EVALUATION/TESTING
EXPERIMENT III

CPSC 544 FUNDAMENTALS IN DESIGNING INTERACTIVE COMPUTATION TECHNOLOGY FOR PEOPLE (HUMAN COMPUTER INTERACTION)

WEEK 11 – CLASS 21

© Joanna McGrenere and Leila Aflatoony
Includes slides from Karon MacLean and Jessica Dawson
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)**

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’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
- Fredericton
- Halifax
- Calgary
- Regina
- St. John's
- Toronto
- Victoria
- Winnipeg
- Kelowna
- Ottawa
- Montreal
- Vancouver
- Quebec City
- Edmonton
- Charlottetown

Static split menu
- St. John's
- Kelowna
- Ottawa
- Montreal
- Fredericton
- Halifax
- Calgary
- Regina
- Toronto
- Victoria
- Winnipeg
- Vancouver
- Quebec City
- Edmonton
- Charlottetown

Most frequent items

[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
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)
---|---|---
Saturn | Ceramic | Canola
Venus | Marble | Sesame
Jupiter | Porcelain | Safflower
Mercury | Granite | Olive
Aquarius | Molson | Cheetah
Gemini | Labatt | Cougar
Taurus | Coors | Tiger
Virgo | Kokanee | Leopard
France | Coupe | Samsung
England | Hatchback | Panasonic
Spain | Minivan | Pioneer
Germany | Sedan | Sanyo
Pecan | Recliner | Cotton
Walnut | Loveseat | Flannel
Almond | Couch | Spandex
Pistachio | Sectional | Linen
Results
(p < .05)

Ephemeral

Color highlighting

Control (static)

Fastest

Preferred

Saturn
Venus
Jupiter
Mercury
Aquarius
Gemini
Taurus
Virgo
Libra
Scorpio
Sagittarius
Capricorn
Pisces

Ceramic
Marble
Porcelain
Granite
Molson
Labatt
Coors
Kekanee
Coke
Hatchback
Minivan
Sedan
Recliner
Loveseat
Couch
Sectional

Canola
Sesame
Safflower
Olive
Cheetah
Cougar
Tiger
Leopard
Samsung
Panasonic
Pioneer
Sanyo
Cotton
Flannel
Spandex
Linen
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???)
• Too much to test in one study (likelihood of success – learning something meaningful – would have been very low)

• At each stage (piloting, Study 1, Study 2) we were able to clarify which variables were important and at which values (i.e., determine factors and factor levels)
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

• 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
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)
FOCUSBING 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?

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
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’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
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!)*
ADJUSTMENTS TO BALANCE OF 544 DELIVERABLES

• Stay tuned for a more detailed piazza post ...

• Idea:
  – Remove Individual research proposal (25%)
  – Add individual experiment report + project level reflections (redistribute the grade across project components)
EXTRA SLIDES
IMPLICATIONS FOR DESIGN

- Try to keep spatial relationships between controls constant to aid learning
- Avoid adaptation schemes that distract the user from performing their usual workflow
- Further evidence to support the importance of pilot testing to address problems early
- Support conclusions about hypotheses from multiple measures if possible
spatial adaptations
  – collects items user likely to need in one place
  – e.g., putting most commonly used fonts at the top of the font selection

  but spatially inconsistency can be a problem
  • e.g., user becomes disoriented when an item moves from expected location
graphical adaptations

- keep items in place, and highlight them using some method
- e.g., highlight commonly used items with colours

• but static, persistent highlighting can be distracting
  • what if the highlighted items aren’t what the user wants right now?

example of colour highlighting used in the experiment
(fig. 5 from Findlater et al.)
ephemeral adaptation
  • maintains spatial consistency
  • highlighting temporary
ephemeral adaptation

- maintains spatial consistency
- highlighting temporary