CPSC 444: ADVANCED METHODS IN HUMAN-COMPUTER INTERACTION

Lecture 8 – Experiments IV

Joanna McGrenere

© 2017 Joanna McGrenere
includes slides from Jessica Dawson
ADMINISTRIVIA

- MSIII Blogs #4 and #5 due Tuesday, March 7 – 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: ANOVA reported in the literature
  – motivations for adaptive highlighting and ephemeral adaptation
  – how is an experiment reported?
  – 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

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 2000 Adaptive

Q1 Q2 Q3 Q4 Q5 Q6 Q7

Word Personal (4 weeks)

Word 2000 Adaptive
**FIELD EXPERIMENT RESULTS**

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

![Graph showing satisfaction levels with Q1: Feature-shy, Q2–Q6: Feature-shy, Q7: Feature-shy.](image)

- Satisfaction levels:
  - Q1: 2
  - Q2–Q6: 3
  - Q7: 2

- Statistical significance: *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

[Findlater and McGrenere, CHI 2004]
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
Presentation Order

Time Elapsed (seconds)

<table>
<thead>
<tr>
<th></th>
<th>Static (S)</th>
<th>Adaptive (Av)</th>
<th>Adaptable (Ab)</th>
</tr>
</thead>
<tbody>
<tr>
<td>S Av-Ab</td>
<td>(1)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Av Ab S</td>
<td>(2)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ab-S-Av</td>
<td>(3)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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

24 participants

Menu selection task

3 conditions (within-subjects)
Results
(p < .05)

Fastest
Preferred

Ephemeral

Color highlighting

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., could this not have all been done in one study???)
Piloting Goals

- Determine reasonable onset delays (250, 500, 1000ms) → note typo in paper
- Get early participant feedback
STUDY 1 GOALS

- Determine if ephemeral adaption improves performance over static menus
  - High prediction accuracy performance should be better
  - Low prediction accuracy performance should be no worse

- Explore how onset delay impacts performance
  - 200-300ms suggested as best in previous research, but task we are using is different
  - Expect a longer delay to improve performance, but not too long
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:

worksheet
STUDY 1 COMPONENTS OF THE EXPERIMENT DESIGN

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

• Dependent Variables:
  • Selection Time (median)
  • Error Rate (counts)
  • Subjective Satisfaction Responses (Likert Scale)
RECALL

• 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
FOCUSBING ON STUDY 1...
HYPOTHESES

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

**H1.2:** For Low accuracy: both Long-Onset and Short Onset will be no worse than Control.

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

**H2.2:** For low accuracy, static will not be preferred to short or low 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 of inferential statistics for H1:

- Omnibus ANOVA, showed sig. (p < 0.05) effect for **menu type** ($F_{2,22} = 3.80$, $p < 0.05$, $\eta^2 = 0.257$)
- Sig. Interaction for **accuracy** and **menu type** ($F_{2,22} = 3.73$, $p < 0.05$, $\eta^2 = 0.253$)

What can we conclude from this?
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
We suggest menu type had an impact on performance, but which one was best?

We suggest the impact of accuracy on performance depends upon menu type, but how?

What sort of information is NOT included in this results section? Why?
REPORTING RESULTS: H2

- Rates a qualitative aspect (preference) on a quantitative scale (1 to 7)
- Why not an ANOVA? (don’t need to know what a Friedman test it)
- What else can you do to make conclusions with this type of data?
  - E.g. Support with field work, questionnaires and illustrative quotes

![Bar chart showing overall satisfaction across different groups and accuracy levels.](image)

_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 static.  
  *Supported - long-onset faster than control*

- **H1.2:** For Low accuracy: both Long-Onset and Short Onset will be 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, static will not be preferred to short or low 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 MICHELLE MAYNARD 2:20 PM ET
A Bush spokesman said that no decision had been made but that a soft landing through a bankruptcy is an option.

The Direct Approach
Obama aides are planning a Wall Street Slides as Oil Falls Below $40 a Barrel
By JACK HEALY 5 minute 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 MICHELLE MAYNARD 2:00 PM ET

A Bush spokesman 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 NEALY 58 minutes ago
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)

• 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
you should be able to identify at least 2 specific threats to validity for the ephemeral study covered today and the C-TOC study (recent prep assignment)
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