” (P15). For these two groups, only one user reported that the explanation were not needed: “ I didn’t really need the Why thing. I didn’t find it useful, I know why the thing recommended all the songs.” (P25).



## Mixed Methods Approach Is Good

- Provide us with richer insights (what + why + how)



## Mixed Methods Approach Is Good

- Provide us with richer insights (what + why + how)
- When your results---based on quantitative data---do not produce any statistically significant result, you may still be able to find interesting insights from qualitative data to publish your research



# Recap

Goals of Experiment 

Where to Start 

Qualitative / Quantitative / Mixed Methods Approach  



**Break?**

---

# Comparative vs Non-Comparative Experimental Design

not comparing anything with another



## Non-comparative study

- How did metacognitive knowledge of students at XYZ high school evolve over the course of the semester?
- How do active users of FitBit use it to maintain healthy lifestyle?





more than 1 condition

## Comparative study

need to compare with baseline

- Does our system A lead to improved productivity?
- Does intervention A motivate users to maintain healthier lifestyle?
- Do millennials have different attitude towards AI?

# Comparative Study

**Between-subjects**  
(between-groups)

**Within-subjects**  
(repeated measures)



# We are distributing conditions ...

**Between**-subjects  
(between-groups)

**Within**-subjects  
(repeated measures)



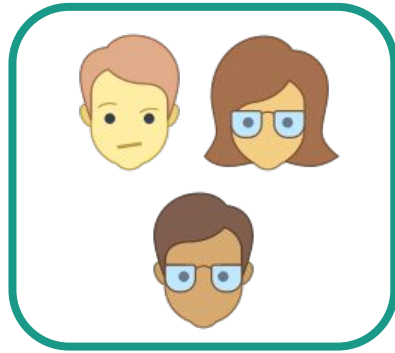
---

# Between-groups

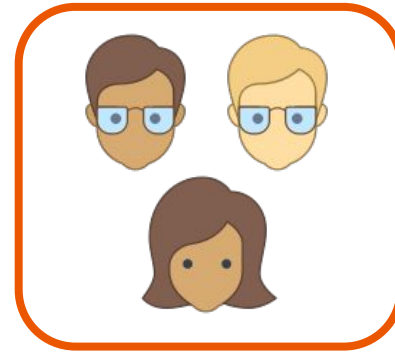
(~~Between-subjects~~)

---

## Between-groups



Condition A



Condition B

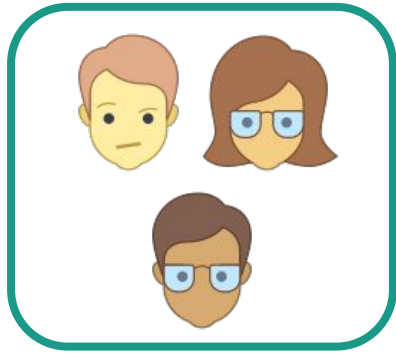


## Between-groups

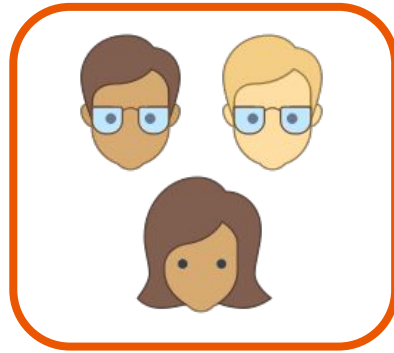
- **Different** participants for each condition
- Participants are assigned to the groups **randomly**

---

## Between-groups

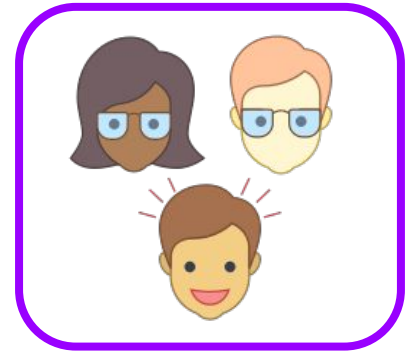


Condition A



Condition B

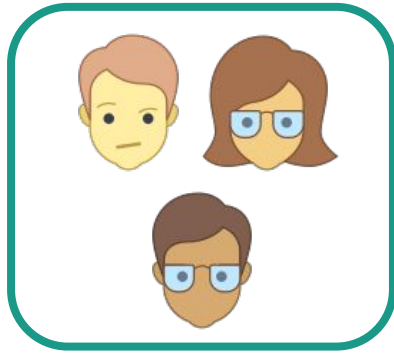
...



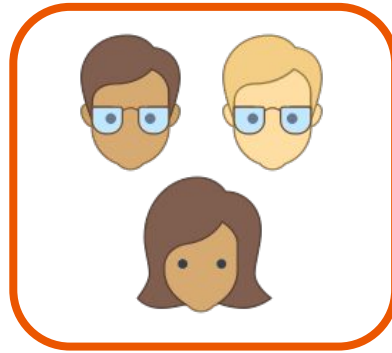
Condition N

---

## Between-groups

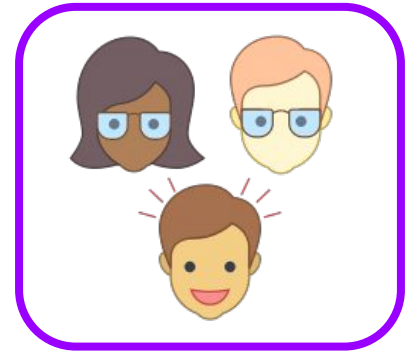


Vaccine A



Vaccine B

...

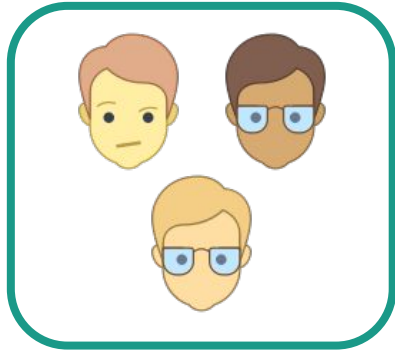


Control  
(Placebo)

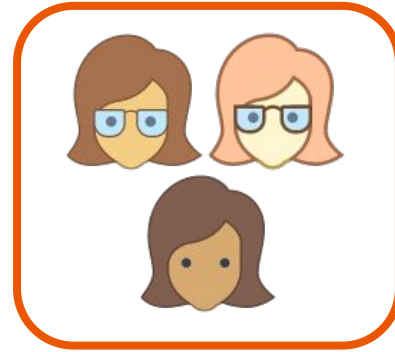


---

## What To Avoid



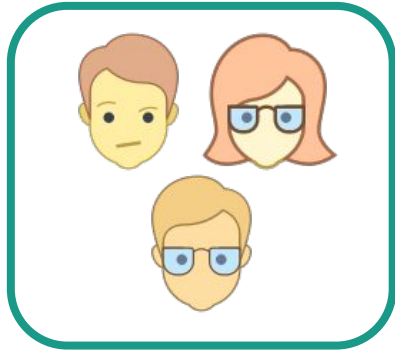
A



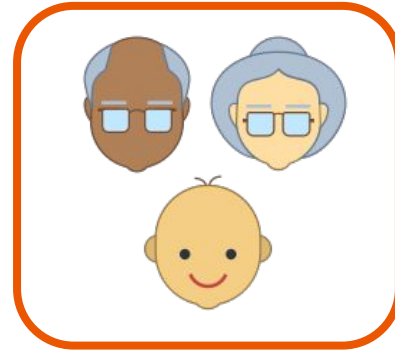
B

---

## What To Avoid



A



B

In between-groups study,  
you should randomly assign  
them to groups but groups will  
need to be balanced if any of  
the variables could influence

---



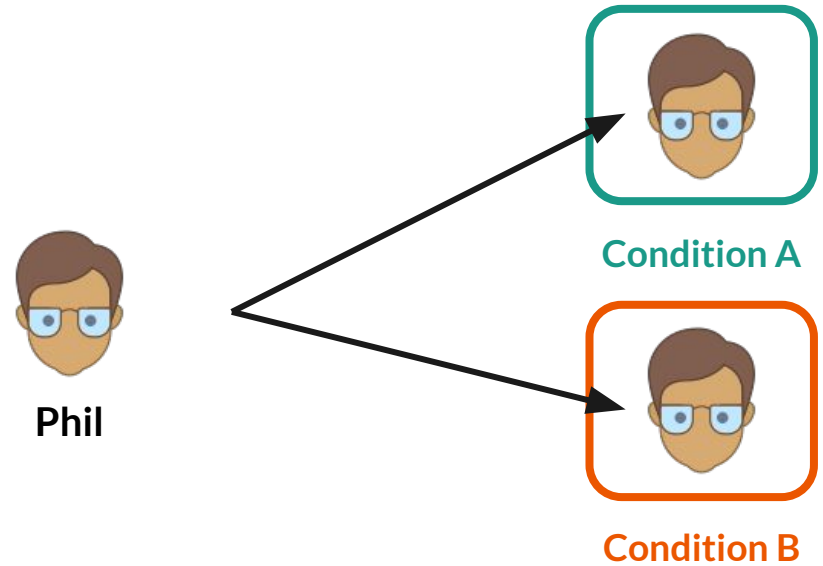
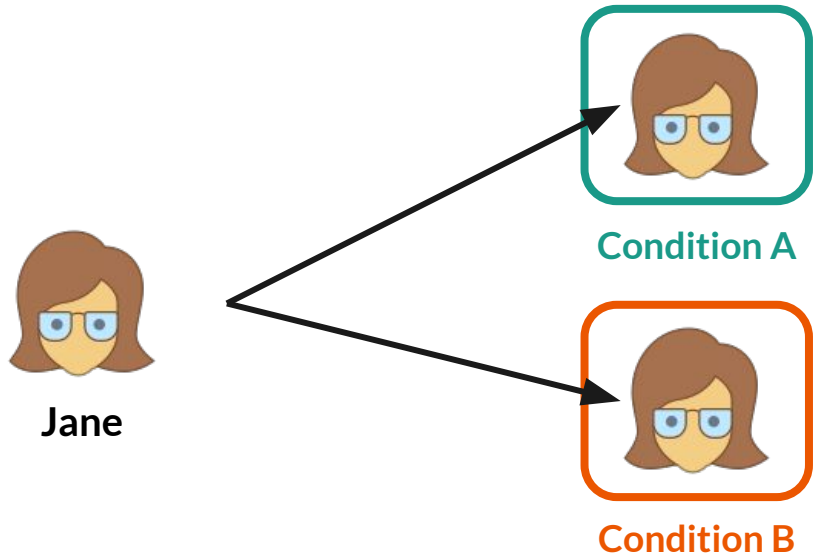
## Exclusion / Inclusion Criteria

- **Who can / cannot participate in the study**
- **Randomly assigning participants to groups can be done if a pool of participants (e.g., undergraduate students who took programming class) has similar attributes**

---

# Within-subjects

# Within-subjects

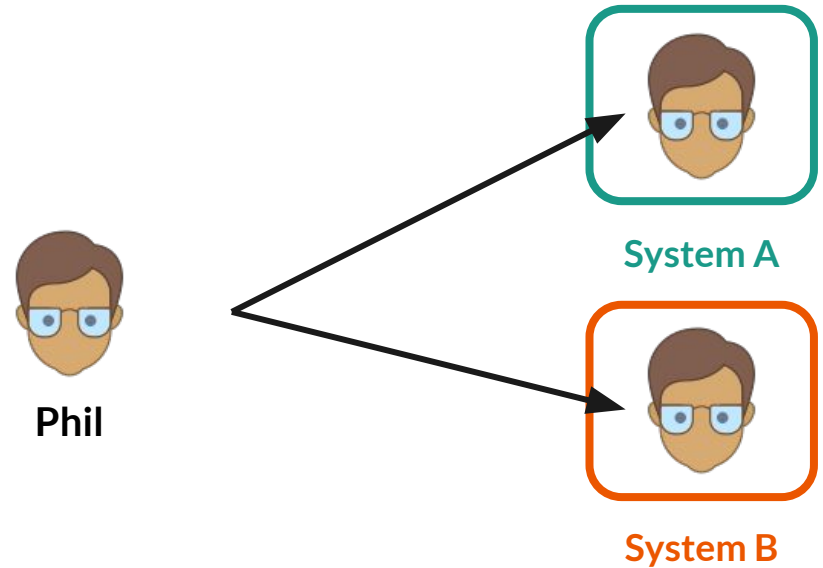
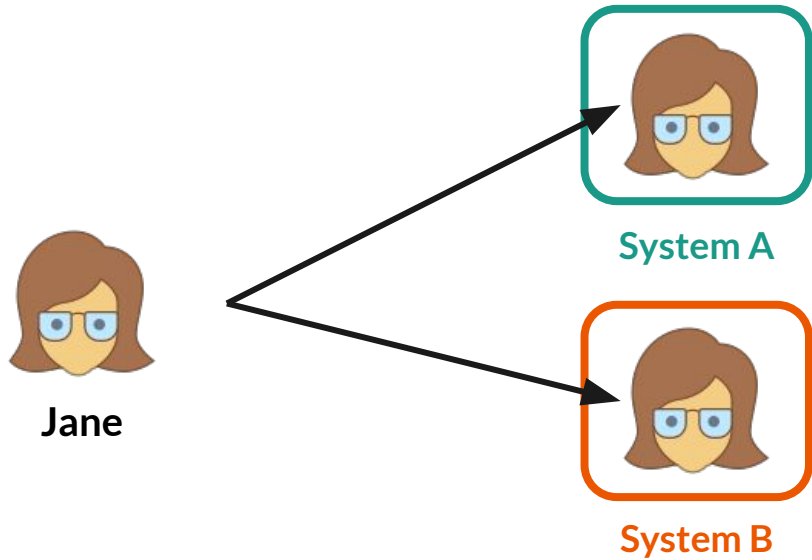




## Within-subjects

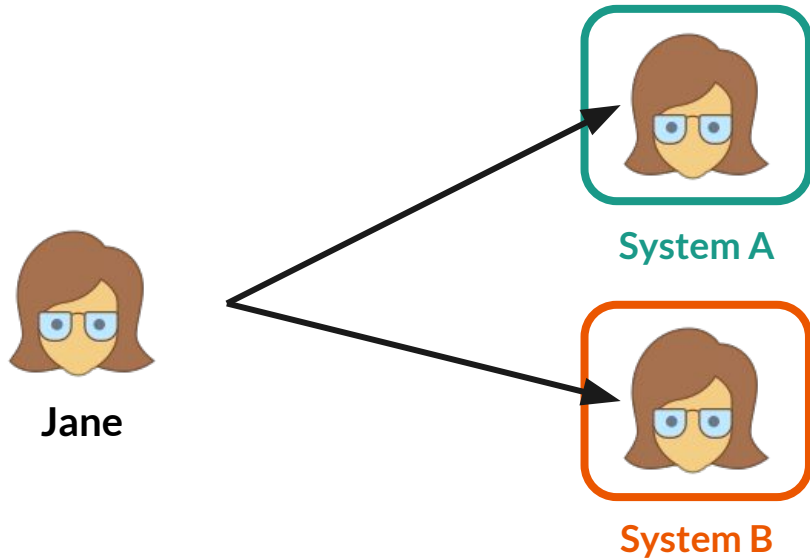
- **Same** participants for each condition

## What To Avoid





## Order Effect



Carry over...

- Fatigue
- Familiarity
- Practice

In within-subjects study,  
randomize the order  
(~~not a must~~)

---



**You May Not Randomize Order If Have Good Reasons**



# You May Not Randomize Order If Have Good Reasons

Interface without  
Creativity Support

vs

Interface with  
Creativity Support

STORY

WRITE YOUR STORY HERE

STORY BOTH CANVAS




# IFFF 







FILTER: 


**CONCEPT**

---

STORY

---


COMIC




TRIGGER CARDS

WHAT THINGS CAN YOU SAY ABOUT THIS CONCEPT?

WHAT IS ITS PURPOSE?



HAS THERE BEEN NEWS ABOUT THIS?




ARE THERE TOOLS TO HELP AVOID SURVEILLANCE?



ARE THERE ANY RISKS INVOLVED? WHAT ARE THEY?



WHAT MAKES IT DIFFICULT TO DELETE YOUR PRIVATE DATA?


















HOW IS COLLECTED? WHAT CHANGES SURVEILLANCE?

CONCEPT


STORY

WRITE YOUR STORY HERE

VIEW MODE  ZEN MODE  GRID MODE

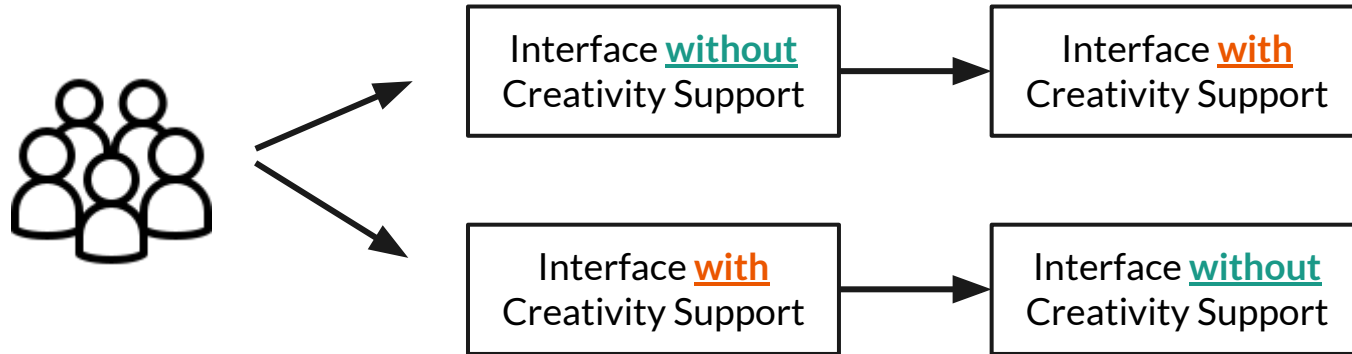
















# FFFF



+ - 🔍 70% 

# You May Not Randomize Order If Have Good Reasons



# You May Not Randomize Order If Have Good Reasons



Interface without  
Creativity Support



Interface with  
Creativity Support

1. Explain **why** you designed the experiment the way you did
2. Acknowledge the **limitations** of this approach





# Counterbalancing (Latin Square)

## Balanced Latin Square Generator

Number of conditions / Square size

	<b>1</b>	<b>2</b>	<b>3</b>	<b>4</b>
<b>1</b>	A	B	D	C
<b>2</b>	B	C	A	D
<b>3</b>	C	D	B	A
<b>4</b>	D	A	C	B

[https://cs.uwaterloo.ca/~dmasson/tools/latin\\_square/](https://cs.uwaterloo.ca/~dmasson/tools/latin_square/)



## **Between**-subjects

(between-groups)

- eliminate order effects (e.g., fatigue, familiarity, practice)

## **Within**-subjects

(repeated measures)

- require fewer participants & less time
- can help reduce errors associated with individual differences



# When Would You Use Within-subjects Design?



## When Would You Use Within-subjects Design?

- A within-subjects design can be a good option if participants or resources are limited



# When Should Within-subjects Design Not Be Used?



## When Should Within-subjects Design Not Be Used?

- If concerned about the potential interferences of practice effects, may want to use a between-subjects design instead
- Within-subjects designs can also take more time to administer in some cases, so it may be helpful to use a between-groups design if many participants are available to quickly conduct data collection sessions

---

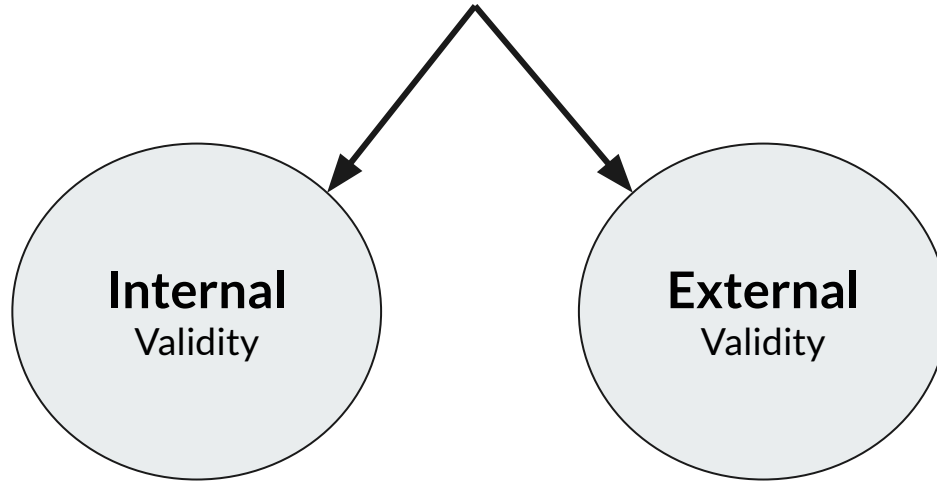
# Validity in Experiment Design

“can we trust these results?”

# Validity in Experiment Design

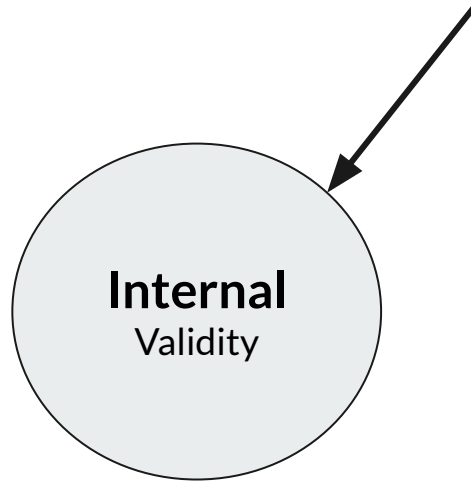


## Validity in Experiment Design

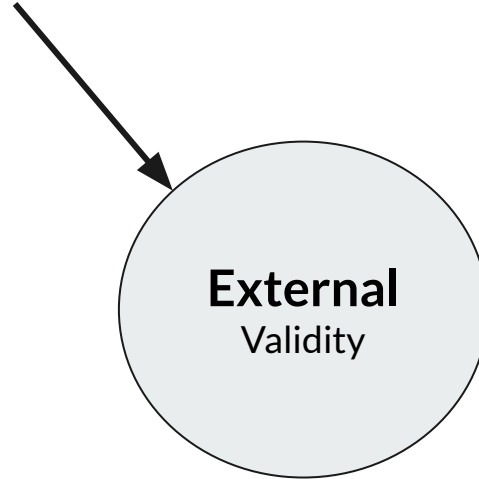


## Validity in Experiment Design

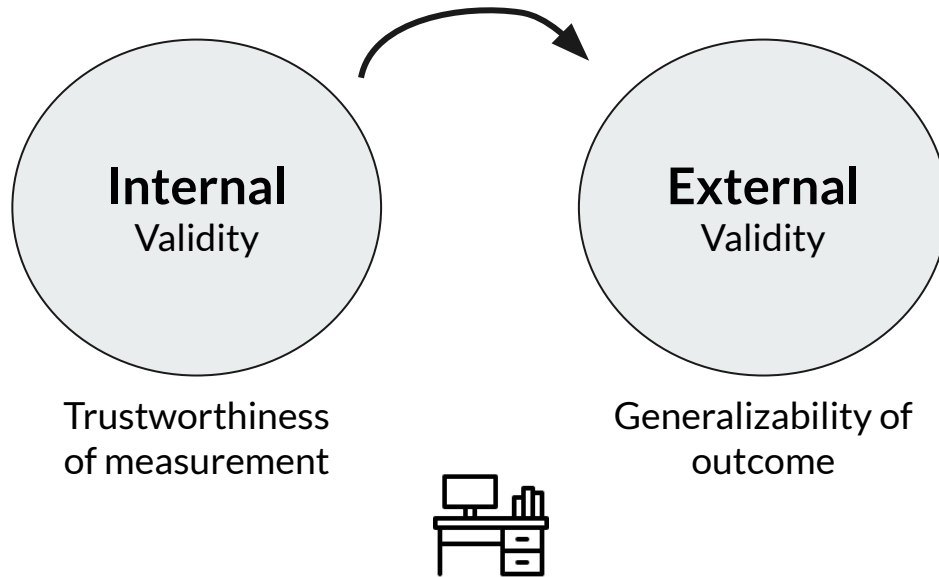
“Can we trust that the results from this study are true?”



## Validity in Experiment Design



“Can we trust that the results from the study are true for people **outside the study** as well?”



“Can we trust the findings from this study and apply them to the world?”



## Threats to External Validity

- **People** - can we generalize to other people (or group of people)?
- **Place** - can we generalize to other places (e.g., country, region)?
- **Time** - can we generalize to other point of time (e.g., different month, year, time period)?



## Threats to Internal Validity

- Influences other than the independent variable that might explain the results of a study



## Threats to Internal Validity

- Confounding variables (e.g., participants with varying prior knowledge, experience, participants have different experience in interview)
- Instrumentation threat (e.g., using different instruments)
- Selection threat (non-equivalent groups for comparison)
- Experimental biases



## Threats to Internal Validity

- Confounding variables (e.g., participants with varying prior knowledge, experience, participants have different experience in interview)
- Instrumentation threat (e.g., using different instruments)
- Selection threat (non-equivalent groups for comparison)
- Experimental biases





## Threats to Internal Validity

- Confounding variables (e.g., participants with varying prior knowledge, experience, participants have different experience in interview)
- Instrumentation threat (e.g., using different instruments)
- Selection threat (non-equivalent groups for comparison)
- **Experimental biases**



# Experimental Biases

- Hawthorne effect
- Experimenter effect
- Pygmalion effect
- Novelty effect



## Hawthorne (Observer) Effect

- Original experiment asked whether lighting change would improve productivity
  - Found that anything they did improved productivity
  - Benefits stopped when studying stopped
- Why?
  - Motivational effect of interest shown to them
- How to reduce
  - Hidden observation / Anonymity

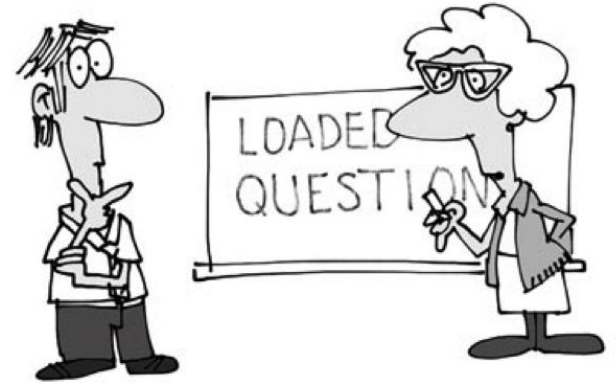


Hawthorne factory in Chicago

## Social Desirability Effect



- Tendency to demonstrate best self
- Problem
  - May not be representative
- Why?
  - People desire to be perceived in the best possible light
- How to reduce
  - Anonymity, carefully formulated instructions and questions



"DO I LOOK FAT IN THIS?"

# Experimenter Effect



- A researcher's bias influences what they see
- Issue
  - If you expect to see sth., maybe sth. in that leads you to see it
- How to reduce
  - Double-blind study



“Clever Hans”  
a horse that can do maths

# Pygmalion Effect



- Self-fulfilling prophecy
- If you place greater expectation on people, they tend to perform better
- Studied teachers and found that they can double the amount of student progress in a year if they believe students are capable
- How to reduce
  - Write and stick to script





## Novelty Effect (“cool”)



- Typically with technology
- Performance improves when technology is instituted because people have increased interest in new technology
- Examples: computer-assisted instruction in secondary schools, computers in the classroom in general, etc



## Novelty Effect (“cool”)



- Do your participants think your system / intervention is cool / useful because of novelty effect?

Did you check if they still feel that way after novelty effect likely went away?





## Recap

- Between-groups vs Within-subjects
  - Counterbalancing for randomizing order in within-subject
- Threats to internal / external validity
- Experimental biases



## Remember

- There is no one absolute formulaic experiment design  
(We did not cover all possible experiment designs)
- The important thing is you explain and provide justification, e.g.,
  - cite accepted papers that use same methodology
  - explain what measures you have taken to reduce potential biases



## Remember

- Find papers you could reference in terms of experiment designs
  - note any explanations/justifications
  - help you be confident about your experiment design



## Find HAI papers covering related topics

- Google “github human-ai interaction papers”
- <https://github.com/manjunath5496/Human-AI-Interaction-Papers>
- <https://github.com/bwang514/awesome-HAI>
- ... more



# Acknowledgement

Some content in the slides are taken from the following:

- Ed Lank. Intro to Experimental Methods. CS889.  
<https://cs.uwaterloo.ca/~lank/CS889/s20/>
- Anastasia Kuzminykh. Experimental Workshop.

---

# Questions?



## Reference

- <https://www.nngroup.com/articles/between-within-subjects/>
- <https://www.verywellmind.com/what-is-a-within-subjects-design-2796014>
- Patino, Cecilia Maria, and Juliana Carvalho Ferreira. "Internal and external validity: can you apply research study results to your patients?." *Jornal brasileiro de pneumologia* 44 (2018): 183-183.



## Images

- <https://cogsresearchmethods.wordpress.com/2017/08/29/social-desirability-s-top-your-self-consciousness/>
- <https://uxplanet.org/implications-of-hawthorne-effect-on-users-behavioral-psychology-5d41d6760ea4>
- [https://commons.wikimedia.org/wiki/File:Ch%C3%A2teau\\_de\\_Versailles,\\_salon\\_des\\_nobles,\\_Pygmalion\\_priant\\_V%C3%A9nus\\_d%27animer\\_sa\\_statue,\\_Jean-Baptiste\\_Regnault.jpg](https://commons.wikimedia.org/wiki/File:Ch%C3%A2teau_de_Versailles,_salon_des_nobles,_Pygmalion_priant_V%C3%A9nus_d%27animer_sa_statue,_Jean-Baptiste_Regnault.jpg)
- <https://www.damninteresting.com/clever-hans-the-math-horse/>





# Icons

- Icons8. <https://icons8.com>