CPSC 444: ADVANCED METHODS IN HUMAN-COMPUTER INTERACTION

Lecture 8 – Experiments IV

Joanna McGrenere
ADMINISTRIVIA

• MSIII Blogs #4 and #5 due Tuesday, March 6 – 9am
• Marking of prototype will depend heavily on what is ready/presented in the blog
  • so majority of prototype should be done for Blog #5 documentation
  • BUT demo isn’t until workshop the following Friday
    • minor improvements or finishing steps between Tuesday/Friday OK

• No prep assignment for next week
  • focus on your project
MSIII: Workplan Recommendation

To allow yourself adequate time for all Parts A, B & C, we suggest the following workplan. You have approximately 3.5 weeks (including reading break):

• **By the end of the 1st week:** Complete Part A (Steps 1 & 2, start Step 3).

• **By the end of the 2nd week:** Complete Step 3, and have Step 4 underway. Complete Step 5, and have Step 6 underway.
  – Decide which team members will be working on refining the experiment and which will be focused on prototype implementation.

• **By halfway through the 3rd week:** Close to completing Step 4, and have significant progress on Step 6.

  Use workshop to get feedback on the experiment design and the plans for the prototype. There should be a clear plan about which team members will be completing which steps in the deliverable. There will be less than a week left in this stage.
EXPERIMENTS IV - 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 two experiments in 444 (individual and team)

• 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

Adaptive Menu

Full Menu

MSWord Smart Menus
WORD PERSONAL
(MULTIPLE INTERFACES: ONE IS CUSTOMIZABLE)
FIELD EXPERIMENT

• experiment: A, B, A design

• 20 participants
  – 10 feature-keen
  – 10 feature-shy

Word Personal (4 weeks)

Q1 Q2 Q3 Q4 Q5 Q6 Q7

Word 2000 Adaptive

Word 2000 Adaptive
FIELD EXPERIMENT RESULTS

Satisfaction

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

Feature-shy
Feature-keen

p < .05
FIELD EXPERIMENT RESULTS

Satisfaction

Control

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

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 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

[Image of computer interface with font options]

[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

\( p < .05 \)

<table>
<thead>
<tr>
<th>Ephemeral</th>
<th>Color highlighting</th>
<th>Control (static)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fastest</td>
<td>Preferred</td>
<td>Preferred</td>
</tr>
<tr>
<td>Pecan</td>
<td>Ceramic</td>
<td>Canola</td>
</tr>
<tr>
<td>Walnut</td>
<td>Marble</td>
<td>Sesame</td>
</tr>
<tr>
<td>Almond</td>
<td>Porcelain</td>
<td>Safflower</td>
</tr>
<tr>
<td>Pistachio</td>
<td>Granite</td>
<td>Olive</td>
</tr>
<tr>
<td></td>
<td>Molson</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Labatt</td>
<td>Cheetah</td>
</tr>
<tr>
<td></td>
<td>Coors</td>
<td>Cougar</td>
</tr>
<tr>
<td></td>
<td>Kokanee</td>
<td>Tiger</td>
</tr>
<tr>
<td></td>
<td>Coupe</td>
<td>Leopard</td>
</tr>
<tr>
<td></td>
<td>Hatchback</td>
<td>Samsung</td>
</tr>
<tr>
<td></td>
<td>Minivan</td>
<td>Panasonic</td>
</tr>
<tr>
<td></td>
<td>Sedan</td>
<td>Pioneer</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Sanyo</td>
</tr>
<tr>
<td></td>
<td>Recliner</td>
<td>Cotton</td>
</tr>
<tr>
<td></td>
<td>Loveseat</td>
<td>Flannel</td>
</tr>
<tr>
<td></td>
<td>Couch</td>
<td>Spandex</td>
</tr>
<tr>
<td></td>
<td>Sectional</td>
<td>Linen</td>
</tr>
</tbody>
</table>
HOW TO DESCRIBE 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:
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: two-factor mixed design: 2 accuracy (low or high; between subjects) x 3 menu types (control, short-onset, or long-onset; within-subjects)
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

• What is the value? Mitigates impact of carry-over effects (such as getting used to the interface) and other issues such as 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
Experiment Designs for Study 1 and Study 2?

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

• Study 1:

• Study 2: single-factor (one-way) design with menu (control, ephemeral, highlight; within subjects)
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 sample 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).
REPORTING INFERENTIAL STATISTICS

- Inferential statistics are techniques that allow us to use samples to make generalizations about the populations from which the samples were drawn. (e.g., menu A is faster than menu B)
- 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?

Figure 4. Satisfaction ratings for Study 1 (N=23). Higher values indicate higher satisfaction. Error bars show 95% CI.
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…
Ephemeral Adaptation: Further Applications
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.

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.

THEATER »

ArtsBeat: Men en Pointe
The hairy Giselles of Les Ballets Trockadero de Monte Carlo are back for the holidays.

Wall Street Slides as Oil Falls Below $40 a Barrel

By JACK HEALY 50 minutes ago
IMPORTANT/NOTEWORTHY FEATURES OF THE REPORT (CAN USE AS A CHECKLIST)

• 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
THINKING ABOUT EXPERIMENT DESIGNS

• in lecture / prep assignments, we’ve now gone through several examples of both t-tests and ANOVA

• you should be able to compare and contrast the richness/complexity of the experimental design and results for the t-test and ANOVA examples