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

**Adaptive**

- MSWord Smart Menus

**Multiple: Word Personal**

- [McGrenere and Moore, GI 2002; McGrenere, Baekker, 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
- Word 2000 Adaptive
**FIELD EXPERIMENT RESULTS**

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**

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.

[Findlater and McGrenere, CHI 2004]
**SPEED RESULTS**

Users need to experience the (potential) value of a personalized interface before personalizing.

<table>
<thead>
<tr>
<th>Time Elapsed (seconds)</th>
</tr>
</thead>
<tbody>
<tr>
<td>280</td>
</tr>
<tr>
<td>Static (S)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Presentation Order</th>
</tr>
</thead>
<tbody>
<tr>
<td>6TFSTOFFEUPFYQFSJFODFUIF QPUFOUJBM WBMVFPG BQFSTPOBMJ[FEJOUFSGBDFCFGPSFQFSTPOBMJ[JOH</td>
</tr>
</tbody>
</table>

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

[Cajos 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

---

**Comparative Experiment**

24 participants

Menu selection task

3 conditions (within-subjects)

<table>
<thead>
<tr>
<th>Ephemeral</th>
<th>Color highlighting</th>
<th>Control (static)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Textured</td>
<td>Highlighted</td>
<td>Static</td>
</tr>
<tr>
<td>Created</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Textured</th>
<th>Highlighted</th>
<th>Control (static)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Created</td>
<td></td>
<td>Static</td>
</tr>
<tr>
<td>Selected</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Menu</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Results

(p < .05)

Does ephemeral adaptation improve performance and user satisfaction?

---

[Findlater, Moffatt, McGrenere, and Dawson, CHI 2009]
**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

FOCUSING 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.

**REPORTING INFERENTIAL STATISTICS**

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
REPORTING RESULTS: H1

• 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 is)
• What else can you do to make conclusions with this type of data?
  • E.g. Support with field work, questionnaires and illustrative quotes

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

IMPLICATIONS FOR DESIGN

- Beyond menus...

LEAVE YOU TO WALK
THOUGH ON YOUR OWN THE
SAME FOR STUDY 2...

Ephemeral Adaptation: Further Applications
Ephemeral Adaptation: Further Applications

**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**

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

**LEFT FOR YOU TO PONDER**

- 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