EXPERIMENTS

CPSC 544: FUNDAMENTALS IN DESIGNING INTERACTIVE COMPUTATIONAL TECHNOLOGY FOR PEOPLE

Material created by Joanna McGrenere, Karon MacLean, Leila Aflatoony, Jessica Dawson & Heather O'Brien

COMING UP

This week (W12: Nov 20, Classes C19 + C20)

- No new readings
- Tues 11/21: Team Deliverable – User Test Report
- Thu-Fri 11/23-24
  - Get started on Medium-Fidelity Prototype. Plan it
    → meet /discuss with your staff Mentor by end of week

Coming Up

- W13:
  - Last Researcher Journal #13 (due Sun 11/26)
  - IP & NDA Workshop (Wed 11/29 – required in-person for DFP students)
- W14: Wed 12/06: Med-Fi Prototype Presentation & Demos

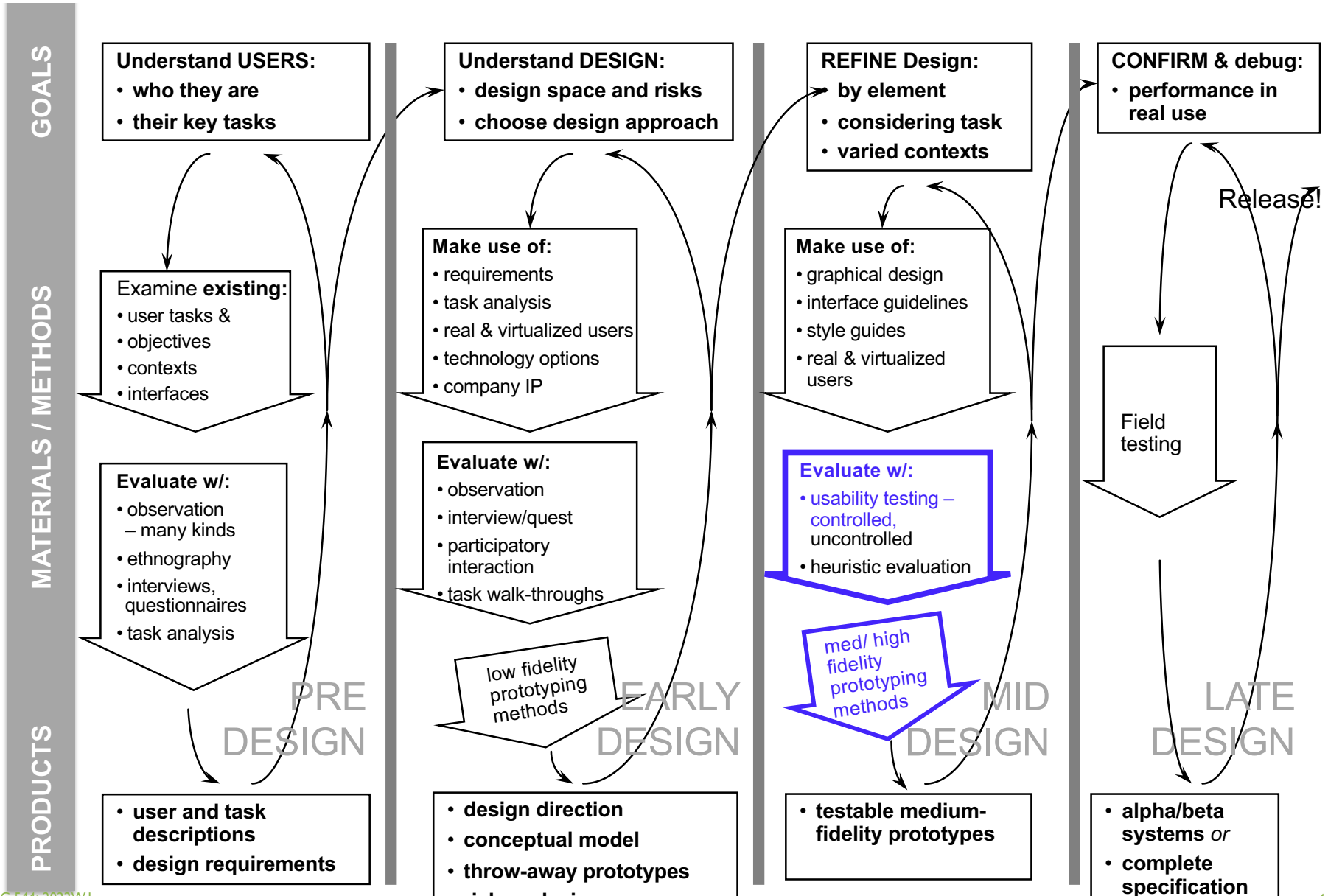
TWO CLASSES ON "EXPERIMENTS"

Experiments I (Nov 20):

- **What's an experiment; basics of designing one**
- **Activity:** *Conduct* a (pseudo) experimental study, collecting data to use in next class.
- **Activity:** *Plan* a (pseudo) experimental study

Experiments II (Nov 22):

- **Data analysis and interpretation**
- **Activity:** Discuss analysis of previous class data



K MacLean - derived from version by Saul Greenberg (U Calgary)

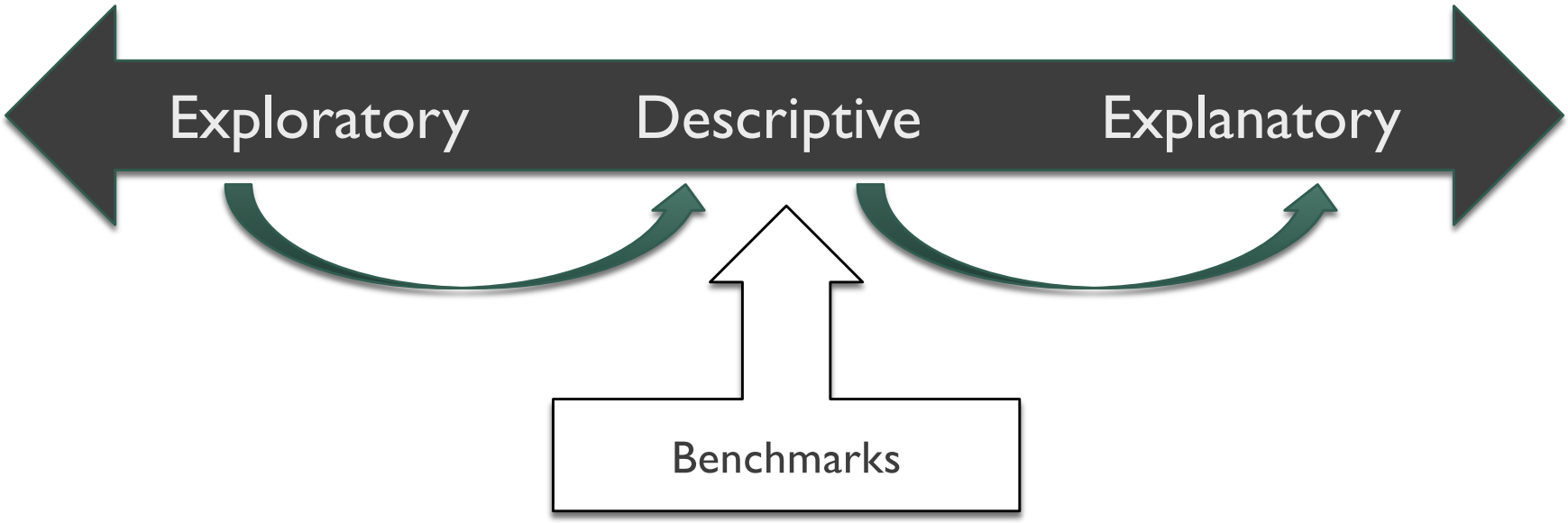
LEARNING GOALS: BE ABLE TO …

- Describe the overall method for user-based experiments
- Understand what an experimental hypothesis is
  - and how it differs from a research or evaluation objective
- Systematically plan a experiment-type user evaluation
- Differentiate within- & between-subject comparisons
- Appropriately apply basic statistics in drawing conclusions
  - And: relate statistical inference to planning (e.g., effect size)
- Discuss significance levels and two types of error
  - Differentiate Type I and type II error
  - How choice of significance levels relates to error types

CPSC 544: Experiments I Basics of Design

Material created by Joanna McGrenere, Karon MacLean, Leila Aflatoony, Jessica Dawson & Heather O'Brien

Learn about a phenomenon	Goal	Make predictions
Unstructured	**Methods**	Structured
Broad, open-ended	**Research questions**	Focused, hypotheses



Kelly (2009). Methods for Evaluating Interactive Information Retrieval Systems with Users.
Hochheiser, Feng & Lazar (2017) Experimental research: Research methods in human computer interaction.

CONTROLLED EXPERIMENTS

- The traditional scientific method
  - Reductionist
    - *Clear convincing result on specific issues*
  - In HCI
    - *Insights into cognitive process, human performance limitations, …*
    - *Allows comparison of systems, fine-tuning of details …*
- Strives for
    - *Lucid and testable hypothesis (usually a causal inference)*
    - *Quantitative measurement*
    - *Measure of confidence in results obtained (inferential statistics)*
    - *Replicability of experiment*
    - *Control of variables and conditions*
    - *Removal of experimenter bias*

DESIRED OUTCOME OF A CONTROLLED EXPERIMENT

- Statistical inference of an event or situation's probability:
  - "Design A is better <in some specific sense> than Design B"

- or, Design A meets a target:
  - "90% of incoming students who have web experience can complete course registration within 30 minutes"

Amount of invested mental effort (AIME) in online searching

Soo Young Rieh *, Yong-Mi Kim, Karen Markey

A B S T R A C T

This research investigates how people's perceptions of information retrieval (IR) systems, their perceptions of search tasks, and their perceptions of self-efficacy influence the amount of invested mental effort (AIME) they put into using two different IR systems: a Web search engine and a library system. It also explores the impact of mental effort on an end user's search experience. To assess AIME in online searching, two experiments were conducted using these methods: Experiment 1 relied on self-reports and Experiment 2 employed the dual-task technique. In both experiments, data were collected through search transaction logs, a pre-search background questionnaire, a post-search questionnaire and an interview. Important findings are these: (1) subjects invested greater mental effort searching a library system than searching the Web; (2) subjects put little effort into Web searching because of their high sense of self-efficacy in their searching ability and their perception of the easiness of the Web; (3) subjects did not recognize that putting mental effort into searching was something needed to improve the search results; and (4) data collected from multiple sources proved to be effective for assessing mental effort in online searching.

ACTIVITY 1A : SEE WORKSHEET (20m)

Rieh, S. Y., Kim, Y. M., & Markey, K. (2012). **Amount of invested mental effort (AIME) in online searching**. *Information Processing & Management*, *48*(6), 1136-1150.

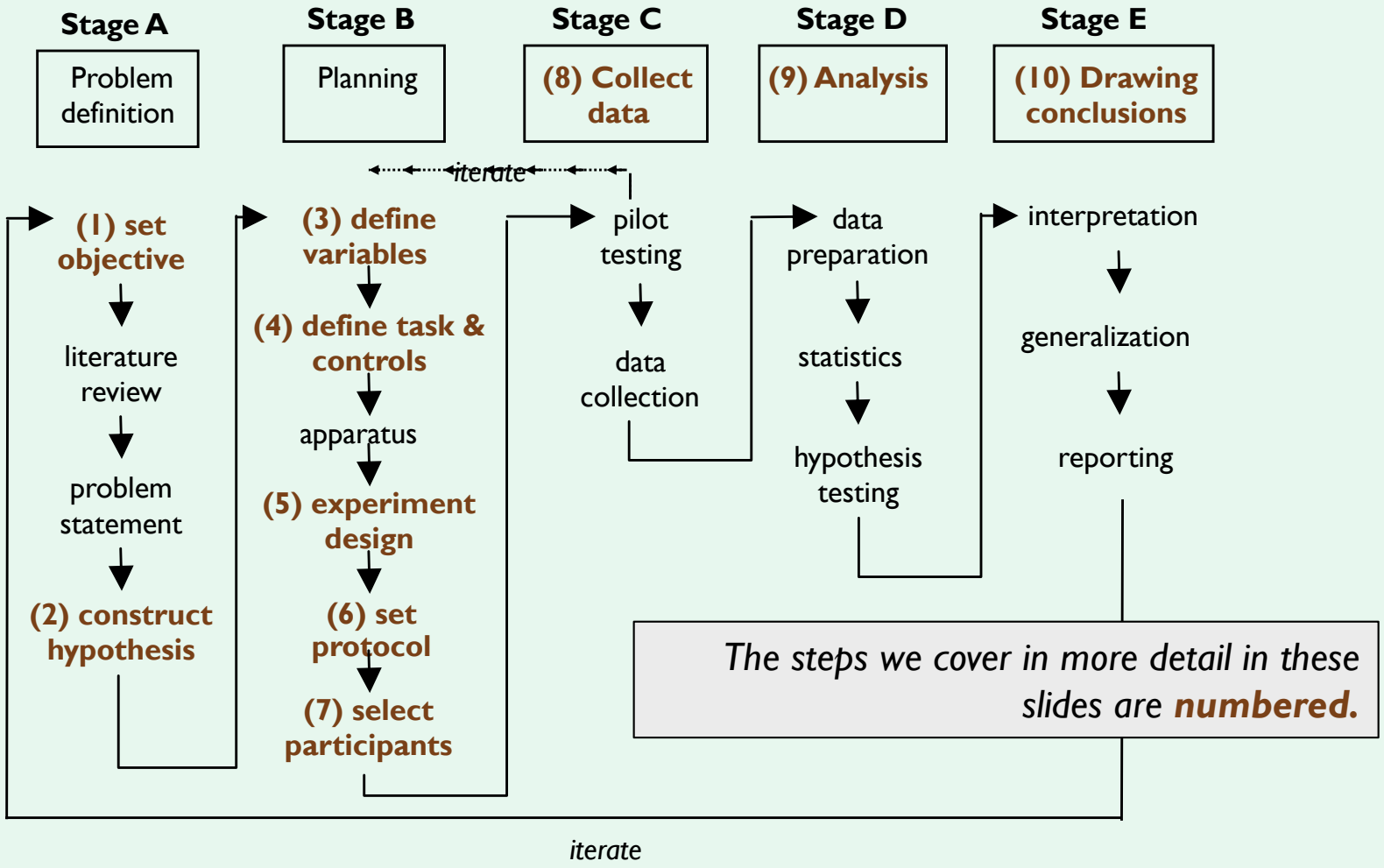
- Research proposition in this paper:

  "...people's perception of information retrieval (IR) systems and their self-confidence in searching are closely related to the extent of mental effort they invest in searching"

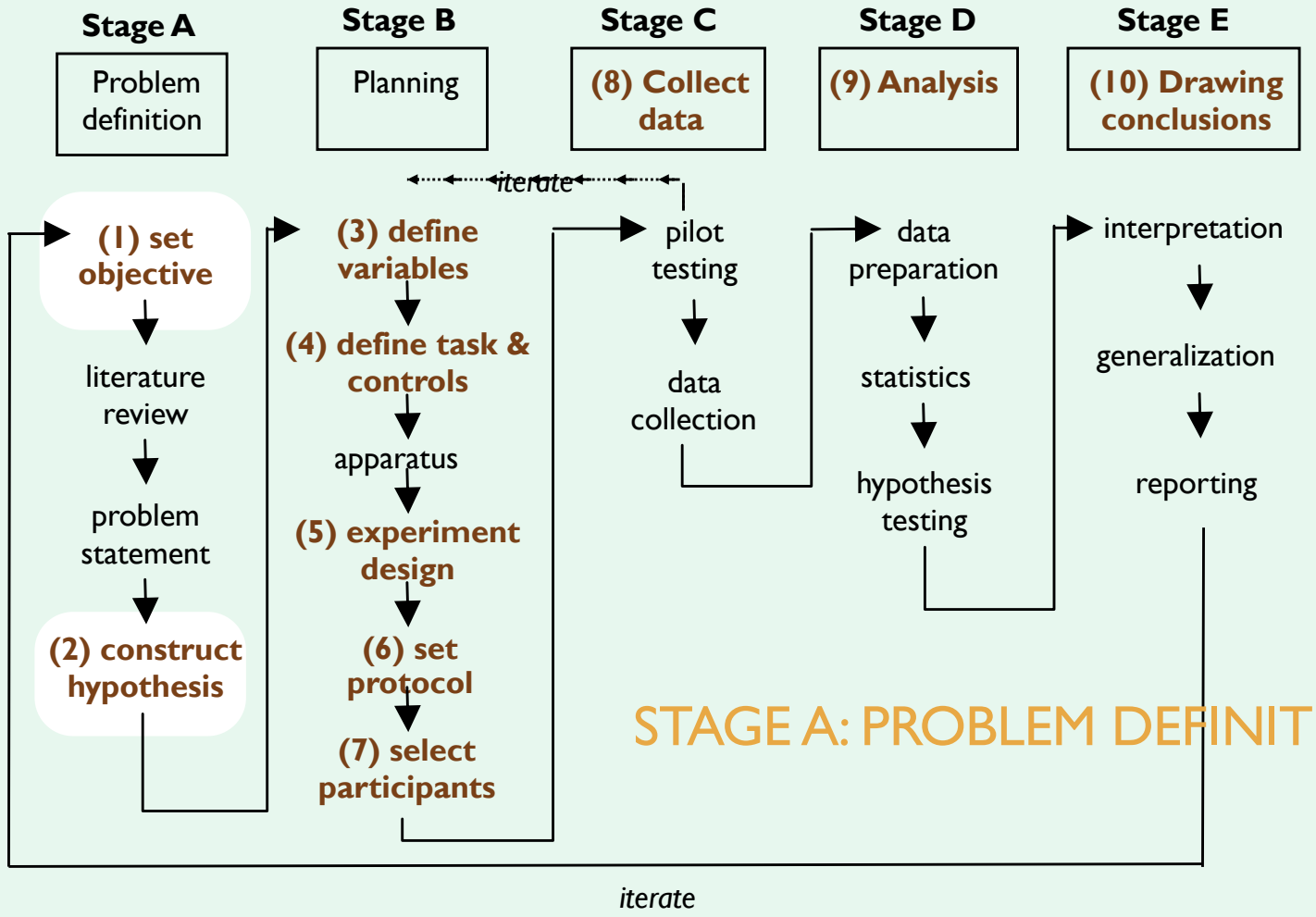
  - Theory: AIME -- Amount of invested mental effort
- Divide into **groups of two**: 1 researcher, 1 participant
- **Study tasks**: for the same two search-target tasks ("research" & "news"),
  - Condition **A**: Search using the **UBC library catalogue** (*half of 2-person groups*)
  - Condition **B**: Search using **Google** (*other half of groups*)
- Pre- and post-task **questions**:
  - Items related to perceived mental effort and participant perceptions of (a) own search skills and (b) system's effectiveness

STAGES IN THE EXPERIMENTAL METHOD

EXPERIMENT PLANNING FLOWCHART



EXPERIMENT PLANNING FLOWCHART



STAGE A: PROBLEM DEFINITION

1. BEGIN WITH AN EXPLICIT **RESEARCH OBJECTIVE** WHICH GENERALLY EXPRESSES WHAT YOU WANT TO LEARN.

Example:

Edwards & Kelly (2017). **Engaged or frustrated? Disambiguating emotional state in search.** In *Proc. ACM SIGIR Conf on Research and Development in Information Retrieval* (125-134).

(1) To what extent do search behaviors differ when people are engaged or frustrated?

(2) To what extent do physiological signals help disambiguate engaging and frustrating search episodes?

2. BEGIN WITH A **LUCID, TESTABLE HYPOTHESIS** WHICH EXPRESSES YOUR RESEARCH QUESTION

For one research question:

- H0: There is no difference in the search behaviours of engaged and frustrated searchers.
- H1: There will be a difference in the search behaviours of searchers who are engaged vs. frustrated.

- For a different research question:
  - H0: There are no differences in the physiological signals of engaging and frustrating search episodes.
  - H1: Physiological signals can disambiguate engaging and frustrating search episodes.

Discuss: what exactly is the difference between RQs and hypotheses?

Edwards & Kelly (2017). Engaged or frustrated? disambiguating emotional state in search. In *Proc. ACM SIGIR Conf on Research and Development in Information Retrieval* (125-134).

GENERAL: HYPOTHESIS TESTING

- **Hypothesis** = prediction of the outcome of an experiment
- Framed in **independent & dependent variables**:
  - A variation in the independent variable will cause a difference in the dependent variable.
- **Aim of the experiment**: prove this prediction
  - By disproving the "null hypothesis"
  - Never by proving the "alternate hypothesis"
- Typical basic hypothesis set (need at least two):
  - H0: Experimental conditions have no effect on performance (to some degree of significance) → **null hypothesis**
  - H1: Experimental conditions have an effect on performance (to some degree of significance) → **alternate hypothesis**

TWO TYPES OF POTENTIAL EXPERIMENT ERRORS

- Type I error: reject the null hypothesis when it is, in fact, true
  - We conclude that there is a genuine effect, when there isn't one (false positive)
  - Confidence level for statistical tests , α-level (e.g., α = .05), is probability of a Type I error

- Type II error: accept the null hypothesis when it is, in fact, false
  - We conclude that there is no effect, when there actually is one (false negative)
  - β -level is probability of a Type II error
    - *related to power (which is defined as 1-β), and which depends on α-level, effect size, and sample size*

TRADEOFFS AND SIGNIFICANCE LEVELS

	Reality	
Outcome of Exp't	H_0 True	H_0 False
Reject H_0	**Type I error** (false positive)	Correct inference (true positive)
Fail to Reject H_0	Correct inference (true negative)	**Type II error** (false negative)

There are trade-offs in planning to minimize these two types of errors

3. DEFINE VARIABLES: OVERVIEW

(a) Independent: *Things you control/manipulate*

- experimental factors
- participant individual differences

(b) Dependent: *Things that you measure*

- *cognitive metrics*
- *affective metrics*
- *behavioral & performance metrics*

(c) "Nuisance": *Things you don't want but can't eliminate*

3. DEFINE VARIABLES: (A) INDEPENDENT – STATE EXPLICITLY!

- Things you **control/manipulate**
  (independently of how a subject behaves)
  to produce different conditions for comparison
- Two main kinds:
  - **Treatment-manipulated ("experiment factors")**
    - *Pin-pointed in experimental design*
    - *Can establish cause/effect to some confidence level*
      → *"true" experiment*
  - **Participant individual differences**
    - *Assess within-participant variation (via structure of experiment)*
    - *Based on (dependent) metrics*
    - *Might control at recruitment stage (include different groups)*
    - *Can never fully establish cause/effect (WHY?)*

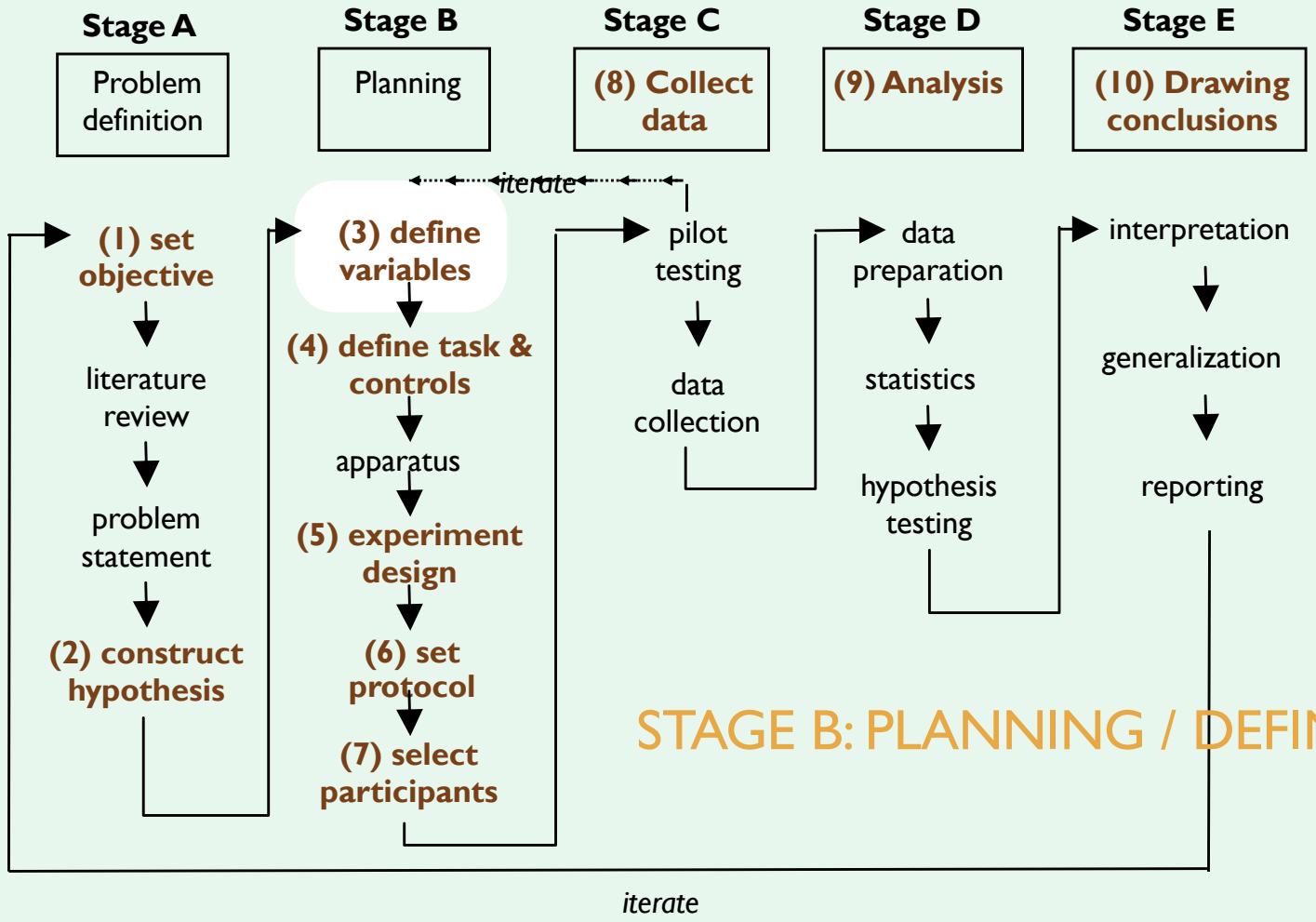
INDIVIDUAL DIFFERENCES

- A study of how people differ and of "central tendency, how well a person can be described in terms of an overall within-person average" (Revelle, 2000, p. 249)
- Ways that people can vary
  - Demographic: age, profession/role, education, expertise...
  - Cognitive: memory, spatial ability, cognitive load, intelligence, reading level...
  - Affective: interest, curiosity, self-efficacy, coping style...
  - Personality...
- To assess experimentally, can recruit samples that vary in any of these aspects, and compare dependent outcomes.

Revelle, W. (2000). Individual differences. In D. E. Leary (Vol. Ed.), Encyclopedia of psychology. Vol. 4. Encyclopedia of psychology (pp. 249–252). Washington, DC: American Psychological Association.

O'Brien, H. L., Dickinson, R., & Askin, N. (2017). A scoping review of individual differences in information seeking behavior and retrieval research between 2000 and 2015. *Library & Information Science Research*, *39*(3), 244-254.

EXPERIMENT PLANNING FLOWCHART



STAGE B: PLANNING / DEFINE VARIABLES

3. DEFINE VARIABLES: (B) DEPENDENT

- Dependent Variables are the things that are measured
  - Also called "metrics"
  - Need to capture the behavior or response you want to understand
- Expectation: DVs depend on the participant's behaviour or reaction to the independent variable
  - (but will be unaffected by other factors)
- What else could we measure?

Using self-report and physiological measures together – what if they contradict?

FOUR BASIC FOCI OF DEPENDENT MEASURES

- **Context**: characterize participants and the interaction situation
- **Interaction:** interaction between user and system
- **Performance**: pertain to the outcome of the interaction
- **Usability:** attitudes and feelings about system/ interaction

Each of these can be viewed through several different facets of human response to the manipulated situation …

WHAT CAN BE MEASURED?
COGNITIVE METRICS

- Cognitive load
- Cognitive volatility
- Comprehension/learning
- Prior knowledge
- Perceived time
- Perceived accuracy

WHAT CAN BE MEASURED?
AFFECTIVE METRICS

- Enjoyment, pleasure
- Satisfaction
- Motivation
- Interest
- User experience

WHAT CAN BE MEASURED?
BEHAVIOURAL AND PERFORMANCE METRICS

- Mouse clicks
- Performance time, i.e. efficiency
- Performance accuracy, i.e. effectiveness
- Number of actions (e.g. queries)

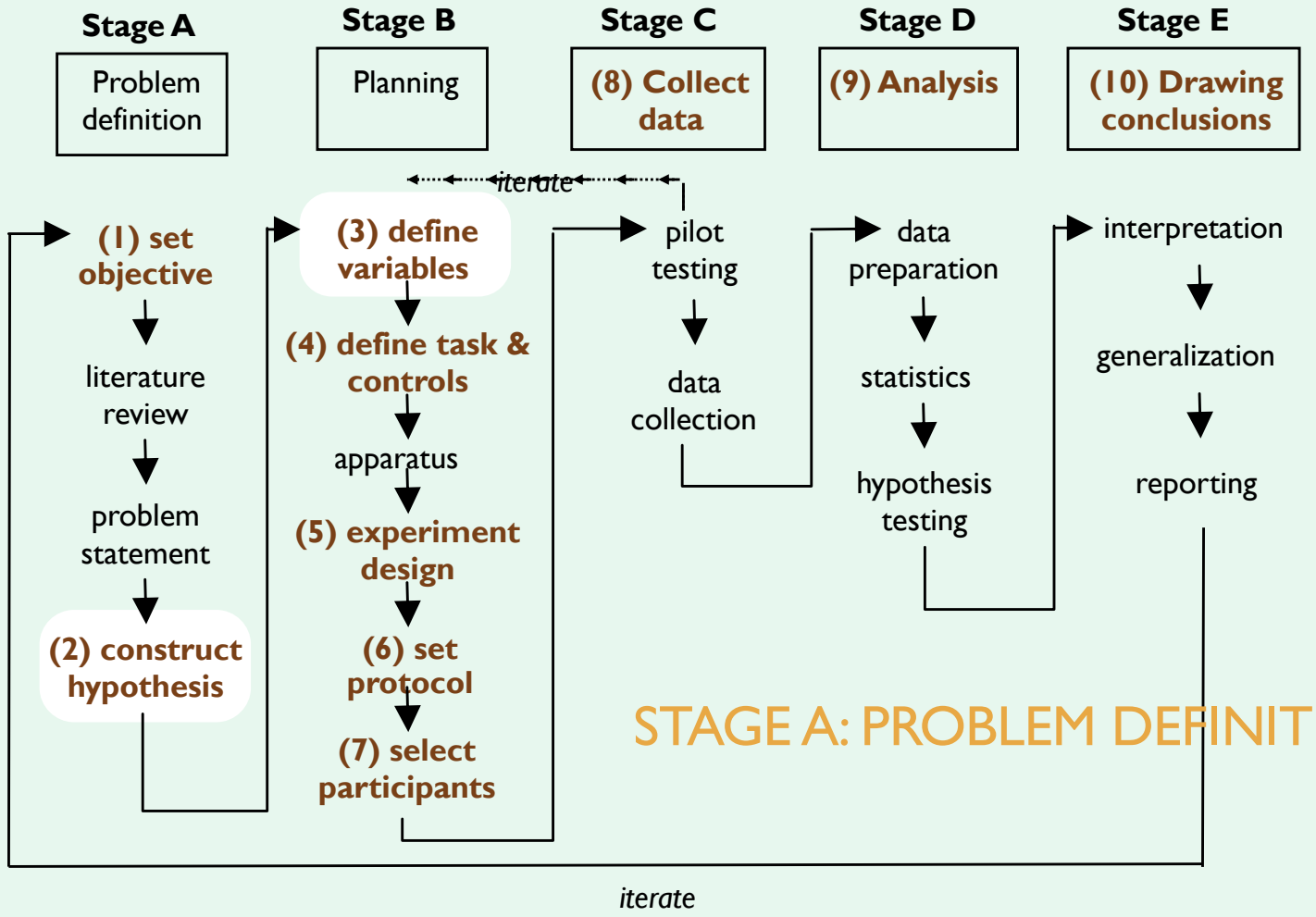
3. DEFINE VARIABLES: (C) "NUISANCE" POSSIBILITIES

- Undesired variations in experiment conditions which cannot be eliminated, but which may affect dependent variables
- Need to identify in advance, and accommodate through experiment design & analysis:
  - Treat as a control variable (if they can be controlled)
  - Randomization (if they cannot be controlled)
- Common nuisance variables:
  - Individual differences (when this is not an explicit focus)
  - Unavoidable variations in experimental conditions

ACTIVITY 2 FOR EXPERIMENTS I:

WORKSHEET – PRACTICE STEPS #2-3

EXPERIMENT PLANNING FLOWCHART



STAGE A: PROBLEM DEFINITION

Engaged or Frustrated? Disambiguating Emotional State in Search

Edwards & Kelly, 2017

One of the primary ways researchers have characterized engagement during information search is by increases in search behaviors, such as queries and clicks. However, studies have shown that frustration is also characterized by increases in these same behaviors. **This research examines the differences in the search behaviors and physiologies of people who are engaged or frustrated during search**. A **2x2 within-subject laboratory experiment** was conducted with 40 participants. **Engagement** was induced by manipulating task interest; and **frustration** was induced by manipulating the quality of the search results. Participants' interactions and physiological responses were recorded, and after they searched, they evaluated their levels of engagement, frustration and stress. Participants reported significantly greater levels of engagement when completing tasks that interested them and significantly less engagement during searches with poor results quality.

For all search behaviors measured, only **two significant differences** were found according to task interest: participants had **more scrolls and longer query intervals when searching for interesting tasks**, suggesting greater interaction with content. Significant differences were found for nine behaviors according to results quality, including queries issued, number of SERPs displayed and number of SERP clicks, suggesting these are **potentially better indicators of frustration rather than engagement**. When presented with poor quality results, participants had significantly higher heart rates than when presented with normal quality results. Finally, participants had lower heart rates and greater skin conductance responses when conducting interesting tasks than when conducting uninteresting tasks. This research provides insight into the differences in search behaviors and physiologies of participants when they are engaged versus frustrated and presents techniques that can be used by those wishing to induce engagement and frustration during laboratory IIR studies.

Edwards & Kelly (2017). **Engaged or frustrated? disambiguating emotional state in search.** In *Proc. ACM SIGIR Conf on Research and Development in Information Retrieval* (125-134).

TO SUMMARIZE:
HOW A CONTROLLED EXPERIMENT WORKS

- Formulate an alternate and a null hypothesis:
  - H1: experimental conditions have an effect on performance
  - H0: experimental conditions have no effect on performance
- Through experimental task, try to demonstrate that the null hypothesis is false (reject it),
  - For a particular level of significance
- If successful, we can accept the alternate hypothesis,
  - and state the probability p that we are wrong (the null hypothesis is true after all) → this is result's confidence level
  - At a .05% CL: 5% chance we got it wrong, 95% confident

CPSC 544: Experiments II Analysis & Interpretation

Material created by Joanna McGrenere, Karon MacLean, Leila Aflatoony, Jessica Dawson & Heather O'Brien

NEXT CLASS (NOV 27):

Special topic: **Design Justice in User Interface Design**

Please try to do at least one of the readings in advance, and/or watch the Constanza-Chock video (from the UIST 2020 conference)

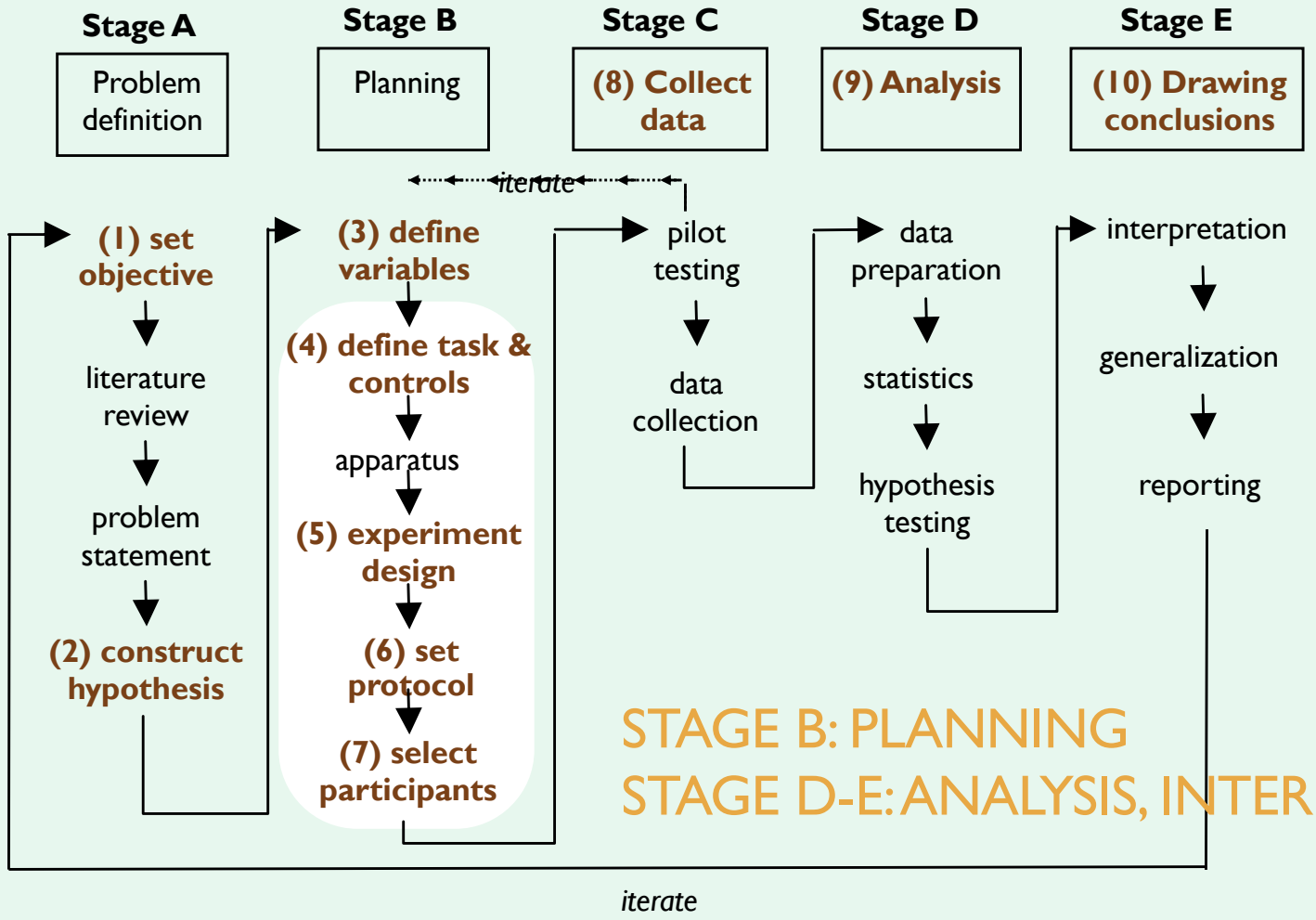
Bring a couple of ideas or questions to discuss

Class will be primarily discussion of these important issues.

REVIEW FROM EXPERIMENTS I CLASS

- Studies may be descriptive, relational or experimental
- Controlled experiments
  - have a testable hypothesis
  - use quantitative methods and inferential statistics,
  - control variables and conditions
  - try to reduce experimenter bias
- Outcome commonly focuses on whether Design A is "better" than Design B, or meets a defined target
- Conducted an in-class study looking at the role of mental effort in people's perceptions of Google or the Library catalogue
  (based on Rieh, Kim & Markey, 2012: IPM, 48(6), 1336-1350

EXPERIMENT PLANNING FLOWCHART



STAGE B: PLANNING

STAGE D-E: ANALYSIS, INTERPRETATION

4. DESIGN THE TASK TO BE PERFORMED
WHAT WILL YOU ASK PARTICIPANTS TO DO?

To generate useful results, study tasks:

- Must be **externally valid:**
  - i.e., generalize to the real-life situation of interest, even though they are abstracted → predict how users would do in real-life
- Prioritize questions of top concern
  - For a large interactive system, controlled experiments can usually only **test a small subset** of all possible tasks.
- **Exercise** the designs, bringing out differences in their support for the task
  - E.g., if a design supports website *navigation*, test task should require participant to *move about the website*.
- Be **feasible**
  - Supported by the system, and executable given study time frame

5. EXPERIMENT DESIGN: MAKE EXPLICIT THE APPLICATION OF INDEPENDENT VARIABLES

- Simplest: 2-sample (2-variable) experiment
- Based on comparison of two sample means:
  - Performance data from using Design A & Design B
    - *e.g., new design & status quo design*
    - *e.g., 2 new designs*
- Comparison of one sample mean with a constant:
  - Performance data from using Design A, compared to performance requirement
  - Determine if Design A meets key design requirement
- Some comparisons involve the idea of a "**control**"
  - When one condition represents the status quo or other benchmark for an innovation
  - In HCI, there isn't always an obvious or relevant benchmark; sometimes you need to compare two innovations.

5. EXPERIMENT DESIGN: MAKE EXPLICIT THE APPLICATION OF INDEPENDENT VARIABLES

- More complex: factorial design
- In Engaged or Frustrated? Experiment (Edwards & Kelly):
  - 2 levels of task (interesting & uninteresting)
  - x2 levels of SERP ("normal" and "bad")
- In AIME experiment:
  - 2 systems
  - x2 task types
- In each: total of 2x2= 4 conditions
  - Within-subject: Each participant does the study task 4 times
  - Implication on resources: can you afford it?

WITHIN/BETWEEN SUBJECT DESIGNS

Within-subject designs

- Each participant exposed to multiple treatment conditions
  → primary comparison internal to each subject
- More control of subjective variation
- Greater statistical power
  → fewer subjects required
- Not always possible, e.g.
  - study session would be too long,
  - one condition contaminates participation in another.

Between-subject designs

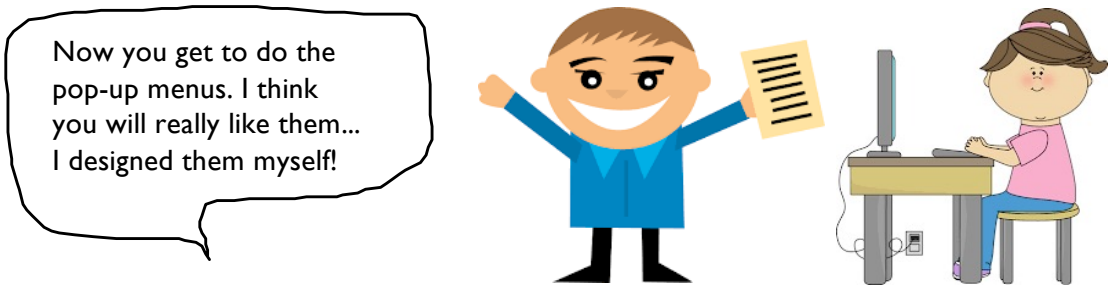
- Participants only exposed to one condition
  - primary comparison is from subject to subject
- Less statistical power, more subjects required
  - why? because greater variability due to more individual differences

Counterbalancing of conditions: Systematic variation of the order in which participants receive the same conditions (e.g. to compare 3 designs: mix up A-B-C, A-C-B, B-A-C, B-C-A, C-B-A, C-A-B)

Many possible variations, including **mixed factorial designs**: Combination of within-subject and between-subject in a factorial design

6. DESIGN EXPERIMENT PROTOCOL
SPECIFY AND DOCUMENT **EXACTLY** HOW THE STUDY WILL RUN

- Prepared and piloted in advance!!
- Includes unbiased instructions + instruments
  - Consent documents, full script, all data collection tools
- Logistics: where (room/online), team & roles, hardware setup
- Timing: estimate → verify with pilots. Must not go overtime.
- Plan out analysis before collection. → data collection formats
- Ideal: Double-blind protocol –
  Researcher unaware of which condition participant receives, thus unable to inadvertently bias them.



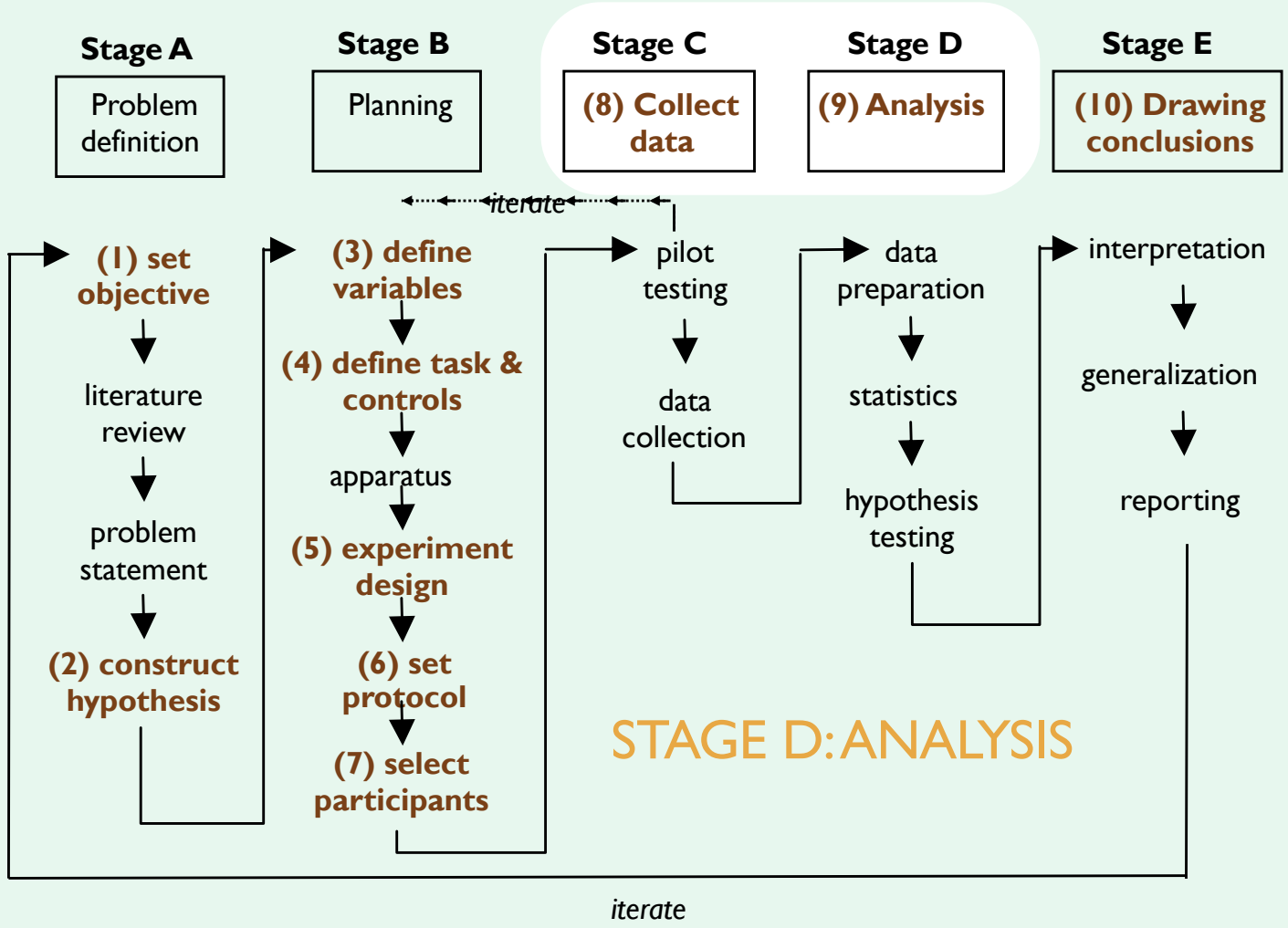
7. RECRUIT AND ASSIGN PARTICIPANTS TO GROUPS

- **Participant pool:**
  - Match expected user population as closely as possible
  - Age, physical attributes, level of education
  - General experience with systems similar to those being tested
  - Experience and knowledge of task domain

- **Assignment:**
  - Only necessary if between-subjects design (different experience)
  - Randomize!

- **Sample size:**
  - Go for "statistical significance"
  - Should be large enough to be "representative" of population
  - Guidelines exist based on statistical methods used & required significance of results
  - Pragmatic concerns may dictate actual numbers

EXPERIMENT PLANNING FLOWCHART



STAGE D: ANALYSIS

8. COLLECT THE DATA

- Varies somewhat by protocol
- A few general guidelines:
  - Take consent process seriously
  - Have backup plan for mishaps
  - For controlled experiments, especially critical to be consistent in instructions
  - Watch closely and take notes -- even if collecting quantitative performance data, there's always lots to learn by watching what people do
  - Take excellent care of your data
    - *Backup your data at every point possible during session (guard against computer crashes, bad synchronizations, etc)*
    - *Look at data after every session, ensure integrity (i.e. is it all there, nothing corrupted?)*

9. ANALYZE THE DATA

- **Prepare:**
  - Organize data from collection format into computable form (e.g., spreadsheet or statistics tool
- **Compute statistics**
  - Examples: t-tests, ANOVA, correlation, regression (more on these in Experiments II)
- **Test hypotheses**
  - Remember: what's possible here is to **reject your null hypothesis** with some degree of confidence
  - **Confidence limits:** the confidence that your conclusion is correct
  - *"The hypothesis that there is no difference in the search behaviours of engaged vs. frustrated searchers **is rejected** at the .05 level" (i.e., null hypothesis rejected)*
  - This means:
    - *95% chance that your finding is correct*
    - *5% chance you are wrong*

"CLEANING UP" THE DATA

- Start with a good stats text! E.g.
  - Tabachnick, B. G., Fidell, L. S., & Ullman, J. B. (2013). Using multivariate statistics (Vol. 6). Boston, MA: Pearson. [in UBC Library]
- Do we have missing data? How should we handle it?
- Have we made any input errors?
- Inspect the data:
  - Basic descriptors
  - Plot the raw data and take a look.
    Does anything look "off"?

For our experiment:
Take a look at **the raw data** (excel file, individual responses)

DEFINITIONS (REFERENCE)

- Valid – number of cases
- Missing – tells us if any of the items were skipped/not completed
- Mean – the average of responses for this item based on the responses of all participants
- Median – the mid-point; half the scores fall above this and half fall below this
- Standard deviation – "A standard deviation (or σ) is a measure of how dispersed the data is in relation to the mean. Low standard deviation means data are clustered around the mean, and high standard deviation indicates data are more spread out."
- Variance – measure of data dispersion; to calculate "each value in data set, subtract the mean and square the difference." Related to st. deviation b/c the square root of the variance is the standard deviation
- Skewness –" Skewness is a measure of symmetry, or more precisely, the lack of symmetry. A distribution, or data set, is symmetric if it looks the same to the left and right of the center point."
- Standard error of skewness: "The ratio of skewness to its standard error can be used as a test of normality (that is, you can reject normality if the ratio is less than -2 or greater than +2). A large positive value for skewness indicates a long right tail; an extreme negative value indicates a long left tail."
- Kurtosis – "Kurtosis is a measure of whether the data are heavy-tailed or light-tailed relative to a normal distribution. That is, data sets with high kurtosis tend to have heavy tails, or outliers. Data sets with low kurtosis tend to have light tails, or lack of outliers."
- Standard error of skewness: "The ratio of kurtosis to its standard error can be used as a test of normality (that is, you can reject normality if the ratio is less than -2 or greater than +2). A large positive value for kurtosis indicates that the tails of the distribution are longer than those of a normal distribution; a negative value for kurtosis indicates shorter tails (becoming like those of a box-shaped uniform distribution)."

VISUAL INSPECTION OF DATA:
IMPORTANT TO DO, BUT BE AWARE OF LIMITS

- There is almost always variation in collected data
- Differences between data sets may be due to:
  - Normal variation
    - *e.g., two sets of ten tosses with different but fair dice*
      - Differences between data and means are accountable by **expected** random variation in the population
  - Real differences between data
    - *e.g., two sets of ten tosses with loaded dice and fair dice*
      - Differences between data and means are not accountable by expected variation; thus more likely due to something in the study (the manipulation, or a problem)
- → Means alone don't tell the story; need to look at variance as well.

PREPARING FOR ANALYSIS

- What are statistics and why are they used?
- What statistics are relevant to HCI?
- How (and why) do we clean and visually inspect the data?
- How do we select an appropriate statistical test?
- How do we describe the data and test for differences in DVs as a result of our IVs?
- What is the difference between statistical and practical significance?

STATISTICAL ANALYSIS

- What is a statistic?
  - A number that describes a **sample**
  - A **sample** is a subset (hopefully representative) of the population we are interested in understanding
- Statistics are calculations that tell us:
  - Mathematical attributes about our data sets (sample)
    - *mean, amount of variance, ...*
  - How data sets relate to each other
    - *whether we are "sampling" from the same or different populations*
  - The probability that our claims are correct
    - *"statistical significance"*

WHAT STATISTICAL TESTS ARE COMMONLY USED IN HCI?

Parametric tests:

Tests group means.

More powerful … use if possible

- Correlation (Pearson)
- Regression
- T-tests and ANOVA

BUT IF any of these apply:

- Small sample size
- Ordinal data
- Outliers
- Data is skewed (non-normal; better represented by median)

Non-parametric tests:

Tests group medians

- Correlation (Spearman)
- Chi-squared
- Mann-Whitney
- Wilcoxon signed-rank
- Kruskal-Wallis
- Friedman's

Learn more about these in EPSE 592 or other statistics courses

Parametric tests are more powerful: More likely to detect significance differences if there are any. But in many cases, they won't be valid.

https://blog.minitab.com/en/adventures-in-statistics-2/choosing-between-a-nonparametric-test-and-a-parametric-test

HOW DO WE SELECT THE RIGHT STATISTICAL TEST?

- Type of Independent & Dependent Variables (IV, DV):
  - **Nominal** (categorical) | **Ordinal** | **Interval** (numerical)
  - What's the difference: https://stats.idre.ucla.edu/other/mult-pkg/whatstat/
- Number of IV levels
  - E.g., *"We are interested in understanding how the amount of education a person has is related to their job satisfaction."*
    - ***IV: Education level****: high school, some college/university, bachelors degree, Master's degree (4 levels, ordinal)*
    - ***DV: Job satisfaction:*** *10 item Likert-scale questionnaire resulting in a job satisfaction score (ordinal or interval)*
- Parametric vs. non-parametric: dictated based on
  - Sample size
  - Normalcy of the distribution, ordinality, outliers

TEST ON DIFFERENCES BETWEEN MEANS, MEDIANS

Example:

- Given: Data for N systems, measuring M conditions
  - E.g., Perceived invested mental effort (collected as a rating scale)
- Question: Is the difference between the means/medians of the data statistically significant? (depends on variance!)
- Null hypothesis:
  - There is no difference between the two means/medians
  - Test can only reject the hypothesis at a certain level of confidence
  - We never actually prove the alternate hypothesis true

→ Look up test type here: https://stats.oarc.ucla.edu/other/mult-pkg/whatstat/

For our experiment:
- How many IVs (**conditions**) do we have? How many **levels** of each?
- What **kind of data** is our **rating scale**? (as opposed to our behavioral data)

→ Should we do a parametric or non-parametric test? Which one?

SELECTING THE TEST

Common choices for simple experiments:

	Parametric test (Means)	**Non-parametric test (Medians)**
Paired observation (within-subjects)	1 sample t-test	1-sample Wilcoxen
Comparing 2 sets of independent observations	2 sample t-test	Mann Whitney U
Compares 3 or more independent groups	One-way ANOVA	Kruskall Wallis

NON-DIRECTIONAL VS. DIRECTIONAL TESTS

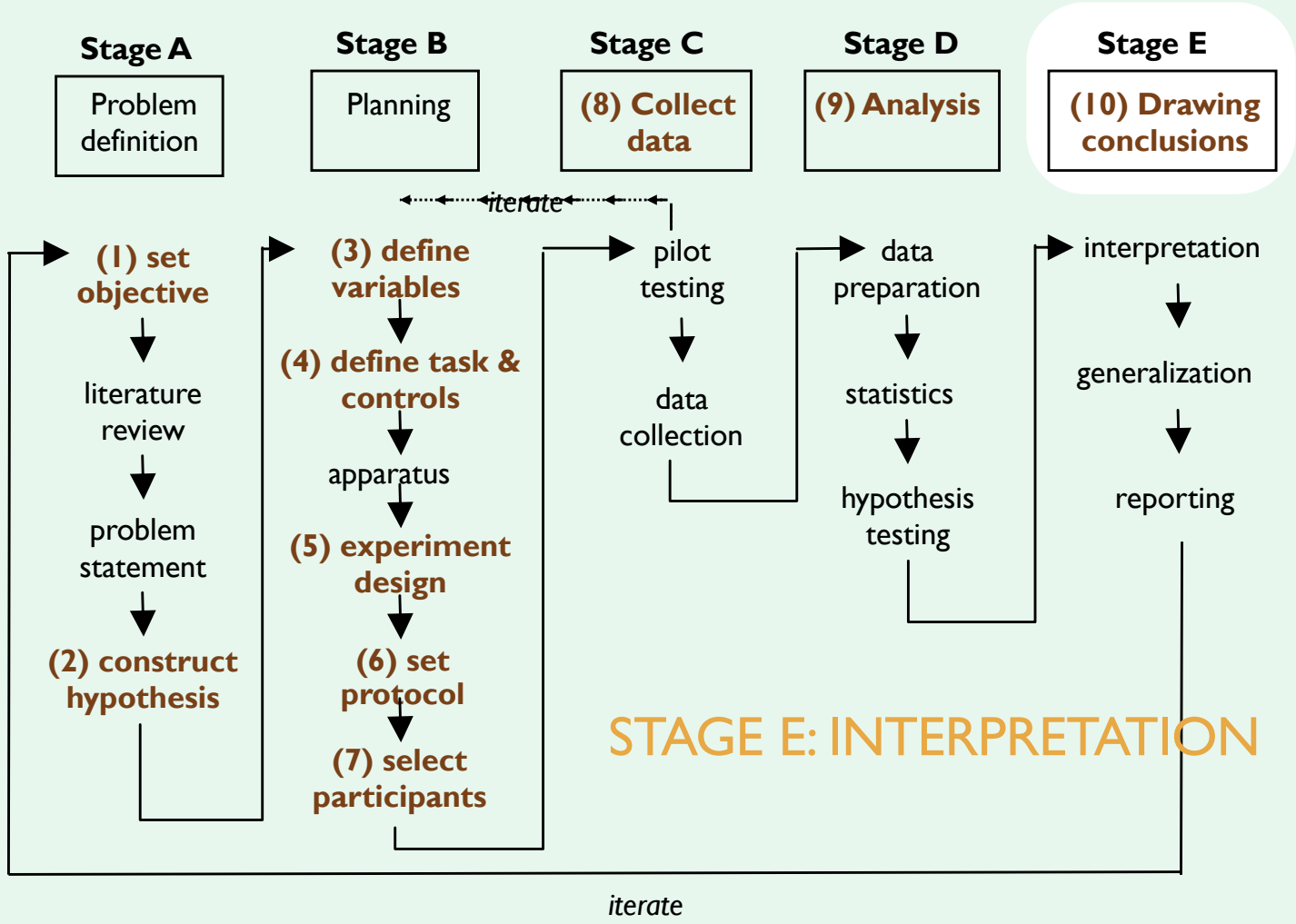
- Non-directional (two-tailed)
  - no expectation that the direction of difference matters
- Directional (one-tailed)
  - only interested if the mean of a given condition is greater than the other

STATISTICAL VS. PRACTICAL SIGNIFICANCE

- When N is large, even a trivial difference (small effect) may be large enough to produce a statistically significant result
- Statistical significance does not imply that the difference is important!
  - A matter of interpretation, i.e., subjective opinion
  - Should always report means to help others make their opinion
- There are measures for effect size
  - Not widely used in HCI research

Kelly, D. (2015, March). Statistical power analysis for sample size estimation in information retrieval experiments with users. In *European Conference on Information Retrieval* (pp. 822-825). Springer, Cham.

EXPERIMENT PLANNING FLOWCHART

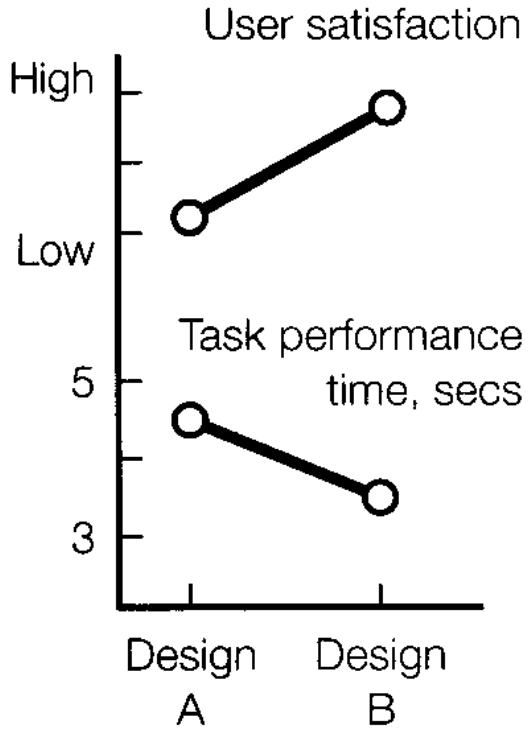


STAGE E: INTERPRETATION

10. DRAW CONCLUSIONS: INTERPRET YOUR RESULTS

- What you believe the results mean, and their implications
- Yes, there can be a subjective component to quantitative analysis!

INTERPRETING RESULTS: VISUALIZE A TEST OUTCOME



This graph illustrates a (probably) **definitive** result

The outcome looks pretty clear

- The two metrics vary in a consistent way
- The differences seem definite

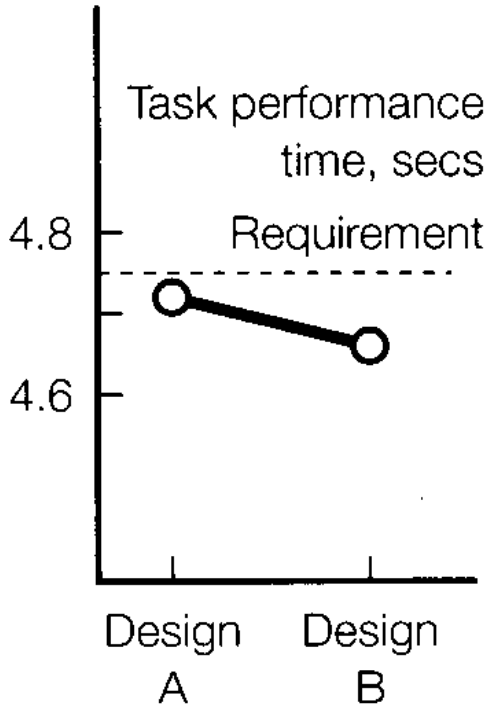
But, be careful:

- This plot is over-simplified – no variance estimate or full scale
- Is the difference statistically significant?
- As always, a result is only informative if the task was well chosen and the study carried out without confound or bias.

POOR EXPERIMENT DESIGN OR RESULTS

Less distinguishable results (inconclusive?)

- There ***might*** be no difference;
- Also possible that the task was poorly chosen, and did not "exercise" the difference in designs



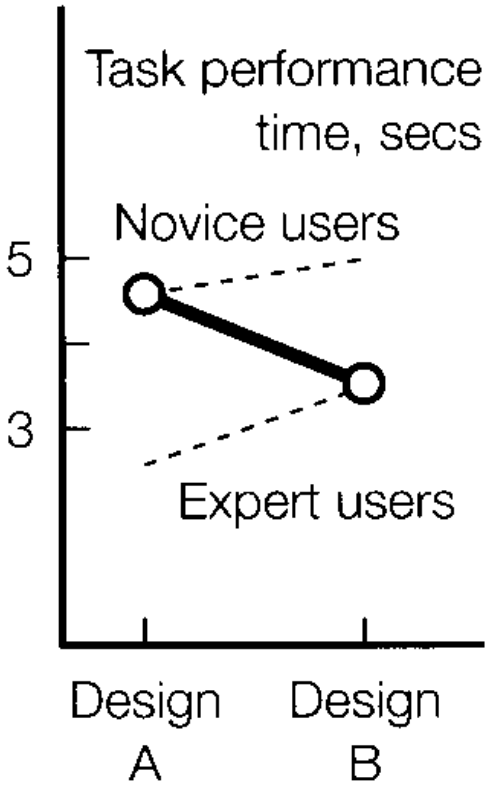
POOR EXPERIMENT DESIGN

Here, the study design introduced a **confound**

The observed difference might be due to at least two different factors. We don't know which.

In this case, *participant assignment* was confounded with *tested design variant*:

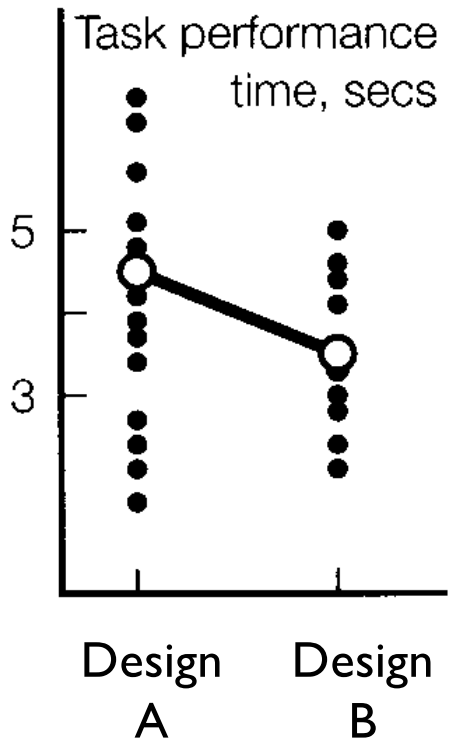
- One design tested on novices, the other on experts,
- Was the difference due to the design, or to the participant group?



POOR EXPERIMENT DESIGN OR RESULTS

High variance also leads to inconclusive results

- This plot finally shows variance.
  - An apparently notable difference might be small compared to the variation within a design.
    The statistical test here probably wouldn't show significance.
- Why?
  - An unidentified "nuisance" variable" including individual differences
  - Sloppy data collection
  - Who knows? But you're not done til you figure it out.



FINDINGS FROM OUR STUDY

Take a look at the results of the in-class experimental study

[Handout with summary and test stats]

First, review the **test results** and ask yourself questions about it.

- Which was our statistical test computed on: Means or Median?
- Which column shows the statistical test result?
- Was anything significant? How can you tell? Which tests were? Circle / mark all the ones you can find. Do any *approach* significance?

- Discuss some of the significant ones we found. What does it mean?
  - Which rating item was the "money question" – the one that the hypothesis about "Perceived invested mental effort" most hinges on?
  - In the ratings data, what exactly does the U test compare?
  - So what was significantly different, in the ratings data? Did anything change, pre to post test?
  - For behavioral data, any patterns?

FINDINGS FROM OUR STUDY

Now, consider at a higher level (draw conclusions):

1. Can we **reject our null hypothesis** that there is no difference in the **perceived amount of search effort** for people searching the **library catalogue vs. Google?** Why or why not?
2. What **trend** do we see in the self-report data with respect to amount of mental effort? What evidence is there for this?
3. Were the **behaviours** of the 2 groups different in any way?

- Think about the quality and adequacy of the data.
  - How much do you trust these results? Why?
- What are some potential confounds or sources for potential error in the study?
  - How might we address these?

TO SUMMARIZE:
HOW A CONTROLLED EXPERIMENT WORKS

- Formulate an alternate and a null hypothesis:
  - H1: experimental conditions have an effect on performance
  - H0: experimental conditions have no effect on performance
- Through experimental task, try to demonstrate that the null hypothesis is false (reject it),
  - For a particular level of significance
- If successful, we can accept the alternate hypothesis,
  - and state the probability p that we are wrong (the null hypothesis is true after all) → this is result's confidence level
  - At a .05% CL: 5% chance we got it wrong, 95% confident