User Evaluation: Controlled Experiments

Human Computer Interaction

Luigi De Russis, Fulvio Corno

Academic Year 2021/2022

Politecnico di Torino 1859

e-Lite

Involving Users: Experimental Methods (recap)

Usability/User Testing

- "Let's find someone to use our app, so that we will get some feedback on how to improve it."
- anecdotal, mostly
- observation-driven

Controlled Experiments

- "We want to verify if users of our app perform task X faster/…/with fewer errors than our competitor's app."
- scientific
- hypothesis-driven

Overview

- Controlled evaluation of **specific** aspects of interactive behavior
  - typically in lab
- The evaluator chooses a **hypothesis** to be tested
  - most appropriately, a null hypothesis to be confuted
- Various experimental **conditions** are considered
  - which differ only in the value of some controlled variables
- Three main steps: *plan, run*, and analyze*

Experimental Design: Planning the Study

1. Choose what you want to study, which **narrow and testable question** you want to answer
2. Choose the **hypothesis** (with variables and measures)
3. Select your **participants**
4. Decide the **experimental method** that you will use
5. Write the **task(s)** you will give participants to (dis-)prove your hypothesis
    - along with the experiment procedure
6. Decide which **statistical tests** you are going to use to analyze the results

Experimental Factors

- Hypothesis
  - the prediction of the outcome of the study, what you would like to demonstrate
  - framed in terms of *variables*
  - in the form of a **null hypothesis**, to be disproved
- Variables
  - things to manipulate and measure, to test the hypothesis
- Subjects (participants)
  - representative, sufficient sample
  - sample size: <u>at least</u> double the number suggested by Nielsen for usability tests
  - vital to the success of any experiment

Variables

Independent Variable (IV)

- Elements of the experiment manipulated or controlled to produce different conditions for comparison
  - e.g., interface style, number of menu items, icon design, …
- Each of these can have different values, called *levels*
- One or more. Also called *factors*

Dependent Variable (DV)

- Characteristics measured in the experiment
  - their values are "dependent" on the changes made to the IV
  - e.g., time taken, number of errors, …
- for usability testing, they were the "measures"

Variables: A Very Simple Example

We want to verify if users of our app perform a task faster/…/with fewer errors than our competitor's app

- "our app… than our competitor's app" -> IV? DV?
- "faster/…/with fewer errors" -> IV? DV?

Variables: Example

We want to test whether selection speed in a menu improves as the number of menu items decreases

Independent Variable (IV)

- It is/They are…
- Each IV has … levels

Dependent Variable (DV)

- It is/They are…

Variables: Example

We want to test whether selection speed in a menu improves as the number of menu items decreases

Independent Variable (IV)

- IV: number of menu items
- If we consider menu items with 3, 5, and 7 items
  -> 3 levels

Dependent Variable (DV)

- Speed of the menu item selection (sec)

Independent Variables and Experimental Conditions

- Experimental condition: e.g., task execution during the experiment
- Each level of an independent variable requires one experimental condition to test
  - 3 menus with 3, 5, and 7 items -> 3 experimental conditions
- More complex experiments may have more than one IV, each with its own levels
  - experimental conditions should account for all combinations of levels

Independent Variables and Experimental Conditions

- Example
- We want to test whether selection speed in a menu improves as the number of menu items decreases AND text or icons are used as labels
  - IVs?
  - Levels?

Independent Variables and Experimental Conditions

- Example
- We want to test whether selection speed in a menu improves as the number of menu items decreases AND text or icons are used as labels
  - 2 IVs:
    1. number of menu items
       - three levels (as before)
    2. label type
       - two levels (text vs. icon)
- How many conditions?

Independent Variables and Experimental Conditions

- Example: we want to test whether selection speed in a menu improves as the number of menu items decreases AND text or icons are used as labels
  - 2 IVs: 1) number of menu items, 2) text vs. icon used in the menu
  - 1) has three levels (as before) and 2) 2 levels
- How many conditions?
  - 6, 3x2
  - 3 levels for the first IV, 2 for the second IV

3-items menu		5-items menu		7-items menu	
textual labels	textual labels + icons	textual labels	textual labels + icons	textual labels	textual labels + icons

Independent Variables: How Many?

- Complex experiments may have multiple IVs
  - is there an upper limit?
- Let's have a look at the *effects* among the variables
  - an experiment with 1 IV includes a main effect on the DVs
  - one with 2 IVs includes 2 main effects and 1 interaction effect (2-way)
  - one with 3 IVs includes 3 main effects and 4 interaction effects (three 2-way and a 3-way)
  - one with 4 IVs includes 14 effects, etc.
    - too many effects, too many variables!
- A good experiment design is one that <u>limits</u> the number of IVs to 1 or 2, three at most!

Other Types of Variables

- Control
  - variables that may influence a dependent variable, but they are not under investigation, can be *controlled*
    - always fixed at a nominal setting during the experiment
  - e.g., display size, mouse cursor speed, chair height, smartphone type, …
- Random
  - instead of trying to control everything, we can allow some variables to vary randomly
  - typically, they pertain to characteristics of participants, e.g., gender, height, hand size, …

Other Types of Variables

- Confounding
  - any circumstance or condition that changes systematically with an IV
  - problematic!
    - is the effect observed due to the IV or the confounding variable?
  - e.g., if you use two different cameras to track a person's eyes in different conditions (near vs. far), the different characteristics of the 2 cameras are the confounding variables

Hypothesis

- Prediction of the study outcome, framed in terms of IVs and DVs
  - a variation in the independent variable will cause a difference in the dependent variable
- This is done by **disproving** (rejecting) the null hypothesis
  - it states that there is no difference in the dependent variable between the levels of the independent variable
- And accepting the alternative hypothesis

Hypothesis

- The difference is evaluated **statistically**
  - some statistical measures produce values that can be compared with various levels of significance
  - if a result is *significant*, at a given level of certainty, the measured differences would not have occurred by chance
    - that is, that the null hypothesis is incorrect

Experimental Methods

- Between-subjects
  - each participant performs under only one condition
  - no transfer of learning
  - more users required; groups must be balanced
  - user variation can bias results
- Within-subjects
  - each participant performs experiment under each condition
  - transfer of learning possible
  - less costly and less likely to suffer from user variation
  - also called *repeated measures*
- When more than one IV is present, it is possible to devise a *mixed design*
  - one IV is placed between-subjects, the other within-in

Within- or Between-Subjects?

- Important trade-offs:
  - a **within-subject design** requires <u>less</u> participants
  - it also exhibits the <u>same</u> participants' predispositions across the different conditions
  - <u>no</u> need to balance groups of participants!
  - however, *transfer of learning* is possible (and not desired)
    - e.g., participants may perform better on the second condition because they benefitted from practice with the first one
  - *fatigue* may also be an issue
- **Counterbalancing** help minimize practice effects
  - divide participants into groups and administer the conditions in a different order for each group

Counterbalancing

- Typically, you counterbalance with a (balanced) **Latin Square**
  - a *nxn* table filled with n different symbols positioned such that each symbol occurs exactly one in each row and each column
    - *n* are levels, typically
- In this case, the number of levels of the IV must divide equally into the number of participants
  - e.g., 1 IV with 3 levels, 12 participants

A	B
B	A

A	B	C
B	C	A
C	A	B

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

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

Counterbalancing: Questions

- Why

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

and not

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

?

- Do not we have the same problem here

A	B	C
B	C	A
C	A	B

?

Tasks and Procedure

- When participants are given a test condition, they are asked to do a task while their "performance" is measured
- A good task should *represent* and *discriminate*
  - representative of the activities people will do with the interface
  - discriminate the test conditions, i.e., to further highlight the different effects between conditions
- Procedure
  - the list of tasks, instructions, demonstrations given to participants
  - any questionnaire
  - …

Statistical Measures

- Disclaimer: before applying *any* statistical tests, you ***must always look*** at data
  - it can expose *outliers*, e.g., a participant took 3 times as long as everyone else to do a task, and you know that that participant had been suffering from a severe flu the day of the experiment
  - we are not going deep on statistics, as this is beyond the scope of this course
- The choice of statistical analysis depends on
  - the type of data
  - the information required
    - is there a difference? how big is it? how is the estimate?
  - the data distribution

Types of Data

- Nominal
  - categorical data
  - arbitrary assign a code to mutually exclusive attributes or categories
  - e.g., car license plate numbers, codes for postal zone, gender, …
- Ordinal
  - provide an order or ranking to an attribute
  - e.g., first choice, second choice, third choice

Types of Data

- Interval
  - data with equal distances between adjacent values
  - no absolute zero
  - e.g., Celsius temperature scale
  - can be continuous or discrete
- Ratio
  - the most sophisticated of the four types
  - have an absolute zero
  - e.g., time, all the physical measurements, age, count, …
  - can be continuous or discrete

Types of Data and Related Statistical Tests

- Non-parametric tests
  - can be applied to any scale of data
    - limited use for ratio data
  - "distribution free"

Types of Data	Appropriate Statistical Tests
Nominal	Non-parametric Tests
Ordinal	
Interval	Parametric Tests Non-parametric Tests
Ratio	

- Parametric tests
  - assume data from a probability distribution
    - e.g., normal or *t*-distribution
  - more *powerful* than non-parametric tests
    - given the same set of data, a parametric test might detect a difference that the non-parametric test would miss

Commonly Used Parametric Tests in HCI

Experiment Design	Independent Variables	Levels for each IV	Type of Test
Between-subjects	1	2	Independent samples t-test
	1	3 or more	One-way ANOVA
	2 or more	2 or more	Factorial ANOVA
Within-subjects	1	2	Paired-samples *t*-test
	1	3 or more	Repeated measures ANOVA
	2 or more	2 or more	Repeated measures ANOVA
Mixed design	2 or more	2 or more	Split-plot ANOVA

When assumptions are not met, the independent samples *t*-test can be "replaced" by the Mann-Whitney U test, the Wilcoxon signed ranks test can be used instead of the paired-samples *t*-test, etc.

Pearson's Chi-Square Test

- It is a significance test used to analyze frequency count among categories
- One of the most used non-parametric test in HCI (for A/B Testing, mainly)
  - it is used with categorical data, to determine whether there is any relationship in your categories
  - i.e., to compare sets of rates (e.g., "% occurrences") to tell whether the percentage differences are statistically significant
    - or happened by change
- It makes two assumptions:
  - data points in the categories must be independent from each others
    - e.g., each participant can only contribute in one category
  - it does not work well with small sample size (<20)

Chi-Square Test: Example

- I toss a coin 20 times and I have "head" for 13 times (and "tail" for 7). I am expecting to have 10 times "head" and 10 "tail", instead.
  - *null hypothesis*: the behavior of the coin does not differ significantly from a "normal" coin
  - *alternative hypothesis*: the behavior of the coin differs significantly from a "normal" coin
- We are going to apply the Chi-square test
  - we would like to reject the null hypothesis
  - and accept the alternative hypothesis

Chi-Square Test: Example

Our experiment

Head	Tail
13	7

Expectation

Head	Tail
10	10

Chi-Square Test: Process

1. Calculate the test statistics, χ^2, a normalized sum of squared deviations between observed and theoretical frequencies
   - $\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$
   - where O_i is the i-th observation and E_i is the expected (theoretical) count of type i

- Coin example:
  - $\chi^2 = \frac{(13-10)^2}{10} + \frac{(7-10)^2}{10} = 1.8$

Chi-Square Test: Process

2. Determine the degrees of freedom, df, of that statistic:
   - the numbers we can freely modify without having other constraints from the others

Head	Tail	
13	7	20

   - here, we have only 1 cell/value that we can freely modify (given the fixed total of 20 trials)
   - $df = 1$

Chi-Square Test: Process

Head	Tail
13	7

2. Determine the degrees of freedom, df, of that statistic:
   - Goodness of fit, $df = (Cols - 1)$
     - how much an entity is a good fit for the general population of those entities
   - Test of independence, $df = (Rows - 1) \times (Cols - 1)$
     - two variables
     - Rows corresponds to number of levels in one variable, and Cols corresponds to number of levels in the second variable

- Coin example: goodness of fit
  - "is our coin part of the population of 'normal' coins?"
  - $df = (2 - 1) = 1$

Chi-Square Test: Process

3. Look for the level of confidence (p-value) related to the χ^2 result (1.8) and df (1) in a Probability Table:

df	0.995	0.99	0.975	0.95	0.90	0.10	0.05	0.025	0.01	0.005
1	---	---	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	9.236	11.070	12.833	15.086	16.750

from https://people.richland.edu/james/lecture/m170/tbl-chi.html

- Coin example:
  - first row, $0.10 < p < 0.25$ ($p \approx 0.20$)

Chi-Square Test: Process

4. **Sustain** or **reject** the null hypothesis
   - we usually reject the null hypothesis at $p < 0.05$ or $p < 0.01$
   - i.e., we are confident that 95% or 99% of the time the test result correctly applies to the entire population

- Coin example:
  - we fail to reject the null hypothesis!
  - so, we cannot say that our coin is "unfair"...
- In the end… is the null hypothesis true?
  - we do not know, but we cannot reject it!
    - the evidence we have is insufficient for rejecting it

References

- Alan Dix, Janet Finlay, Gregory Abowd, Russell Beale, Human Computer Interaction, 3rd Edition
  - Chapter 9: Evaluation Techniques
- I. Scott MacKenzie, Human-Computer Interaction – An Empirical Research Perspective
  - Chapter 5: Designing HCI Experiments
- Jonathan Lazar, Jinjuan Heidi Fend, Harry Hochheiser, Research Methods in Human-Computer Interaction, 1st Edition
  - Chapter 4: Statistical Analysis, page 73

License