UNIVERSITY OF BIRMINGHAM

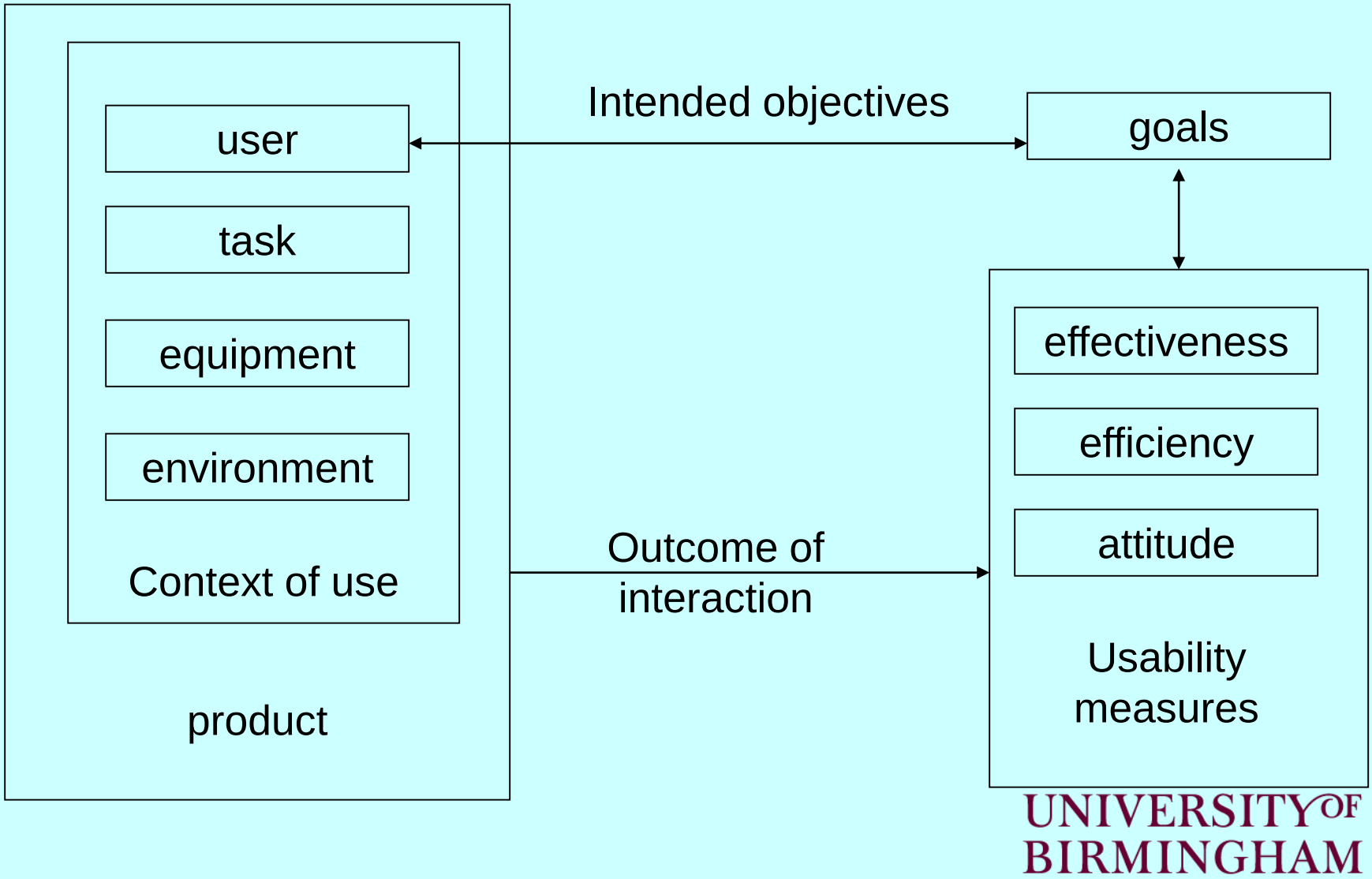
School of Computer Science

EVALUATION METHODS AND STATISTICS

Prof. Chris Baber

Chair of Pervasive and Ubiquitous Computing

ISO 9241 Context of Use



UNIVERSITY OF BIRMINGHAM

EXERCISE#2

- measurements on each scale for Usability…

Usability Measure	What is measured?	Nominal	Ordinal	Interval / Ratio
Effectiveness	How well has a task be performed?			
Efficiency	How well has performance used the resources available?			
Attitude	How good do users think the performance has been?			

Possible answers to Exercise#2

	Nominal	Ordinal	Interval	Ratio
Efficiency	Helpful / Not	Likert scale of 'helpfulness'	Proportion of task completed	Time completion time
Effectiveness	Easy / Hard	Subjective Workload		Number of errors; physiological measures of effort/ attention
Attitude	Like / Dislike	Software Usability Scale; Software Usability Inventory Metric		

EXERCISE#3

name the colour of the ink in which these words are printed as quickly as possible

Yellow

UNIVERSITY OF BIRMINGHAM

Blue

UNIVERSITY OF BIRMINGHAM

Red

UNIVERSITY OF BIRMINGHAM

Green

UNIVERSITY OF BIRMINGHAM

Blue

UNIVERSITY OF BIRMINGHAM

Green

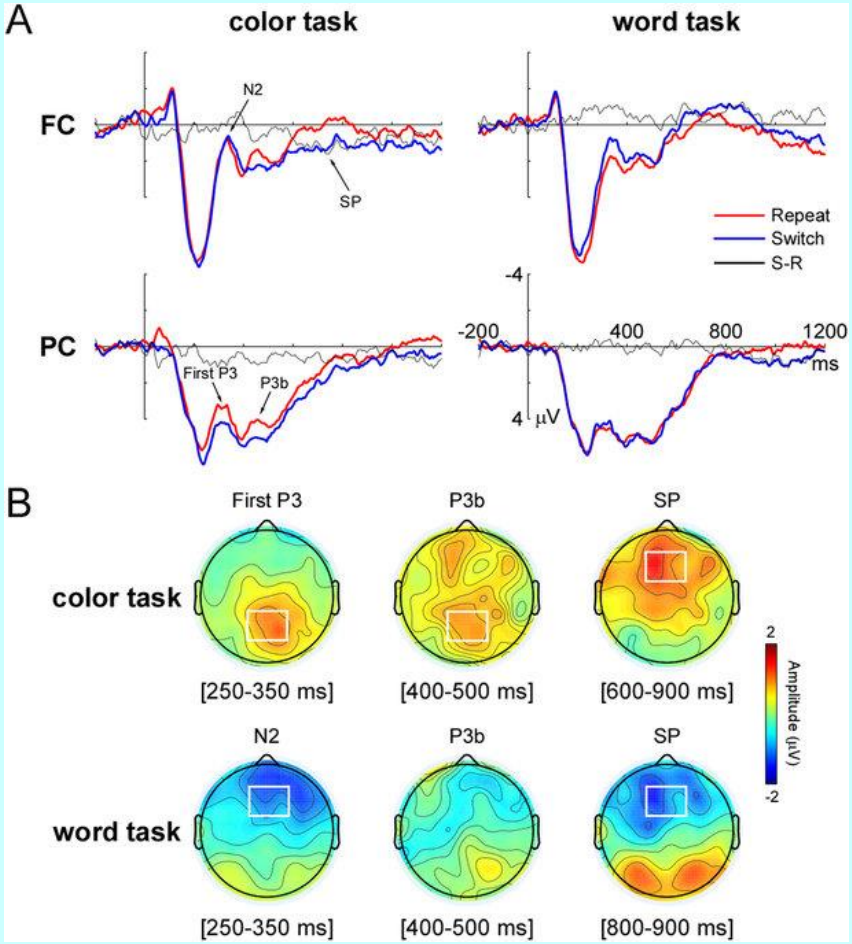
UNIVERSITY OF BIRMINGHAM

Stroop Task

Stroop (1935):

- Congruent: word name equals ink colour
- Incongruent: word name does not equal ink colour

Stroop, J.R. (1935). Studies of interference in serial verbal reactions. J. Exp. Psychol., 18:643-662.



Wu et al., 2015, ***Nature*, http://www.nature.com/articles/srep10240**

UNIVERSITY OF BIRMINGHAM

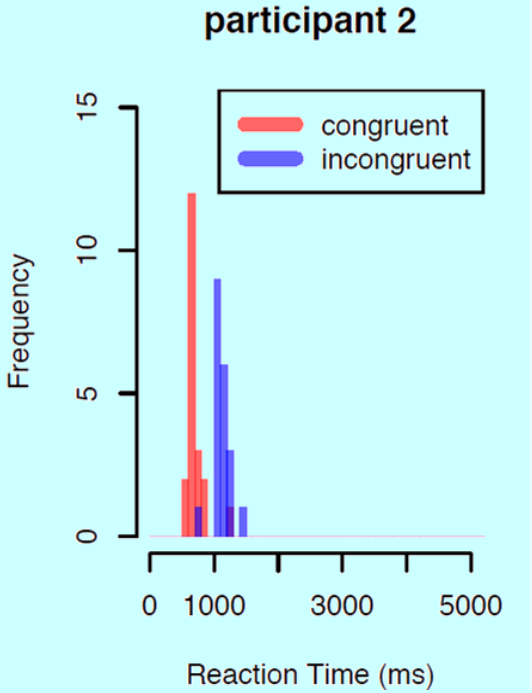
Warrant (Theory / Assumptions)

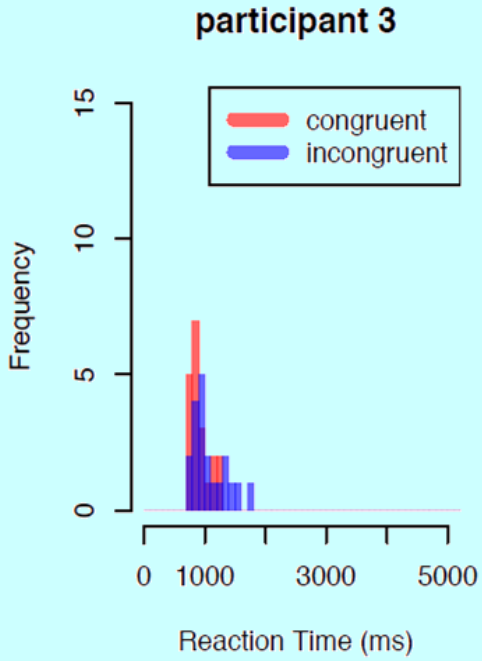
- Top-down control of human information processing is limited
- (Adult) Humans do not seem to capable of ‘switching off’ word reading
- Words can be read more quickly than colours can be named

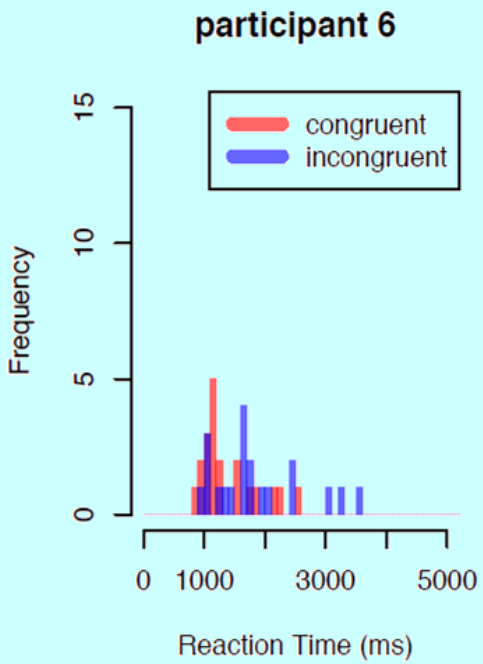
Example of Data from the Stroop Task

UserID	T	Cond	Word	Color	Response	Time
74229	15	IncW	YELLOW	G	G	896
74229	16	IncW	GREEN	B	B	1472
74229	17	IncW	YELLOW	R	R	1008
74229	18	IncW	BLUE	Y	B	1023
74229	19	IncW	GREEN	R	R	1056
74229	20	IncW	BLUE	Y	Y	1040
74229	21	ConW	YELLOW	Y	Y	1548
74229	22	ConW	RED	R	R	840
74229	23	ConW	YELLOW	Y	Y	640
74229	24	ConW	RED	R	R	752
74229	25	ConW	GREEN	G	G	815
74229	26	ConW	YELLOW	Y	Y	800
74229	27	ConW	RED	R	R	736

Looking at the data…







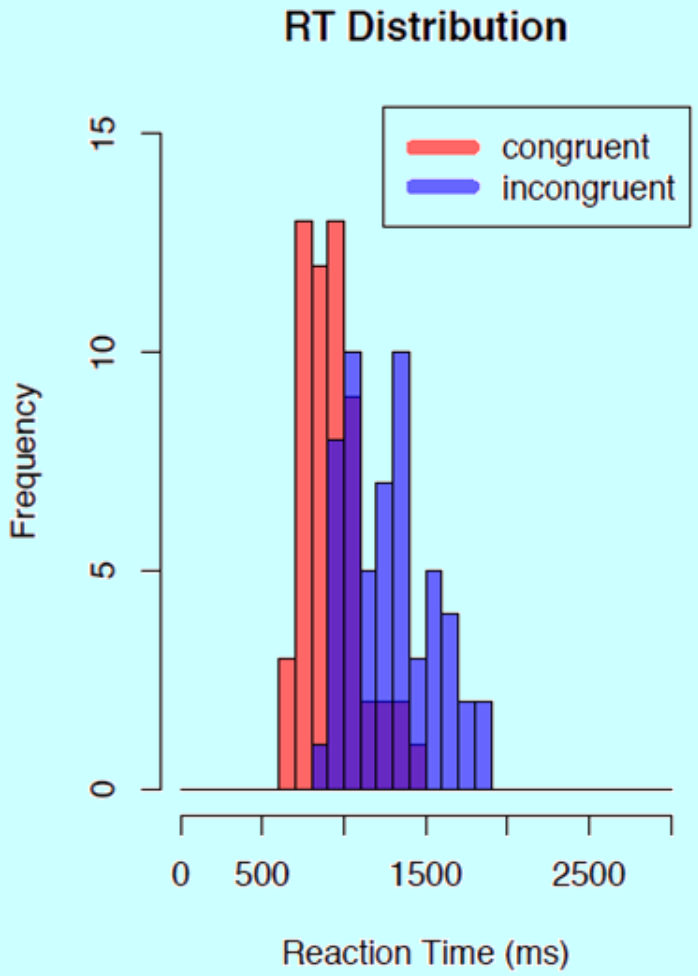
UNIVERSITY OF BIRMINGHAM

Features of the Frequency Plots

- There are more values in the middle than at the extremes.
- The data are is **noisy**. While there is a pattern the curves are not perfectly smooth.
- The plots are **skewed**. The frequency distributions have a long-tail to the right (i.e., there is a limit on how fast you can be (to the left) but no limit on how slow.)
- Plots for 4 and 6 have **outliers.**
- There are general properties of human reaction time curves though for tasks that take a longer duration the curves become progressively less skewed.

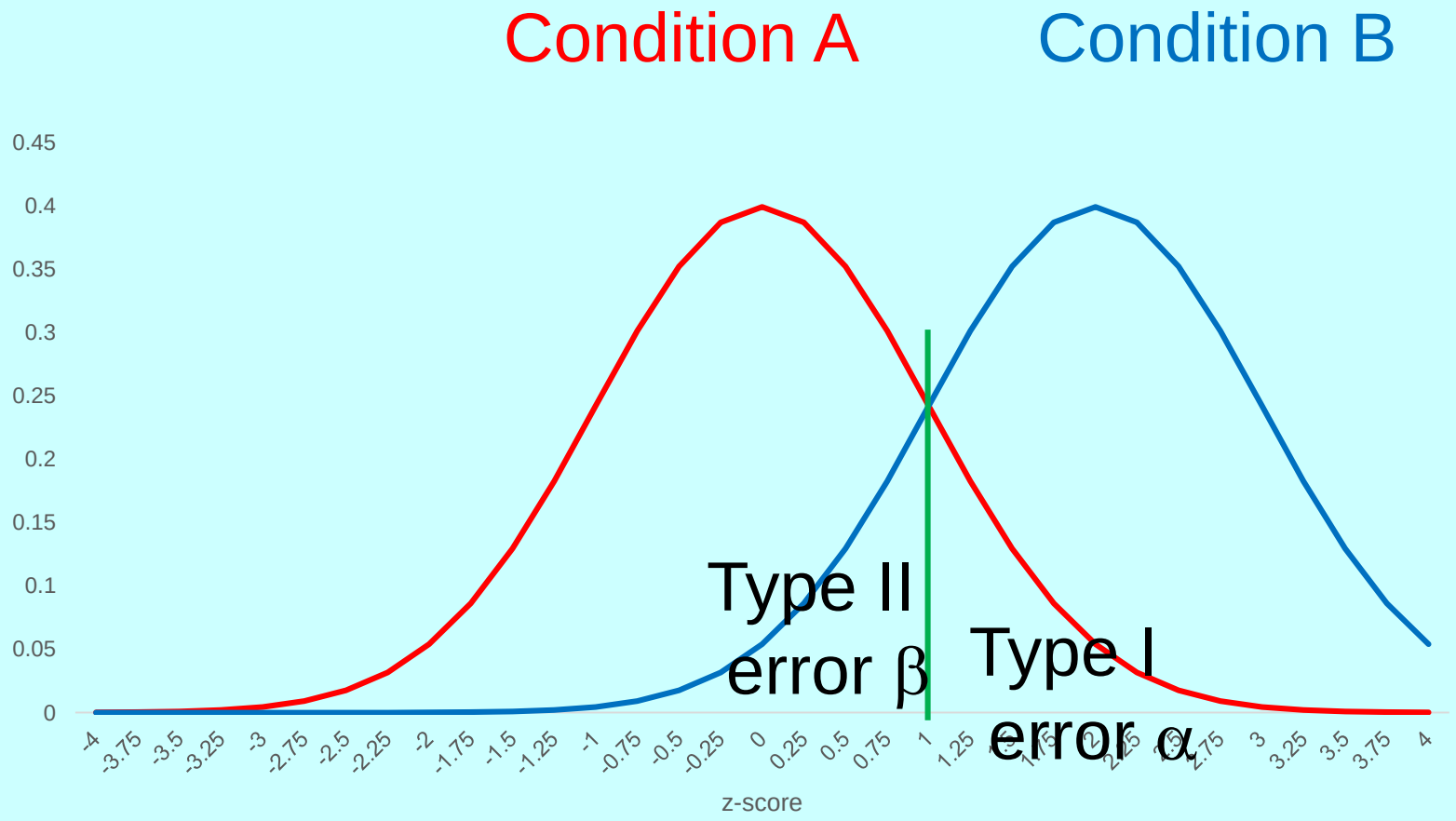
Distribution of means from Stroop Experiment

- n = 57 participants
- Skew is (somewhat) reduced for Congruent and Incongruent conditions



UNIVERSITY OF BIRMINGHAM

Type I and Type II Errors



UNIVERSITY OF BIRMINGHAM

False Claims and Errors

- **Type I error**
  - We could accept the Alternative hypothesis when it is false (false positive).
  - Many statistics tests are designed to minimise this error.
  - Type I errors define the significance level (α) that the experimenter will accept (conventionally 5%)

False Claims and Errors

- **Type II error**
  - We could accept the Null hypothesis (fail to reject it) when it is false (false negative).
  - The probability of a Type II error is defined as β
  - The probability of correctly rejecting a false null hypothesis is defined as 1- β, which called Power.

Hypothesis Definition and Testing

- A good hypothesis is one that can be rejected.
- That is, a good hypothesis is **falsifiable**.
- We define Null and Alternative Hypotheses when planning an experiment. These can be regarded as claims (in an argument).
- If we present claims, we should also present warrant (to explain why we make the claim) and data

Which hypothesis can be falsified?

H0 (null): There are no birds of prey on the University of Birmingham campus.

H1 (alternative): There are birds of prey on the University of Birmingham campus.

Absence of evidence is not evidence of absence.

But, we can **reject** H0 as soon as we see (or hear) a bird of prey.

UNIVERSITY OF BIRMINGHAM

Statistical testing can help us reject the null hypothesis

The results indicated significant correlation between Facebook use and social capital, $r(267) = .29$, $p<0.001$.

There was a significant effect on incongruence on reaction times [$t(56) = 15.58$, $p<0.001$].

These state that the likelihood of accepting the null hypothesis is less that 1 in 1000, so we can reject it (and accept the alternative hypothesis)

Designing Experiments

- **Independent** Variables define the conditions of the experiment
  - In the Stroop task, there are two Levels of the independent variable ‘ink colour and name of word’: congruent and incongruent
- **Dependent** Variables defines what is measured
  - In the Stroop task, the dependent variable is Reaction Time
- **Confounding** Variables define what *might* expect the results of the experiment if they are not taken into account
  - These could include the Order in which each condition is performed or the Order in which stimuli are presented
  - These could include characteristics of people taking part in the experiment

Designing Experiments

- **Independent** Variables define the conditions of the experiment
  - In the Stroop task, there are two Levels of the independent variable ‘ink colour and name of word’: congruent and incongruent
- **Dependent** Variables defines what is measured
  - In the Stroop task, the dependent variable is Reaction Time
- **Confounding** Variables define what *might* expect the results of the experiment if they are not taken into account
  - These could include the Order in which each condition is performed or the Order in which stimuli are presented
  - These could include characteristics of people taking part in the experiment

Experimental Design for Stroop Task

- The **Independent Variable** for the Stroop experiment has two conditions: One with congruent stimuli and the other with incongruent stimuli.
- The **Dependent Variable** is Reaction Time
- The claim concerns the relative effect of congruent and incongruent colour words on **Reaction Time** (**RT**) and applies to a **population**, i.e., all adult humans.
  - To test this claim, we define an Hypothesis
    - Two-tailed: There will be a difference in reaction time to congruent and incongruent words.
    - One-tailed: Reaction time to congruent words will be faster than reaction time to incongruent words.

Variation in Dependent Variables

- Systematic: due to change in Independent Variable
- Unsystematic: due to confounding variables

Experimental Design

Hypothesis: Reaction time to congruent words will be faster than reaction time to incongruent words	
Independent Variable: Congruent Words (colour of ink = name of word), Incongruent Words (colour of ink ≠ name of word)	
Control Condition: Congruent Words	**Experimental Condition:** Incongruent Words
Dependent Variable(s): Reaction Time	
Task: participants will be asked to read, as quickly as possible, single words on a display. The words will be the names of colours and will be presented either in the same colour as the word's name or in a different colour	
Confounding Variables: performance could be affected by ability to perceive colour ('colour-blindedness') and knowledge of the names of colour ('language skills')	

UNIVERSITY OF BIRMINGHAM

Types of Study Design

- Post-test
  - Control versus experimental group complete task: outcome is measured and compared
- Pretest-Post-test
  - Control versus experimental group complete task and outcome is measured and compared; experimental group treatment and both groups tested again
- Solomon Four Group
  - 2 control groups and 2 experimental groups; pretest-post-test and post-test only.
- Factorial Design
  - 2 or more independent variables manipulated.
- Crossover (repeated measures) Design
  - Participants randomly allocated to perform both control and experimental conditions complete task and outcome is measured and compared

Experimental Design for Stroop Task

- From the population (say, all adult humans in the world), we take a **sample** of **N participants**.
- Stroop experiments typically use a **within-participant design**, where all participants take part in all **conditions**.

Participants (Subjects)

	Source of Unsystematic DV variation	Control
Between Subjects	Individual differences	Match participants on key characteristics
		Random allocation to condition
Within Subjects	Practice / Order effects	Modify order of stimuli
	Boredom / fatigue effects	Design in breaks
	Asymmetric transfer effects	Counterbalance conditions

Assigning Participants to Conditions

- We want to reduce the Confound of Order Effects
  - That is, if every participant does the same tasks in the same order, how do we know that they are not simply getting better with practice (or worse through fatigue)?
  - So, we randomise the Order in which people complete the tasks
  - For two conditions, this is easy to manage…

Participant	Trial 1	Trial 2
1	A	B
2	B	A

  - …and we assign participants to either group 1 or group 2 (either by Appearance (when they turn up to do the study) or by Random number generation)

UNIVERSITY OF BIRMINGHAM

Latin Squares

- We *could* simply put in more conditions, offsetting to cycle around…
- …making sure that each condition appears once in each row and column

Participant	Trial 1	Trial 2	Trial 3	Trial 4
1	A	B	C	D
2	D	A	B	C

- …BUT this introduces a confound of unbalanced Order Effects because, sometimes B follows A, C follows B…

Balanced Latin Squares

- …so we ensure that each condition appears once in each row and column
- AND that each condition follows each other condition.

Group	Trial 1	Trial 2	Trial3	Trial4
1	A	B	D	C
2	B	C	A	D
3	C	D	B	A
4	D	A	C	B

- There is a simple algorithm here…
  - Column 1 has the conditions in alphabetic order
  - Column 2 is the same as column 1 with wraparound…
- Notice also, that for four conditions, we need multiples of 4 groups (assuming that we'd need at least 8 participants per group, we can estimate the size of the sample we require for such an experiment)

UNIVERSITY OF BIRMINGHAM