Experimenting: experiment design

Per Ola Kristensson

Research Methods
M.Phil. Advanced Computer Science
University of Cambridge

Michaelmas Term, 2009

Scientific method in one minute

1. Use experience and observations to gain insight about a phenomenon
2. Construct a hypothesis
3. Use hypothesis to predict outcomes
4. Test hypothesis by experimenting
5. Analyse outcome of experiment
6. Go back to step 1

Typical computer science scenario

- A particular task needs to be solved by a software system
- This task is currently solved by an existing system (a baseline)
- You propose a new, in your opinion, better system
- You argue why your proposed system is better than the baseline
- You support your arguments by providing evidence that your system indeed beats the baseline

Running example in this lecture

- Text entry on a Tablet PC

A. Handwriting recognition
B. Software keyboard

Why experiments?

- Substantiate claims
  - A research paper needs to provide evidence to convince other researchers of the paper's main points
- Strengthen or falsify hypotheses
  - "My system/technique/algorithm is [in some aspect] better than previously published systems/techniques/algorithms"
- Evaluate and improve/revise/reject models
  - "The published model predicts users will type at 80 wpm on average after 40 minutes of practice with a thumb keyboard. In our experiment no one surpassed 25 wpm after several hours of practice."
- Gain further insights, stimulate thinking and creativity

Back to our example

- Why this experiment?
  - Despite decades of research there is no empirical data of text entry performance of handwriting recognition
  - An inappropriate study of handwriting (sans recognition) from 1967 keeps getting cited in the literature, often through secondary or tertiary sources (handbooks, etc.)
  - Based on these numerous citations in research papers, handwriting recognition is perceived to be rather slow
  - However, there is no empirical evidence that supports this claim

Different kinds of experiments

- Surveys
- Field studies
- Simulations and computational experiments
- Controlled experiments
- … and quasi-experiments, and many more…

Controlled experiments and hypotheses

- A controlled experiment tests the validity of one or more hypothesis
- Here we will consider the simplest case:
  - One method vs. another method
  - Each method is referred to as a *condition*
- The null hypothesis H0 states there is no difference between the conditions
- Our hypothesis H1 states there *is* a difference between the conditions
- To show a statistically significant difference the null hypothesis H0 needs to be rejected

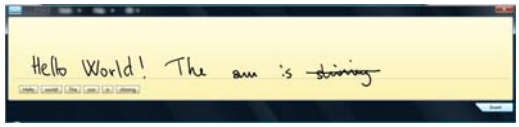
Choice of baseline

- A baseline needs to be accepted by your readers as a suitable baseline
- Preferably the baseline is the best method that is currently available
- In practice a baseline is often a standard method which is well-understood but often not representative of the state-of-the-art

Our example, two conditions

1. Software keyboard (baseline)

2. Handwriting recognition



Why this baseline?

- The software keyboard is well understood
  - Many empirical studies of their performance
  - Also exists expert computational performance models
- The software keyboard is the *de-facto* standard text entry method on tablets
- The literature compares handwriting recognition text entry performance against measures of the software keyboard

Aim of controlled experiment

- To measure effects of the different conditions
- To control for all other confounding factors
- To be internally valid
- To be externally valid
- To be reproducible

Experimental design

- Dependent and independent variables
- Within-subjects vs. between-subjects
- Mixed designs
- Single session vs. longitudinal experiments

Dependent and independent variables

- Dependent variable:
  - What is measured
  - Typical examples (in CS): time, accuracy, memory usage
- Independent variable
  - What is manipulated
  - Typical examples (in CS): the system used by participants, feedback to participant (e.g. a beep versus a visual flash)

Deciding what to manipulate and what to measure

- This is a key issue in research
- Boils down to your hypothesis:
  - What do you believe?
  - How can you substantiate your claim by making measures?
  - What can you measure?
  - Is it possible to protect internal validity without sacrificing external validity?

Our example

- We let participants write phrases using either:
  - Software keyboard (baseline)
  - Handwriting recognition
  - That is, we manipulate the *input method*
- We measure:
  - Entry rate in words-per-minute
  - Error rate in number of written characters that do not match the stimulus

Between-subjects design

- Each participant is exposed to only one condition
- One of the simplest experimental designs
- Advantages:
  - No risk of confounds or skill-transfer from one condition to the other
  - Therefore no need to do counter-balancing or check for asymmetrical skill-transfer effects
- Disadvantages:
  - Variance is not controlled within the participant
  - Therefore demands more participants than a within-subjects design to show a statistically significant difference

Within-subjects design

- Each participant is exposed to all conditions
- One of the most common experimental designs in practice
- Advantages:
  - Variance is controlled within the participant
  - Therefore requires fewer participants than a between-subjects design
- Disadvantages:
  - More involved, requires counter-balancing of start condition to avoid transfer effects
  - Risk of asymmetrical skill transfer

Mixed designs

- It is also possible to combine within- and between-subjects experimental designs
- Such designs are called mixed designs
- These are difficult to design because they are more difficult to control
- A mixed design can be a symptom of no clear set of hypotheses, or lack of ability to prioritise among them
- Often a mixed design can be broken down into smaller studies that study isolated phenomena separately

Single session vs. longitudinal

- Do you believe participants will improve significantly over time?
- If so, how much will they improve?
- How are previous related studies set up in the literature?

Pilot study

- A pilot study (sometimes called "formative study") is a small study conducted before the actual controlled experiment
- A pilot study may be designed as the controlled experiment and typically requires much fewer participants (perhaps only one participant)
- A pilot study is important for many reasons:
  - Provides some idea of the feasibility that the null hypothesis will be rejected
  - Enables you to ensure the apparatus and software is working correctly
  - Enables you to ensure instructions to participants are clear
  - Can inform certain parameters of the controlled experiment, such as appropriate session length

Explaining what you did

- An experiment needs to be reproducible by others
- It is your responsibility to ensure that you explained your experimental procedure in enough detail
- Choices made in the experimental design needs to be motivated
- This part of a research paper is typically referred to as the *Method* section

Method

- Participants
- Apparatus
- Procedure

Participants

- How many?
- How is the sample constructed?
  - Is it representative of the population we believe will use the interface?
  - Are potential problematic confounds taken care off?
- Did participants receive any compensation?
- Was the study approved by the university ethics committee? [if applicable]

Participants, our example

We recruited 12 volunteers from the university campus. We intentionally wanted a rather broad sample and recruited participants from many different departments with many different backgrounds. Six were men and six were women. Their ages ranged between 22-37 (mean = 27, sd = 4). Participants were screened for dyslexia and repetitive strain injury (RSI). Seven participants were native English speakers and five participants had English as their second language. No participant had used a handwriting recognition interface before. One participant had used a software keyboard before. No participant had regularly used a software keyboard before. Participants were compensated £10 per session.

Participants, our example

We recruited **12 volunteers** from the **university campus**. We intentionally wanted a rather broad sample and recruited participants from many different departments with many different backgrounds. **Six were men and six were women**. Their **ages ranged between 22-37** (mean = 27, sd = 4). Participants were **screened** for dyslexia and repetitive strain injury (RSI). **Seven** participants were **native English speakers** and **five** participants had **English as their second language**. No participant had used a handwriting recognition interface before. One participant had used a software keyboard before. No participant had regularly used a software keyboard before. Participants **were compensated £10 per session**.

Apparatus

- Which equipment and which software?
  - Needs to be described in sufficient detail to enable other researchers to replicate your experiment
- Typical information:
  - Physical and logical screen size
  - Sensor device characteristics
  - CPU clock speed
  - Computer brand/model
- Choices that are not obvious need to be motivated

Apparatus, our example

We used a Dell Latitude XT Tablet PC running Windows Vista Service Pack 1. The 12.1" color touch-screen had a resolution of 1280 × 800 pixels and a physical screen size of 261 × 163 mm. Participants used a capacitance-based pen to write directly onto the screen in both conditions.

…

Both the handwriting recognizer and the software keyboard were docked to the lower part of the screen. The dimensions of the software keyboard were 1266 × 244 pixels and 257 × 50 mm. The dimensions of the handwriting recognizer writing area measured 1266 × 264 pixels and 257 × 55 mm.

Apparatus, our example

We used a Dell Latitude XT Tablet PC running Windows Vista Service Pack 1. The 12.1" color touch-screen had a resolution of **1280 × 800 pixels** and a physical screen size of **261 × 163 mm**. Participants used a **capacitance-based** pen to write directly onto the screen in both conditions.

…

Both the handwriting recognizer and the software keyboard were docked to the **lower part of the screen**. The dimensions of the software keyboard were 1266 × 244 pixels and 257 × 50 mm. The dimensions of the handwriting recognizer writing area measured 1266 × 264 pixels and 257 × 55 mm.

Apparatus, motivating your choices

The handwriting recognizer was configured to learn and adapt to participants' handwriting style (the default setting on Windows Vista). Each participant performed the experiment in a separate user account on the machine to ensure handwriting adaptation was carried out on an individual basis. There was a potential confound in enabling handwriting adaptation since it caused the system, as well as the user, to learn as a function of usage. In the interest of external validity we enabled adaptation since in actual use users would most likely have adaptation turned on.

Apparatus, motivating your choices

The handwriting recognizer was **configured to learn and adapt** to participants' handwriting style (the **default setting** on Windows Vista). Each participant performed the experiment in a **separate user account** on the machine to ensure handwriting adaptation was carried out on an individual basis. There was a **potential confound** in enabling handwriting adaptation since it caused the system, as well as the user, to learn as a function of usage. In the interest of **external validity** we enabled adaptation **since in actual use** users would most likely have adaptation turned on.

Procedure

- Describes how the experiment was carried out
- Needs to be described in sufficient detail for other researchers to be able to replicate your experiment
- Again, choices need to be motivated

Procedure, our example

The experiment consisted of one introductory session and ten testing sessions. In the introductory session the experimental procedure was explained to the participants. Participants were shown how to use the software keyboard and the handwriting recognizer, including demonstrations of how to correct errors.

…

Each testing session lasted slightly less than one hour. Testing sessions were spaced at least 4 hours from each other and subsequent testing sessions were maximally separated by two days. In each testing session participants did both conditions (software keyboard and handwriting recognition). The order of the conditions alternated between sessions and the starting condition was balanced across participants. Each condition lasted 25 minutes. Between conditions there was a brief break. Participants were also instructed that they could rest at any time after completing an individual phrase.

Procedure, our example

The experiment consisted of **one introductory** session and **ten testing** sessions. In the introductory session the experimental procedure was explained to the participants. Participants were shown how to use the software keyboard and the handwriting recognizer, including demonstrations of how to correct errors.

…

Each **testing session lasted slightly less than one hour**. Testing sessions were **spaced at least 4 hours** from each other and subsequent testing sessions were **maximally separated by two days**. In each testing session participants **did both conditions** (software keyboard and handwriting recognition). The **order of the conditions alternated between sessions** and the **starting condition was balanced across participants**. **Each condition lasted 25 minutes**. Between conditions there was a brief break. Participants were also instructed that they could rest at any time after completing an individual phrase.

Procedure, our example

In each condition participants were shown a phrase drawn from the phrase set provided by MacKenzie and Soukoreff [8]. Each participant had their own randomized copy of the phrase set. Participants were instructed to quickly and accurately write the presented phrase using either the software keyboard or the handwriting recognizer. Participants were instructed to correct any mistakes they spotted in their text. In the handwriting condition we instructed participants to write using their preferred style of handwriting (e.g. printed, cursive or a mixture of both). After they had written the phrase they pressed a Submit button and the next phrase was displayed. The Submit button was a rectangular button measuring 248 × 16 mm. It was placed 9 mm above the keyboard and handwriting recognizer writing area.

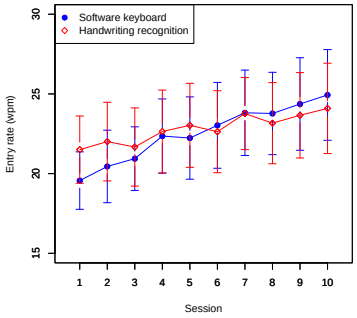
Procedure, our example

In each condition participants were **shown a phrase** drawn from the **phrase set** provided by MacKenzie and Soukoreff [8]. Each participant had their **own randomized copy of the phrase set**. Participants were instructed to **quickly and accurately** write the presented phrase using either the software keyboard or the handwriting recognizer. Participants were **instructed to correct any mistakes they spotted in their text**. In the handwriting condition we instructed participants to **write using their preferred style of handwriting** (e.g. printed, cursive or a mixture of both). After they had written the phrase they pressed a Submit button and the next phrase was displayed. The Submit button was a rectangular button measuring 248 × 16 mm. It was placed 9 mm above the keyboard and handwriting recognizer writing area.

After the experiment

- Results
- Limitations and implications

Our example



Summary

- A well-designed controlled experiment provides you empirical evidence that your new method is better [in some aspects] than some previous method in the literature (a baseline)
- Important to consider the experimental design early
  - Within vs. between
  - Dependent and independent variables
  - Internal and external validity
- Pilot study often a good idea (perhaps your method has a fatal flaw)
- Important to point out limitations and implications
- Experiments must be reproducible