

What makes one experiment good, and another not?

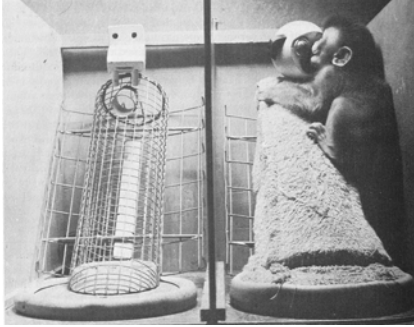
Four characteristics of sound experiments

- Internal validity
- Reliability
- Sensitivity
- External validity

Experimental control

- Taking care to investigate factors one by one
- When you do an experiment, there are many outside factors that you are not interested in that could influence your results. Part of doing good research is trying to control for these factors.

Harlow affect experiment



From Harlow, reprinted in Schaughnessy & Zechmeister, *Research methods in psychology*, 2nd ed.

Experimental control

- In order to make a valid comparison, all other factors must be controlled for.
- The problem is that we only choose to control those factors that we think are relevant.
- You need to be alert to the possibility of confounding factors in experiments whose influence wasn't controlled for.

Subject control

- Imagine I was examining whether people did better at recognizing speech in noise when the voice was familiar.
- The group that knew the talker made fewer errors.
- But the group that knew the talker consisted of 8 females & 2 males. The other group had 3 females and 7 males.
- Was it really knowing the voice that mattered, or was it that females are just better at the task overall?
- Was it really the gender difference that led to the effect?

How to control for outside factors

- Manipulation
- Holding constant
- Balancing

Manipulation

- If you purposely change a variable and find no effect, then you need not worry about that variable being a confound.
- This does not rule out interactions...

Holding constant

- Select groups such that variable is held constant.
- Holding constant a subject variable also means that your results may not generalize.
- Generally, you want to hold constant factors about the testing situation.

Balancing

- Balancing the influence across the different conditions.
- Balancing is not the same as holding constant – the variability is still there, but is equivalent in the two groups.

Examples

- On a CI study, testing only kids with a Nucleus 22 implant is holding the implant constant.
 - This reduces variability.
 - But results may not generalize to other implants.
- Having 8 boys, 2 girls per group is balancing.
 - If boys and girls behave differently, there would be a lot of within-group variability, making it less likely that the between-groups variability would be larger.

Subject factors

- Subject factors generally cannot be held constant.
- Ex: testing the effect of alcohol use on errors made in video game playing.
 - People differ in their experience playing video games.
- Ex: In a language study, people naturally vary vocabulary size, reading skill, etc...
 - These differences are not avoidable.
- When a factor cannot be manipulated or held constant, you try and control for it by balancing the factor across the different conditions.

Random assignment

- The most common technique for balancing is random assignment

Random assignment, cont.

- In a descriptive study, where your IV is a subject factor, random assignment is not possible.
- Ex: comparing elderly to young adults on a hearing task.
 - Age is the IV you are studying
 - But young and elderly adults differ in other ways
 - different experiences
 - different vocabulary
 - different motivations to do the task
- In cases like this, you may need to match the participants on the characteristics of interest

Nuisance factors

- Potential IVs that are not of interest but can confound
- Ex: in a subject pool, differences between Ss who participate early in semester vs. late
 - Those that volunteer early tend to be more academically oriented and on-the-ball
- Ex2: Differences in experimenter
 - Ss are likely to respond differently to different experimenters, even when experimenters try and treat Ss the same way

Nuisance factors, cont.

- Generally, holding nuisance factors constant leads to more sensitive experiments with less generality, while balancing leads to lower sensitivity but increased generality.

Things to control for

- Confounds due to differences between groups in testing conditions or procedures
 - You want differences in testing location, procedure, experimenter to be held constant or balanced
- Confounds due to assignment of subjects to conditions
 - Ex: sex, height, weight, attitude, personality, motor ability, mental ability...
 - If assigning subjects to comparison groups results in unequal distribution of subject-related variables, there is a possible threat to internal validity.

Experimenter biases

- Biases that make the experimenters behave differently
- Ex: If a study examines the effect of alcohol, the experimenter might read the instructions more slowly to people who have had alcohol, or might be more likely to notice unusual motor movements or slurred speech among the “drinkers”
- Experimenter should be blind to trial type

Subject biases

- Subject generally know they are in an experiment, and might guess what behavior is expected
- Demand characteristics
- Ex: if Ss knew they've been given alcohol, they might expect certain effects as being relaxed or giddy

Expectancy effects can influence results

- If the Ss know what the prediction is, they may try and follow it.
- The placebo effect
 - In a medical drug test, one groups of subjects actually gets the drug; another group thinks they are getting it, but actually gets a sugar pill (a placebo)
 - Most of the time, Ss getting a placebo report feeling better
 - And they almost always report adverse drug effects, like headache & insomnia

The placebo effect

- If Ss don't know that there is a placebo group, placebo response rates are as high as 70%
- Likelihood of getting a placebo influences results.
 - Study examined the efficacy of acetaminophen for postpartum pain.
 - Ss had a 1/2 chance of getting the drug vs. a placebo (and knew this).
 - Second study compared acetaminophen vs. naproxen (and Ss knew there was a 1/3 chance of getting a placebo).
 - The reported efficacy of the acetaminophen was lower in the first study than the second.

Experimenter bias effects

- Intons-Peterson compared perceptual scanning of maps with imaginal scannings of maps.
- Ss were asked to estimate distances between two cities while looking at a map or imagining the map.
- One group of experimenters were led to believe that Ss' estimates would be more accurate when imagining the map.
- The other experimenters were led to expect better estimates when the Ss were looking at the map.
- Ss' estimates varied with the experimenter's beliefs.

Experimental control

- For factors like testing rooms, you can either balance or hold constant.
- But for subject factors, balancing is really your only option.

Types of experimental design

- There are generally two types of experimental designs you can use to deal with these factors:
 - Between-subjects designs (independent groups)
 - Within-subjects designs

Between subjects

- Different subjects in different conditions
- IV is manipulated between groups of subjects

Within subjects

- Repeated measures designs
- Same subjects in all conditions of experiment
- IV is manipulated within a single group of subjects

Between vs. Within

- Between

- Within

Between subjects design

- Main control issue is ensuring equivalent groups of subjects (balancing)
- Nonequivalent groups are when there is some difference between the subjects in diff groups
- If the characteristics of subjects differ across the groups and this affects the measured values of DV, there is a confound
- The confound arises from how the subject characteristics are divided between conditions

How do you create equivalent groups?

- Most common way is random assignment
- Each subject has an equal chance of being assigned to each condition
- Subject characteristics should get spread evenly across various conditions

Random assignment versus random selection

- Selection has to do with how you choose samples of subjects from the population
- Random selection is thus a type of probability sampling where each member of the population is equally likely to participate in the experiment
- This is ideal statistically, in that you want your groups to represent the population as a whole.

Random selection, cont.

- If you don't randomly select them, there is no way to guarantee that your subjects do represent the population.
- In practice we almost never use random selection – we use convenience sampling.

Random assignment

- Assignment has to do with how subjects in a sample are divided into experimental conditions.
- Even if you didn't select the subjects randomly, you can at least assign them randomly to your different groups, in order to control for differences between groups

Simple random assignment

- Examples: flipping a coin, or using a random number table
- Problem: Different groups are unlikely to come out with equal numbers
 - If you flip a coin, not half will be heads.
- With small numbers, you want the number of observations per condition to be the same.

Block random assignment

- A block consists of one instance of each condition
- Subjects are randomized among them.
- Ex: with 2 conditions, each pair of subjects is a block – it is random which subject of the pair goes to which condition.
- Randomization of assignment within blocks results in equal numbers per condition.

Brady (1958) "executive monkey"

- Monkeys were trained to avoid an electric shock by pressing a lever whenever a light came on.
- Each press of the lever postponed the shock for 20 seconds.
- Monkeys were paired such that one monkey had access to the lever, but both received shocks together.
- If the "executive" failed to press the lever, both monkeys got shocked; the yoked monkey had no control over the shocks.

Executive monkey, cont.

- **Results:** All four executive monkeys developed ulcers, while none of the yoked monkeys did.
- **Conclusion:** The stress of making decisions caused the ulcers.

BUT....

Executive monkey, cont.

- How were the monkeys select for the executive positions?
Not randomly!
- Monkeys that were easy to train (high responders) were selected as executives.
- High responders are more prone to ulcers than low responders.
- When the study was repeated with animals equated for activity level, the animals that had control had less severe ulcers than those that do not have control.

Problems with random assignment

- Random assignment reduces the likelihood of systematic differences, but does not eliminate the possibility of chance differences
- It works better as sample size increases
- Problematic situation when there is a subject characteristic that will affect the DV and the effect is strong enough that slight differences in distribution will bias results

Problems with random assignment, cont.

- Ex: The effect of alcohol consumption on reaction times
 - There are 20 subjects, of whom 8 are on the tennis team (faster responses)
 - A 5/3 split may be sufficient to cause a confound
- Solutions
 - Use large # subjects
 - Use matched random assignment (matching)

Matching

- Rather than trust randomization to make comparable groups, the experimenter makes the groups comparable by matching individual members
- Examples: split-litter technique, siblings, homozygotic twins

Matching, cont.

- But most matching is done by determining factors that are likely to be relevant, and then selecting a pretest that equates the groups on that dimension.
- Ideally, pretest is the same task used in the experiment
- Example: if you want to see whether a blood test medicine works, you match subjects on their initial blood test and then one member of each pair takes the drug.

Matching example

- Study: effect of background noise on problem solving.
- 2 conditions, with & without noise, 5 subjects each.
- We expect overall academic ability to effect problem-solving scores, so it is critical to equate overall academic ability across groups.
- Thus, deliberately match subjects on academic ability, using GPA

Matching example, cont.

- Obtain scores on matching variable (GPA)
 - S1 3.24
 - S2 3.91
 - S3 2.78
 - S4 2.05
 - S5 2.62
 - S6 2.45
 - S7 3.85
 - S8 3.12
 - S9 2.91
 - S10 2.21

Matching example, cont.

- Sort in increasing order
 - Create pairs of adjacent scores
 - For each pair, randomly assign members to conditions.
- S2 3.91
 - S7 3.85
 - S1 3.24
 - S8 3.12
 - S9 2.91
 - S3 2.78
 - S5 2.62
 - S6 2.45
 - S4 2.05
 - S10 2.21

Groups are now matched

- | Group 1 | | Group 2 | |
|---------|------|---------|------|
| • S7 | 3.85 | • S2 | 3.91 |
| • S8 | 3.12 | • S1 | 3.24 |
| • S9 | 2.91 | • S3 | 2.78 |
| • S5 | 2.62 | • S6 | 2.45 |
| • S4 | 2.05 | • S10 | 2.21 |
- Average = 2.91
 - Average = 2.90

Difficulties in using matching

- Have to identify all Ss in advance
- Often difficult to find matches for all Ss
- Usually need 2 lab visits per subject (pretest needed to measure matching variable)
- Matching process may so severely restrict people included that it is not generalizeable

Difficulties in using matching, cont.

- Which variables should you match on? Are you sure you got them all?
- Matching ensures comparable groups only on the dimension being measured in the matching task.

Difficulties in using matching, cont.

- In practice, researchers try to avoid matching by using large enough groups of subjects
- When to use matching:

Other problems with a between-subjects design

- Subject loss

Types of subject loss

- Mechanical subject loss

- Selective subject loss

Example

- You want to test the effectiveness of a language training program for kids with SLI.
- You test SLI kids initially so as to match groups.
- One group participates in the intervention program
- You retest both groups to see if the group that participated improved.
- You start with 25 kids per group.
- But, a bunch of the kids drop out of the rigorous intervention program, and you're left with 15.
- Is it ok to compare these kids with the group that didn't participate?

No.

- The kids who dropped out may have done so because the tough program was frustrating.
- If so, the most likely to drop out would be the kids who had the most problems to begin with.
- So the ones remaining may have simply been the better-functioning kids among the 25.
- It then wouldn't be surprising if they did better on the final test.

- The problem isn't that the control group had more people, but that they had a different type of person.
- *(A control group is a group that does not get the intervention – it allows you to control for changes between the pretest and posttest that are not a result of the treatment.)*

How do you avoid this?

- Match the children in the two groups ahead of time. If a subject drops out of the experimental group, exclude that subject's match from the control group as well.
- This restores the comparability across groups, although the results will not be as generalizable.

Within-subjects design

- All treatments or conditions are given to each subject
- The same subject participates as a member of each group

Advantages

- Fewer subjects needed
- For short experiments, requires less time overall (combining multiple conditions in one session)
- Some areas of research require its use (longitudinal or learning studies)
- Best control of subject factors – the subjects in the different groups are most similar on other factors, since they are actually identical
- No possibility of subject factors being confounded with condition

Disadvantages

- Cannot be used to compare subject variables (effect of hearing loss, gender)
- Practice effects

Practice effects

- Subjects may change across repeated testings, even within a condition.
 - Skill development/learning
 - Fatigue or boredom
- Practice effects **MUST** be balanced in order to have a valid study

Example

- Do children perceive color as adults do?
- Task: arrange 15 colored caps in order.
- Several colors were tested.
- In early reports, 50% of 3-year-olds made errors on the color blue, but only 11% of 10-year-olds did.
- Suggested that the ability to detect differences in the color blue increased with age.
- But, the experimenters always tested blue last – and children quickly become bored with a repetitive task.
- When the order of colors were balanced, the effect disappeared.

Controlling for practice effects

- Randomization
 - Most common method
 - Risk of slight unevenness
 - With enough trials & subjects, fluctuations likely will wash out
- Block randomization
 - Ex: If you have 3 conditions, and 10 trials each, have the first 3 trials be one of each type (randomly ordered), then the next 3 be one of each type...

ABBA counterbalancing

- Conditions are presented in one order and then the order is reversed.
- Advantages
- Problems

Incomplete within-subjects design

- Conditions are not repeated
- For each subject, the order is confounded with the condition
- Must balance order effects ACROSS subjects (since you're not balancing them within)
- Each condition must occur equally often in each ordinal position

Three ways

- All possible orders
- Latin square design
- Random start point with rotation

All possible orders

- For three different items, there are 6 possible orders:
 - ABC, ACB, BAC, BCA, CAB, CBA
- The number of orders increases exponentially with the number of conditions.
- For N conditions, there are N! orders
 - $N! = N(N-1)(N-2)\dots$
 - So for 3 conditions, there are $3 \times (3-1) \times (3-2)$ or $3 \times 2 \times 1 = 6$ orders
 - For 5 conditions, there are $5 \times 4 \times 3 \times 2 \times 1$ or 120 orders!
- So the use of all possible orders is usually restricted to cases where you don't have many conditions.

Latin square design

- A Latin square is a means of controlling order effects while still using only a subset of the orders
- It is a particular design involving as many orders as conditions

Building a Latin Square

- The first order: 1, 2, n, 3, n-1, 4, n-2, 5, n-3, etc.
- Each following row is generated by adding 1 to each number in the previous row
- For 6 conditions, the Latin square would be:

1	2	6	3	5	4
2	3	1	4	6	5
3	4	2	5	1	6
4	5	3	6	2	1
5	6	4	1	3	2
6	1	5	2	4	3

Random starting point with rotation

- Begin with a random order
- Systematically rotate this sequence to create other orders
- Disadvantage:

Characteristics

- All possible orders
 - Each condition occurs at each ordinal position equally often
 - Each condition precedes and follows every other condition equally often
 - Each condition precedes and follows every other condition equally often at each ordinal position
- Latin square
 - Each condition occurs at each ordinal position equally often
 - Each condition precedes and follows every other condition equally often
- Random starting point with rotation
 - Each condition occurs at each ordinal position equally often

Limitations of within subject designs

- Cannot be used to study subject variables
- Practice effects
- Differential transfer

Example of differential transfer

- Study examining the effect of familiar voices.
- Three different conditions:
 - Novel voice
 - Familiar voice, but subjects not told this (implicit familiarity)
 - Familiar voice, explicit knowledge

Differential transfer

- Different from progressive carryover effects, such as learning or fatigue
 - Progressive effects are approximately equal whichever condition comes first
- Counterbalancing can only control for progressive carryover effects, not differential ones

Within-subjects designs

- More powerful statistically
 - because you are doing a better job of controlling for variables such as subject characteristics
- Has the risk of order effects
- If you suspect differential transfer, you can't do within-subjects experiments at all

When to use which design

Use within subjects if

- Differential carryover not expected
- It is not possible to create equivalent groups

Use between subjects if

- Differential carryover is expected
- Equivalent groups can be created

When to use which design, cont.

- Must use within subjects if learning is the focus
- Must use between subjects if the IV is a subject variable

Within or between?

- Psychologist hypothesizes that spaced practice of verbal info will lead to greater retention than massed practice.
- Neuroscientist believes that damage to primary visual cortex is permanent in older animals but not younger animals.
- Researcher predicts students use more slang when talking with a peer than an older adult.
- Child psychologist predicts that cloth diapers lead to faster toilet training than paper diapers (beginning with day-old infants).

Single-subject designs

- Each individual subject is studied in depth.
- Often called time-series studies.
- Do not generally have only one subject.
- The individual results of each participant are analyzed, rather than just group averages.

Single-subject designs, cont.

- Usually involves measuring behavior over time without intervention first, to allow for measurement of variability/stability.
- Then the subject's behavior is measured across time during treatment.
- Sometimes this will then be followed by another baseline condition (to evaluate whether the treatment effect lasts while treatment is not being continued), or may be followed by a baseline and then a different treatment.

Single-subject designs, cont.

- In most tasks, people will not always respond identically
 - emotional state, hunger, motivation, etc. can influence how a person does on any one test
- In group designs, you average subjects so the data is reliable.
- In single subject designs, you do not, so you instead need to have multiple samples from the same condition.

Single-subject designs, cont.

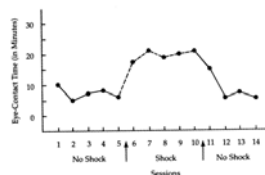
- Advantages
- Concerns

Single-subject designs, cont.

- Why do it?
 - Sometimes group data is very misleading - no person may really behave anything like the group average
 - The reason psychologists started using groups was because single subjects' were often very variable; but sometimes the variability is not inherent in the participant, but is a result of the experimenter's failure to control all the variables. If the experimenter does a better job controlling outside variables, this is less of an issue.

How to do single-subject designs

- First, establish a baseline, or steady state
 - You want to continue to test until response rate remains constant across sessions
- Then add experimental manipulation
 - Continue with that variable until you reach a stable transition steady state
- Then remove experimental manipulation
 - Should return to baseline (should be reversible)
- Can repeat it to make it more convincing.



Martin, D.W. (1985) Doing Psychology Experiments, 2nd Ed., p. 89

Single-subject designs, cont.

- Logic of method:
 - Once you have reached baseline, it is unlikely that any uncontrolled variable would suddenly change the person's behavior at the exact time that the experimenter added the manipulation
 - Even if it did, the likelihood that it also ended at the exact time that the experimenter ended the manipulation is very unlikely.

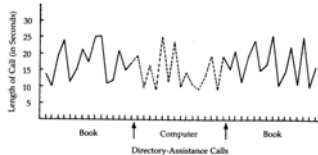
Single-subject designs, cont.

- Advantages:

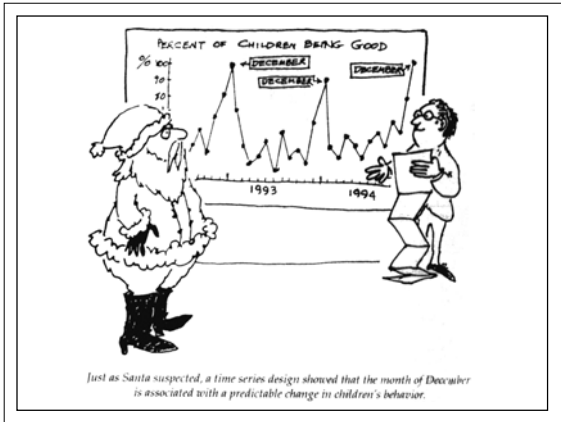
- Disadvantages

Example of variability masking an effect

- Imagine you want to see whether the time for directory assistance is reduced by using a computerized system over a telephone book.
- You record the length of each call with one system, then switch, then switch back.
- There looks like there is no effect here - but the average time with the computer is really 3 seconds less - this could have been significant in a standard experiment (and is important, since it could save the phone company lots of \$\$).



Example from on Martin, D.W. (1985) Doing Psychology Experiments, 2nd Ed.



Which one to use

Single-subject designs

- Tell you about typical responses from a given individual

Group designs

- Tell you about behavior of a typical group member.
- Best if you want to know about how well a something works.
- Only make sense when differences among people are of degree, not of kind.

Examples of poor selection

- Cats falling from tall buildings
- Data came from 132 victims admitted to the Animal Medical Center
 - The longer the fall, the greater the chance of survival
- Why did cats from higher floors fare better than those on lower ones?

Example 2

- Minneapolis Star Tribune (March 6, 2001)
 - study shows that adults with hobbies that exercise their brains are 2.5 times less likely to have Alzheimer's disease, while leisure limited to TV watching can increase risk
 - investigators surveyed people in their 70's to get information about leisure activities in young adulthood--ages 20 to 39--and middle adulthood--ages 40-60.
 - 193 Alzheimer's patients (cases) and 358 people who did not have Alzheimer's disease (controls).

Example 3

- Every year ETS puts out a listing of average SAT scores for each state.
- Which state almost always has the highest average?
IOWA
- Why?
