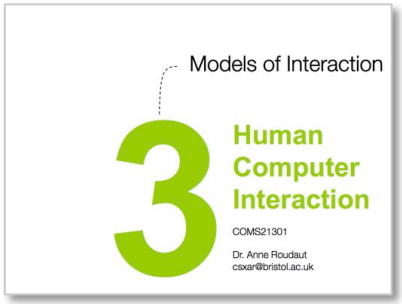
Designing an experiment

5 Human Computer Interaction

COMS21301

Dr. Anne Roudaut
csxar@bristol.ac.uk



guess

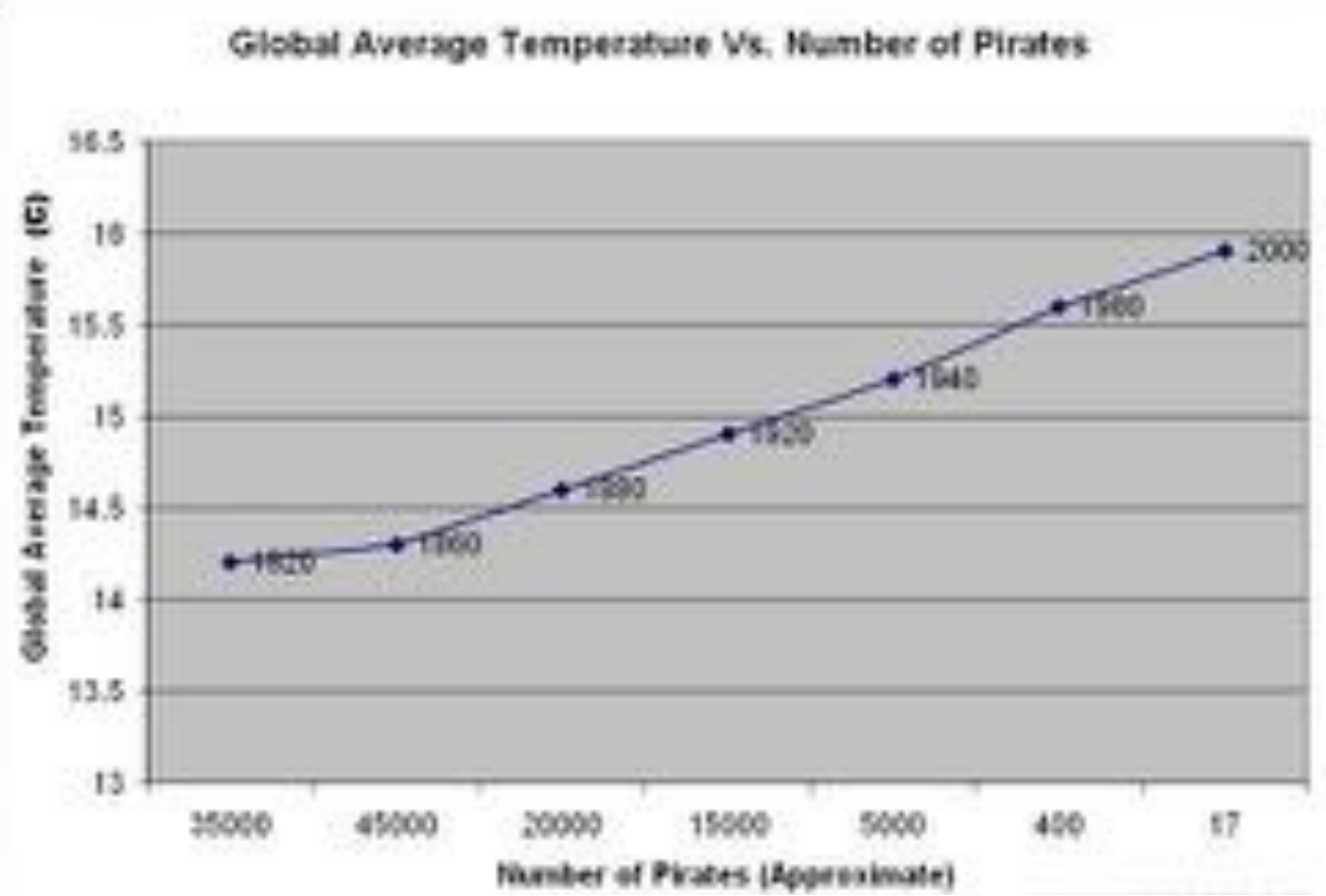
(repeated)
observation

theories
models

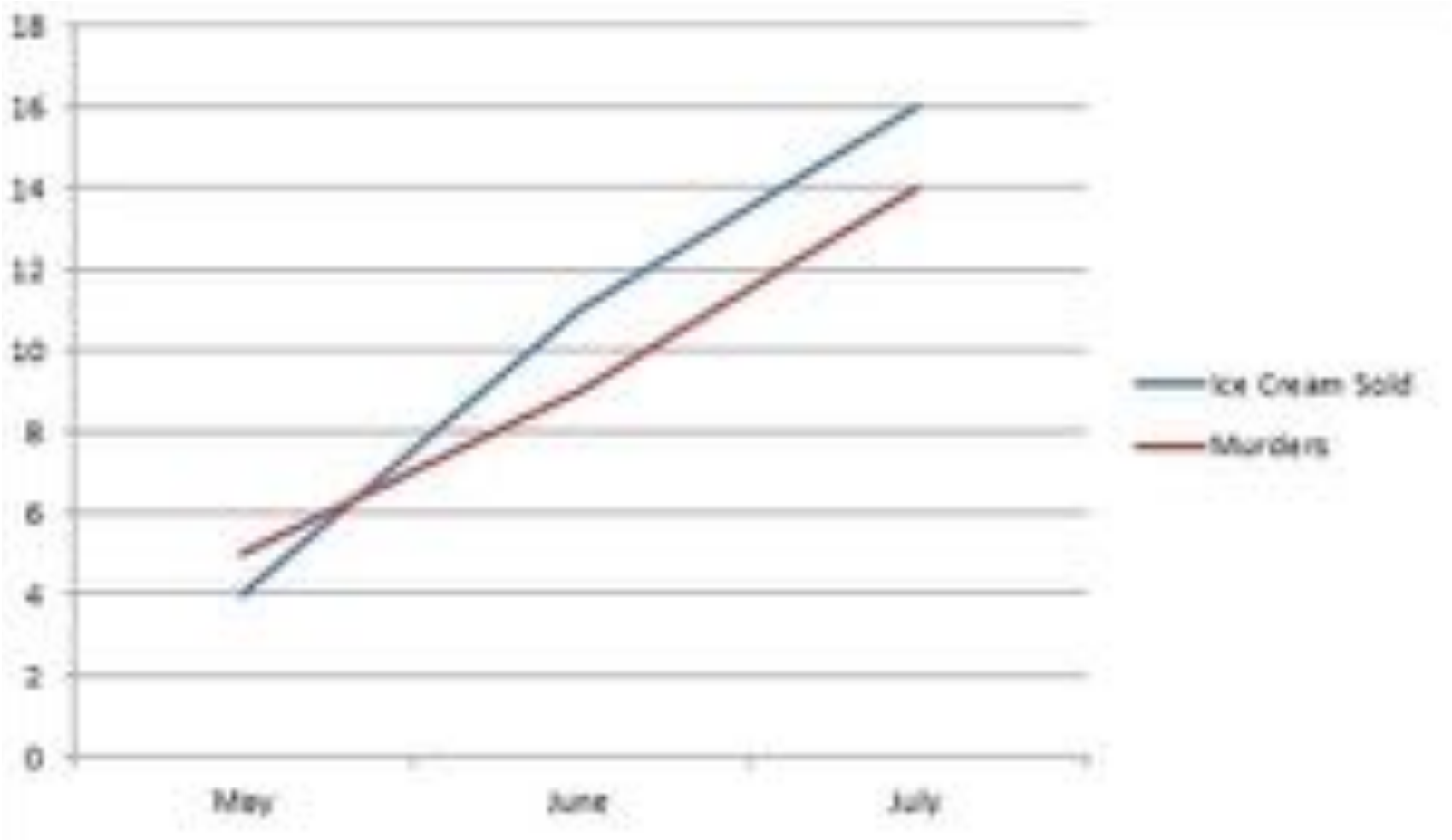
coursework
session 2

derive a prediction
= hypothesis

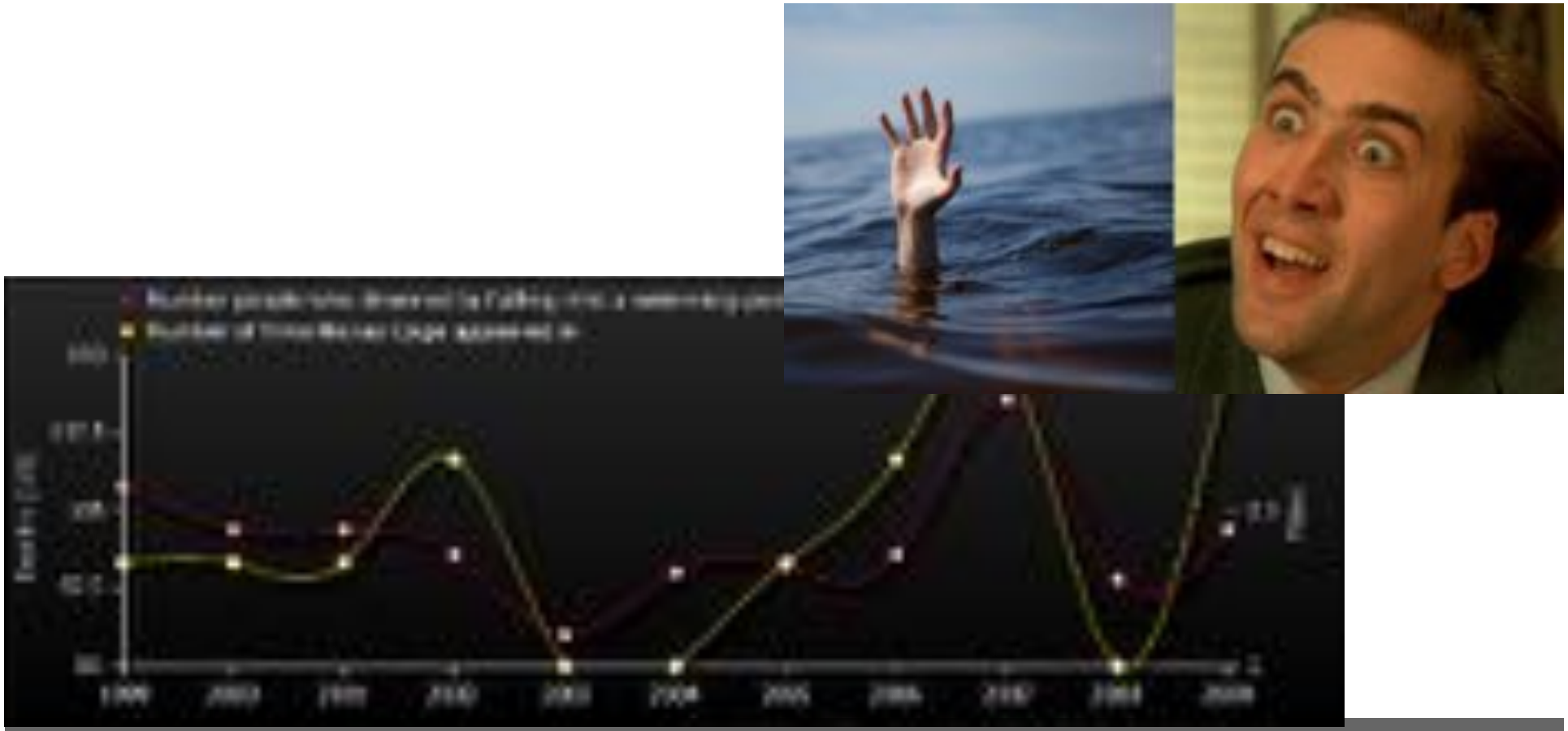
(implement & conduct
experiment
= compare to nature



a pirate shortage caused global warming



ice cream consumption leads to murder



number of people drowned by falling into a swimming-pool correlates with number of films Nicolas Cage appeared in

this lecture is not about **correlation**

this lecture is about how to show **causality,**
i.e., that **some A causes some B**

example

A

a company is offering a new set of herbal supplements they claim to **help with depression**

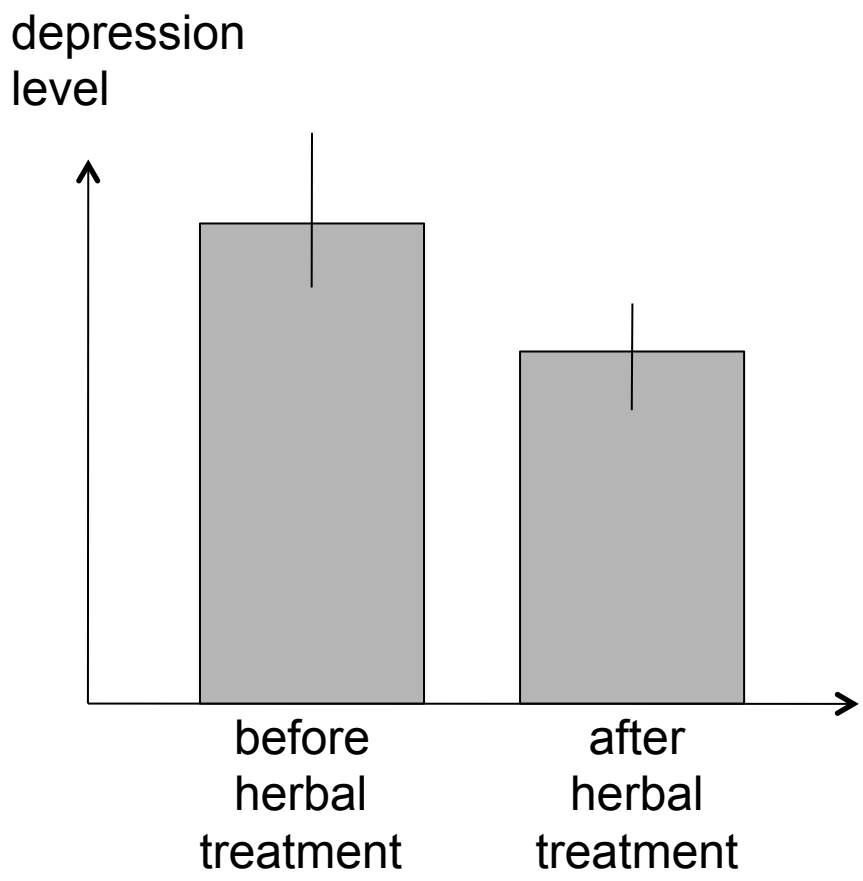
B

design an experiment that tests this claim

<60sec brainstorming> (wait, don’t tell me)

how about...

1. recruit participants
2. **measure** their depression level (e.g., self-assessed using questionnaire)
3. administer the supplement
4. **measure** their depression level again



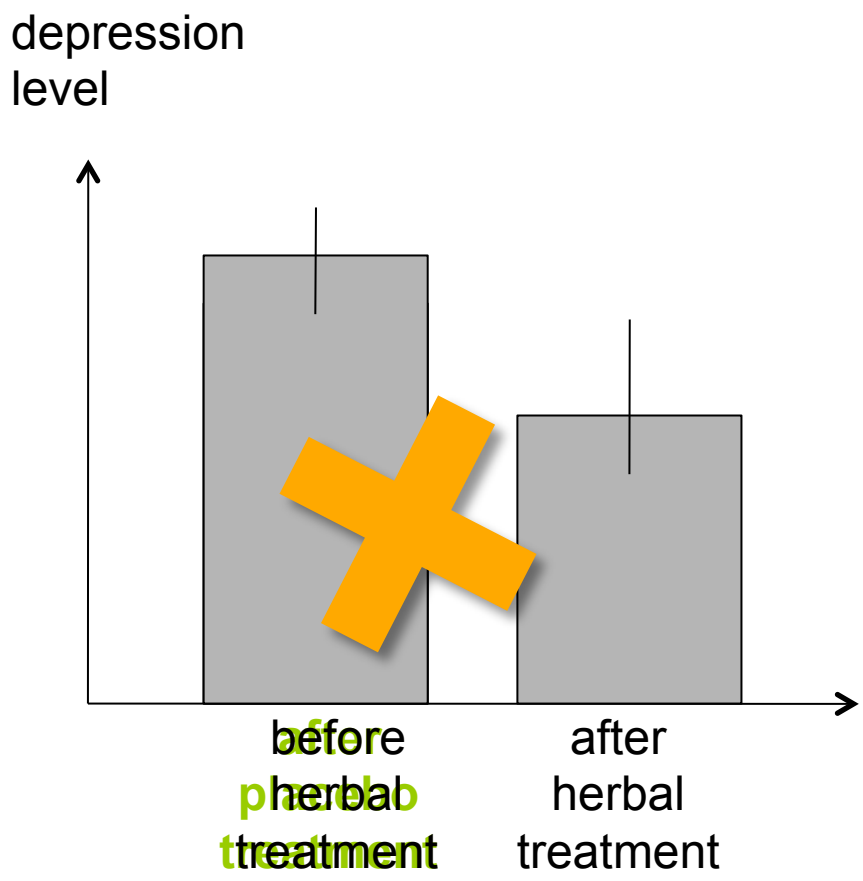
nope.
ok, so what is the **challenge** here?

placebo effect ::

the tendency of any medication or treatment, even an inert or ineffective one, to exhibit results simply **because the recipient believes** that it will work

Instead of comparing with depression level before, compare with a **control condition** with participants that use a **placebo**.

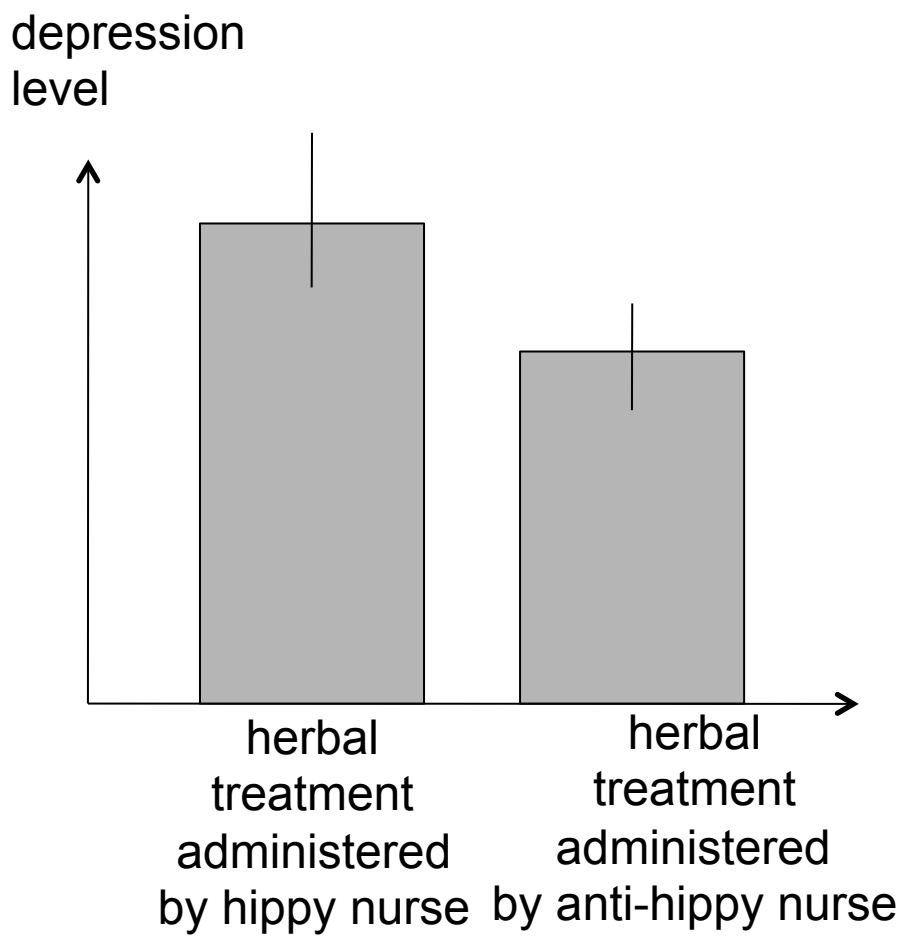
to get the placebo effect participants must not know what experimental condition they are in **("blind" experiment)**



blind ::

a scientific experiment where some of the persons involved are **prevented from knowing certain information** that might lead to conscious or subconscious bias on their part, invalidating the results.

what if the nurse administering the treatments (subconsciously) has a **preferred outcome** (either loves herbal medication or hates it)?



nurse must not know either what is being administered

→ **double blind experiment**

double blind ::

an especially stringent way of conducting an experiment, usually on human subjects, in an attempt to eliminate subjective bias on the part of both experimental subjects and the experimenters.

neither the individuals nor the researchers know who belongs to the control group and the experimental group.

only after all the data have been recorded (and in some cases, analyzed) do the researchers learn which individuals are which.

random assignment of the subject to the experimental or control group is a critical part of double-blind research design.

The key that identifies the subjects and which group they belonged to is kept by a third party and not given to the researchers until the study is over.

showing causality

so we want to show that **A causes B**

here is what to do:

1.correlation: show that a change in A occurs with a change in B

2.order: show that A takes place before B

3.no hidden cause: show that there is no C with C→A and C→B

the reason for the complexity of experimental design is #3, i.e., to show **that there is “no hidden cause”**

that’s why we (1) run studies in the lab without external influences, (2) assign participants randomly to conditions, and so on

experiments
have three types
of variables

vary A → make A
an **independent variable**

so we want to show that **A causes B**

measure B → make B
a **dependent variable**

(in)dependent variable ::

the **dependent variable** is the event studied and expected to change whenever the **independent variable** is altered

everything else should be a…

controlled variable ::

the variables that are kept constant to prevent their influence on the effect of the independent variable on the dependent.

avoid…

confounding variable ::

extraneous statistical variable in a statistical model that **correlates with both** the dependent variable and the independent variable

the goal of a quantitative study is to find **a signal** in **a lot of noise**

e.g., “A→B” or “A != B”

variance, outliers, sequence effects…

experimental design:

aims at maximizing your chances of **finding the signal** and not the noise

1. need to absolutely **avoid systematic biases** (e.g., the motivation factor, fatigue). They can give you **false results!**

2. **avoid random noise.** It makes your results non-significant. Clever experimental design is all about keeping the noise down

within vs. between subjects design

you want to compare your new mouse design with last year’s model

the new model has a new feature (dual-speed switch or whatever), so you **hypothesize that participants perform faster using the new model.**

You have 24 participants. How do you proceed?

which design do you prefer?

[ ] 12 participants use old mouse, 12 new mouse

[ ] all 24 participants use both mice

between subjects

within subjects

why?

<30 sec brainstorming>

“within subjects” gives you additional information and get significance with fewer participants

however not all the time possible, e.g. right handed vs. left handed

Within groups design

- Each subject performs experiment under each condition.
- Less costly and less likely to suffer from user variation
- Statistical power with smaller number of participants
- Demands more time from each subject
- Transfer of learning possible

Between groups design

- Each subject performs under only one condition
- No transfer of learning
- More users required
- User variation can bias results

telling within/between subjects to your t-test…

= TTEST(B1:B100,C1:C100,2,2)

data set 1 data set 2

1: paired t-test = **within subjects**

2: unpaired t-test, a two-sample equal variance test **(between subjects)**

one vs. two things to test (tails)

= TTEST(B1:B100,C1:C100,2,2)

#tails

one-tailed (non-directional) := A cannot be slower than B; is at least as fast or faster (no need to test A slower B; you are only testing B slower A)

two-tailed (directional) := A faster B possible, as well as B faster A

you will almost always use this one

you are developing a new drug for hair loss.
for completeness you also check side-effects on cholesterol

one-tailed or two-tailed?

two-tailed. Who knows this might increase hair loss.

<30sec brainstorming>

2

you are checking if a person responds to a (previously tested) cholesterol-lowering drug

you are already confident that it does not raise cholesterol

one-tailed or two-tailed?

dealing with sequence effects

so you are comparing reaction times with your neighbor

sequence effects:

1. training → better go late in class
2. fatigue → better go early in class
3. second participant knew time to beat→ better go second

how do you prevent these from influencing results?

<30sec brainstorming>

when using between subjects, we can

1. avoid sequence effects, i.e., create identical conditions for all users (in this case: don’t let them know about each other)

but when using within subjects

2. counterbalancing: make sure each interface sees the same amount of sequence effects

counterbalancing ::

a method of avoiding confounding among variables. Counterbalancing is performed by placing participants in groups and **presenting conditions to each group in a different order.**

one approach to counterbalancing is to use a…

A	B	C
C	A	B
B	C	A

Latin square ::

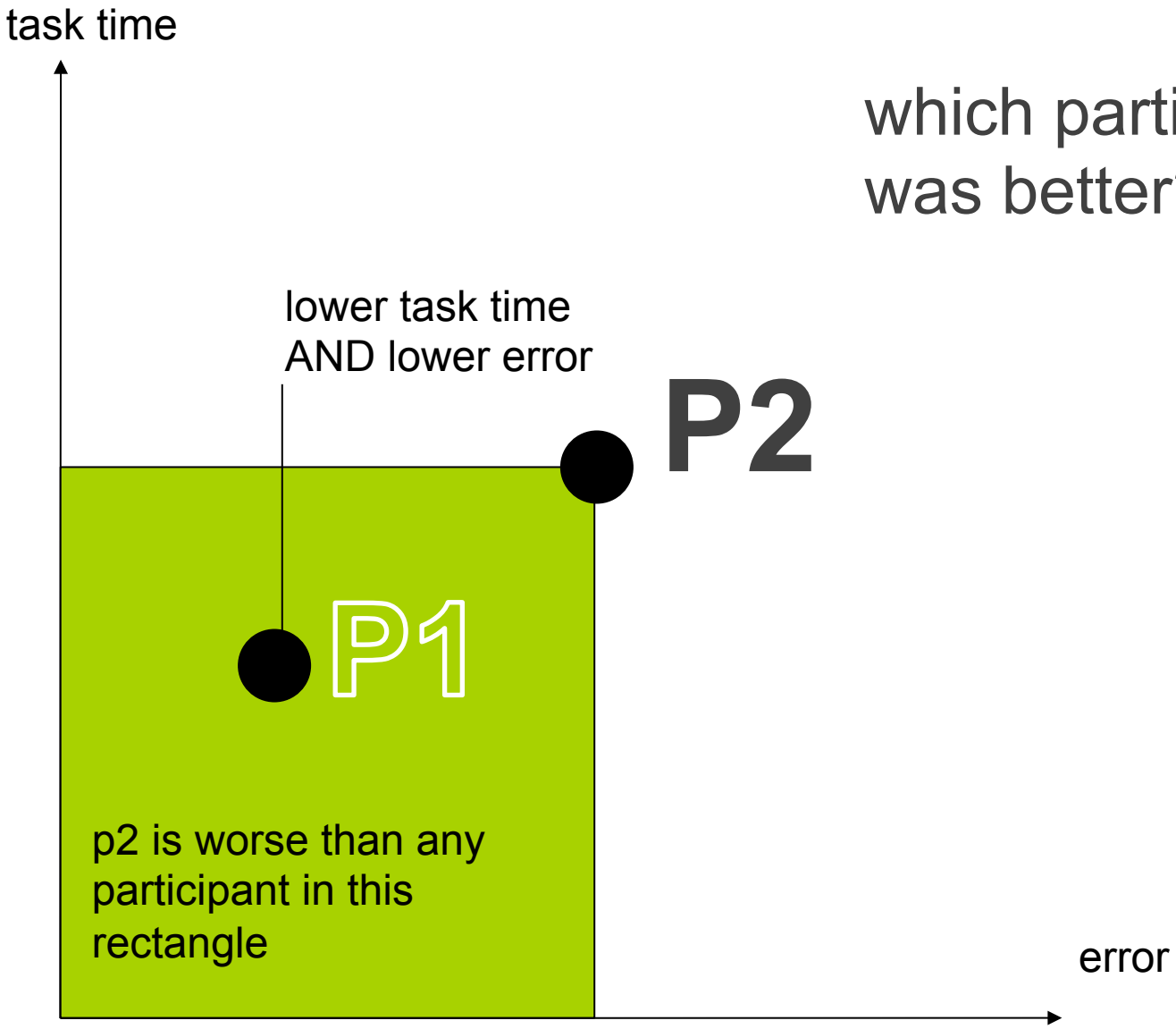
an $n \times n$ array filled with n different Latin letters, each occurring exactly once in each row and exactly once in each column.

task time, but what about error?

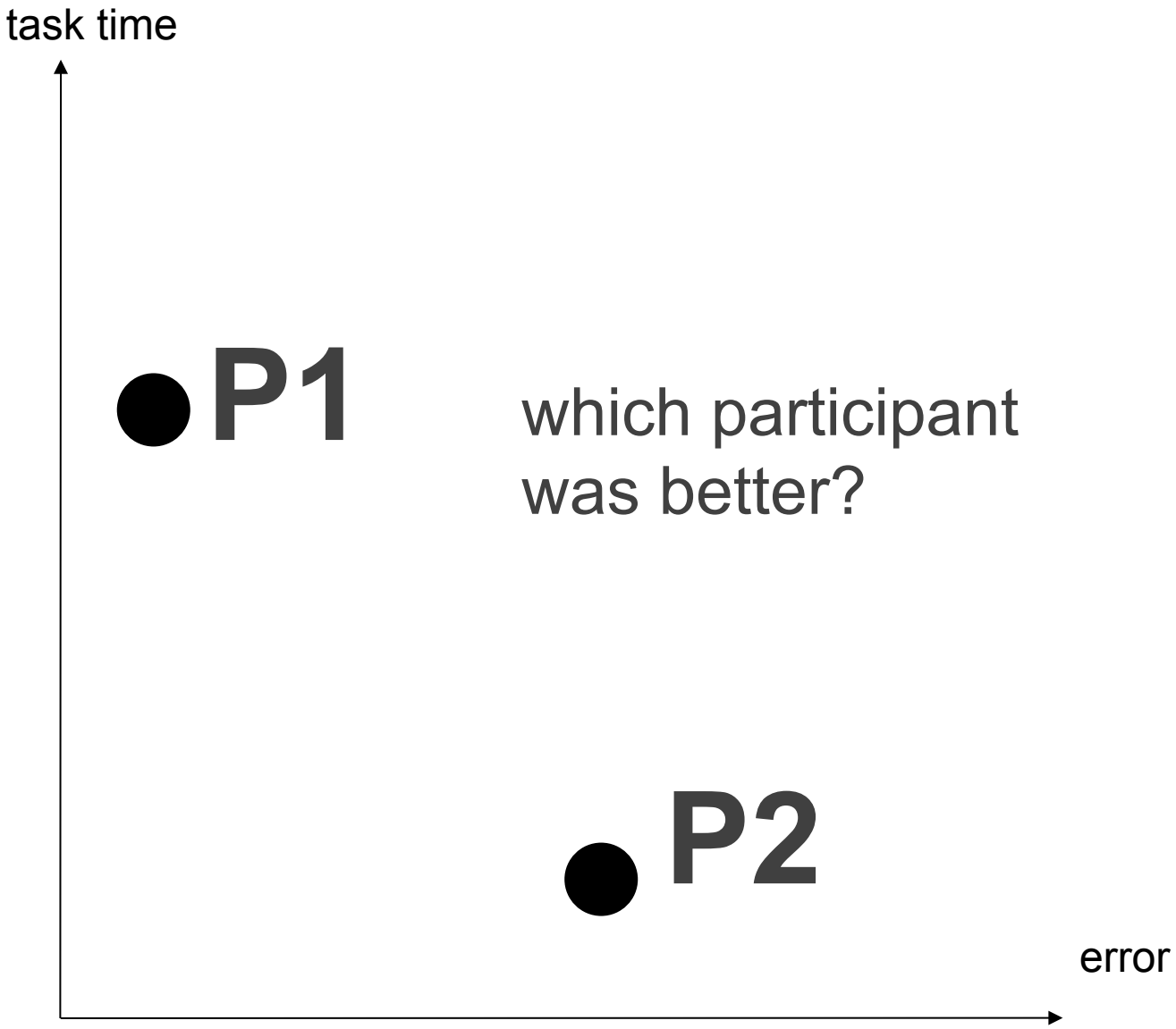
so far, we have **only** talked about task time.
however in most case you will also want to know about **errors**

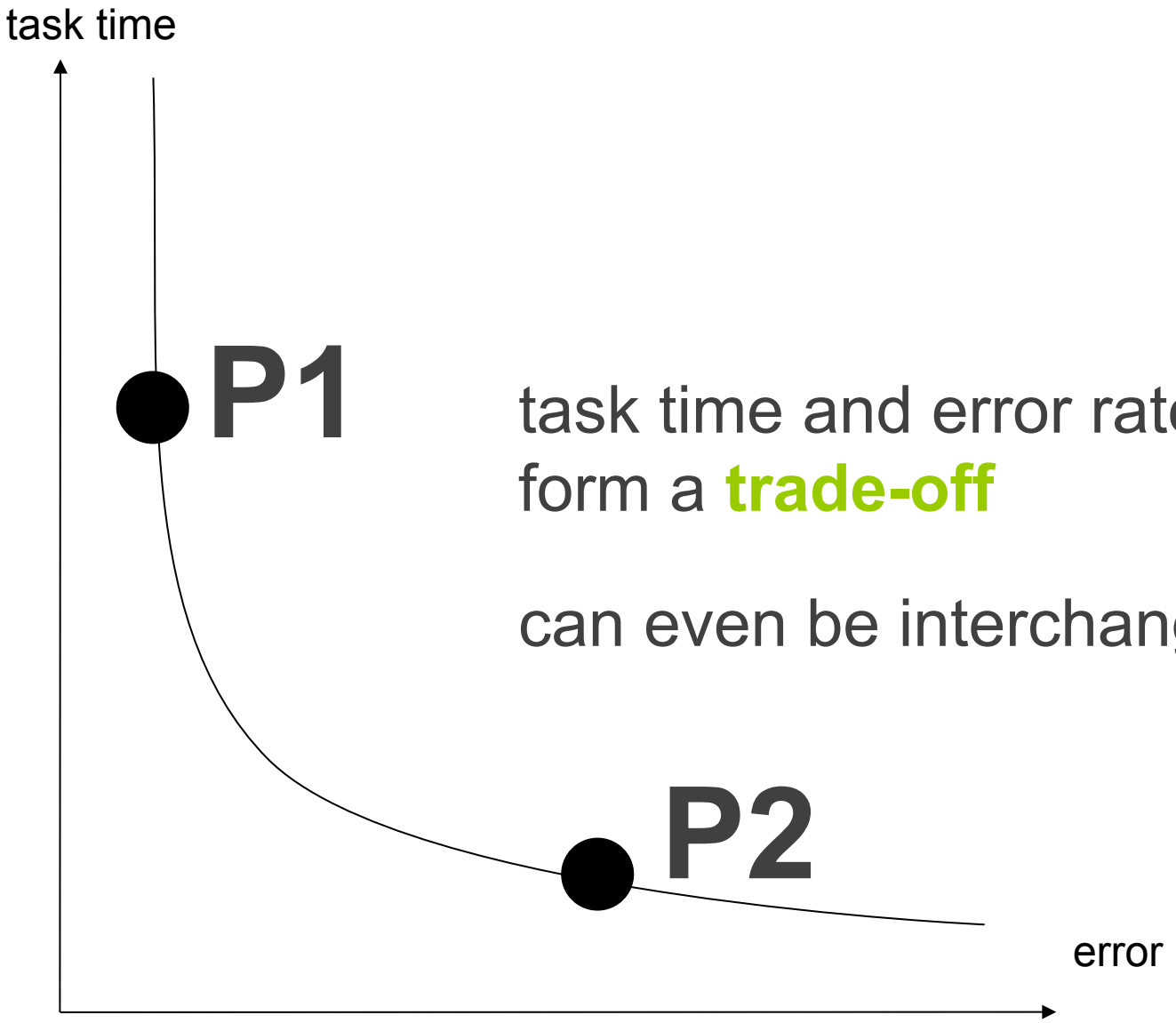
if there is no penalty for error, participants can improve their task time by slamming the keyboard randomly

→ we also **need to consider error rate.**
→ each trial is effectively a (task time, error) pair



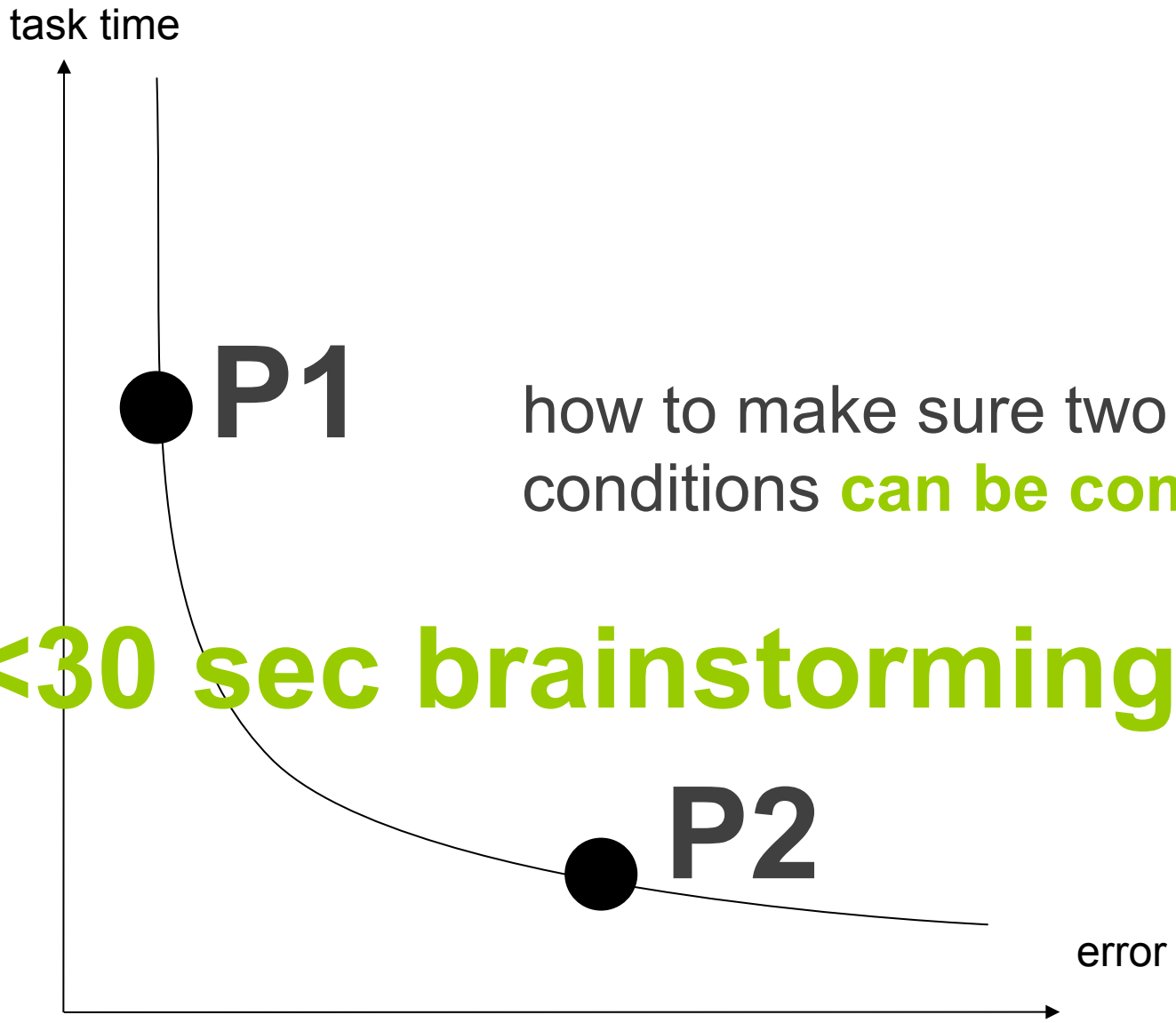
which participant
was better?


task time
P1
which participant
was better?
P2
error



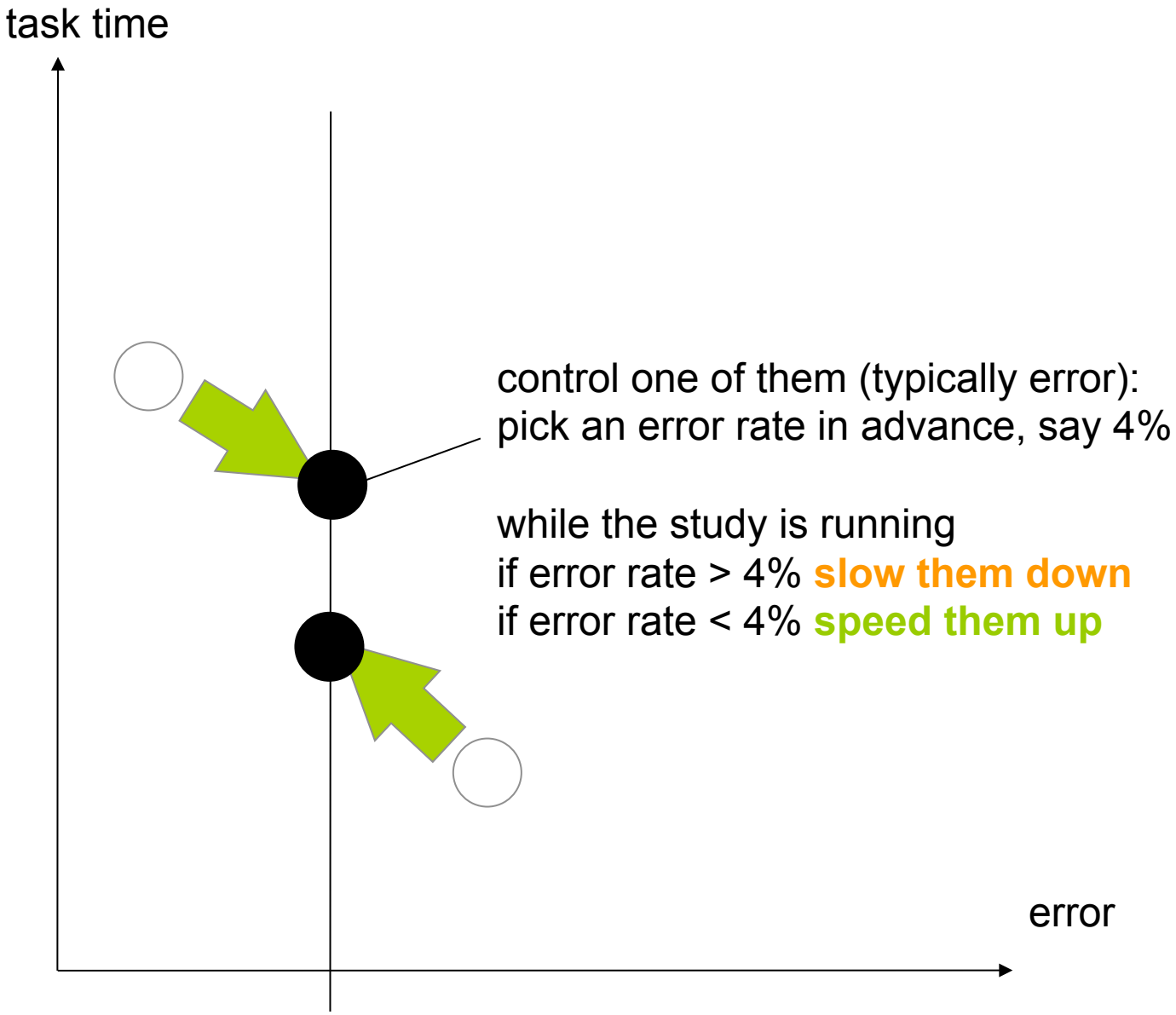
task time and error rate
form a **trade-off**

can even be interchangeable



how to make sure two conditions **can be compared?**

<30 sec brainstorming>


task time
error
control one of them (typically error):
pick an error rate in advance, say 4%
while the study is running
if error rate > 4% slow them down
if error rate < 4% speed them up

less common: **keep task time constant**
(e.g., metronome study [Wobbrock & Cutrell])

ok you should now be ready to run you own xp but let’s do a

quick summary

1. define your **research question**, e.g. “I believe that X”

ideomotor actions can express nonconscious knowledge

1. define your **research question**, e.g. “I believe that X”

2. derive **hypothesis** If X is true then we should observe a,b,c

H1. when they are not sure of the answer

participants with the Ouija >> participants without

H2. when they know the answer

participants with the Ouija == participants without

cannot show things are similar

Consider the following questions about a new technique:

Is it viable?
Is it as good as or better than current practice?
What are its performance limits and capabilities?
What are its strengths and weaknesses?
Does it work well for novices, for experts?
How much practice is required to become proficient?

The preceding questions, while **unquestionably relevant and interesting**, are **not good empirical research questions**

Empirical - capable of being verified or disproved by observation or experiment. (Websters dictionary)

Very weak (in an empirical sense)

Is the new technique any good?

Weak

Is the new technique better than X?

Better

Is the new technique faster than X?

Better still

Is the new technique faster than X within 1h of use?

Even better

If error rates are kept under 2%, is the new technique faster than X within one hour of use?

Hypotheses::

A statement of the predicted relationship between at least two experimental variables. **A provisional answer to a research question.**

Question: How does having information on the context of a caller affect whether the receiver picks up the call?

Hypothesis: Receivers will be more likely to pick up when they have information on callers' context than they will be when they do not.

Good Hypothesis Formation

Testable: the means for manipulating the variables and/or measuring the outcome variable must potentially exist

Falsifiable: must be able to disprove the hypothesis with data

Parsimonious: should be stated in simplest adequate form

Precise: should be specific (operationalized)

Useful: relate to existing theories and/or “point” toward new theories. It should lead to studies beyond the present one (often hard to determine in advance)

1. define your **research question**, e.g. “I believe that X”

2. derive **hypothesis** If X is true then we should observe a,b,c

3. create your **experimental design**

what are you variables (ind/dep)

what is your task

what is your design (within/between)

1. define your **research question**, e.g. “I believe that X”

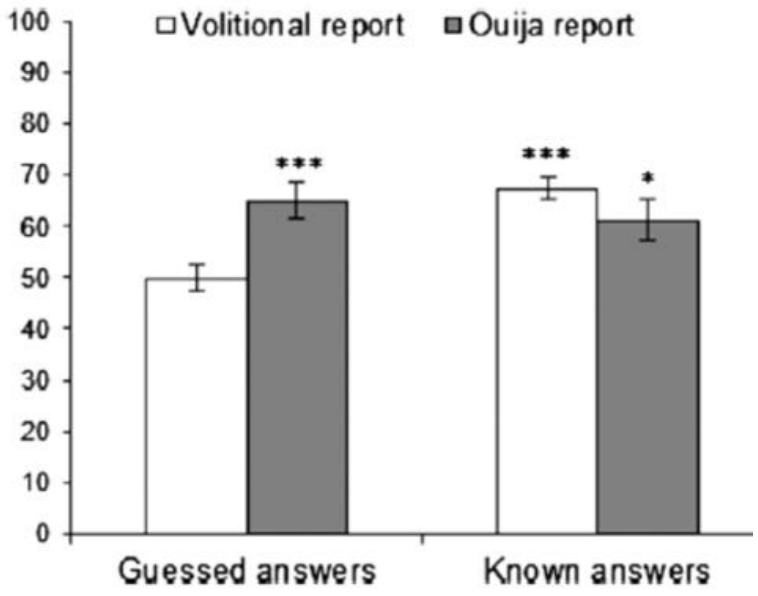
2. derive **hypothesis** If X is true then we should observe a,b,c

3. create your **experimental design**

what are you variables (ind/dep)

what is your task

what is your design (within/between)



4. **analyze** your data

look a raw data and check for bugs

compare things (ttest / anova / bonferroni)

models things (eventually)

last tips

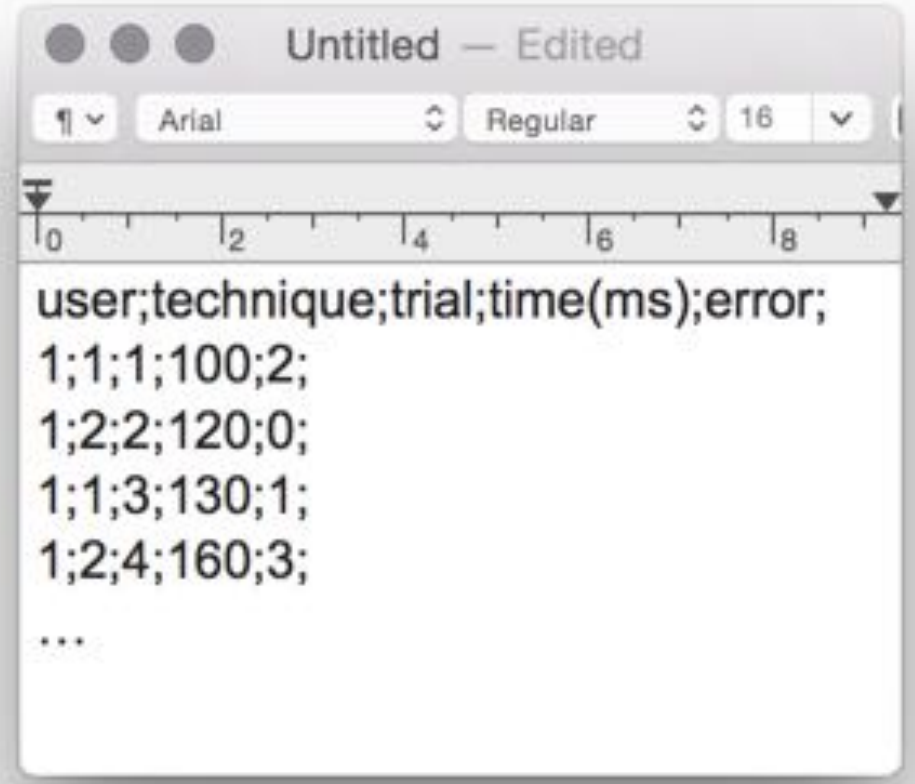
Design the experiment in such way that the results will be easy to analyze (if possible!).

Simple experiments need simple analysis.

Before performing a complex experiment, be sure you will be able to perform the appropriate statistical analysis.

Be sure the method used in your analysis can handle the type of the data of the dependant variable.

to gather your data use a **csv file format** (comma separated values). Used in most statistical soft and can be opened in excel.



	Interval/Ratio (Normality assumed)	Interval/Ratio (Normality not assumed), Ordinal	Dichotomy (Binomial)
Compare two unpaired groups	Unpaired t test	Mann-Whitney test	Fisher's test
Compare two paired groups	Paired t test	Wilcoxon test	McNemar's test
Compare more than two unmatched groups	ANOVA	Kruskal-Wallis test	Chi-square test
Compare more than two matched groups	Repeated-measures ANOVA	Friedman test	Cochran's Q test
Find relationship between two variables	Pearson correlation	Spearman correlation	Cramer's V
Predict a value with one independent variable	Linear/Non-linear regression	Non-parametric regression	Logistic regression
Predict a value with multiple independent variables or binomial variables	Multiple linear/non-linear regression		Multiple logistic regression

http://yatani.jp/HCIstats/HomePage

coursework

12th October

form teams, discuss study topics

19th/21th October

presentations of ideas (5 slides) & feedback

16th/18th November

building complete (software, procedure), run studies

30 November/1st December

presentations of results (graph) & feedback

you goal is to design and run a controlled experiment with human participants which tests the role of physicality in an interactive setting

your control group will experience a virtual setting (assuming you hypothesise that physicality is more valuable, not less)

you can exactly replicate an existing study to verify its results, or you can design an innovative study based on an existing one

Meet in your groups to create an experimental design
Your submission should total 5 slides:

a. hypothesis (1 slide)
What is your hypothesis? Is it precise enough that it can be tested?
If your hypothesis is vague/non-incremental it will be hard to verify

b. independent Variable(s) (1 slide)
What are you testing, and what are you comparing it against? Is this the most stringent/appropriate comparison you could run? The more IVs you have the more complex to control your procedure is likely to be

c. dependent Variable(s) (1 slide)
How are you measuring your test? What form(s) of data are you collecting? Can you directly sample data or do you need to calculate it from multiple samples (e.g. pre- and post- tests). Do not gather data which does not address your hypothesis, random searches for patterns which are not covered by your hypothesis confound your data.

d. procedure/experimental design (2 slides)
Are you measuring between- or within-subjects? Do you need to worry about counterbalancing? Is your procedure *valid* (i.e. could someone else replicate your results consistently)? Is your procedure *reliable* (i.e. do the data sufficiently address your hypothesis)? Is your procedure ethical? What kind of an environment do you need to build/configure in order to ensure your procedure can be followed? What kind of statistical tests will suit analysis of your data (this last question does not have to be answered at this stage, it can wait till stage 2).

Complicated procedures tend to require more participants and may be less likely to find statistically significant results. In general, control is better than measurement for unimportant factors.

Submit your slides (pdf or ppt) by Monday 19/10/2015 9am at csxar@bristol.ac.uk

Slides available at https://goo.gl/B1SZTe

end