The 5-Step Approach to Design Controlled Experiments

Shengdong Zhao
NUS-HCI Lab
National University of Singapore

Outline

The 5 Step Approach to Experiment Design

2. Define the research question
3. Determine variables
4. Arrange conditions
5. Decide blocks and trials
6. Set instruction and procedures

Let's Start with an Example Problem

earPod vs. iPod


PayPal Mobile
05:06
1 Send
2 Balance
3 Cancel
4 Text to Give
5 Help
Select
Back
SAMSUNG


ProTrader Menu
Watch list
Subscribe
Positions
Orders
Alerts
General
News
Messages
Reconnect
Ok
Exit
VIDEO CALL


DELL
AXIM
Start
12:25
Today
ActiveSync
Calendar
Contacts
Home
Inbox
Internet Explorer
Notes
Tasks
Windows Media
Programs
Find
Help
X51


iPod
Music
Photos
Extras
Settings
Shuffle Songs
MENU

Menu Selection on Mobile Devices

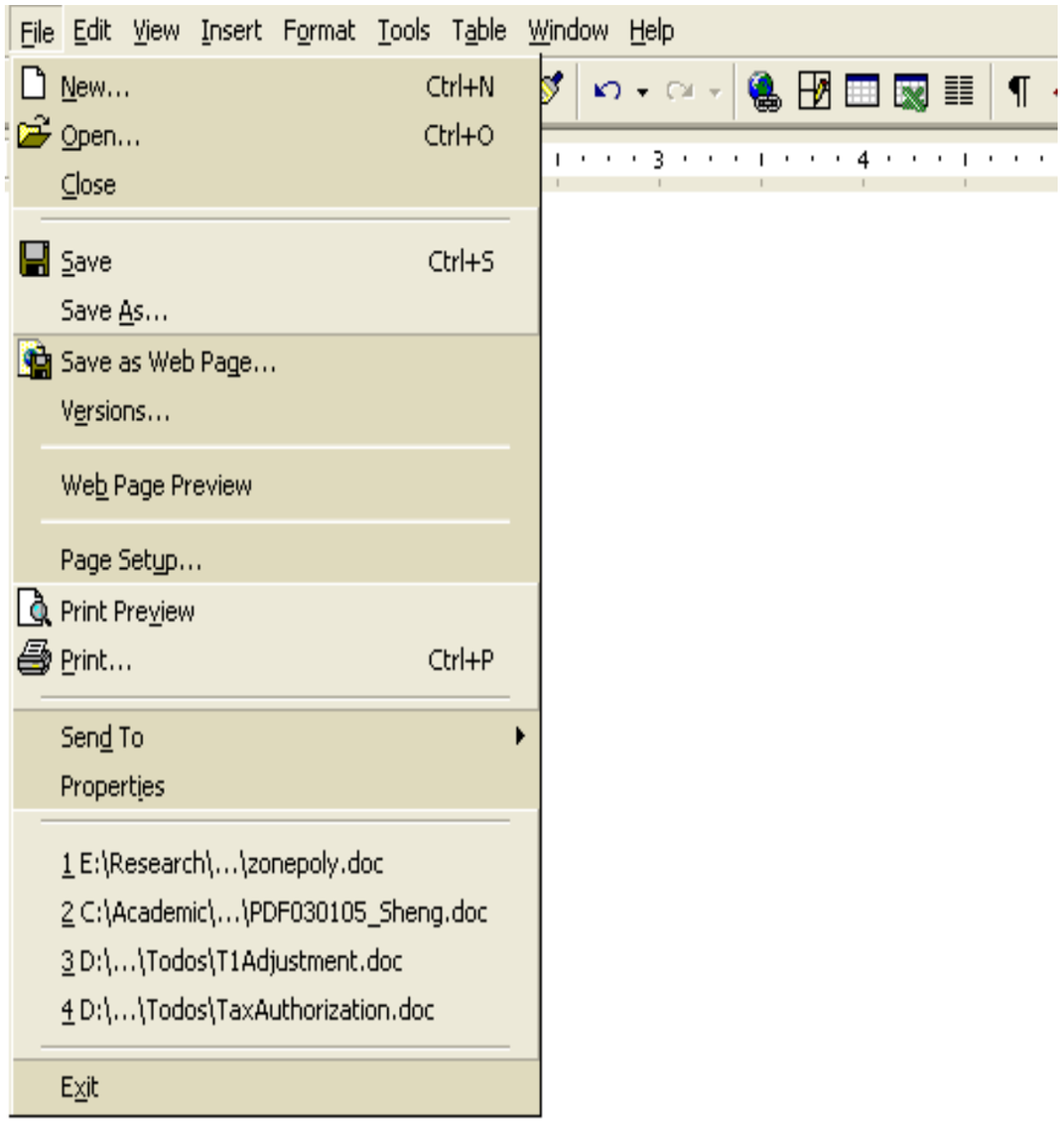
Often Requires

Visual Feedback

Why Eyes-free?

Why eyes-free?

Visual vs. auditory menu

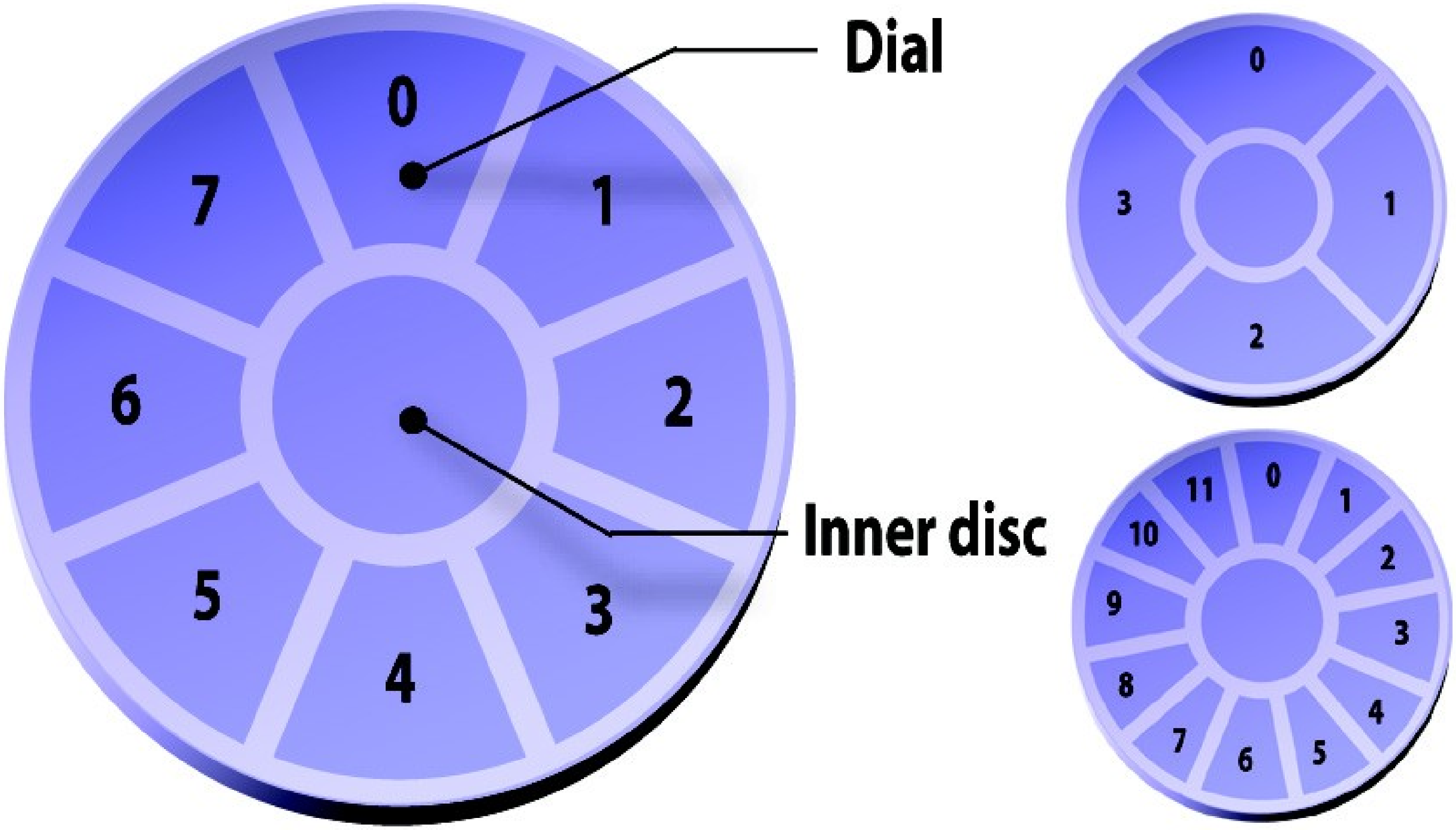


Visual Linear Menu

IVR System

earPod

earPod design



Video

- http://www.youtube.com/watch?v=bATkA0Usoio

Paper

- Shengdong Zhao, Pierre Dragicevic, Mark H. Chignell, Ravin Balakrishnan, Patrick Baudisch (2007). earPod: Eyes-free Menu Selection with Touch Input and Re*. Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI)*. pp. 1395-1404

Question: earPod vs. iPod

The 5 Step Approach to Experiment Design

- **Define the research question**
- Determine variables
- Arrange conditions
- Decide blocks and trials
- Set instruction and procedures

The 5 Step Approach to Experiment Design

1. **Define the research question**
   - Step 1.1 Start with a general question
   - Step 1.2 Define target population
   - Step 1.3 Define task(s)
   - Step 1.4 Define measure(s)
   - Step 1.5 Define factor(s)
2. Determine variables
3. Arrange conditions
4. Decide blocks and trials
5. Set instruction and procedures

Step 1.1: Start with a General Question

How does *earPod* compare with iPod's menu in terms of performance?

Step 1.2: Define Target Population

General question: How does *earPod* compare with iPod's menu in terms of performance?

Target population?

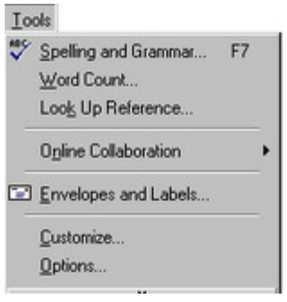
Question: we designed earPod for whom?

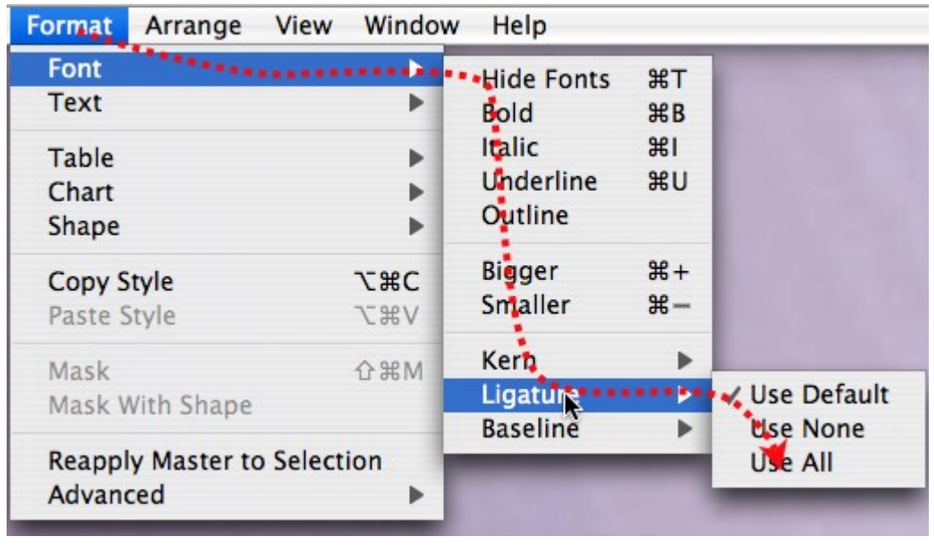
Step 1.3: Define Task(s)

General question: How does *earPod* compare with iPod's menu in terms of performance?

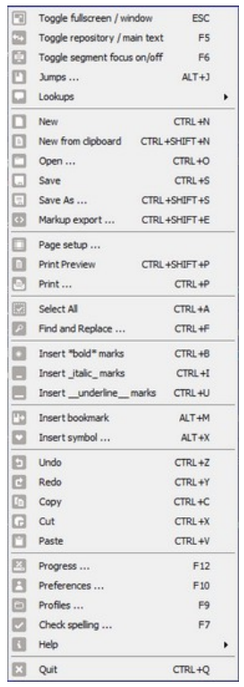
Task(s): menu selection

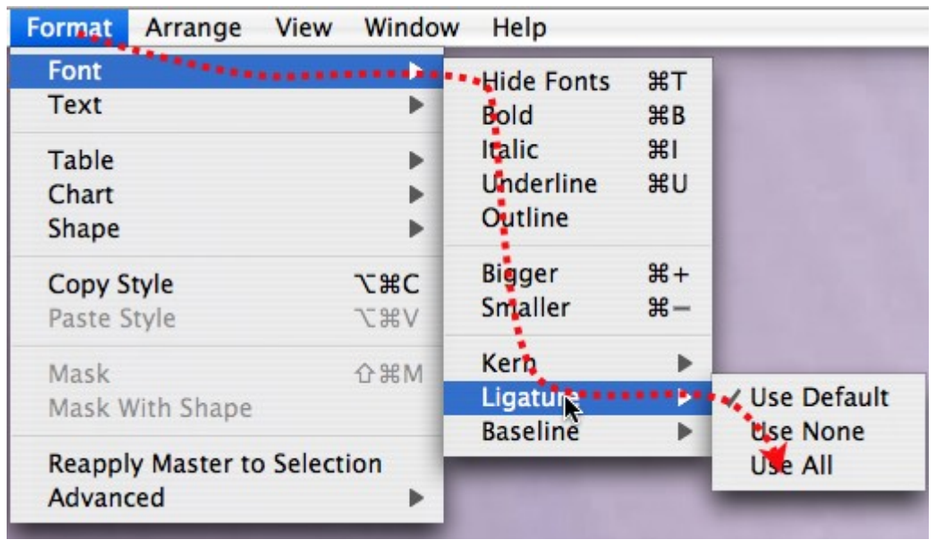
However, the menu selection task has endless possibilities: single short menu, single long menu, hierarchical menus





Step 1.3: Define Task(s)





Key insight: experiment design need to decide what subset of tasks is appropriate to test.

Question: how do you choose the subset?

Step 1.4: Define Measures

Question: How does *earPod* compare with iPod's menu in terms of performance?

Measures: performance

In HCI, we typically use three measures to quantify performance:

- Speed
- Accuracy
- Learnability

Key insight: need to define "testable" measures

Step 1.5: Define (other) Factors

Question: How does *earPod* compare with iPod's menu in terms of performance?

Factors: other than the different type of tasks, what other factors can influence the measures?

Again: the number of factors are unlimited …

- Scenario of use
- Input device
- Background of the user
  - Educational level
  - Gender
  - Ethnic background
  - Age
  - …

Key insight: experiment design need to determine a subset of factors to test.

Question: how to choose the factors?

Let's Review Step 1

Step 1: Define the research question

- Step 1.1: Start with a general question
- Step 1.2: Define target population
- Step 1.3: Define task(s)
- Step 1.4: Define measure(s)
- Step 1.5 Define factor(s)

Let's Practice

Example 1: "earPod vs. iPod"

1.1: General Question

- How does *earPod* compare with iPod's menu in terms of performance?

1.2: Target Population

- Young generation

1.3: Task(s)?

- e.g., Menu selection for three types of breadth (4, 8, 12) and two types of depth (1, 2), content of the menu is from common categories

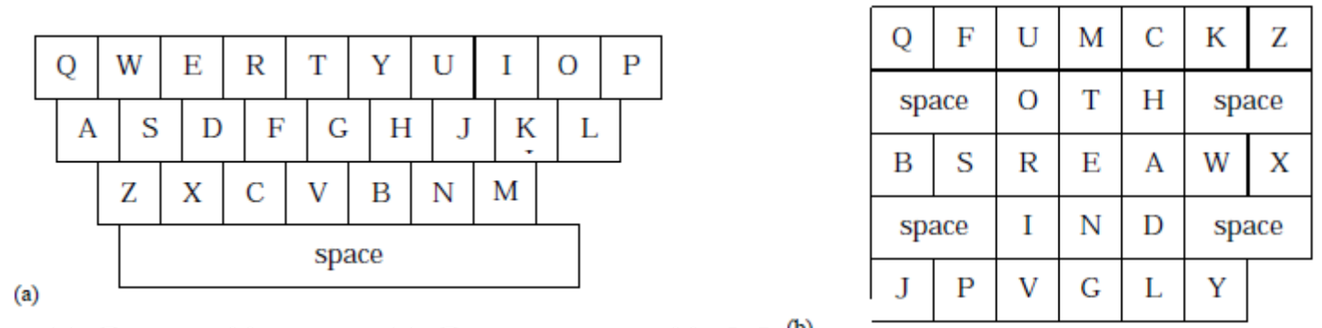
1.4: Measures?

- Speed, accuracy, learning

1.5: (Other) Factors?

- Single-task vs. multi-tasking
- …

Let's Practice Again



Example 2: "Opti" vs. "Qwerty" Keyboard

1.1: General question
- how does the "opti" keyboard layout compare with the "qwerty" keyboard in performance?

1.2: Target Population?
- Computer users?

1.3: Task(s)?
- Type "the quick brown fox jumps over the lazy dog"

1.4: Measure(s)?
- Speed, accuracy, learning

1.5: (Other) Factors?
- Device: Touch typing vs. stylus?
- Screen size: different screen size?

The 5 Step Approach to Experiment Design

- Define the research question
- **Determine variables**
- Arrange conditions
- Decide blocks and repetitions
- Set instruction and trials

Step 2: Define Variables

Type of variables

- Independent variable (IV)
  - Factors that are manipulated in the experiment
  - Have **multiple levels**
- Dependent variable (DV)
  - Factors which are measured
- Control variable
  - Attributes that will be fixed throughout experiment
  - Confound – attribute that varied and was not accounted for
    - Problem: Confound rather than IV could have caused change in DVs
  - Confounds make it difficult/impossible to draw conclusions
- Random variable
  - Attributes that are randomly sampled
  - Increases generalizability

Type of Independent Variables

- Primary
  - The most important independent variable(s) that you want to investigate
  - For the question: "How does the earPod compare with iPod's menu in terms of performance, the primary focus of interest is the different type of device/technique (earPod vs. iPod), so the primary IV is device/technique
- Secondary
  - The other interesting factors you want to manipulate in the experiment. They help to answer the main question in a richer way. For example, a secondary IV for the earPod vs. iPod experiment will be the scenario of use (stationary vs. mobile). This variable helps to answer the primary question "how does earPod vs. iPod" in a richer way: earPod may work better in mobile while iPod better in stationary, etc.

Task Type & Factor → Independent variables

Measures → Dependent variables

Everything else → Control/Random Vars.

Let's Try

Example 1: earPod vs. iPod

- Independent variables
  - Technique
    - 2 levels (earPod vs. iPod)
  - usage scenario
    - 2 levels (single-task vs. dual-task)
  - menu breadth
    - 3 levels (4, 8, 12)
  - menu depth
    - 2 levels (1, 2)

- Dependent variables
  - Speed (measured in completion time)
  - Accuracy (measured in percentage of errors)
  - Learning (measured in speed & accuracy change over time)

Let's Try

Example 1: earPod vs. iPod

- Control variables
  - Same computer, experiment time, environment, instruction, etc.

- Random variables
  - Attributes of participants: age, gender, background, etc.

Let's Try Again

Example 2: "Opti" vs. "Qwerty" keyboard layout

- Independent variables
  - Type of keyboard
    - 2 levels (opti vs. qwerty)
  - Input method
    - 2 levels (touch vs. stylus)
  - Screen size
    - 3 levels (watch, mobile phone, tablet)

- Dependent variables
  - Speed (measured in word per minute)
  - Accuracy (measured in?)
  - Learning (measured in speed & accuracy change over time)

Example 2

Example 2: “Opti” vs. “Qwerty” keyboard layout

- Control variables
  - Same computer, experiment time, environment, instruction, etc.

- Random variables
  - Attributes of participants: age, gender, background, etc.

Confounding Variable

Any variable other than the independent variables that can possibly explain the change in measures

Example 1 – two techniques are compared (earPod, iPod)

- All participants are tested on earPod, followed by iPod
  - Performance might improve due to practice
  - The order of presenting the technique is a confounding variable (because it explains the changes in measures but it is not an IV)

Example 2 – two software interfaces are compared (Microsoft Word vs. new)

- All participants have prior experience with Microsoft Word, but no experience with the new interface
  - "Prior experience" is a confounding variable

Note: Order of presentation & Prior experience are two important confounding variables we need to control. More on this later …

The 5 Step Approach to Experiment Design

- Define the research question
- Determine variables
- **Arrange conditions**

  From Independent Variables to Experimental Conditions

4. Decide blocks and trials
5. Set instruction and procedures

What is a condition?

- Let's start with an example
- A particular **independent variable** "Technique" has two **levels**: earPod and iPod.
  - If it is the only independent variable considered, this experiment has two conditions
- However, an experiment rarely only has 1 independent variable, suppose there is another independent variable "Menu Breadth" with 3 levels (4, 8, 12).
  - There are 2 (Techniques) x 3 (Menu Breadth) = 6 experimental conditions
  - The each unique combination of the different levels of the various independent variables (such as earPod, 4) is an experimental condition

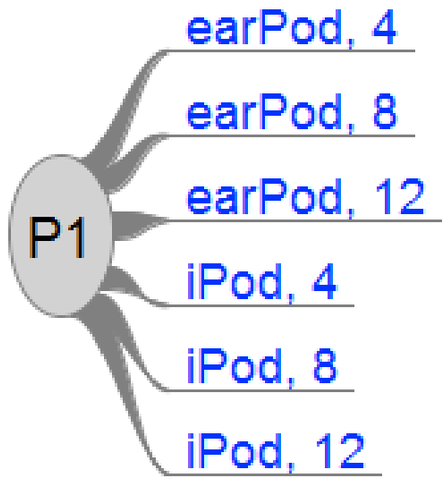
How can We Test These Conditions?

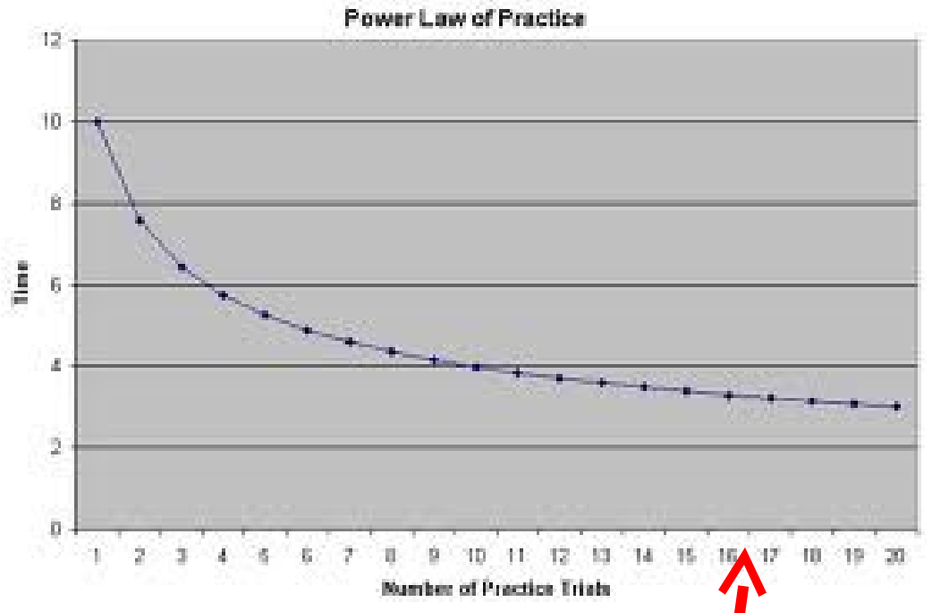
- **Method 1**:
  - Recruit 6 participants, one for each condition (this is also called **between-subject design**, which means the conditions are tested between different subjects)
    - P1: earPod, 4
    - P2: earPod, 8
    - P3: earPod, 12
    - P4: iPod, 4
    - P5: iPod, 8
    - P6: iPod 12
  - What's the problem of this approach?
    - What about individual differences?
    - To balance individual differences, we need lots of participants

- **Key insight:** this method is expensive

How can We Test It?

- **Method 2**:
  - Recruit the same participants to test all 6 conditions (this is also called **within-subject design** since all conditions are tested within the same subject)





  - This method is much more economical
  - What's the problem of this approach?
    - Practice (or order effect) as a confounding variable
    - However, in many cases, this effect can be controlled

Control Order Effect using Counter-balancing

If we assume the order effect is symmetric, which means A -> B = B->A, and is linear, which means the increment between different conditions is about the same, we can use **counter-balancing** to cancel the effect out.

E.g., we assume the transferring effect between (A after B) and (B after A) are both 10

Participant 1: A followed by B (A B)
Participant 2: B followed by A (B A)

Observation: the order effect equally affects both A and B, so the absolute relationship between A and B is not changed.

However, a minimum number of participants is needed for counter-balancing to work

Counter-balancing with 3 Levels

What if an IV has 3 levels? A B C

If the same assumption holds: assume effects are symmetric, and equal in size. We need to counter-balance as follows.

P1: A B C
P2: A C B
P3: B A C
P4: B C A
P5: C A B
P6: C B A

What about 4 levels, 5 levels, 6 levels, …?
4 level = 4! (24), 5 levels = 5! (120), …

Introducing Partial Counter-balancing: Latin Square

Latin square:

- Ensures each level appears in every position in order equally often:

A B C
B C A
C A B

Assume A-B = B-A = A-C = C-A = B-C = C-B = 10

P1: a + (b+10) + (c+20)
P2: b + (c+10) + (a+20)
P3: c + (a+10) + (b+20)
Average A = (3a +30)/3 = a + 10
Average B = (3b+30)/3 = b + 10
Average C = (3c+30)/3 = c + 10

However, A-B = B-A = 10
A-C = C-A = 20
B-C = C-B = 30

P1: a + (b+10) + (c+50)
P2: b + (c+30) + (a+30)
P3: c + (a+20) + (b+40)
Average A = (3a +50)/3 = a + 50/3
Average B = (3b+50)/3 = b + 50/3
Average C = (3c+80)/3 = c + 80/3

Steps for Arranging Conditions for Within-Subject Design

3.1: List all Independent Variables and their levels

3.2: Decide counter-balancing strategy for each variable

3.3: Determine the minimum No. of participants 3.4: Arrange the overall design

3.5: Determine detailed arrangement for each participant

Example 1: earPod vs. iPod

Assume we have three IVs

Step 3.1: list the IV and their levels

- Technique (2 levels: earPod, iPod)
- Scenario of use (2 levels: single-task, multi-task)
- Menu depth (2 levels: 1, 2)

Step 3.2: determine counter-balancing strategies for each IV

- Choices: 1) fully counter-balancing, 2) Latin-square, 3) no counter-balancing (sequential)
- Question: how to decide which strategy to use?
  - It depends on how interesting is the independent variable
  - It depends on how much resource we have

Example 1: earPod vs. iPod

Step 3.2: Counter-balancing strategies

- Technique **(fully counter-balance)**
- Scenario of use **(fully counter-balance)**
- Menu depth **(no counter-balance, sequential)**
- Why?

Step 3.3: Determine the minimum No. of participants

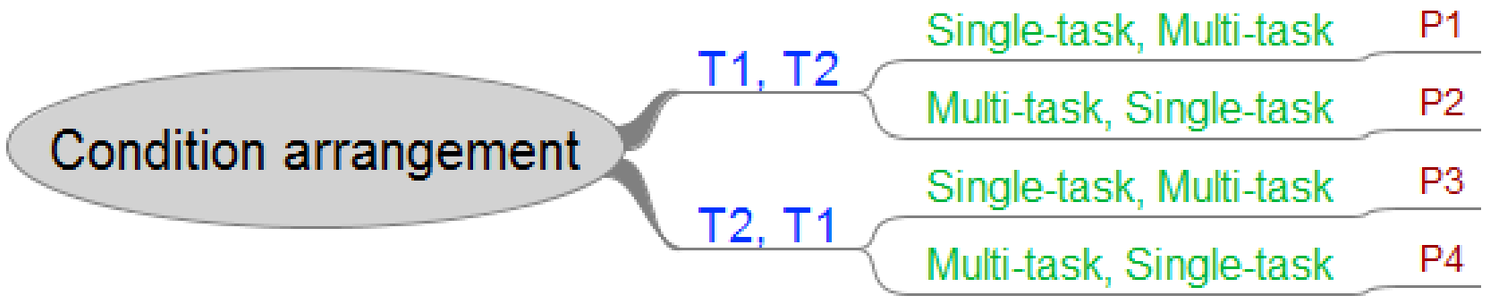
- Minimum No. = 2 Tech. conditions X 2 Scenario conditions X 1 Menu depth arrangement = 4
- Question: if Menu depth is also fully counter-balanced, how many participants we need?
- Question: if Technique has 3 levels and it is fully counter balanced (assume menu depth is not counter-balanced), how many participants we need?

Step 3.4: Determine the overall arrangement

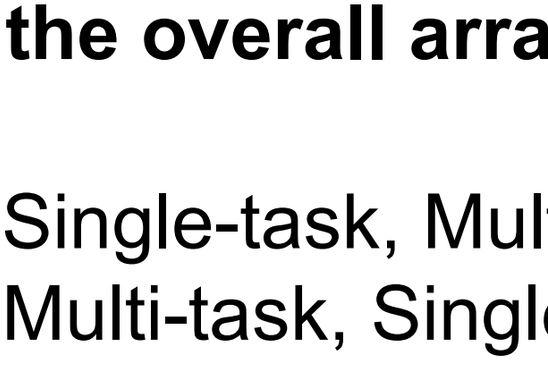
T1, T2
T2, T1

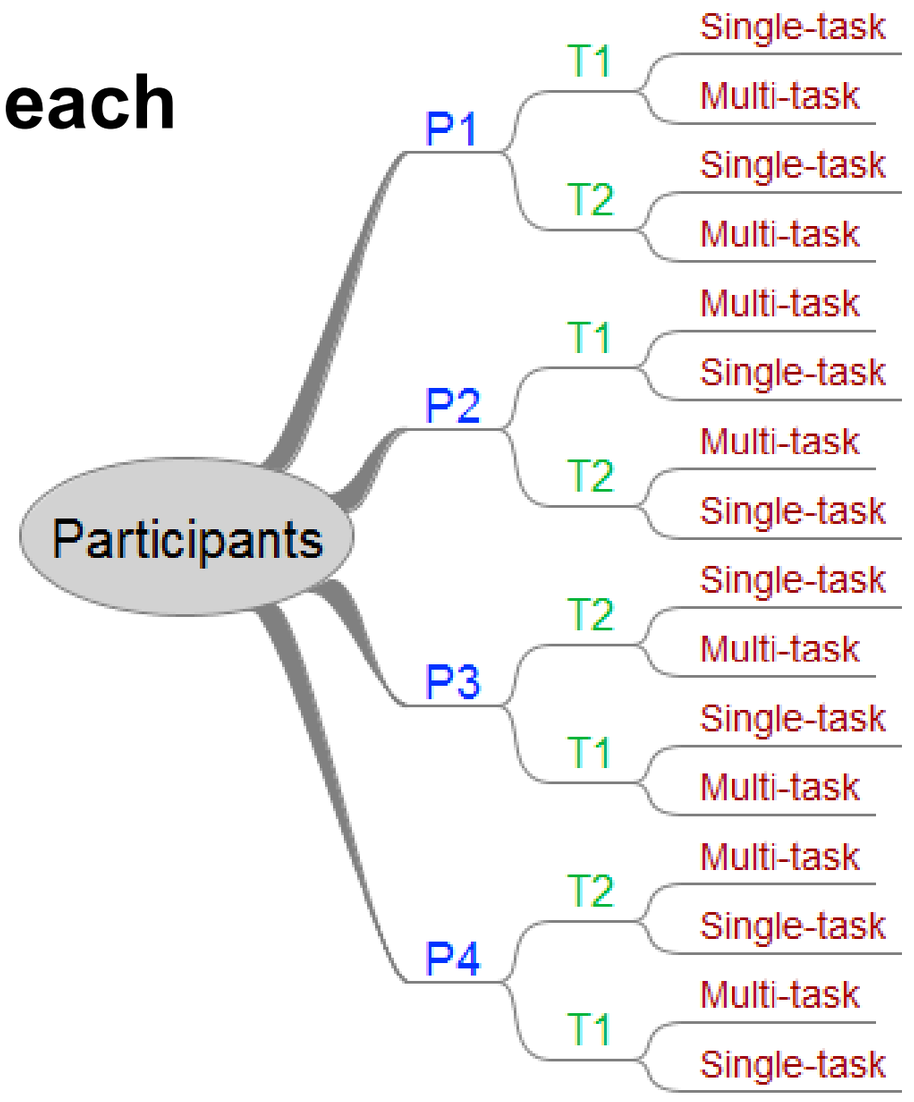
x

Single-task, Multi-task
Multi-task, Single-task



Step 3.5: Determine arrangement for each participant





In-class Exercise: Example 2

Step 3.1: List IVs

- Technique (3 levels: A, B, C)
- Scenario of use (2 levels: single-task, multi-task)

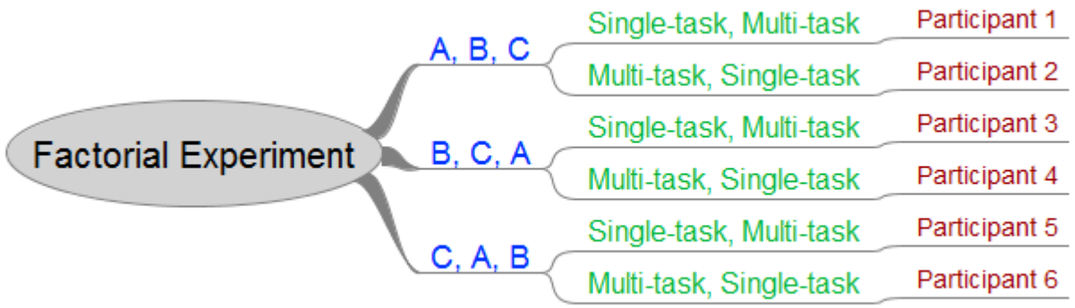
Step 3.2: Decide counter-balancing strategy

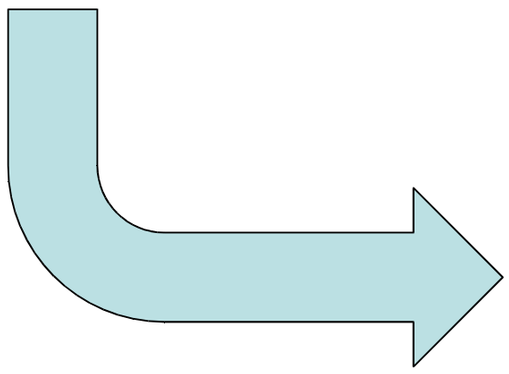
Step 3.3: Determine Minimum No. of Participants

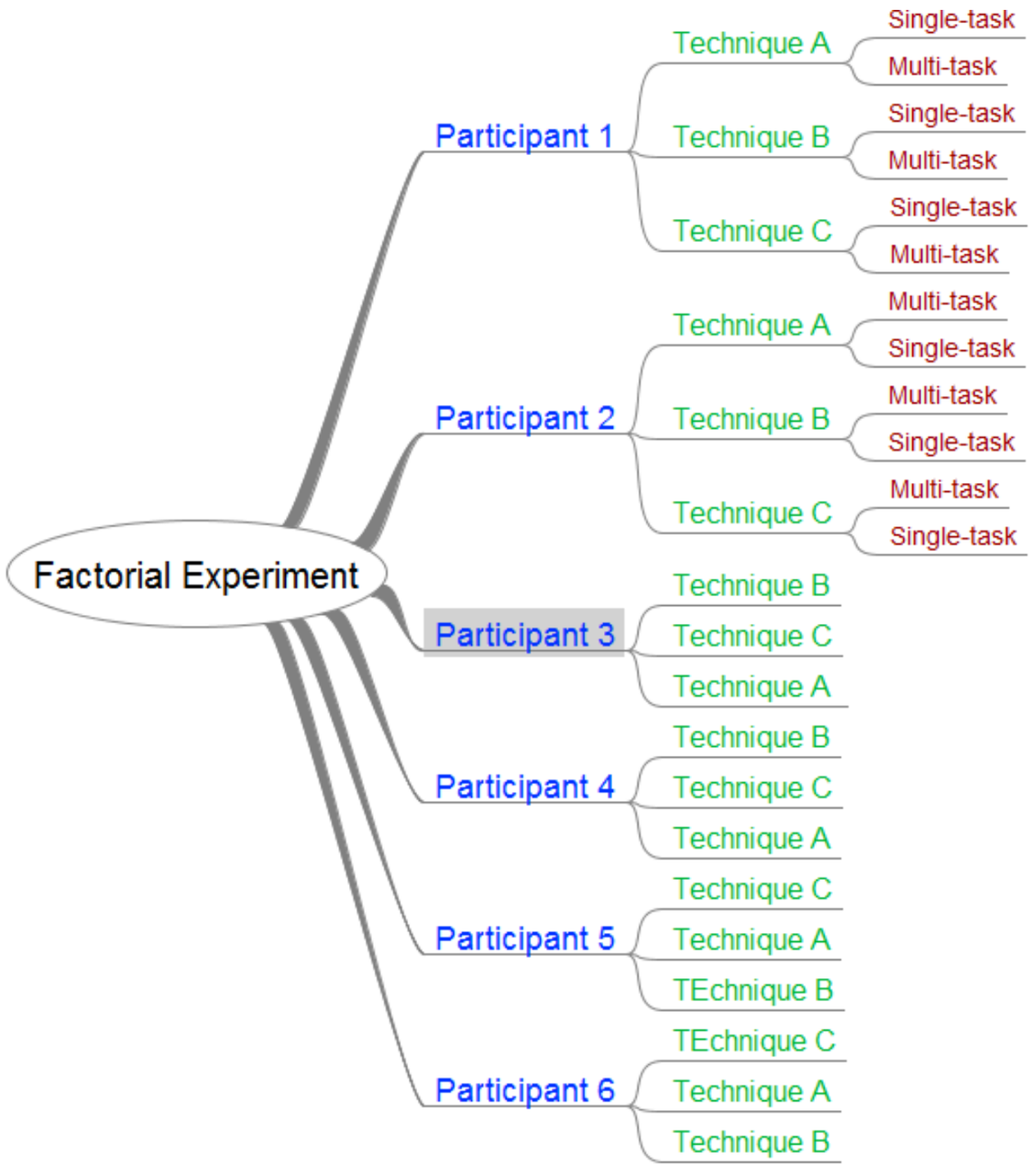
Step 3.4: Determine the overall arrangement

Step 3.5: Determine individual arrangement for each participant

Possible Answer







However, Counter-balancing May not Always Work

Counter-balancing Assumes symmetric transfer and linear increment

- A-B transfer == B-A transfer
- A-B transfer == B-C transfer

If asymmetric transfer and non-linear increment

- i.., A-B transfer > or < B-A transfer or A-B <> B-C
- Have to use **Between-subjects design**
- In addition, some factors have to be between-subject
  - Age, Gender, etc.

No. of Condition Reduction Strategies

In experiment design, one major problem we often face is there are many possible relevant factors. It's important for experiment designers to pick the most important/interesting factors to test.

Run a few independent variables at a time

- If strong effect, include variable in future studies
- Otherwise pick fixed control value for it

Not all within-subject IVs need to be counter-balanced

- If we are not interested in the absolute difference among different levels, we don't need to counter-balance. E.g., Menu Breadth, Menu Depth, etc.

Exercise: earPod vs. iPod

- Independent variables
  - Technique (2 levels: earPod vs. iPod)
  - Usage scenario (2 levels: single vs multi-tasking)
  - Menu breadth (3 levels: 4, 8, 12)
  - Menu depth (2 levels: 1, 2)
- Question: which of these factors need counter-balancing?

Let's Review: Between- vs. Within-Subject Design

- Method 1: use a lot of participants, randomly assign them to each technique (between-subject design)
  - Drawback: costly

- Method 2: use the same participant to test both techniques (within-subject design)
  - Drawback: practice effect

For Between-subject Design, there is no need for counter-balancing, just assign different users to different conditions!

Steps for Arranging Conditions for Within-Subject Design

3.1: List all Independent Variables and their levels

3.2: Decide counter-balancing strategy for each variable

3.3: Determine the minimum No. of participants 3.4: Arrange the overall design

3.5: Determine detailed arrangement for each participant

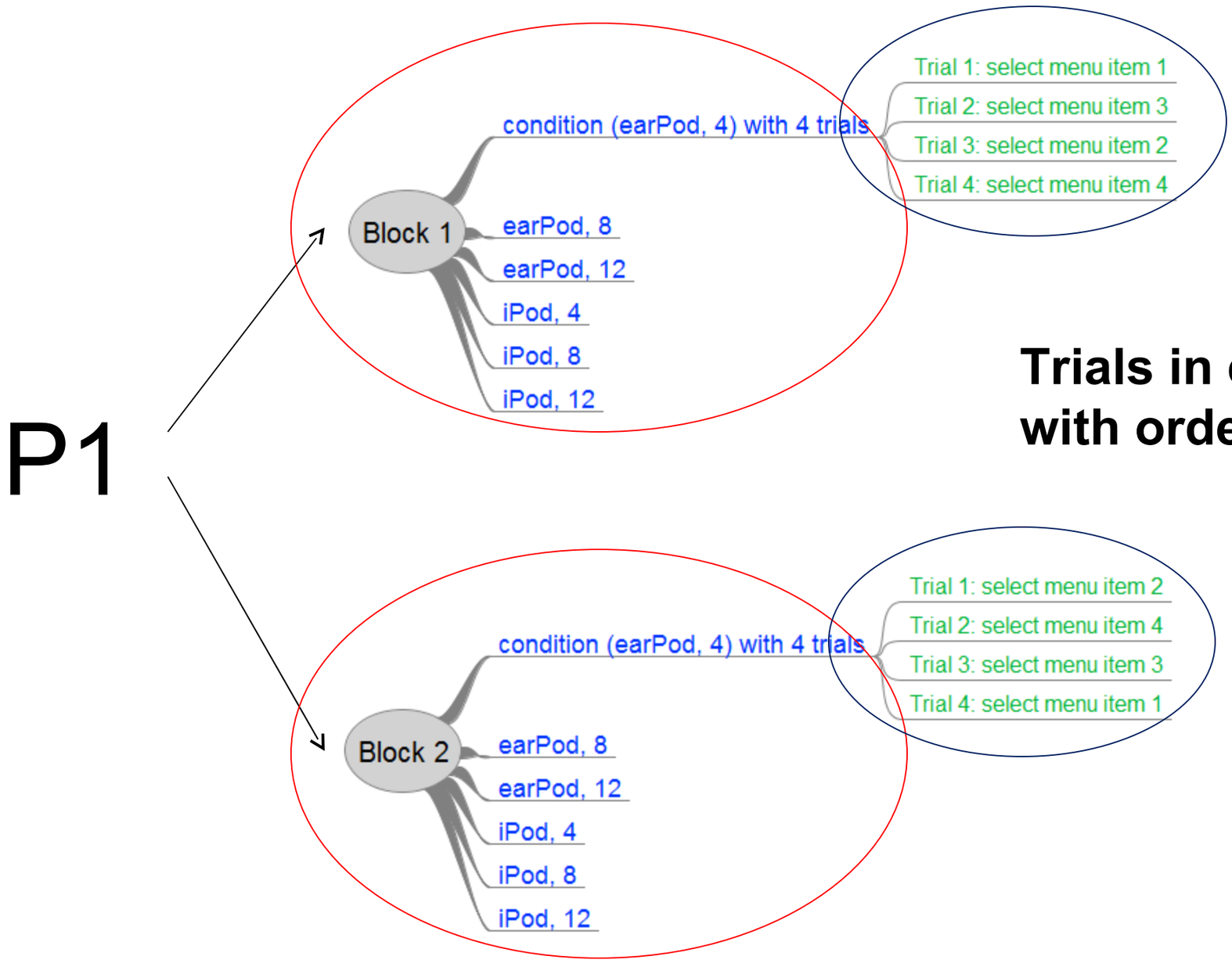
The 5 Step Approach to Experiment Design

- Define the research question
- Determine variables
- Arrange conditions
- **Decide blocks and trials**
- Set instruction and procedures

Definitions

- Trial
  - A single repetition of a single condition/cell
  - A number of trials are used to increase reliability
- Block*
  - An entire section of the experiment
  - Repeated to analyze learning

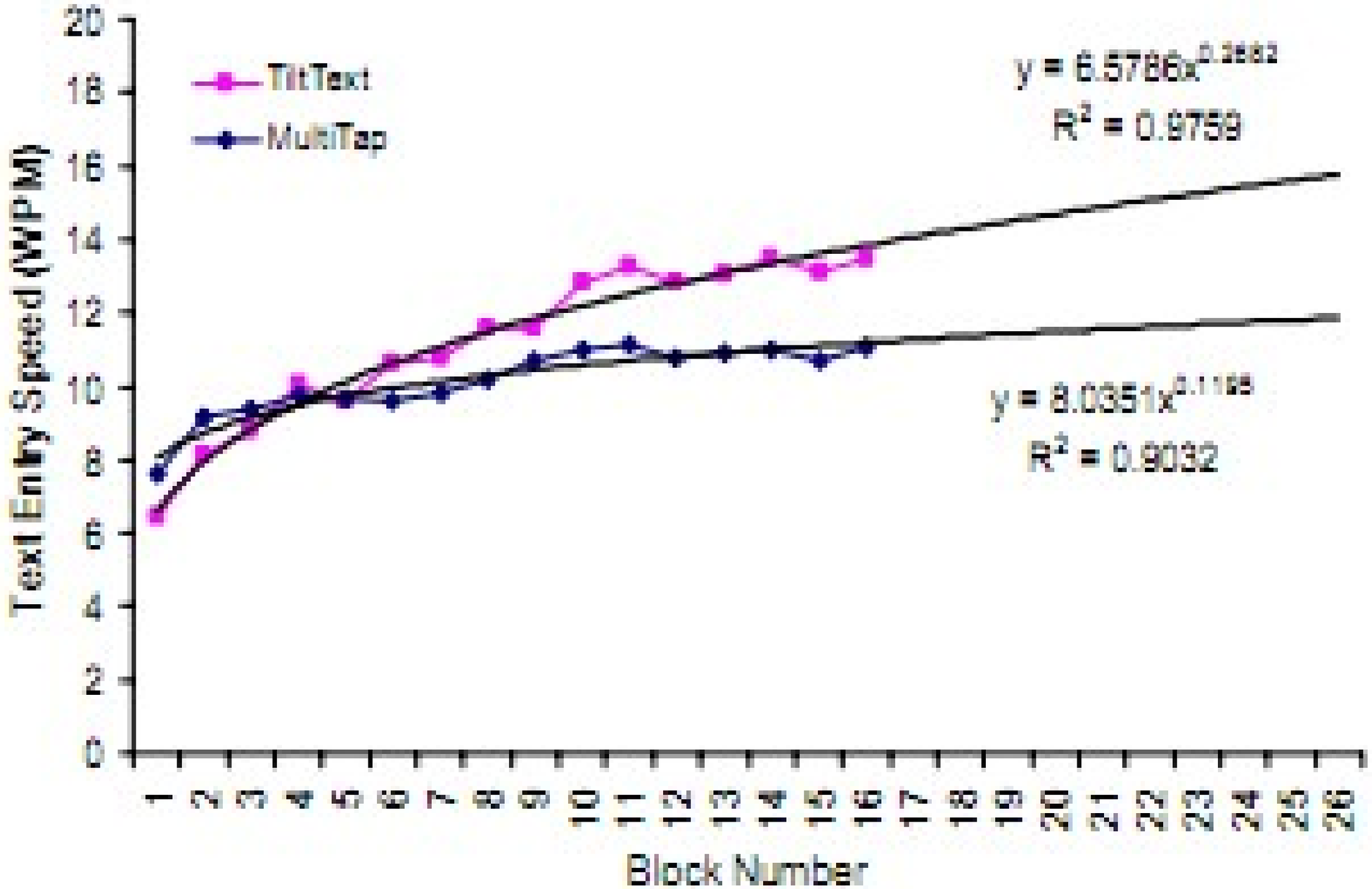
* Block has other definitions. This is a simplified definition for the purpose of this assignment.



Trials in each block: same content, with order randomized

Block: same arrangement, repeated

Block Indicates Learning



Adapted from the TiltText paper by Widgor & Balakrishnan

Determine Number of Blocks/ Repetitions

- Reasonable experiment duration
  - Time Constraint and Fatigue
  - Typically within 1 hour
    - However, minus pre- and post-experiment interviews, only left with 45 minutes
  - In some cases, up to 2 hours
- Enough data points for significant effects

Step 4: Determine Blocks and Trials

- Step 4.1: estimate the time for each trial (typically at least 3 trials per condition)
- Step 4.2: estimate the time for each block
- Step 4.3: balance the trials and blocks so that the main part of the experiment is within 45 minutes
- Step 4.4: combine with the condition arrangement

Exercise

Full experiment design: earPod vs. iPod

Independent variables

- Technique (2 levels: earPod vs. iPod)
- Usage scenario (2 levels: single vs multi-tasking)
- Menu breadth (3 levels: 4, 8, 12)
- Menu depth (2 levels: 1, 2)

Block = 3

Trials per condition = 4

Each trial takes roughly 10 seconds to finish

Question: how is the experiment arranged?

Question: how long will the experiment take?

The 5 Step Approach to Experiment Design

- Define the research question
- Determine variables
- Arrange conditions
- Decide blocks and trials
- **Set instruction and procedures**

Detailed Steps

Step 5.1: Recruit participants (determine target users and randomize)

Step 5.2: Consent form and pre-experiment questionnaire

Step 5.3: Instructions

Step 5.4: Practice trials

Step 5.5: Main experiment with breaks

Step 5.6: Post-experiment questionnaire and interview

Step 5.7: Debriefing

Conducting the Experiment

Before the experiment

- Have them read and sign the consent form
- Explain the goal of the experiment
  - In a way accessible to users
  - Be careful about the demand characteristic
    - Participants biased towards experimenter's hypothesis
- Answer questions

During the experiment

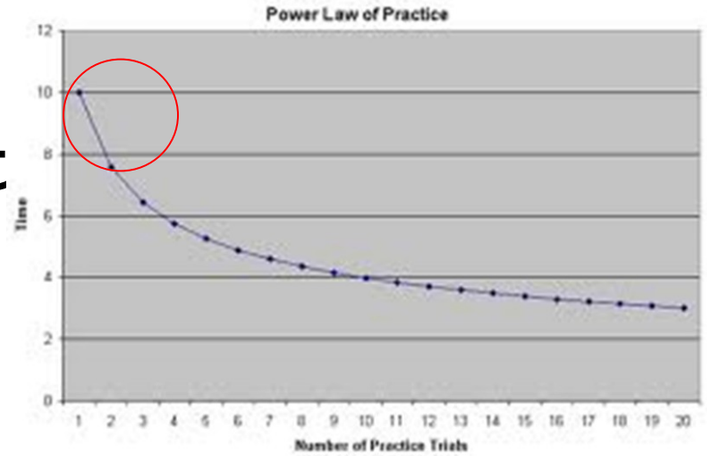
- Stay neutral
- Never indicate displeasure with users performance

After the experiment

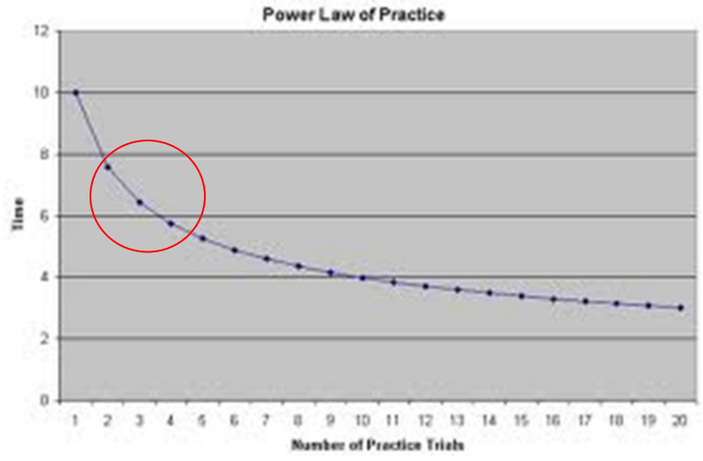
- Debrief users
  - Inform users about the goal of the experiment
- Answer any questions they have

The Importance of Practice Trials

- earPod
  - New technique, no one has seen it



- iPod
  - Existing technique, many people used or seen it



- **Question**: how do we control this?

Pilot Study and Protocols

Always pilot it first!

- Reveals unexpected problems
- Can't change experiment design after starting it

Always follow same steps – use a checklist

Get consent from subjects

Debrief subjects afterwards

Let's Review the Entire Process

1. Define the research question
2. Determine variables
3. Arrange conditions
4. Decide blocks and trials
5. Set instruction and procedures

Step 1: Define the Research Question

Define the research question has 4 sub-steps

Step 1.1 Start with a general question

Step 1.2 Define the target population

Step 1.3 Define task(s)

Step 1.4 Define measure(s)

Step 1.5 Define factor(s)

Step 2: Define Variables

Task Type & Factor → Independent variables

Measures → Dependent variables

Everything else → Control/Random Vars.

Spot and remove confounding variables

Step 3: Arranging Conditions for Within-Subject Design

3.1: List all Independent Variables and their levels

3.2: Decide counter-balancing strategy for each variable

3.3: Determine the minimum No. of participants

3.4: Arrange the overall design

3.5: Determine detailed arrangement for each participant

Step 4: Determine Blocks and Trials

- Step 4.1: estimate the time for each trial (typically at least 3 trials per condition)
- Step 4.2: estimate the time for each block
- Step 4.3: balance the trials and blocks so that the main part of the experiment is within 45 minutes
- Step 4.4: combine with the condition arrangement

Step 5: Set Introduction & Procedure

Step 5.1: Recruit participants (determine target users and randomize)

Step 5.2: Consent form and pre-experiment questionnaire

Step 5.3: Instructions

Step 5.4: Practice trials

Step 5.5: Main experiment with breaks

Step 5.6: Post-experiment questionnaire and interview

Step 5.7: Debriefing

Final Note

One of the challenges in experiment design is to decide what to test (independent variables) and what to measure (dependent variables). With the 5-Step approach, things become much easier, but you still have to make those tough decisions. When you face difficulties, it's always important to go back to your research question and ask what exactly you want to test in the experiment. This will help you to prioritize your choices.

End