Experimental Design Handout

Catherine Arnett

February 21, 2023

Introduction

Experimental design usually comes into the research process after the research question and hypotheses have been identified. The main goal of an experiment should be to find a piece of evidence that disambiguates at least two hypotheses relating to the research question.

When designing an experiment, some of the main things you have to think about are the experimental task, experimental variables and controls, experimental materials, and the statistical analysis. This handout will introduce key terminology and point out some of the most important considerations in experimental design relating to these topics.

Experimental Tasks

One of the first steps in designing an experiment is picking the experimental task. One of the biggest divides in the taxonomy of experimental tasks is whether the experiment uses offline or online measures. **Offline measures** focus on observing behaviors, such as answer choice given options or what kind of sentence a participant produces in an experiment. **Online measures** observes a process as it's occurring, for example using EEG (HSP Blog).

Within each of these types of measures, there are a lot of different ways to run experiments. Individual experimental tasks can be grouped into **paradigms**.

Offline experiments often collect **behavioral** data, or data about what people do and say. Some examples of offline experimental paradigms are:

- Limited Response Measures
  - **Forced Choice Task**: in this kind of experiment, a participant will be presented with a question and multiple choice answers. Usually in a forced-choice task, the participant will only have two options.
  - **Likert Scale Choice**: a similar experimental paradigm requires participants to rate something on a scale from 1 to 7 (or sometimes 1 to 5), which is referred to as the Likert scale. This is often used for acceptability judgements and confidence ratings.

  - **Discrimination Task**: this kind of task requires participants to detect whether two things, usually phones, are the same or different. Usually this takes the form of an ABX task, where the participant has to say whether A or B is most similar to X (ScienceDirect).

- Open Ended Response Measures
  - **Priming**: a common production paradigm presents participants sentences with a particular feature (certain words, syntactic constructions) and then measures what kinds of words or constructions the participants use in their next utterance.
  - **Speech Errors**: another way of looking at language produced by participants is by measuring the number of errors they produce in speech, whether they are speaking naturalistically or responding to prompts.
  - **Recall Tasks**: in this paradigm, participants are asked to repeat or recall information or language presented earlier in the experiment.
  - **Shadowing**: this refers to a task where participants listen or watch a narrative and then repeat the story back verbatim in the same language. Experimenters will then compare the original narrative with the participant's retelling.

Some examples of online experimental paradigms are:

- Reaction Time as a measure
  - **Lexical Decision Task**: in this experimental paradigm, participants see a word (often by itself), and have to decide whether or not it is a real word.
  - **Self-Paced Reading**: these kinds of experiments have participants read a text one word at a time. Usually the participants will press the space bar to move to the next word. Reaction time is measured for each word.
  - **Picture Naming**: in production studies, a common paradigm is a picture naming study, where participants see a picture and have to say the word it represents outloud. The reaction time is measured as **speech onset latency**, or the amount of time between the presentation of the image and the time when the participant begins to speak.

- Eye Tracking
  - **Preferential Looking Paradigm**: this paradigm consists of participants listening to a recording and then their eye movements are recorded to see whether they look at objects that reflect what kinds of predictions they are making about the sentence they are listening to.
  - **Visual World Paradigm**: similarly, in this paradigm, participants see an image with objects that could be relevant to the language they are listening to. Their eye movements are tracked while they listen to the sentence to see how what they are looking at relates to the sentence they are listening to.

- **Pupilometry**: this task measures participant pupil diameter and how it changes as they are exposed to different sentences, usually aurally.

Offline and online measures are not independent from one another. For instance, in many experiments using online measures, experimenters will collect behavioral or offline measure data. In a Lexical Decision Task, for example, you may only want to look at reaction time data in trials where participants actually chose the right answer, because participants may answer more quickly when they answer incorrectly. Having both of these measures might result in more accurate and clear experimental results.

Variables, Factors, and Controls

Once you've picked an experimental task, the next step is to think about how to set up the experiment so you're measuring what you want to be measuring.

Types of Variables

The key variables in the experiment are the independent and dependent variables. In an experiment, you want to find the relationship between these variables. Taking for example an experiment where you want to know X, so you ask participants to do Y, the independent variable is A and the dependent variable is B. Generally the **independent variable** is the variable you know in advance of the experiment (e.g. whether the participants are in the **control group** or not) and the **dependent variable** is what you're measuring in the experiment (e.g. answer choice, reaction time). This is sometimes also referred to as a **control variable**.

There also may be other variables, or **factors**, that you want to pay attention to in your experiment. One type of variable is a **random variable**. This is a variable that you don't manipulate or explicitly measure, but you collect data on to include in the statistical analysis. Examples of random variables are individual differences (operationalized as participant ID; (Kliegl et al., 2011)), which trial number (e.g., question 17/48; Schuman et al. (1981)), or time and day of an experiment (see Oakhill and Davies (1989); Knight and Mather (2013)). The current approach to random variables is collect information about them and take them into account in the statistical analysis.

A **confounding variable** is something that is not measured in the experiment but may explain some of the observed relationship between the independent and dependent variables. An example of this can be seen Fig. 1. In this example, there is a third variable (temperature) that is causally related to the dependent variable. So, even though there may be a correlation between ice cream sales and shark attacks, the confounding variable, temperature, may actually be the cause of increased shark attacks. Additionally, increased temperature also explains increased ice cream sales. If you do not make sure to reduce the possibility of confounds in an experiment, your experiment will likely not provide accurate inferences.

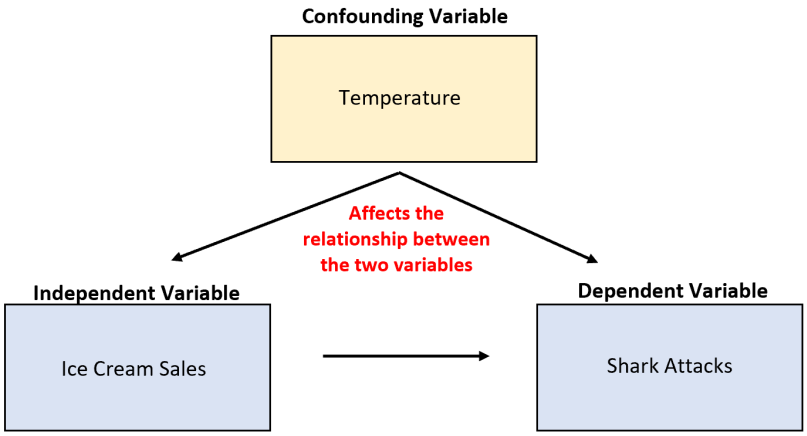

Figure 1: Example of a confound (image source)

In a psycholinguistic context, if you are measuring reading time, you may find an increase in reading time for a certain type of sentence. A confounding variable could be word length or word frequency. In order to prevent this kind of confound, we can **control** for word length or word frequency by making sure that the sentences have words of similar length or frequency to eliminate these factors from affecting the experiment.

Within- and Between-Subjects

Another consideration is how each participant will be exposed to the different conditions you've created. For each variable, each value of the variable constitutes a different **condition**.

If you set up an experiment where each participant sees trials from more than one condition of a given (independent) variable, then this is a **within-subjects** experiment. Sometimes this is also referred to as **repeated measures** (Salkind, 2010). This reflects the fact that we are looking for effects that you can see within each individual subject. This is helpful because it minimizes the possibility that your results are due to individual differences between treatment groups, as the same participants complete trials in all conditions.

A **between-subjects** experiment consists of splitting participants into groups based on the conditions. For example, half of participants will only see trials from Condition A, and the other half will only see trials in Condition B. In this case, the results show the effect of the different conditions, comparing different groups of participants, hence the name *between subjects*. Between-subjects testing is common in medical experiments. Drawbacks for this type of experiment are that you may need a larger sample size.

Controls for Within-Subjects

When doing a within-subjects or repeated measures experiment, certain **controls** need to be implemented to reduce confounds. To do this, psycholinguists also counterbalance their

stimuli. Counterbalancing refers to the practice of arranging the experimental conditions in a way that minimizes carryover effects from one condition to the other, order effects, and participant fatigue (APA). For example, if there is an experiment where each participant sees 16 trials, exactly 8 of the trials must be in Condition A and exactly 8 must be in Condition B (Science Direct). An example of counterbalancing would be to present half the participants with Condition A trials followed by Condition B trials and present the other half with Condition B trials followed by Condition A trials. This minimizes the likelihood that effects seen in the results are due to the order of exposure of the conditions.

A more complex version of counterbalancing is Latin Square design. **Latin Square** designs are used to distribute trials of different conditions of the same variable equally across participants, especially when there are more than two conditions. Fig. 2 shows example trials for different subjects, each with different order of trials and conditions.

Subject	Consecutive tests (or study periods)			
	1	2	3	4
#1	A	B	C	D
#2	B	C	D	A
#3	C	D	A	B
#4	D	A	B	C

Figure 2: Latin Square design (image source)

Multiple Dependent Variables

In some experiments, it might be necessary to manipulate more than one independent variable. We use **factorial design** to talk about the design of an experiment with multiple variables (aka factors; APA). Factorial design is talking about the experiment on an abstract level, rather than talking about the order of stimuli like Latin Square design.

The most common factorial design is a **2x2 design**. In this format of talking about factorial design, the number of numbers represents the number of independent variables. So a 2x2 study has two independent variables, a 2x2x2 study has three independent variables, and so on. The actual numbers themselves represent the number of **levels**, or values, of each variable. In a 2x2 design, then, there are two independent variables each with two levels.

Fig. 3 shows an example of this. The two independent variables in this case are sunlight amount and watering frequency. The two levels for sunlight amount are high and low. The two levels of watering frequency are daily and weekly.

		Watering Frequency	
		Daily	Weekly
Sunlight	Low	Plant Growth	Plant Growth
	High	Plant Growth	Plant Growth

Figure 3: 2x2 design (image source)

Materials and Participants

When reporting an experiment, in addition to a description of the methods or procedure, the authors generally provide information about the participants and the experimental materials.

Participants

Typically experiments report the number of participants recruited and the number of participants included in the analysis, if those numbers are different. Sometimes **N** is used to represent the number of participants, e.g. 'N=100' represents a study with 100 participants. Other information that is generally provided about participants includes:

- How participants were recruited (SONA, MTurk, Prolific) and compensated (academic credit, money)
- Was the experiment conducted in person?
- Participant language background (English native speaker, balanced Spanish-English bilingual, L2 learners of Dutch)
- Participant demographic information, e.g. sex or gender, age, etc.
- Other information relevant to the study (e.g. no eyesight or hearing impairment, right/left handedness)
- What exclusions were done (e.g. participants were excluded who did not complete all trials or who answered practice trials incorrectly)

Materials and Stimuli

Next, stimuli are normally described in a 'Materials' section. There are two types of stimuli in an experiment. **Critical items** are the stimuli that are testing the research question.

These are the questions included in the analysis. **Filler trials** are trials that usually consist of a slightly different task than the critical items (for example, asking participants to read active sentences instead of passive sentences). Filler trials make it more difficult for participants to figure out what the experiment is about and help reduce fatigue by mixing up the experimental task.

An experimental paper will often also report how stimuli were created and whether or not they were normed. Many papers will say whether stimuli were taken from a different paper, translated into a different language, or created with a particular method. Sometimes an experimenter will **norm** stimuli, which usually refers to showing all the stimuli to native speakers (somewhere between one and ten) and asking them to check the stimuli for naturalness or to make sure there are no errors. Sometimes norming may also include getting some kind of rating, like how positive/negative a word meaning is.

n is used to refer to the number of stimuli. An experiment with 48 critical items will be described with 'n=48.'

As discussed earlier with counterbalancing, the order in which the trials happen is important. Usually trials are **pseudorandomized**, which means that they are mixed up by an algorithm, so each participant sees the trials in a different order. The reason the term is pseudorandomize and not randomize is because randomization algorithms are never truly random (APA).

peudorandomization

Summary

When designing an experiment, it's important to take the following things into consideration:

- Based on your research question, pick the experimental task and measure carefully based on the question
- With the research question in mind, think about what your independent and dependent variables are. What is the factorial design of your experiment?
- What are your random variables? How will you control for possible confounds?
- When preparing your experimental materials, consider the number of stimuli and participants you need. These may be influenced by the experimental task, whether the experiment is repeated measures or not, and other aspects of the experiment?
- Should you norm your stimuli? What are the filler trials? What order are you presenting the trials in?
- How are you recruiting participants? In person or online? How are they being compensated? Do you need to add in attention checks?

Understanding these terms can also help you better engage with and evaluate experimental journal articles.

Acknowledgements

I used Kaiser (2013) as reference for a list of different types of experiments. Thank you to Rachel Miles and Ben Lang for feedback on this handout.

References

Kaiser, E. (2013). Experimental paradigms in psycholinguistics. *Research methods in linguistics*, pages 135–168.

Kliegl, R., Wei, P., Dambacher, M., Yan, M., and Zhou, X. (2011). Experimental effects and individual differences in linear mixed models: Estimating the relationship between spatial, object, and attraction effects in visual attention. *Frontiers in psychology*, 1:238.

Knight, M. and Mather, M. (2013). Look out—it's your off-peak time of day! time of day matters more for alerting than for orienting or executive attention. *Experimental aging research*, 39(3):305–321.

Oakhill, J. and Davies, A.-M. (1989). The effects of time of day and subjects' test expectations on recall and recognition of prose materials. *Acta Psychologica*, 72(2):145–157.

Salkind, N. J. (2010). *Encyclopedia of research design*. sage.

Schuman, H., Presser, S., and Ludwig, J. (1981). Context effects on survey responses to questions about abortion. *Public Opinion Quarterly*, 45(2):216–223.