

Outline

- Introduction
- Variables
- Design
 - Within-Subjects VS Between-Subjects
 - Counterbalancing
- Procedure
- Participants
- Questionnaires
- Longitudinal Studies

Methodology

- Learning to conduct and design an experiment is a skill required of all researchers in HCI
- Experiment design is the process of deciding participants, apparatus, tasks, order of tasks, procedures, variables, data collected, and so on
- Methodology is critical:

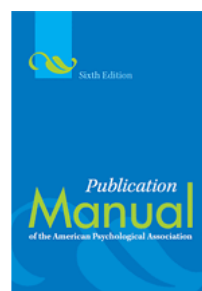
Science is method. Everything else is commentary.¹

- What methodology?
 - Don't just make it up because it seems reasonable
 - Follow standards for experiments with human participants (next slide)

¹ This quote from Allen Newell was cited and elaborated on by Stuart Card in an invited talk at the ACM's SIGCHI conference in Austin, Texas (May 10, 2012).

APA

- American Psychological Association (APA) is the pre-dominant organization promoting research in psychology – the improvement of research methods and conditions and the application of research findings (<http://www.apa.org/>)
- *Publication Manual of the APA*¹, first published in 1929, teaches about the writing process and, implicitly, about the methodology for experiments with human participants
- Recommended by major journals in HCI



¹ APA. (2010). *Publication manual of the American Psychological Association* (6th ed.). Washington, DC: APA.

Ethics Approval

- *Ethics approval* is a crucial step that precedes every HCI experiment
- HCI experiments involve humans, thus...

Researchers must respect the safety, welfare, and dignity of human participants in their research and treat them equally and fairly.¹

- Proposal submitted to ethics review committee
- Criteria for approval:
 - research methodology
 - risks or benefits
 - the right not to participate, to terminate participation, etc.
 - the right to anonymity and confidentiality

¹ <http://www.yorku.ca/research/students/index.html>

Getting Started With Experiment Design

- Transitioning from the creative work in formulating and prototyping ideas to experimental research is a challenge
1. Formulate a research questions:

Can a task be performed more quickly with my new interface than with an existing interface?

2. Begin with...

What are the experimental variables?

Properly formed research questions inherently identify experimental variables (can you spot the independent variable and the dependent variable in the question above?)

Independent Variable

- An *independent variable* (IV) is a circumstance or characteristic that is manipulated in an experiment to elicit a change in a human response while interacting with a computer.
- “Independent” because it is independent of participant behavior (i.e., there is nothing a participant can do to influence an independent variable)
- The terms *independent variable* and *factor* are synonymous

7

Test Conditions

- An independent variable (IV) must have at least two levels
- The levels, values, or settings for an IV are the *test conditions* aka *experimental conditions*
- Name both the factor (IV) and its levels (test conditions):

Factor (IV)	Levels (test conditions)
Device	mouse, trackball, joystick
Feedback mode	audio, tactile, none
Task	pointing, dragging
Visualization	2D, 3D, animated
Search interface	Google, custom

8

Test Conditions

- An experiment with 2 independent variables, each with m and n levels respectively, is called an
 - m x n factorial design
- Example: TouchTap experiment
 - Design: within-subjects
 - Independent Variables (factors):
 - Input Method in {TouchTap, widget-based technique};
 - Font Size in {1.75, 3.25, 4.75};
 - **2 x 3 within-subjects factorial design**
 - 6 test (experimental) conditions
- Experiments with >2 factors are also possible

9

Human Characteristics

- Human characteristics are *naturally occurring attributes*
- Examples:
 - Gender, age, height, weight, handedness, grip strength, finger width, visual acuity, personality trait, political viewpoint, first language, shoe size, etc.
- They are legitimate independent variables, but they cannot be “manipulated” in the usual sense
- Causal relationships are difficult to obtain due to unavoidable confounding variables

10

How Many IVs?

- An experiment must have at least one independent variable
- Possible to have 2, 3, or more IVs
- But the number of “effects” increases rapidly with the size of the experiment:

Independent Variables	Effects					Total
	Main	2-way	3-way	4-way	5-way	
1	1	-	-	-	-	1
2	2	1	-	-	-	3
3	3	3	1	-	-	7
4	4	6	3	1	-	14
5	5	10	6	3	1	25

- Advice: Keep it simple (1 or 2 IVs, 3 at the most)

11

Dependent Variable

- A *dependent variable* is a measured human behaviour
- “Dependent” because it depends on what the participant does
- Examples:
 - task completion time, speed, accuracy, error rate, throughput, target re-entries, task retries, presses of backspace, etc.
- Dependent variables must be clearly defined
 - Research must be reproducible!

12

Unique DVs

- Any observable, measurable behaviour is a legitimate dependent variable
 - So, feel free to “roll your own”
- Example: *negative facial expressions*¹
 - Application: user difficulty with mobile games
 - Events logged included frowns, head shaking
 - Counts used in ANOVA, etc.
 - Clearly defined → reproducible

¹ Duh, H. B.-L., Chen, V. H. H., & Tan, C. B. (2008). Playing different games on different phones: An empirical study on mobile gaming. *Proceedings of MobileHCI 2008*, 391-394, New York: ACM. 13

Data Collection

- Obviously, the data for dependent variables must be collected in some manner
- Ideally, engage the experiment software to log timestamps, key presses, button clicks, etc.
- Planning and *pilot testing* important
- Ensure conditions are identified, either in the filenames or in the data columns
- Better to have two logs:
 - **High-level**: directly reports the values of dependent variables.
 - **Low-level**: reports all salient events of the trials

High-Level Log

0_0_A_stats.tema

2	[#]presented	transcribed	presented	characters	transcribed	characters	input_time(sec)	pause_time(sec)	total_time(sec)
	wpm	msd	numBksp	numDelChars	total_error	cor_error	uncor_error		
3	Mary had a little lamb	Mary had a little lamb	22	22	15.228	0.0	15.267	14.184397163120567	0 2 8
	0.2666666666666666	0.2666666666666666	0.0						
4	suburbs are sprawling up everywhere	suburbs are sprawling up everywhere	35	35	11.287	0.0	11.308	29.768760520953307	0
	0 0 0.0 0.0 0.0								
5	pay off a mortgage for a house	Pay off a mortgage for a house	30	30	9.106	0.0	9.123	35.58093564682627	1 0 0
	0.03333333333333333	0.0	0.03333333333333333						
6	taking the train is usually faster	taking the train is usually faster	34	34	9.148	0.0	9.166	36.729339746392654	0
	0 0 0.0 0.0 0.0								
7	exceed the maximum speed limit	exceed the maximum speed limit	30	30	12.513	0.0	12.54	23.01606329417406	0 1 5
	0.14285714285714285	0.14285714285714285	0.0						
8	the laser printer is jammed	the laser printer is jammed	27	27	7.998	0.0	9.268	36.00900225056264	0 0 0
	0.0 0.0 0.0								
9	a big scratch on the tabletop	a big scratch on the tabletop	29	29	8.49	0.0	8.507	39.57597173144876	0 0 0
	0.0 0.0 0.0								
10	Mary had a little lamb	Mary a little lamb	22	18	13.056	0.0	13.075	13.786764705882353	4 2 7
	0.2413793103448276	0.13793103448275862						0.3793103448275862	
11	the music is better than it sounds	the music is better than it sounds	34	34	9.57	0.0	9.599	38.87147335423197	0
	0 0 0.0 0.0 0.0								
12	microscopes make small things look big	microscopes make small things look big	38	38	10.461	0.0	10.49		
	30.972182391740752	0 0 0	0.0 0.0 0.0						

Low-Level Log

0_0_A_events.tema

1	#Log opened: Fri Apr 12 09:16:35 CEST 2013
2	#just like it says on the can good
3	#1365751074947
4	0,<Entr>,pos@0
5	220,<Entr>,pos@1
6	#REJ,
7	#just like it says on the can good
8	#1365751188459
9	0,j,pos@0
0	452,jus,pos@0
1	1219,just,pos@0
2	1970,just,pos@0
3	1981,<Sp>,pos@4
4	2561,l,pos@5
5	2828,li,pos@5
6	3567,lilo,pos@5
7	7975,lilo,pos@5
8	7988,<Entr>,pos@9
9	8451,<Entr>,pos@10
0	#REJ,just lilo
1	#just like it says on the can good

Control Variable

- A *control variable* is a circumstance (not under investigation) that is kept constant while testing the effect of an independent variable
- More control means the experiment is less generalizable (i.e., less applicable to other people and other situations)
- Research question: Is there an effect of font color or background color on reading comprehension?
 - Independent variables: font color, background color
 - Dependent variable: comprehension test scores
 - Control variables
 - Font size (e.g., 12 point)
 - Font family (e.g., Times)
 - Ambient lighting (e.g., fluorescent, fixed intensity)
 - Etc.

17

Random Variable

- A *random variable* is a circumstance that is allowed to vary randomly
- More variability is introduced in the measures (that's bad!), but the results are more generalizable (that's good!)
- Research question: Does user stance affect performance while playing *Guitar Hero*?
 - Independent variable: stance (standing, sitting)
 - Dependent variable: score on songs
 - Random variables
 - Prior experience playing a real musical instrument
 - Prior experience playing *Guitar Hero*
 - Amount of coffee consumed prior to testing
 - Etc.

18

Control vs. Random Variables

- There is a trade-off which can be examined in terms of internal validity and external validity (see below)

Variable	Advantage	Disadvantage
Random	Improves external validity by using a variety of situations and people.	Compromises internal validity by introducing additional variability in the measured behaviours.
Control	Improves internal validity since variability due to a controlled circumstance is eliminated	Compromises external validity by limiting responses to specific situations and people.

- Remember:
 - Internal validity: The extent to which the effects observed are due to the test conditions
 - External validity: The extent to which results are generalizable to other *people* and other *situations*

19

Confounding Variable

- A *confounding variable* is a circumstance that varies systematically with an independent variable
- Research question: In an eye tracking application, is there an effect of “camera distance” on task completion time?
 - Independent variable: Camera distance (near, far)
 - Near camera (A): inexpensive camera mounted on eye glasses
 - Far camera (B): expensive camera mounted above system display
 - Dependent variable: task completion time
 - But, “camera” is a confounding variable: camera A for the near setup, camera B for the far setup
 - Are the effects due to camera distance or to some aspect of the different setups?

20

Design

Within-subjects VS Between-subjects

- Two ways to assign conditions to participants:
 - Within-subjects* → each participant is tested on each condition
 - Between-subjects* → each participant is tested on one condition only
 - Example: An IV with three test conditions (A, B, C):

Within-subjects

Participant	Test Condition		
1	A	B	C
2	A	B	C






Between-subjects

Participant	Test Condition
1	A
2	A
3	B
4	B
5	C
6	C

21

Within-subjects, Between-subjects

Pros and Cons

	Within-Subjects	Between-Subjects
Participants	Fewer 	More
Variation due to participants	Less 	More
Balance groups	No need 	Needed
Order effects	There may be	Not present 
Amount of trials per participant	High	Low 

22

Within-subjects, Between-subjects

- Sometimes...
 - A factor must be assigned within-subjects
 - Examples: input method, ...
 - A factor must be assigned between-subjects
 - Examples: gender, handedness
- With two factors, there are three possibilities:
 - both factors within-subjects
 - both factors between-subjects
 - one factor within-subjects + one factor between-subjects (this is a *mixed design*)

23

Order Effects, Counterbalancing

- Only relevant for within-subjects factors
- The issue: *order effects* (aka *learning effects*, *practice effects*, *fatigue effects*, *sequence effects*)
- Order effects avoided by *counterbalancing*:
 - Participants divided into groups
 - Test conditions are administered in a different order to each group
 - Possible counterbalancing schema: Latin square
 - Distinguishing property of a Latin square → each test condition occurs precisely once in each row and column (next slide)

24

Latin Squares

2 x 2

A	B
B	A

3 x 3

A	B	C
B	C	A
C	A	B

4 x 4

A	B	C	D
B	C	D	A
C	D	A	B
D	A	B	C

5 x 5

A	B	C	D	E
B	C	D	E	A
C	D	E	A	B
D	E	A	B	C
E	A	B	C	D

25

Balanced Latin Square

- With a balanced Latin square, each condition precedes and follows any other condition an equal number of times
- Only possible for even-orders
- Top row pattern: A, B, n , C, $n - 1$, D, $n - 2$, ...

4 x 4

A	B	D	C
B	C	A	D
C	D	B	A
D	A	C	B

6 x 6

A	B	F	C	E	D
B	C	A	D	F	E
C	D	B	E	A	F
D	E	C	F	B	A
E	F	D	A	C	B
F	A	E	B	D	C

26

Test Conditions with multiple factors

- The different test conditions can be different levels of the same factor.
- They can also be a tuple of levels for more factors.
- A Factorial Design with multiple factors
 - A: A1, A2, ..., An
 - B: B1, B2, ..., Bm
- leads to a number of possible test conditions given by:
 - $A \times B$ (cartesian product: size = $n \times m$)
- A subset of all the possible test conditions can be chosen

27

Example

- An experimenter seeks to determine if three editing methods (A, B, C) differ in the amount of time to do a common editing task:

Replace one 5-letter word with another, starting one line away.

- Conditions are assigned within-subjects
 - Method A: arrow keys, backspace, type
 - Method B: search and replace dialog
 - Method C: point and double click with the mouse, type
- Twelve participants are recruited and divided into three groups (4 participants/group)
- Methods administered using a 3×3 Latin Square

28

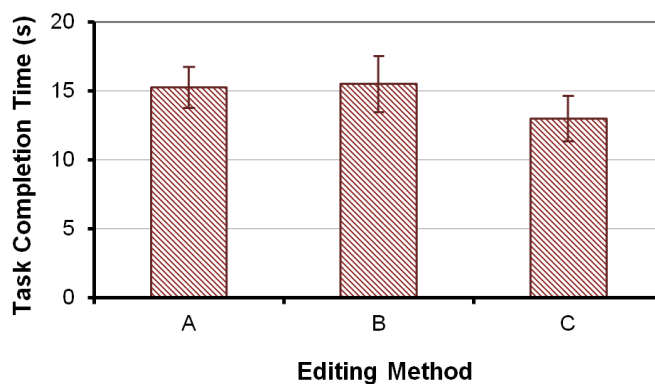
Results - Data

Participant	Test Condition			Group	Mean	SD
	A	B	C			
1	12.98	16.91	12.19	1 ABC	14.7	1.84
2	14.84	16.03	14.01			
3	16.74	15.15	15.19			
4	16.59	14.43	11.12			
5	18.37	13.16	10.72	2 BCA	14.6	2.46
6	15.17	13.09	12.83			
7	14.68	17.66	15.26			
8	16.01	17.04	11.14			
9	14.83	12.89	14.37	3 CAB	14.4	1.88
10	14.37	13.98	12.91			
11	14.40	19.12	11.59			
12	13.70	16.17	14.31			
Mean	15.2	15.5	13.0			
SD	1.48	2.01	1.63			

Group effect is small
 \therefore Counterbalancing worked!

29

Results - Chart



30

Example: TouchTap

- Design: within-subjects
- Independent Variables (factors):
 - Input Method in { TouchTap, widget-based technique };
 - Font Size in { 1.75, 3.25, 4.75 };
- **2 x 3 within-subjects factorial design**
- 6 test conditions

Method\Font	small	medium	large
TouchTap	T-s (A)	T-m (B)	T-l (C)
Widget-based	W-s (D)	W-m (E)	W-l (F)

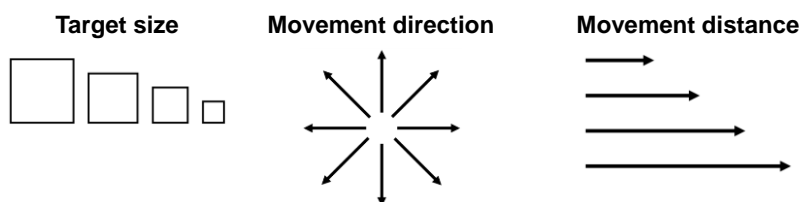
- We can counterbalance with a balanced latin square
- If necessary, to avoid confusing participants, we only change the input method once (see the paper)

31

Other Techniques

- Instead of using a Latin square, all orders ($n!$) can be used; 3 case →
- Conditions can be randomized
- Randomizing best if the tasks are brief and repeated often; examples (Fitts' law, see below)

A	B	C
A	C	B
B	C	A
B	A	C
C	A	B
C	B	A



32

Challenge (homework)

- Work in project groups
- Modify the table summarizing your literature
- Look at the main factorial experiment described in each of the 3 papers and add the following columns to the table:
 - Factorial design: Within-subjects / Between subjects
 - Participants number
 - Independent variables: report levels and factors
 - Counterbalancing schema
 - Dependent variables: for each of them also report unit of measure, gathering method (computer log, manual annotation, ...)

33

Experiment Task

- Recall the definition of an independent variable:
 - a circumstance or characteristic that is manipulated in an experiment to *elicit a change in a human response* while interacting with a computer
- The experiment task must “elicit a change”
- Qualities of a good task: *represent, discriminate*
 - Represent activities people do with the interface
 - Discriminate among the test conditions

34

Task Examples

- Usually the task is self-evident (follows directly from the research idea)
- Research idea → a new graphical method for entering equations in a spreadsheet
 - Experiment task → insert an equation using (a) the graphical method and (b) the conventional method
- Research idea → an auditory feedback technique for programming a GPS device
 - Experiment task → program a destination location into a GPS device using (a) the auditory feedback method and (b) the conventional method

35

Knowledge-based Tasks

- Most experiment tasks are *performance-based* or *skill-based* (e.g., inserting an equation, programming a destination location; see previous slide)
- Sometimes the task is *knowledge-based* (e.g., “Use an Internet search interface to find the birth date of Albert Einstein.”)
- In this case, participants become contaminated (in a sense) after the first run of task, since they have acquired the knowledge
- Experimentally, this poses problems (beware!)
- A creative approach is needed (e.g., for the other test condition, slightly change the task; “...of William Shakespeare”)

36

Procedure

- The *procedure* encompasses everything that occurs with participants
- The procedure includes the experiment task (obviously), but everything else as well...
 - Arriving, welcoming
 - Signing a consent form
 - Instructions given to participants about the experiment task (next slide)
 - Demonstration trials, practice trials
 - Rest breaks
 - Administering of a questionnaire or an interview

37

Instructions

- Very important (best to prepare in advance; write out)
- Often the goal in the experiment task is “to proceed as quickly and accurately as possible but at a pace that is comfortable”
- Other instructions are fine, as per the goal of the experiment or the nature of the tasks, but...
- Give the same instructions to all participants
- If a participant asks for clarification, do not change the instructions in a way that may cause the participant to behave differently from the other participants

38

Participants

- Researchers want experimental results to apply to people not actually tested – a population
- Population examples:
 - Computer-literate adults, teenagers, children, people with certain disabilities, left-handed people, engineers, musicians, etc.
- For results to apply generally to a population, the participants used in the experiment must be...
 - Members of the desired population
 - Selected at random from the population
- True random sampling is rarely done (consider the number and location of people in the population examples above)
- Some form of *convenience sampling* is typical

39

How Many Participants?

- Too few → experimental effects fail to achieve statistical significance
- Too many → statistical significance for effects of no practical value
- The correct number... (drum roll please)
 - Use the same number of participants as used in similar research¹
- Also consider the counterbalancing schema

¹ Martin, D. W. (2004). *Doing psychology experiments* (6th ed.). Pacific Grove, CA. Belmont, CA: Wadsworth.

40

Questionnaires

- Questionnaires are used in most HCI experiments
- Two purposes:
 - Collect information about the participants
 - Demographics (gender, age, first language, handedness, visual acuity, etc.)
 - Prior experience with interfaces or interaction techniques related to the research
 - Solicit feedback, comments, impressions, suggestions, etc., about participants' use of the experimental apparatus
- Questionnaires, as an adjunct to experimental research, are usually brief

41

Information Questions

- Questions constructed according to how the information will be used

Which browser do you use? _____ **Open-ended**

Which browser do you use?

☐ Mozilla *Firefox* ☐ Google *Chrome*

☐ Microsoft *IE* ☐ Other (_____)

Closed

Please indicate your age: _____ **Ratio-scale data**

Please indicate your age?

☐ < 20 ☐ 20-29 ☐ 30-39

☐ 40-49 ☐ 50-59 ☐ 60+

Ordinal-scale data

42

Participant Feedback

- Using NASA Task Load Index (TLX):

Frustration: I felt a high level of insecurity, discouragement, irritation, stress, or annoyance.

1	2	3	4	5	6	7
Strongly			Neutral			Strongly
disagree						agree

- ISO 9241-9:

Eye fatigue:

1	2	3	4	5	6	7
Very						Very
high						low

43

System Usability Scale

- 10 Questions in 5 (or 7) point Likert scale
- Odd items are positive, even are in negative form;
- Scores are in range 0-100:
 - Sum the score contributions from each item (0 to 4).
 - For items 1,3,5,7,and 9 the score contribution is the scale position minus 1.
 - For items 2,4,6,8 and 10, the contribution is 5 minus the scale position.
 - Multiply the sum of the scores by 2.5 to obtain the overall value of SU.

Brooke, J. (1996). SUS-A quick and dirty usability scale. Usability evaluation in industry, 189(194), 4-7.

44

Challenge

- Design SUS questionnaire for Tapping VS Gesture Writing experiment:
 - Search the questions in the literature
 - Prepare a spreadsheet with score calculation
 - Populate with participant's data (your colleague)
 - Report the score in the experiment spreadsheet
- Homework: Design the initial questionnaire with demographics and previous experience for your project work.

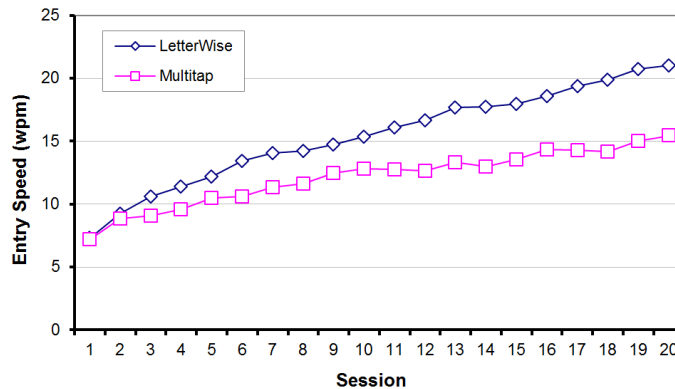
45

Longitudinal Studies

- Sometimes instead of “balancing out” learning effects, the research seeks to promote and investigate learning
- If so, a *longitudinal study* is conducted
- “Practice” is the IV
- Participants are practiced over a prolonged period of time
- Practice units: blocks, sessions, hours, days, etc.
- Example on next slide

46

Longitudinal Study – Results¹

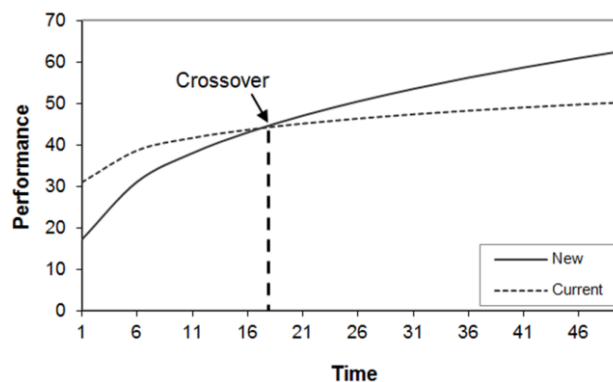


¹ MacKenzie, I. S., Kober, H., Smith, D., Jones, T., & Skepner, E. (2001). LetterWise: Prefix-based disambiguation for mobile text entry. *Proceedings of the ACM Symposium on User Interface Software and Technology - UIST 2001*, 111-120, New York: ACM.

47

The New vs. The Old

- Sometimes a new technique will initially perform poorly in comparison to an established technique
- A longitudinal study will determine if a crossover point occurs and, if so, after how much practice (see below)



48

Running the Experiment

- **Pilot testing** with 1 or 2 participants to: smooth the protocol, check the amount of time for each participant, etc.
- The experiment begins...
 - Participants sign the consent form fill questionnaire
 - Instructions are given
- The experimenter is the *public face* of the experiment:
 - Must portray him/her self as neutral
 - Should avoid to be overly attentive or conveying indifference

49

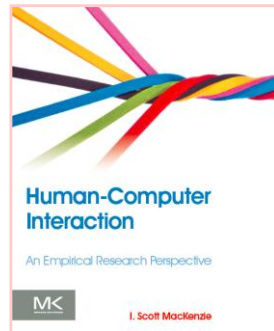


Connect to: <http://join.quizizz.com>

QUESTION TIME

50

Thank You



51