CPSC 544:
Experiments III

© 2017 Joanna McGrenere
includes slides from Leila Aflatoony, Karon MacLean and Jessica Dawson
TODAY

• Admin [5 min]
• Experiments III lecture [1h 15min]
  • Running worksheet

• Type I and Type II error (*still* unfinished from last lecture)
**Project Questions?**

- **Evaluation goals**
  - Sally and I read and discussed yesterday
  - want brief follow-up with each group after class

- **Test-1 presentations + report next Tuesday**

**Winding down...**

- One reading for next Thursday
- Then guest lecture from UILO
- Then working class – LAST class
LEARNING GOALS

• example: experiment and ANOVA reported in the literature
  – What are the motivations for adaptive highlighting and ephemeral adaptation?
  – How is an experiment reported?
  – Inferential vs. descriptive statistics?
  – What is the value of pilot testing?
  – How are hypotheses tested?
  → you will be writing up your project experiment

• types of validity
  – what are the different forms of validity?
  – how are they related, if at all?
  – what are examples of each form of validity?
CASE STUDY: EPHEMERAL ADAPTATION

FIRST, SOME BACKGROUND
MOTIVATION...
GUIs: Increasing in Size/Complexity

For many users

Frustration
Decreased performance

How can a personalized interface mitigate the complexity?
How?

• Adaptable
• Adaptive
• Mixed-initiative
ADAPTABLE (CUSTOMIZABLE)
Adaptive Menu

Full Menu

MSWord Smart Menus
MULTIPLE: WORD PERSONAL

[McGrenere and Moore, GI 2002; McGrenere, Baecker, and Booth CHI 2002]
FIELD EXPERIMENT

• experiment: A, B, A design
• 20 participants
  – 10 feature-keen
  – 10 feature-shy

Word Personal (4 weeks)

Word 2000 Adaptive
FIELD EXPERIMENT RESULTS

Satisfaction

Q1: Word 2000
Q2 – Q6: Word Personal
Q7: Word 2000

Feature-shy
Feature-keen

p<.05
Feature-shy’s satisfaction and sense of control increased, feature-keen’s remained flat

Majority of all users preferred Word Personal

But were they more efficient with Word Personal?
EFFICIENCY: ADAPTABLE VS ADAPTIVE VS STATIC

Traditional menu

Static split menu

Most frequent items

[Findlater and McGrenere, CHI 2004]
LAB EXPERIMENT

1. **static**: most frequent items (*designed to be optimal*)
2. **adaptive**: algorithm using *recency* and *frequency*
3. **adaptable**: simple user-controlled mechanism

27 subjects, within-subjects design
Users need to experience the (potential) value of a personalized interface before personalizing
Majority preferred adaptable

Optimal performance can be reached with an easy to customize split menu

How can we nudge the user?

Can we build a mixed-initiative system?

(Yes! But no time to tell you about it today)
Are there designs that can improve the overall benefits (mitigate costs) of adaptive personalization?
Spatial  Inconsistent results  Graphical  Lack of evaluation  Temporal  Underexplored

[Gajos et al., 2006]
Ephemeral Adaptation

**Approach**
- Abrupt onset of predicted items
- Gradual onset of non-predicted items

**Design Benefits**
- Temporary adaptive support
- Maintains spatial consistency
- Based on literature in visual attention

[Findlater, Moffatt, McGrenere, and Dawson, CHI 2009]
Does ephemeral adaptation improve performance and user satisfaction?
Comparative Experiment (Study 2)

24 participants

Menu selection task

3 conditions (within-subjects)

Ephemeral  Color highlighting  Control (static)
Results
(p < .05)

Ephemeral
Fastest
Preferred

Color highlighting
Preferred

Control (static)

WHAT IS EPHEMERAL ADAPTATION?

• an adaptive method of highlighting menu items that reduces visual search time while maintaining spatial consistency
HOW IS AN EXPERIMENT DESIGN REPORTED?

• how easy/difficult was this paper to read?

• what were the elements that made it
  • easy?
  • difficult?
VALUE OF PILOTING AND 2 STUDIES

• what was the benefit of piloting and having two separate studies (study 1 and study 2)?

  (i.e., why not just do one BIG study???)
PILOTING GOALS

• Determine reasonable onset delays (250, 500, 1000ms)
• Get early participant feedback
STUDY 1 GOALS

- Determine if ephemeral adaption improves performance over static menus
- Explore how onset delay impacts performance
STUDY 2 GOALS

• To compare the best onset delay from Study 1 (long-onset) to adaptive highlighting

• To compare adaptive highlighting to a control condition
EXPERIMENT DESIGNS FOR STUDY 1 AND STUDY 2?

• experimental design language: repeated measures, ANOVA, one-way/two-way, between-subjects, within subjects, mixed design, factorial design, latin square

• Study 1:

• Study 2:
STUDY 1 COMPONENTS OF THE EXPERIMENT DESIGN

• Independent Variables:
  ● Menu (Control, Short-Onset, Long-Onset)
  ● Prediction Accuracy (low 50%, high 79%)

• Dependent Variables:
  ● Selection Time (median)
  ● Error Rate (counts)
  ● Subjective Satisfaction Responses (Likert Scale)
• Mixed design – Each participant saw only one prediction accuracy, but all menu types
  – Why?
• Fully counterbalanced presentation order of menu – each possible ordering seen the same number of times
  – Why?
• A 3-way ANOVA was used?
  – Why?
COUNTERBALANCING

• Why? getting used to the interface, getting tired, getting bored

• Methods:
  – Full factorial – Test every order equally, good for smaller experiments (not many factor levels)
  – Latin square – Test a subset of orders (judiciously chosen), best for larger experiments
  – Randomized – Good compromise for extremely large experiments
FOCUSING ON STUDY 1...
HYPOTHESES

**Performance**

**H1.1:** For high accuracy, at least one Short or Long-Onset condition will perform better than Control.

**H1.2:** For low accuracy, both Long-Onset and Short-Onset will be (perform) no worse than Control.

**Preference**

**H2.1:** For high accuracy: at least one of Long-Onset or Short-Onset will be preferred to Control.

**H2.2:** For low accuracy, Control will not be preferred to Short or Long-Onset conditions.
PICKING APART A RESULTS SECTION

• what do all the numbers and symbols mean?
  • Why do these matter to readers?

• descriptive vs. inferential statistics
  • Which are which?

• F, alpha level, p value, effect size (i.e. eta squared), confidence interval
REPORTING DESCRIPTIVE STATISTICS

- Describes the data without directly inferring any conclusions (do first!)
- Includes means, medians, deviations, etc.

Figure 2. Average selection time per trial for Study 1 (N = 23). Error bars show 95% confidence intervals (CI).
What counts as an inferential statistic?
REPORTING RESULTS: H1

Reporting of inferential statistics for H1:

- Omnibus ANOVA, showed sig. (p < 0.05) effect for **menu type** ($F_{2,22} = 3.80$, $p < 0.05$, $\eta^2 = 0.257$)
  - Suggests **menu type** had an impact on performance, but which one was best?

- Sig. Interaction for **accuracy** and **menu type** ($F_{2,22}=3.73$, $p < 0.05$, $\eta^2 = 0.253$)
  - Suggests the impact of **accuracy** on performance depends upon **menu type**, but how?
What do the symbols mean?

Note statistics summarized as:

\[(F_{2,22} = 3.80, \ p < 0.05, \ \eta^2 = 0.257)\]

- 2 = Condition DOF = var levels - 1
- 22 = Participants DOF = participants - 1
- Alpha level of 0.05 denotes significance
- Eta squared measures effect size, roughly how much of variance attributed to condition differences, > 0.14 large
REPORTING RESULTS: H2

- Rates a qualitative aspect (preference) on a quantitative scale (1 to 7)
- Why a Friedman test and not an ANOVA? What test was used for pairwise comparisons?

![Graph showing overall satisfaction ratings for Study 1 (N=23). Higher values indicate higher satisfaction. Error bars show 95% CI.](image)
TRENDS, QUOTES, AVERAGES

- 10 out of 11 high accuracy participants preferred one of the adaptive conditions
- 9 out of 12 low accuracy participants preferred one of the adaptive conditions
- For high accuracy preference skewed towards long onset (7 versus 3)

• What can we conclude from this?
RESULTS BY HYPOTHESES

**H1.1:** For high accuracy, at least one Short or Long-Onset condition will perform better than Control
*Supported* – Long-Onset faster than Control

**H1.2:** For low accuracy, both Long-Onset and Short-Onset will be (perform) no worse than Control.
*Supported* – no difference for speed in low accuracy condition

**H2.1:** For high accuracy: at least one of Long-Onset or Short-Onset will be preferred to Control.
*Somewhat supported* - users seemed to prefer ephemeral but more tests needed

**H2.2:** For low accuracy, Control will not be preferred to Short or Long-Onset conditions
*Somewhat supported* - not disproved, but needs more study
CONCLUSIONS

• Ephemeral Adaption may improve menu selection performance over static menus
• No data to suggest that less accurate predictions degrade performance more than static menus
• Participants may prefer ephemeral adaption to static menus
LEAVE YOU TO WALK
THOUGH ON YOUR OWN THE
SAME FOR STUDY 2...
IMPLICATIONS FOR DESIGN

- Beyond menus…
Moguls and Arab States Are Big Donors to Clinton Charity

BY PETER BAKER and CHARLIE SAVAGE 20 minutes ago

Lifting a cloak of secrecy, former President Bill Clinton disclosed the names of more than 200,000 donors to his foundation as part of a deal with the Obama transition team.

Bush Weighs ‘Orderly’ Bankruptcy for Automakers

BY DAVID STOUT and MICHELINE MAYNARD 3:20 PM ET

A Bush spokeswoman said that no decision had been made but that a soft landing through a bankruptcy is an option.

Helene Cooper

ON THE WHITE HOUSE

The Direct Approach

Obama aides are planning a

Wall Street Slides as Oil Falls Below $40 a Barrel

BY JACK HEALY 50 minutes ago
Moguls and Arab States Are Big Donors to Clinton Charity

BY PETER BAKER and CHARLIE SAVAGE 20 minutes ago

Lifting a cloak of secrecy, former President Bill Clinton disclosed the names of more than 200,000 donors to his foundation as part of a deal with the Obama transition team.

Post a Comment | Read (130)

Bush Weighs ‘Orderly’ Bankruptcy for Automakers

BY DAVID STOUT and MICHELINE MAYNARD 3:20 PM ET

A Bush spokeswoman said that no decision had been made but that a soft landing through a bankruptcy is an option.

Helene Cooper
ON THE WHITE HOUSE

The Direct Approach

Obama aides are planning a

Wall Street Slides as Oil Falls Below $40 a Barrel

BY JACK HEALY 50 minutes ago

A New Chapter for Baghdad Book Market

BY ERIC OWLES 2:32 PM ET

Resurrecting Mutanabi Street, an intellectual center of Baghdad, has been a pet project for the Iraqi leadership.

Iraqi Arrests Extend Beyond Key Ministry

BY CAMPBELL ROBERTSON and TAREQ MAHER 2:05 PM ET

The Iraqi Ministry of the Interior confirmed that 23 of its officials had been arrested and also said the arrests extended into other security ministries.

Shoe-Hurling Journalist Said to Ask for Pardon

BY 12:04 PM ET

The Latest on Deal Makers and Deal Breakers

nytimes.com/dealbook
IMPORTANT/NOTEWORTHY FEATURES OF THE REPORT

• image/diagram of system in use/being examined, with a descriptive caption
• related work section divided into subsections according to topic area
• experimental methodology section
  – participants, conditions, design, procedure, task (incl. image of task being performed, w/ caption), measures, apparatus, hypotheses
• results: quantitative (F-stats, p-values, effect size) and qualitative (subjective response), means/SDS, bar/line charts w/ confidence intervals, validation of hypotheses
• limitations
• discussion - relating to other research, generalizability
• conclusion and future work
• references
THREATS TO VALIDITY
Threats to Validity

How do you make sure your data is good? And that your conclusions hold?

Construct validity
- Are we measuring what we think we are measuring?
- E.g., create a questionnaire to assess early “adopter-ness”, but in fact it assesses financial ability to buy new technology instead.

Internal validity
- Is there a causal relation between independent & dependent variables?
- E.g., nuisance variable causing the change in the dependent variable.
- E.g., Hawthorne effect – subjects change their behavior because they know they are being studied.
THREATS TO VALIDITY (CONT’D)

statistical validity
  – could the results be a fluke?
  – e.g., were the statistical tests used appropriate? (e.g., many tests assume a normal distribution)

external validity
  – do the results generalize?
  – e.g., sample not representative of true population
  – e.g., insufficient description of experiment protocol

ecological/face validity (form of external validity)
  – e.g., tasks in experiment not representative of real tasks
Left for you to ponder

• you should be able to identify at least 2 specific threats to validity for the ephemeral study covered today
This concludes Experiments topic

Highly recommend, if you will be designing and running an experiment for your graduate research:

– **EPSE 592** Experimental Designs and Analysis in Educational Research (*register early!*)


ON DECK...

• Test-1 presentations + report next Tuesday