Human Communication
Lecture 19: Experimental Design
e.g. evaluating STANDUP and a 2LL system

Research Methodology

Week 1 2 3 4 5 6 7 8 9
Formal testing Video observation Formal testing

4. Evaluation of Standup

The evaluation study
1. 9 participants from independent special school
2. 14 sessions c. 30 minutes over 9 weeks (April/May/June),
3. Consent obtained from parents and children
4. Pre-testing with standardised tests
5. Children shown how to use the software weeks 1 and 2
6. Intervention period exploring software weeks 3 to 6
7. Level of support and guidance reduced, and task complexity increased, as sessions went on
8. Use of system video-recorded for study
9. Favourite jokes stored in paper folder and on AAC devices
10. Evaluation period weeks 7 and 8
11. Further standardised testing
12. Structured interviews and questionnaires for feedback from staff and parents
13. Talking mats to collect feedback from children

Use with typically-developing children

Evaluation Instruments

Standardised tests - normalised to intended population
CELF Clinical Evaluation of Language Fundamentals
(Semel, Wиг, Secord, 1995)
CELF Linguistic concepts: participants are asked to point to...
"the blue line", "the line that is not yellow";
(participants must point to a stop sign if they think they cannot
do what they are asked to do.)
CELF Sentence structure e.g. show me...
"The girl is not climbing", "The dog that is wearing a collar is eating a bone"
CELF Oral directions e.g. point to...
"The black circle", "The last white triangle and the first black square"
CELF Word classes: participants choose two related items from a
set of four, e.g. "girl boy car table", "slow nurse doctor rain"

PIPA Preschool and primary inventory of phonological awareness
(Frederickson, Frith and Reason, 1997)

Evaluation Instruments: The KMT

Keyword Manipulation Task (O’Mara, 2005): standardised across 57 children, including language impaired children; 5 – 12 years.

Stimulus: How can you tell there has been an elephant in your fridge?
Footprints in the butter.

Keyword Alternates:

Stimulus: What do you get when you cross a car and a sandwich?
A traffic-jam.

Keyword Alternates:
The Standard deviation $\sigma$ is the square root of the variance:

$$\sigma = \sqrt{\left(\frac{1}{N}ight) \sum (X_i - \mu)^2}$$
### Statistical Comparison: T-test
#### Performance on CELF Test

- **Pre-intervention:**
  - Mean = 12.1
  - Standard Deviation = 7.87
- **Post-intervention:**
  - Mean = 16.2
  - Standard Deviation = 9.76
- **Difference:**
  - Mean = -4.11
  - Standard Deviation = 3.30

The results of a paired t-test:
\[ t = -3.74 \]
\[ \text{degrees of freedom} = 8 \]

The probability of this result, assuming the null hypothesis, is 0.006.

*So cannot assume the null hypothesis*

### Statistical Comparison: T-test
#### Performance on PIPA Test

- **Pre-intervention:**
  - Mean = 8.11
  - Standard Deviation = 4.01
- **Post-intervention:**
  - Mean = 8.67
  - Standard Deviation = 3.39
- **Difference:**
  - Mean = -0.556
  - Standard Deviation = 2.60

The results of a paired t-test:
\[ t = -0.640 \]
\[ \text{degrees of freedom} = 8 \]

The probability of this result, assuming the null hypothesis, is 0.540.

*So no reason NOT to accept the null hypothesis*

### Unexpected Outcomes

- Impact on school curriculum
  - Questionnaires with parent, teachers and Classroom assistants (not significant issues raised but all positive)
  - Semi-structured interviews with SLTs

### Preliminary Results: Feedback

#### Feedback using Talking Mats

- **Good:** Jester character, way screen changes, way of telling jokes
- **OK:** Touchscreen
- **Bad:** Scanning, jokes, voice

### Participant Feedback using Talking Mats

#### Feedback using Talking Mats

- **Good:** Jester character, way screen changes, way of telling jokes
- **OK:** Touchscreen
- **Bad:** Scanning, jokes

### STANDUP: some initial conclusions

- Interfaces CAN be designed which provide children with CCN with successful access to complex underlying technology
- Using STANDUP:
  - the generative capabilities allows opportunity for natural language development, cf DA choosing punchline first
  - the generative capabilities allows novel explorative learning, cf NI searching subjects
- All children benefited
  - enhanced desire to communicate
  - knock on effect on other AAC usage
  - Illustrated children’s abilities and potential of AAC
- Illustrated use of technology within a wider environment
STANDUP: some initial conclusions

Issues with interface design
- scanning
- voice output
- improved appropriateness of vocabulary

The telling of the joke is important - what is the impact of STANDUP:
- on interactive conversation
- on joke comprehension and vocabulary acquisition

Do we want better jokes? (yes)

Use with speaking children with language impairment and other user groups

Hypothesis Formation

Typical hypothesis: factor X affects behaviour Y
Typical Null hypothesis: no effect of X on Y

What will we measure about X and Y?

Observation v Manipulation
- Observation studies: look at the population to see if X correlates with Y
- Manipulation experiments: change X and see what happens to Y

But we need to be sure that any change in Y is due only to the differences in X...

Collecting, Analysing and Interpreting Data

1. What questions are we asking that we need data to answer?
2. What data would provide the answers to these questions?
3. What methods would enable us to collect this data?
4. How would we analyse the data?
5. How would we interpret it?

2LL Tutoring System Example

Goal: develop intelligent computer tutor:
- for children to improve second language learning
- to better understand how learners learn
- to decide what forms of feedback work best

When designing the system, we might consider:
1. What errors do students typically make?
2. What should the system do when students make errors?

Having developed the system, we might use it to:
- better understand the learner,
- see what errors they make,
- to assess effectiveness of different feedback strategies

What errors do students typically make?

1. Interview teachers about errors that target users frequently make (error types and examples)
2. Devise a set of language test examples
3. Give target user group test set and observe, collect log of their interaction (example errors)
4. Analyse results to see most frequent errors
5. Give questionnaire to teachers with example errors and ask what feedback they would give (feedback types in relation to each error)
6. Observe tutor teaching student through chat interface + record interaction (example errors)
7. Analyse interaction in relation to student errors and actions taken by teacher (feedback types)
8. Cognitive walkthrough by tutor (when feedback type given and general feedback strategies)
What should the system do when students make errors?

Using these methods you find that human tutors usually use one of the following feedback options:
1. give feedback immediately
2. just flag to the student that they have made an error
3. let the student realise they have made a mistake and ask for help

You want to see which works best...

Do some experiments with the tutoring system, with some students....

[Based loosely on a experimental study described in Corbett, A.T. and Anderson, J.R., 1990]

Testing Hypotheses

"Immediate Feedback is best!"

Hard to test - we need to be more specific

"Differences in performance on a specific test will be shown between students given no feedback and students given immediate feedback."
= the experimental hypothesis

"There will be no difference in performance shown by students given immediate feedback or no feedback."
= the null hypothesis

Possible Variables

* Whether or not feedback is given
* When it is given -- immediately? after 3 errors? at the end?
* What is given as feedback: correct or incorrect; detailed explanation; further examples
* How much control does student have over feedback?
* How long does the student take to complete an exercise?
* What is the student's level of performance?
* How does the student feel about the different types of feedback - which do they prefer?

Qualitative v. Quantitative Data

Qualitative
- Descriptive data
- Based on system behaviour or user experience
- Obtained from observation, questionnaires, interviews, protocol analysis, cognitive and post task walkthrough
- Subjective

Quantitative
- Numerical data
- Based on measures of variables relevant to performance or user experience
- Obtained from empirical studies, e.g. experiments, also questionnaires, interviews
- Amenable to statistical analysis
- Objective

(see Dix et al, 2004, chapter on evaluation)

Formative v. Summative Evaluation

Formative Evaluation:
- throughout design and implementation
- incremental
- assessing impact of changes
- frequently qualitative

Summative Evaluation:
- on completion of each stage
- assessing effectiveness
- frequently quantitative
Experimental Design

Experimental conditions:
1. immediate error feedback and correction
2. immediate error flagging but no correction
3. feedback on demand

Control condition: to eliminate alternative explanations of the data obtained
4. no feedback

Experimental Variables

Independent Variable - manipulated by experimenter
Dependent Variable - not manipulated, but look to see if manipulating the independent variable has an effect on it (but not necessarily a causal relationship)

Independent Variable: type of feedback
Dependent variable: time to complete the exercises; post-test performance

Control condition: to eliminate alternative explanations of the data obtained
4. no feedback

Controlling for Extraneous Variables (1)

1. Make the extraneous variable an independent one, and include it in the experiment (if possible)
   i.e. vary the value of it together with that of the independent variable

2. Partition the test cases such that the extraneous variable effects cancel out
   e.g. “effect of gender on language performance” - collect a large number of pairs of 1 male + 1 female such that each pair is matched on age, socio-economic class, training, etc. so differences within each pair is solely attributable to gender

Controlling for Extraneous Variables (2)

3. Random sample of the population of individuals with each of the values of the independent variable, compare the behaviour of these samples
   e.g. Run 100 randomly different runs of algorithm for each chosen set of algorithm parameters

Effects of other, extraneous, variables should appear as random variation in the dependent variable - effects of independent variable will not be random - a statistical test can distinguish them

Be careful than samples are really random with respect to the extraneous variables - if there is a cause-effect relationship we do not know about, effects of the extraneous variables may compound instead of cancelling out

Have to be very careful in selecting random samples

Participants

Use the same subjects for the different conditions?
- or different groups of subjects for each condition?
- or matched subjects?

A. Different subjects (= between group comparison):
   • different subjects undergo different conditions
   • assume all from the same population

Pros
- less order effects
- simpler design

Cons
- individual differences
- needs more subjects

Within group design

A. Same subjects (= within group comparisons):
   • each subject uses the tutor under all conditions
   • vary order of conditions to avoid order effects
   • use isomorphic problems of equivalent difficulty, and vary these also across conditions

Pros
- needs fewer subjects
- avoids individual differences

Cons
- more complex design
- need isomorphic problems
- may still get order effects
- testing v. learning issues
- fatigue/boredom
Matched Subjects

C. Matched subjects (between groups, where pairs of subjects across groups are matched):

Could match on:
• intelligence
• previous number of years 2LL experience
• previous performance in language courses

Pros

Cons

+ as between groups plus reduces individual differences appropriate matches

So could use between groups design:
55 students from the same class. Assumed roughly the same experience

Choosing Between Experimental Study Designs (Ainsworth, 2003)

Validity

Construct validity
Is it measuring what it is supposed to?

External validity
Is it valid for this population?

Ecological validity
Is it representative of the context?

Reliability
Would the same test produce the same results if:
Tested by someone else?
Tested in a different context?
Tested at a different time?

Results: Test Scores and Completion Time

Say we measured mean post-test scores (% correct) and mean exercise completion times (minutes) for the 4 versions of the tutor.

We could then compare the sets of scores across conditions to see if the differences are statistically significant...

<table>
<thead>
<tr>
<th></th>
<th>Immediate feedback</th>
<th>Error flagging</th>
<th>Demand feedback</th>
<th>No tutor</th>
</tr>
</thead>
<tbody>
<tr>
<td>Post-test Scores</td>
<td>55%</td>
<td>75%</td>
<td>75%</td>
<td>70%</td>
</tr>
<tr>
<td>Exercise Times</td>
<td>4.6</td>
<td>3.9</td>
<td>4.5</td>
<td>4.5</td>
</tr>
</tbody>
</table>

Discussion and Conclusions

We might find in this case that the effect of tutor type, as measured by post-test scores and mean exercise completion times, is not statistically significant.

- So there would be no evidence in this case that feedback manipulation affected learning [though other research may show that there is].

Writing-up empirical studies 1

Abstract: Short summary of the problem, results and conclusion.

Introduction: What is the problem? What related work have other people done?
[Go from general statement of problem to a succinct and testable statement of hypothesis].

Method:
Participants: state number, background and any other relevant details of participants
Materials: exactly what test materials, teaching materials, etc. were used, giving examples
Procedure: clear and detailed description of what happened at each stage in the experiment
[Someone reading should be able to duplicate it. Should clearly indicate what data was collected and how.]

Writing-up empirical studies 2

Results:
- Give actual data, or a summary of it.
- Provide an analysis of data, using statistical tests where/if appropriate.
- Use tables and graphs to display data clearly.
[Interpretation of results does not go here, but in discussion section].

Discussion:
- Interpretation of results; restating of hypothesis and implications of results; discussion of methodological problems e.g. weaknesses in design, unanticipated difficulties, confounding variables
- Wider implications of the work considered here, and perhaps further studies suggested.

Conclusion:
Statement of overall conclusion of the study.
Reading

Chapter 9: Evaluation Techniques pp 318 - 364