PARTE I

Design and Application of Experiments and User Studies

Victor Adriel de Jesus Oliveira
INF - UFRGS

.inf INSTITUTO DE INFORMÁTICA UFRGS

GRAPHICS VISUALIZATION INTERACTION | LAB

About Me

Victor Adriel de J. Oliveira

PhD Candidate in Computer Science - UFRGS
- *Design and Assessment of Haptic Interfaces*

Masters in Computer Science - UFRGS (2014)
- *Designing Tactile Vocabularies for Human-Computer Interaction*

Computer Scientist - UESC (2012)
- *Acessibilidade em Sites e Sistemas Web*

Member of the EuroHaptics Society (EHS), of the Technical Committee on Haptics (TCH), of the Institute of Electrical and Electronics Engineers (IEEE), and of the Sociedade Brasileira de Computação (SBC)

GRAPHICS VISUALIZATION INTERACTION LAB .Inf INSTITUTO DE INFORMÁTICA UFRGS UFRGS

Summary

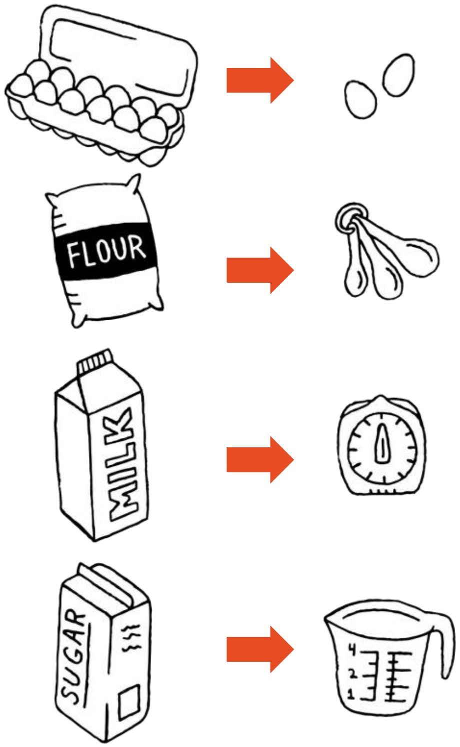
- PART I
  - Introduction to the Design of Experiments (DOX)
  - Designing User Studies
  - Hands-on (Design of Experiment)
- PART II
  - Applying User Studies
  - Analysis and Report of Results
  - Hands-on (Analysing Data)
  - Conclusions

Introduction to DOX

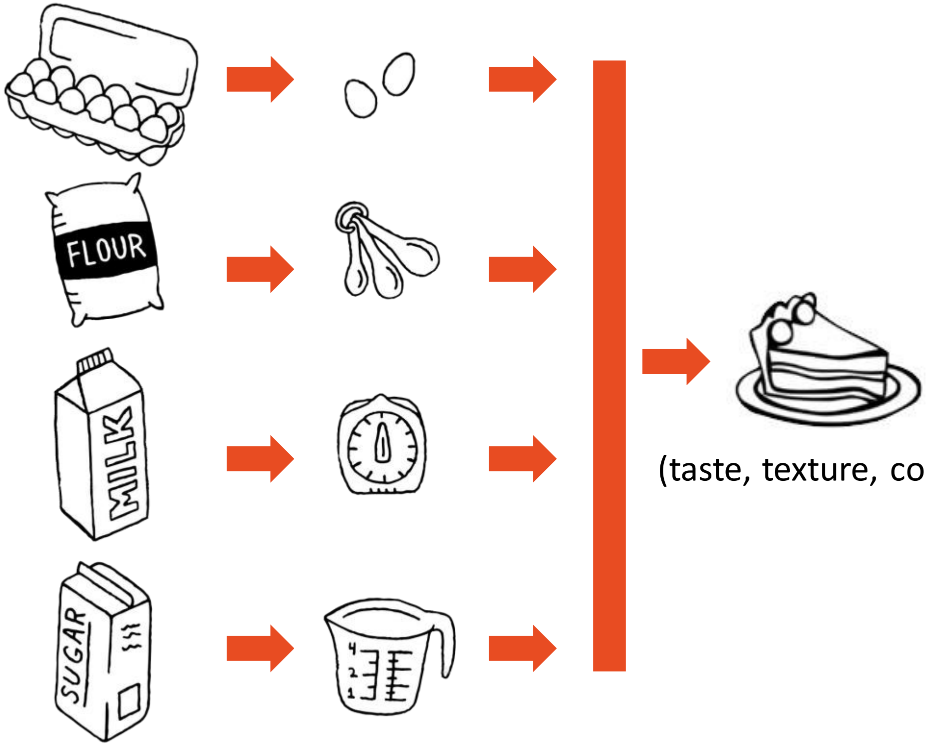
Purpose of Experimentation

https://youtu.be/HIonKbKM-tE

Purpose of Experimentation

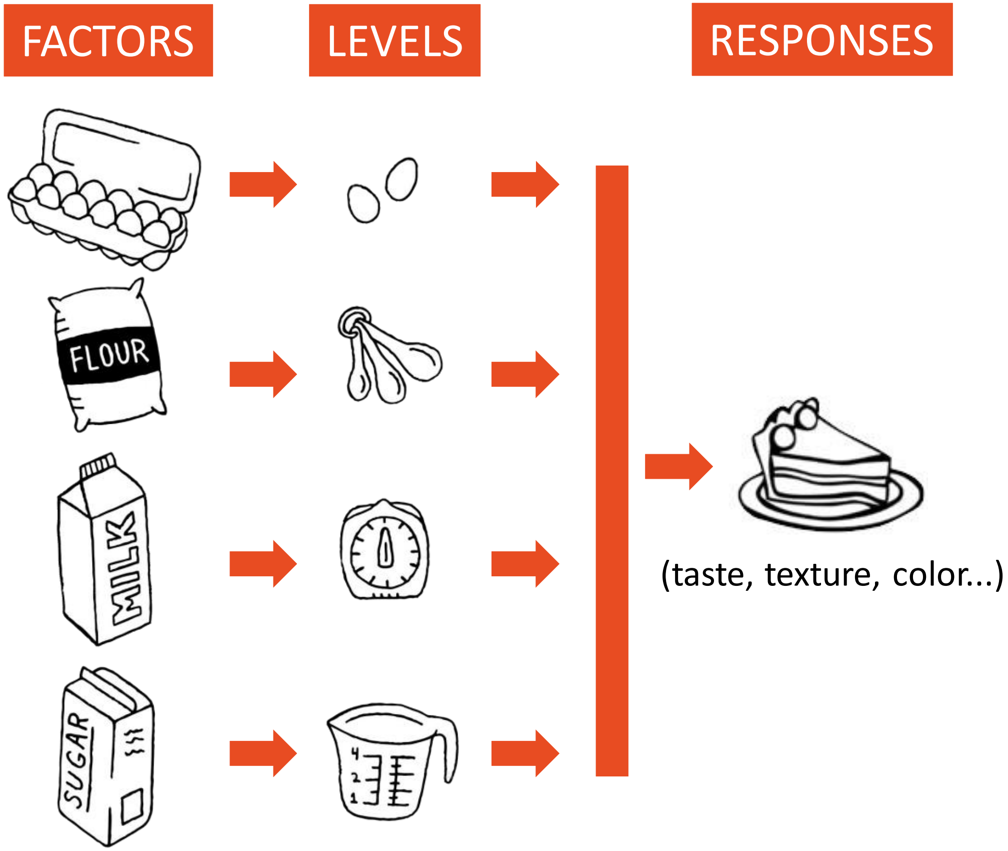


Purpose of Experimentation



(taste, texture, color...)

Purpose of Experimentation



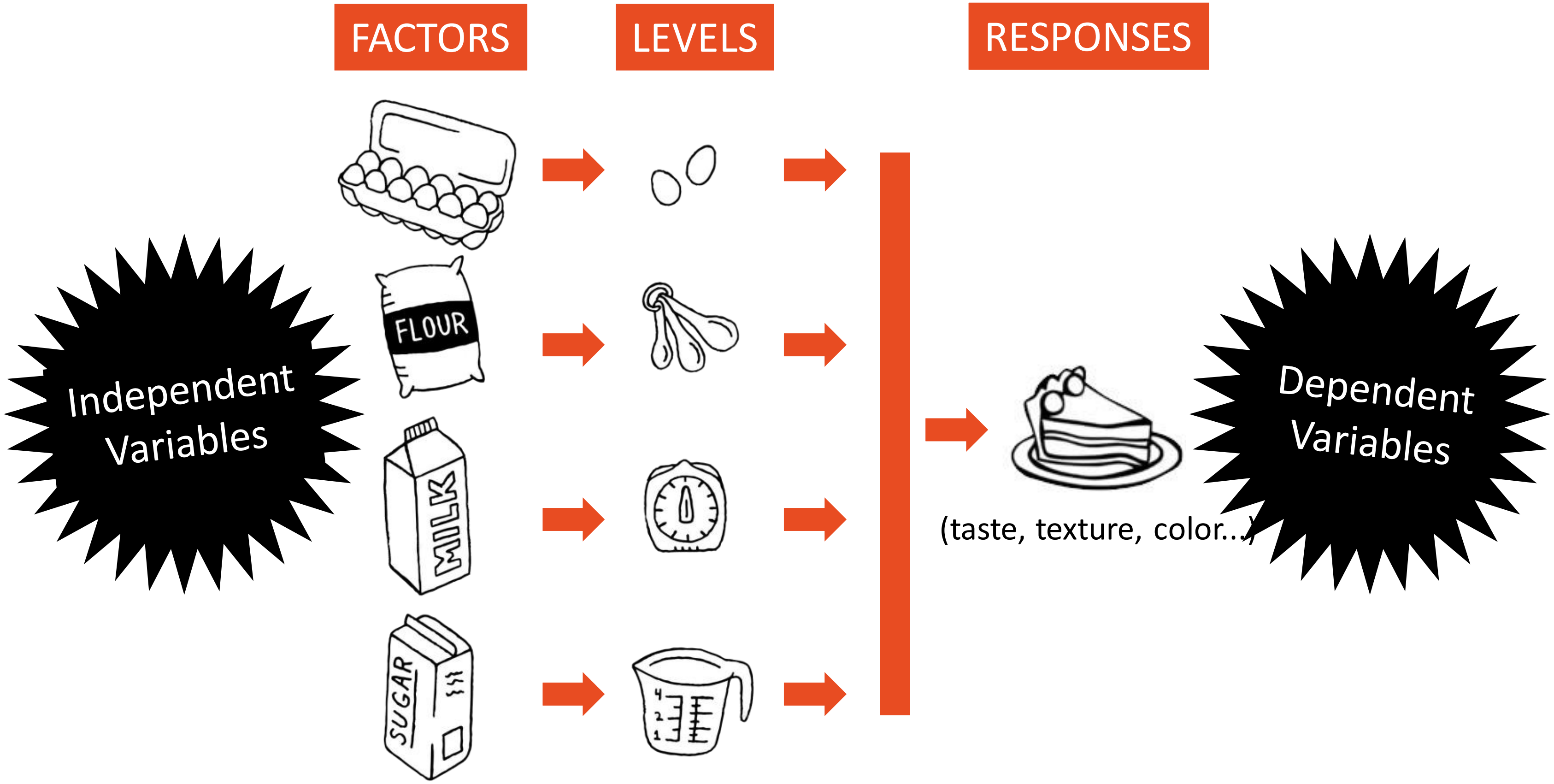
FACTORS

LEVELS

RESPONSES

(taste, texture, color...)

Purpose of Experimentation



FACTORS

LEVELS

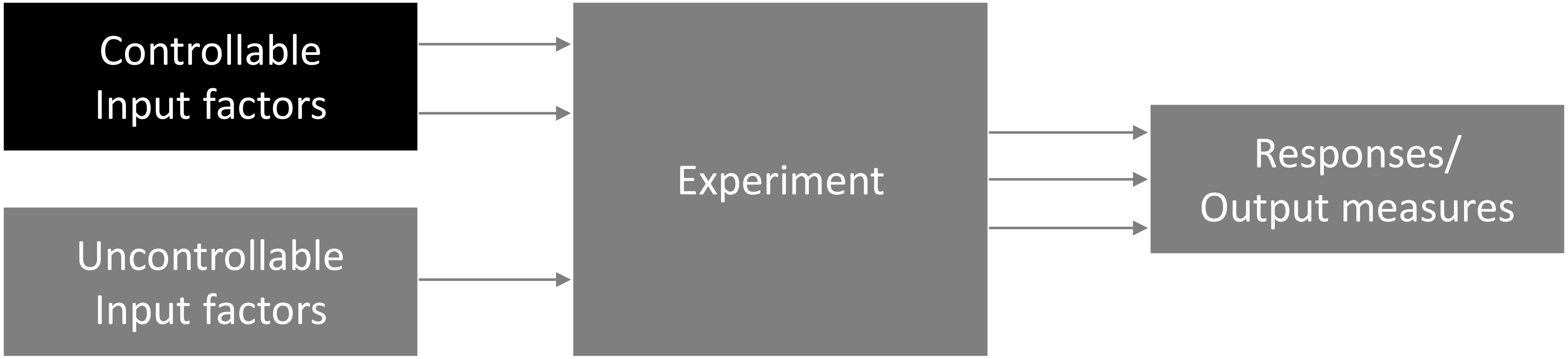
RESPONSES

Independent Variables

Dependent Variables

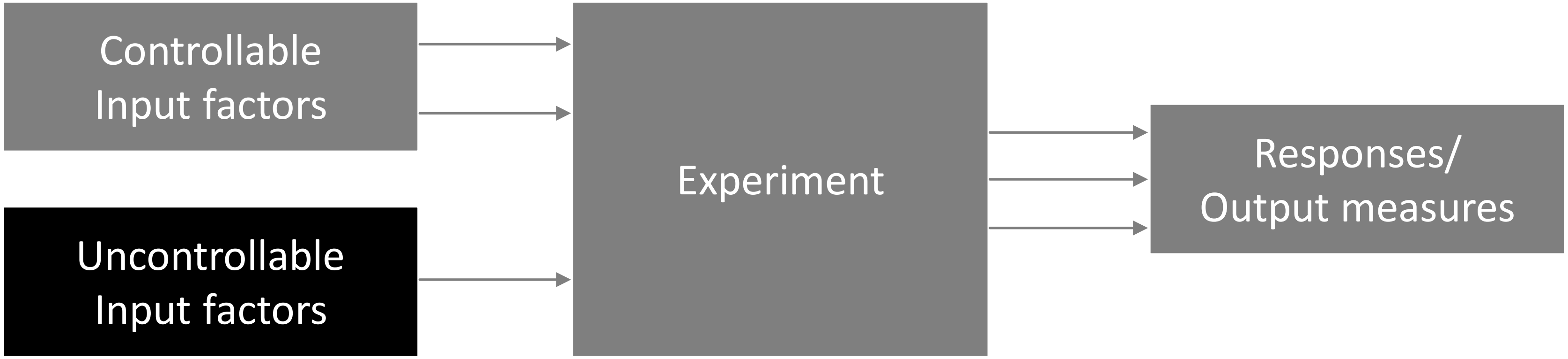
(taste, texture, color...)

Components of Experimental Design



- **Controllable input factors**: are those input parameters that can be modified in an experiment or process
- In our baking example, these factors include the quantity and quality of the flour and the temperature of the milk

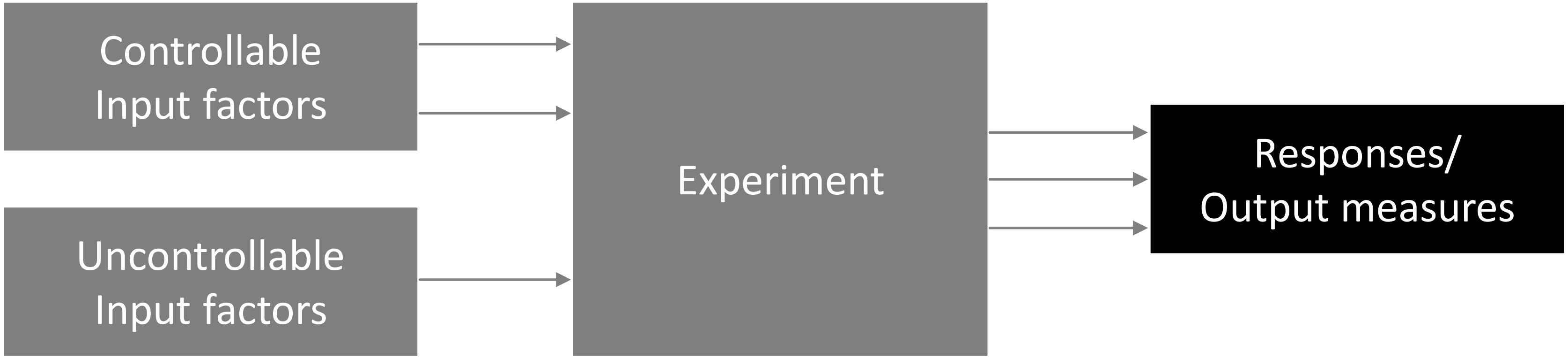
Components of Experimental Design



- **Uncontrollable input factors:** are those parameters that cannot be changed
- In our example, this may be the temperature in the kitchen

*These factors need to be recognized to understand how they may affect the response

Components of Experimental Design



- **Responses:** are the elements of the process outcome that gage the desired effect
- In the baking example, the taste and texture of the cake are the responses

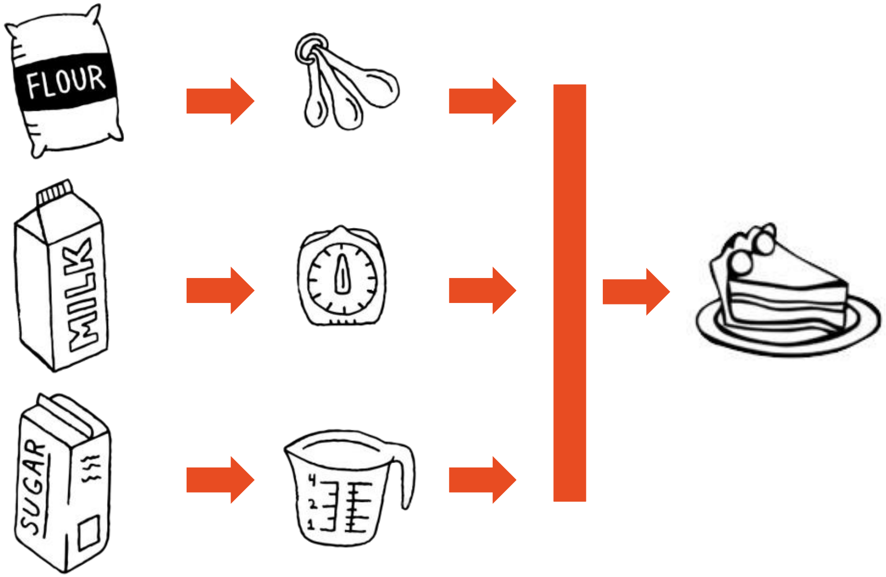
Fisher’s Principles

Ronald Fisher

The Design of Experiments (1935)

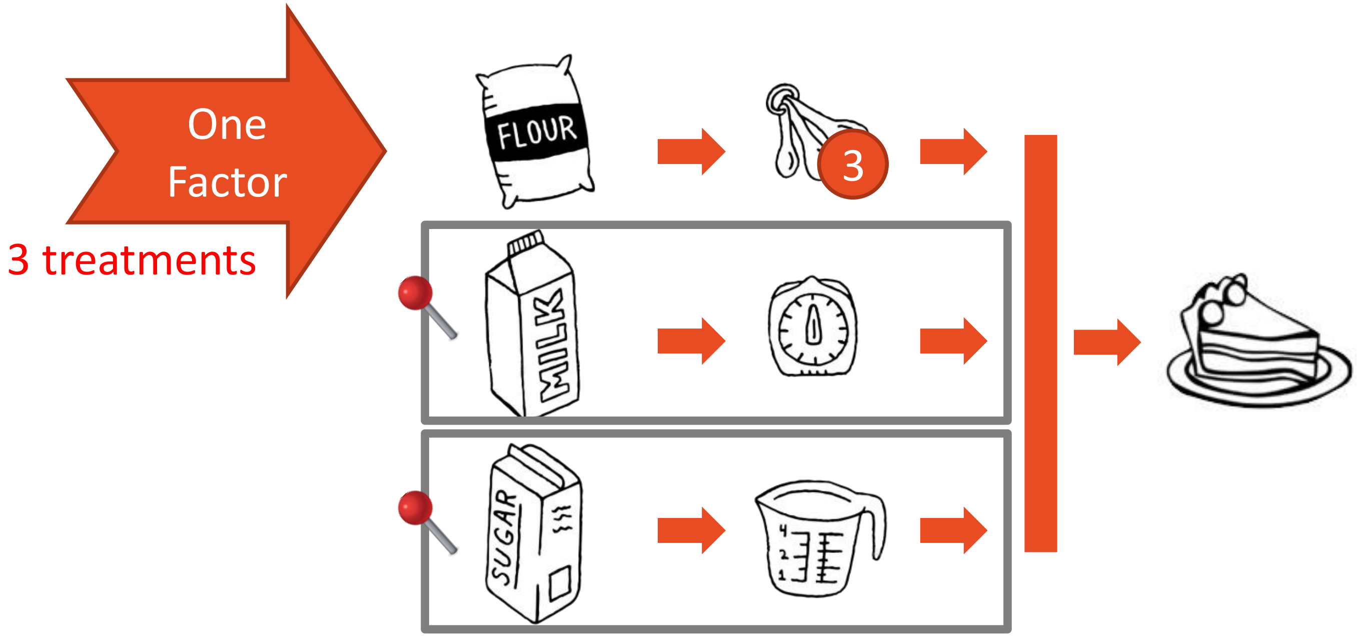
Fisher's Principles

- Factorial Experiments
  - Factorial experiments are efficient at evaluating the effects and possible interactions of several factors (independent variables)



Fisher's Principles

- Factorial Experiments
  - Factorial experiments are efficient at evaluating the effects and possible interactions of several factors (independent variables)

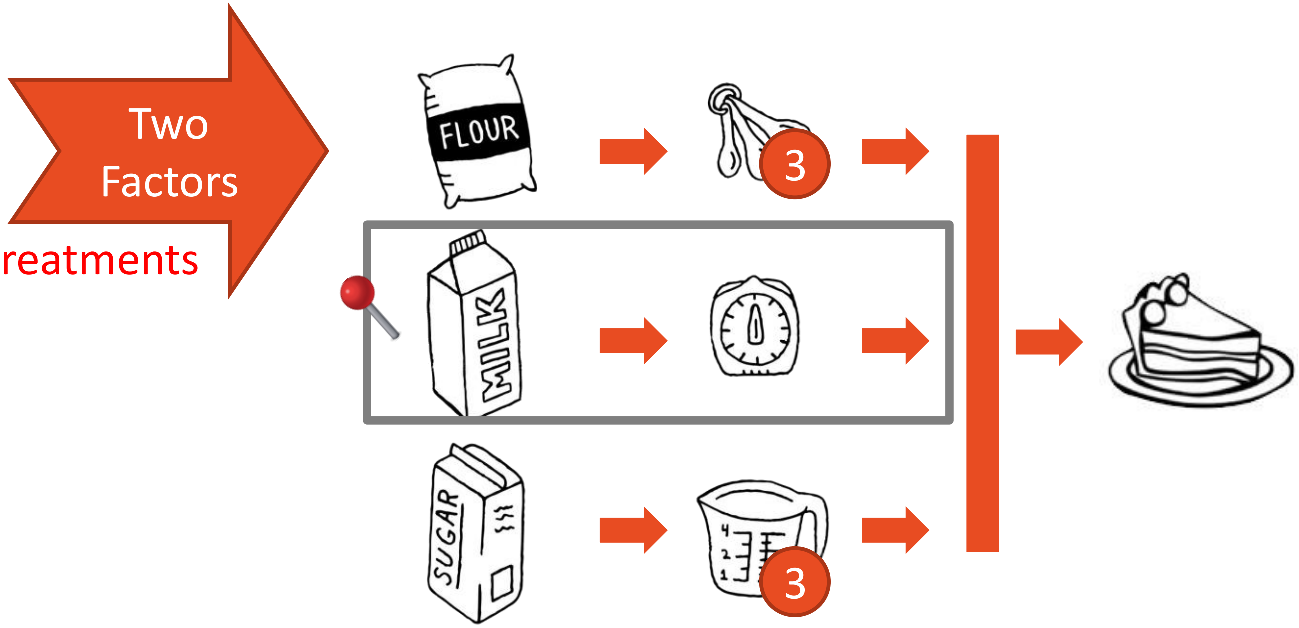


One Factor

3 treatments

Fisher's Principles

- Factorial Experiments
  - Factorial experiments are efficient at evaluating the effects and possible interactions of several factors (independent variables)



Two Factors

3 x 3 treatments

Fisher’s Principles

- Comparison
  - Comparisons between treatments are much more valuable and are usually preferable, and often compared against a scientific control or traditional treatment that acts as baseline

Baseline

Test A

Test B

Fisher’s Principles

- Randomization

Baseline

Test A

Test B

Fisher’s Principles

- Randomization

Baseline

1

Test A

2

Test B

3

1	2	3
2	3	1
3	1	2

Normalized Latin Square

Fisher’s Principles

- Randomization
  - Random assignment means assigning individuals at random to groups in an experiment, so that each individual of the population has the same chance of becoming a participant in the study

Fisher’s Principles

- Blocking
  - The non-random arrangement of experimental units into groups (blocks/lots). Blocking reduces known but irrelevant sources of variation between units and thus allows greater precision in the estimation of the source of variation under study.

Fisher’s Principles

- Statistical Replication
  - Experiments are replicated to help identify the sources of variation, to better estimate the true effects of treatments, to further strengthen the experiment's reliability and *validity*, and to add to the existing knowledge of the topic

Experimental Validity

- Internal Validity
  - It is an inductive estimate of the degree to which conclusions about causal relationships can be made, based on the measures used, the research setting, and the whole research design
  - Good experimental techniques, in which the effect of an independent variable on a dependent variable is studied under highly controlled conditions, usually allow for higher degrees of internal validity than, for example, single-case designs

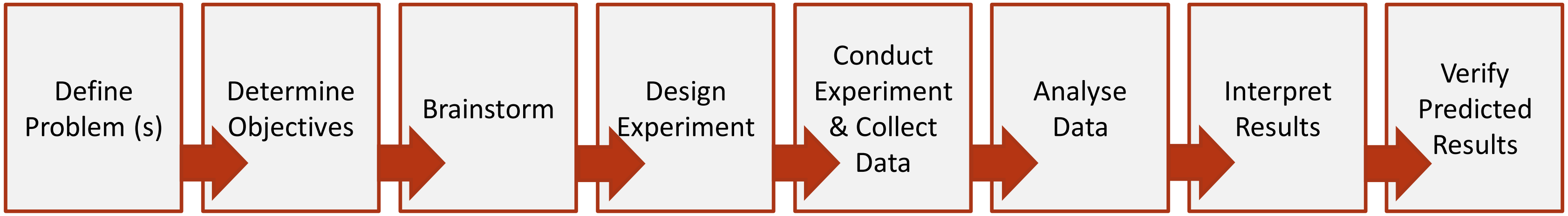
Experimental Validity

- External Validity
  - It concerns the extent to which the (internally valid) results of a study can be held to be true for other cases, for example to different people, places or times. In other words, it is about whether findings can be validly generalized
  - A major factor in this is whether the study sample (e.g. the research participants) are representative of the general population along relevant dimensions

Experimental Validity

- Statistical Conclusion Validity
    - It is the degree to which conclusions about the relationship among variables based on the data are correct or reasonable
    - Statistical conclusion validity involves ensuring the use of adequate sampling procedures, appropriate statistical tests, and reliable measurement procedures

Experiment Design Process



Define Problem (s)

Determine Objectives

Brainstorm

Design Experiment

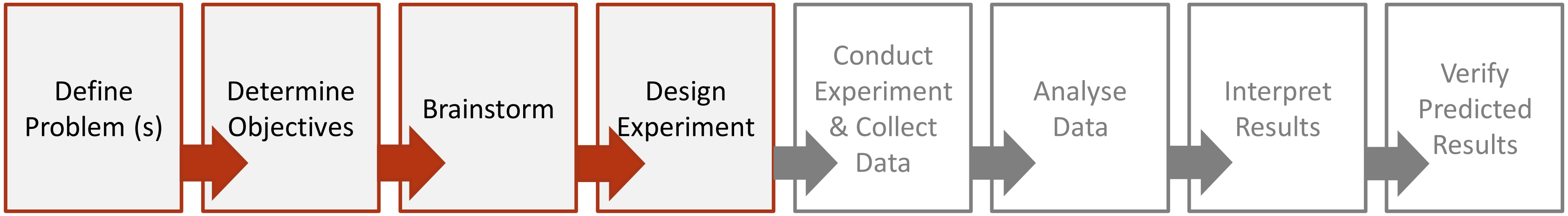
Conduct Experiment & Collect Data

Analyse Data

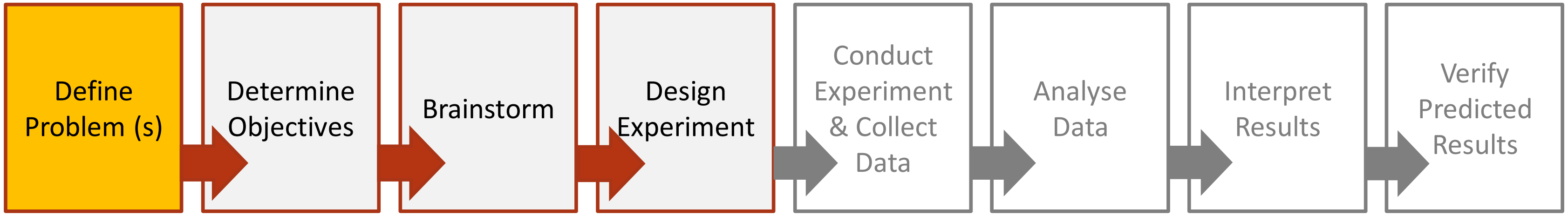
Interpret Results

Verify Predicted Results

Experiment Design Process

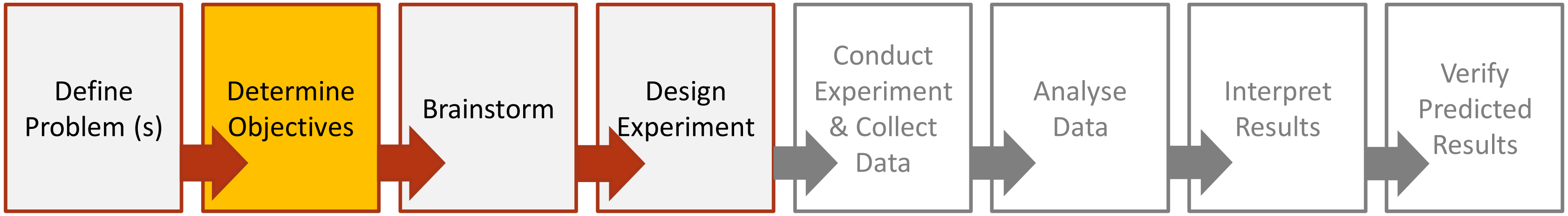


Experiment Design Process



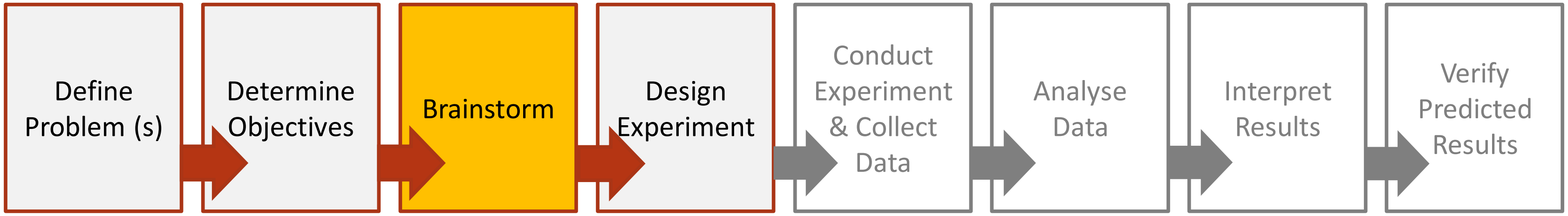
- Training employees is important but it also can be very expensive
  - Training costs can be reduced by using virtual reality

Experiment Design Process



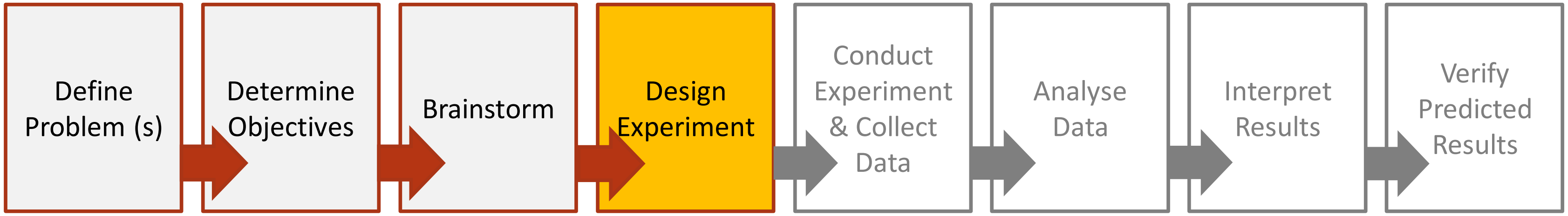
- Training employees is important but it also can be very expensive
  - Training costs can be reduced by using virtual reality
- Goal
  - General: To design effective VR applications for training
  - Specific: To assess how different input and output techniques affect learning during VR training

Experiment Design Process



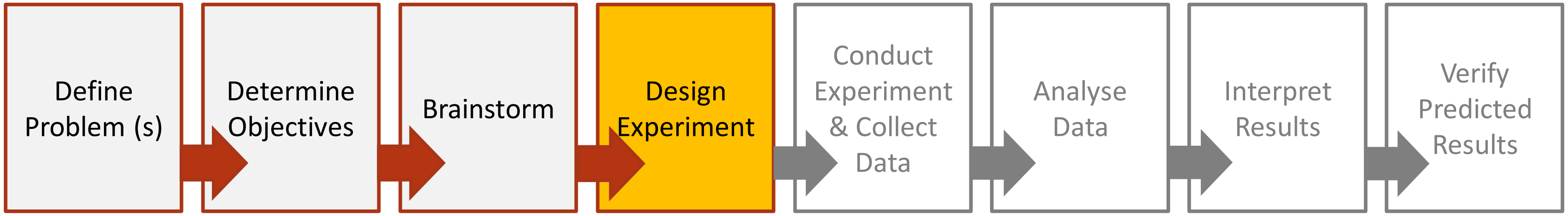
- Previous experience
- Related work
  - What are the main output techniques for VR?
  - How are they classified?
  - What is the baseline?
  - What means ”to learn” in this context?

Experiment Design Process



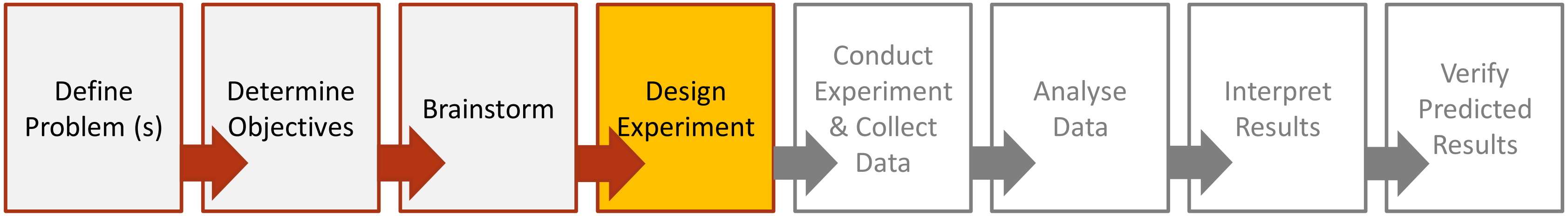
- Experimental question
  - Does different output techniques affect learning during VR training?
- Hypothesis
  - I predict that output techniques and learning will be related
  - I predict that learning will increase according the level of immersion caused by the technique

Experiment Design Process


Define Problem (s)
Determine Objectives
Brainstorm
Design Experiment
Conduct Experiment & Collect Data
Analyse Data
Interpret Results
Verify Predicted Results

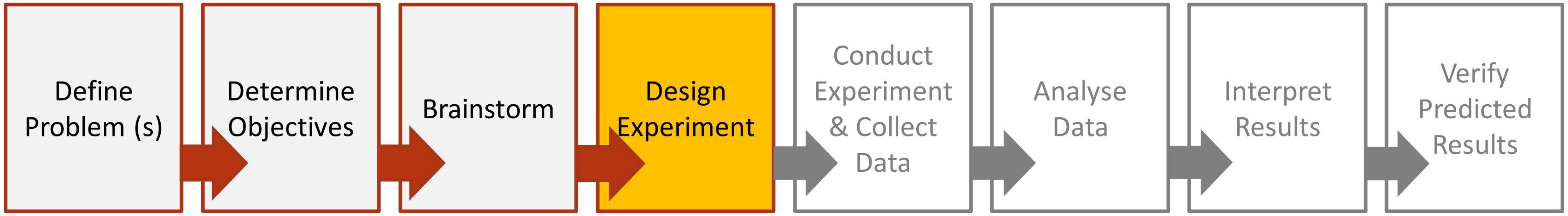
- H: I predict that learning will increase according the level of immersion caused by the technique

Experiment Design Process



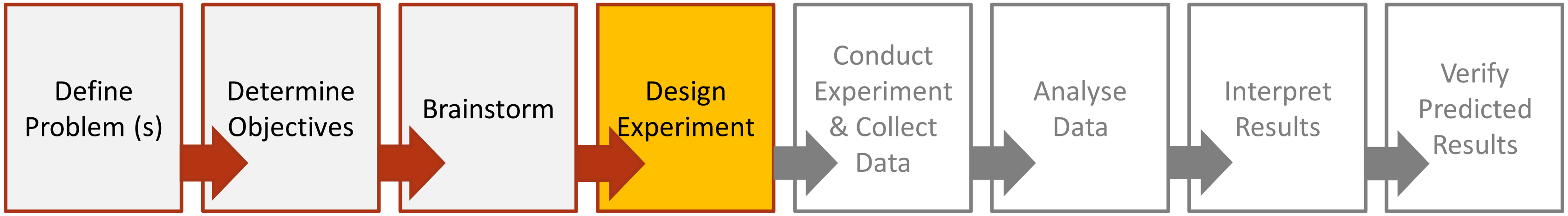
- H: I predict that learning will increase according the level of immersion caused by the technique
  - Factor: Immersion
    Levels: Non-immersive to fully-immersive
    Response: Learning

Experiment Design Process



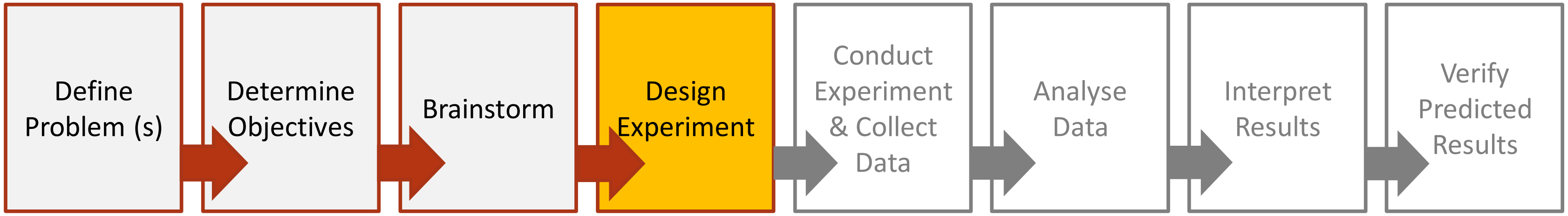
- H: I predict that learning will increase according the level of immersion caused by the technique
  - Factor: Immersion
    Levels: Non-immersive to fully-immersive
    Response: Learning
  - Blocking: Age

Experiment Design Process



- H: I predict that learning will increase according the level of immersion caused by the technique
  - Factor: Immersion
    Levels: Non-immersive to fully-immersive
    Response: Learning
  - Blocking: Age
  - Randomization: population and treatments

Experiment Design Process



But what about the users?

Design of User Studies

Psychological Tests

- It is a systematic procedure for obtaining samples of behavior, relevant to cognitive, affective, or interpersonal functioning, and for scoring and evaluating those samples according to ***standards*** (Urbina, 2014)

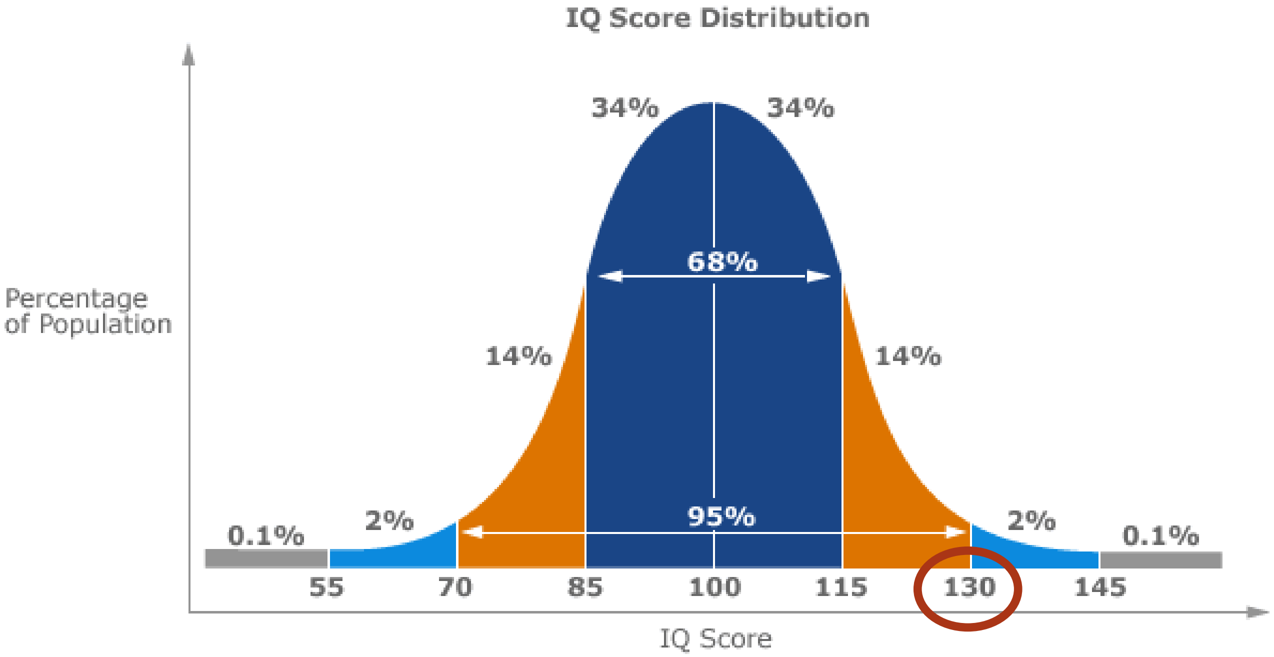
Psychological Tests

- It is a systematic procedure for obtaining samples of behavior, relevant to cognitive, affective, or interpersonal functioning, and for scoring and evaluating those samples according to ***standards*** (Urbina, 2014)

- Example: IQ 115

Psychological Tests

- It is a systematic procedure for obtaining samples of behavior, relevant to cognitive, affective, or interpersonal functioning, and for scoring and evaluating those samples according to ***standards*** (Urbina, 2014)



Psychological Tests

- **Classical Test Theory** (CTT) is a body of related psychometric theory that predicts outcomes of psychological testing such as the difficulty of items or the ability of test-takers
- It is a theory of testing based on the idea that a person's observed or obtained score on a test is the sum of a true score (error-free score) and an error score

$$\boxed{X = T + E}$$

Observed score   True score   error

Psychological Tests

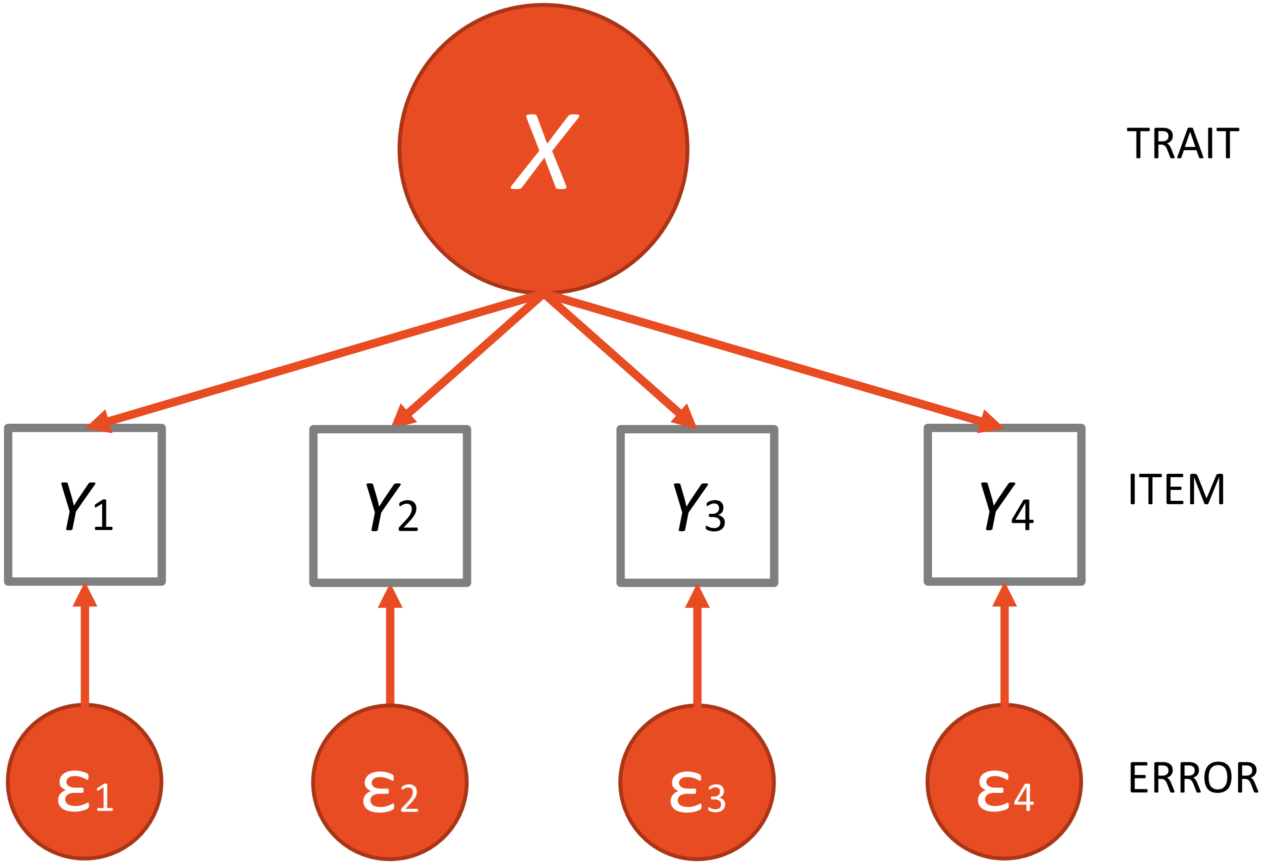
- **Reliability** is the overall consistency of a measure
- It is the characteristic of a set of test scores that relates to the amount of random error from the measurement process that might be embedded in the scores
- Scores that are highly reliable are accurate, reproducible, and consistent from one testing occasion to another
- Researchers use a measure of internal consistency known as Cronbach's α

Psychological Tests

- **Item Response Theory** (IRT) is a paradigm for the design, analysis, and scoring of tests, questionnaires, and similar instruments measuring abilities, attitudes, or other variables
- IRT models are often referred to as *latent trait models*
- The term *latent* is used to emphasize that discrete item responses are taken to be *observable manifestations* of hypothesized traits, constructs, or attributes, not directly observed, but which must be inferred from the manifest responses

Psychological Tests

- **Latent trait model**

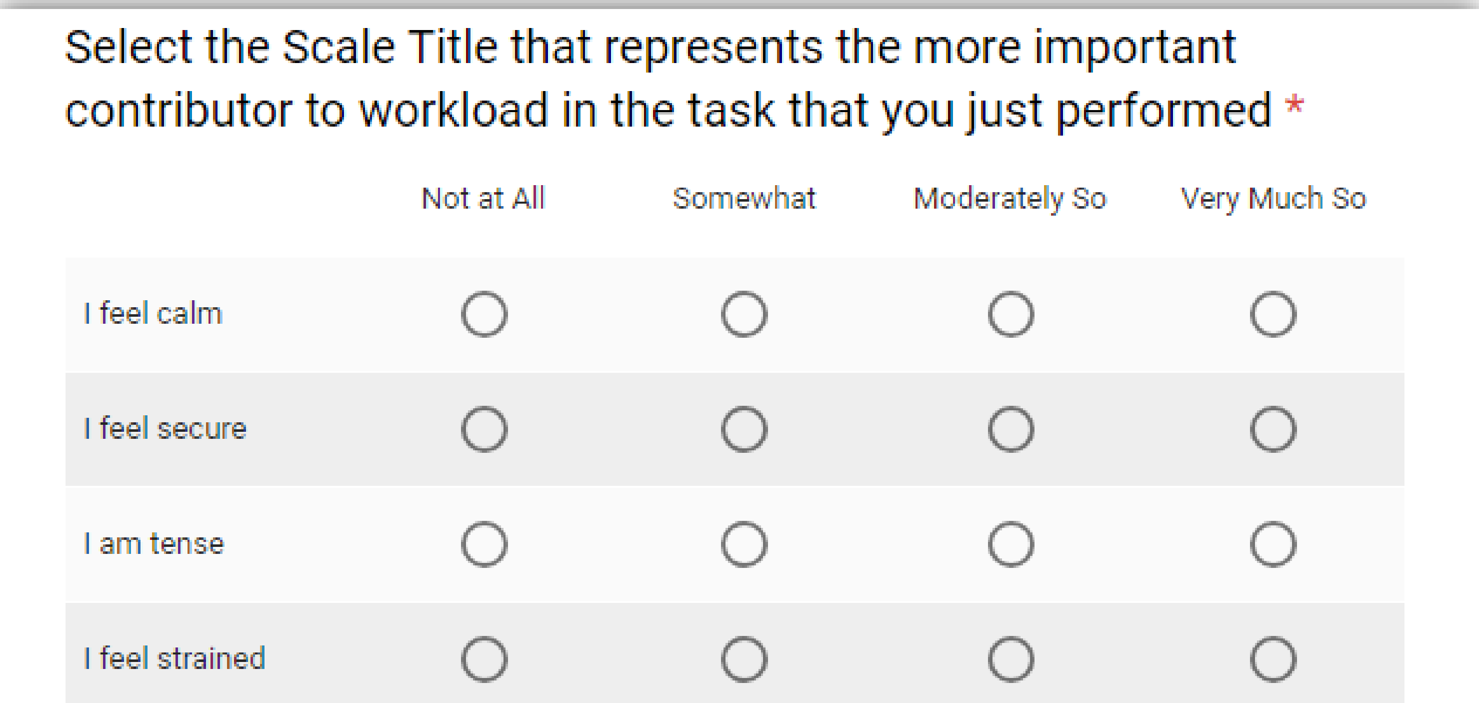


Psychological Tests

- **Test validity** is the extent to which a test accurately measures what it is supposed to measure
- Validity refers to the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests
  - **Construct validity** is the degree to which a test measures what it claims, or purports, to be measuring
  - **Content validity** refers to the extent to which a measure represents all facets of a given construct
  - **Criterion validity** is the extent to which a measure is related to an outcome

Anxiety

- The State-Trait Anxiety Inventory (STAI) is a commonly used measure of trait and state anxiety
- The Form Y, its most popular version, has 20 items for assessing trait anxiety and 20 for state anxiety



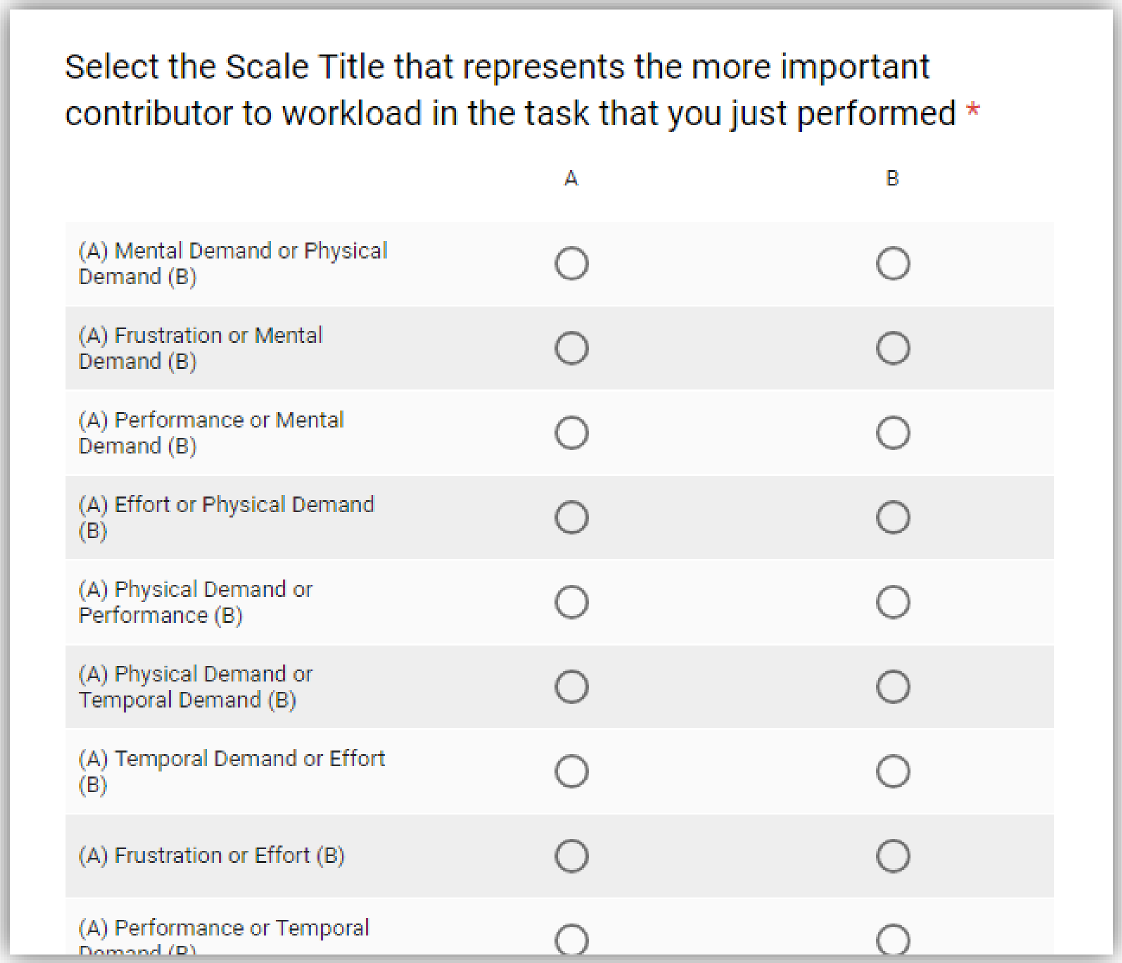
Select the Scale Title that represents the more important contributor to workload in the task that you just performed *

	Not at All	Somewhat	Moderately So	Very Much So
I feel calm	○	○	○	○
I feel secure	○	○	○	○
I am tense	○	○	○	○
I feel strained	○	○	○	○

Workload

- There are several questionnaires for assessment of perceived workload, but the most popular seems to be the **NASA-TLX** (NASA Task Load Index)
- The NASA TLX has been developed by NASA to assess the relative importance of six factors in determining how much workload the subject experienced: Mental Demand, Physical Demand, Temporal Demand, Effort, Performance, and Frustration.

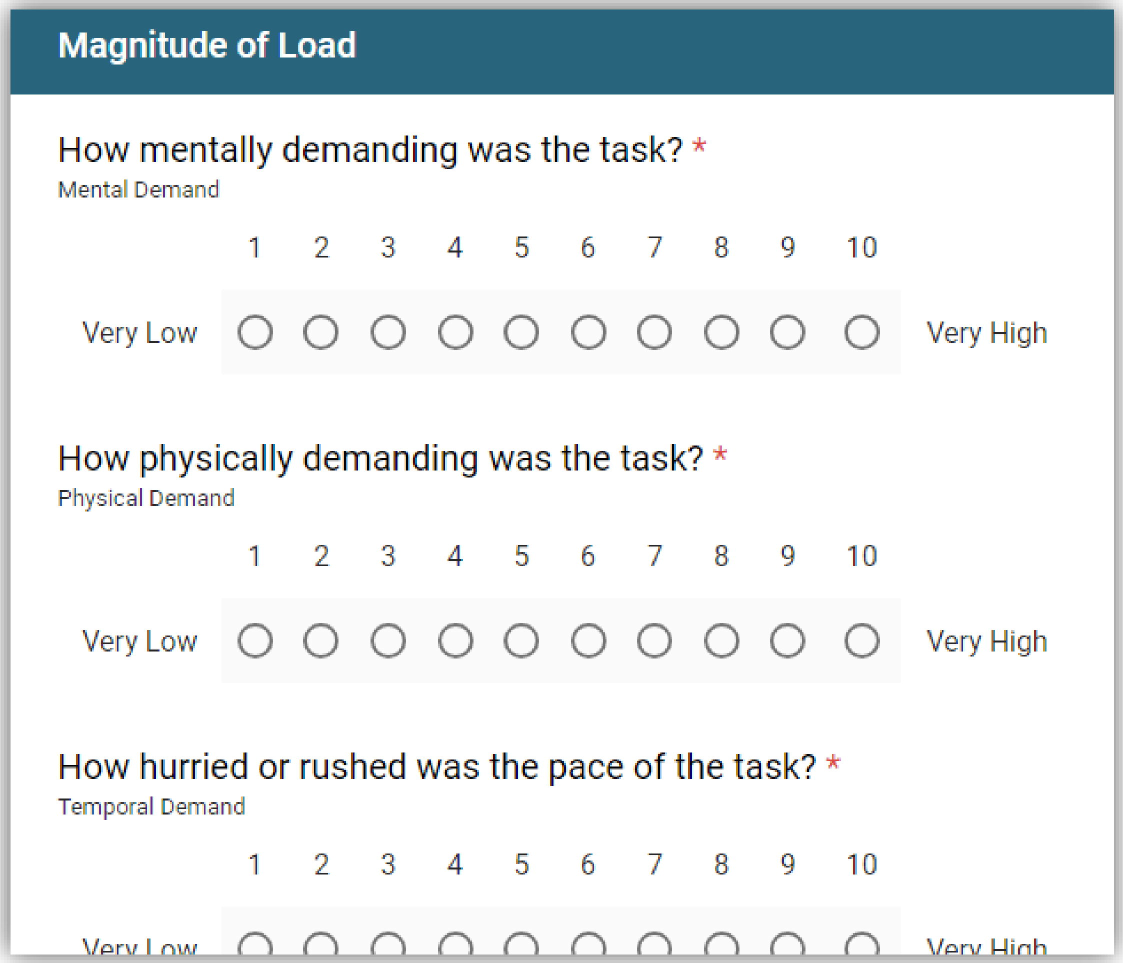
Workload



Select the Scale Title that represents the more important contributor to workload in the task that you just performed *

	A	B
(A) Mental Demand or Physical Demand (B)	○	○
(A) Frustration or Mental Demand (B)	○	○
(A) Performance or Mental Demand (B)	○	○
(A) Effort or Physical Demand (B)	○	○
(A) Physical Demand or Performance (B)	○	○
(A) Physical Demand or Temporal Demand (B)	○	○
(A) Temporal Demand or Effort (B)	○	○
(A) Frustration or Effort (B)	○	○
(A) Performance or Temporal Demand (B)	○	○

Workload



Magnitude of Load

How mentally demanding was the task? *

Mental Demand

	1	2	3	4	5	6	7	8	9	10	
Very Low	○	○	○	○	○	○	○	○	○	○	Very High

How physically demanding was the task? *

Physical Demand

	1	2	3	4	5	6	7	8	9	10	
Very Low	○	○	○	○	○	○	○	○	○	○	Very High

How hurried or rushed was the pace of the task? *

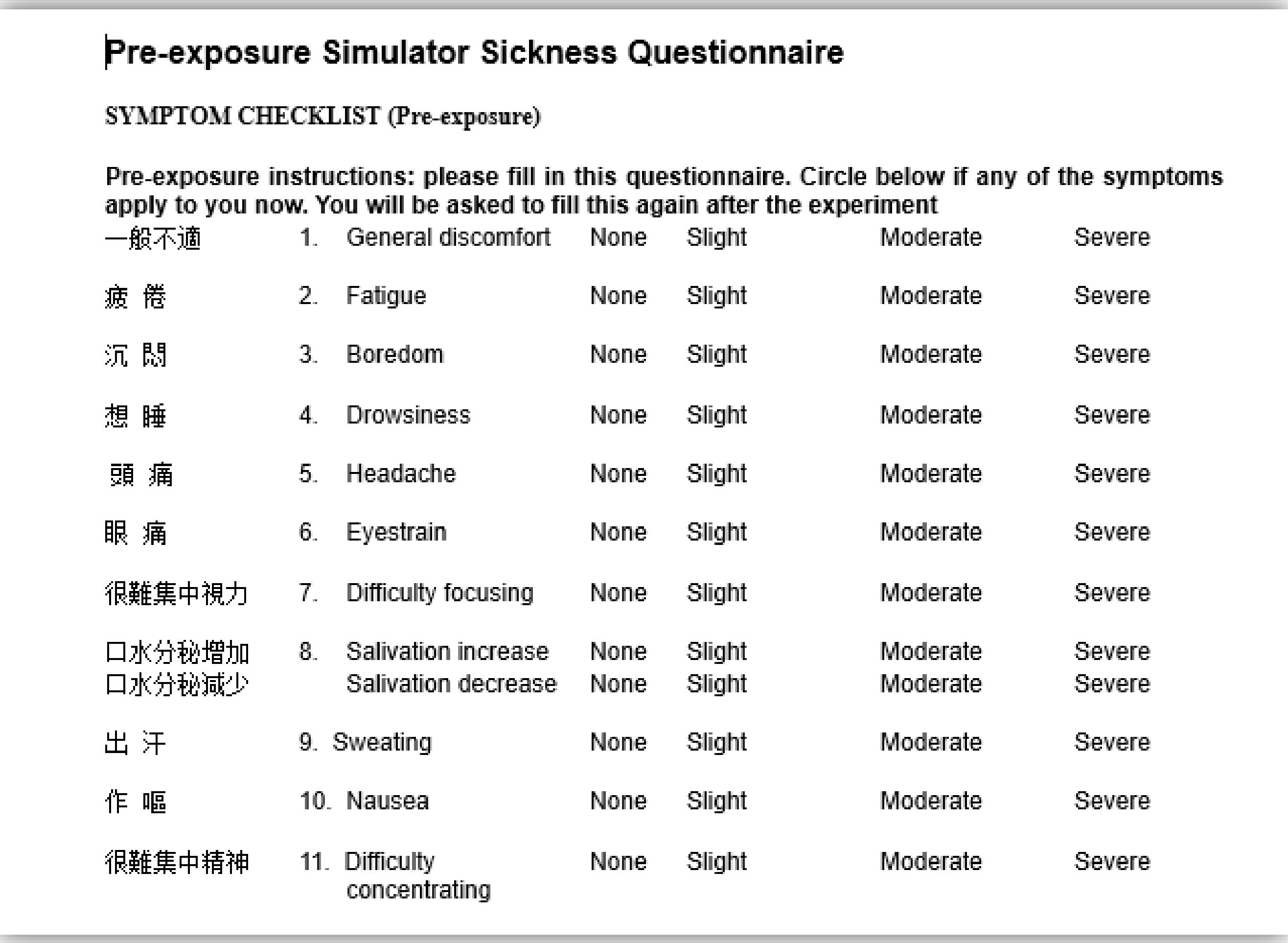
Temporal Demand

	1	2	3	4	5	6	7	8	9	10	
Very Low	○	○	○	○	○	○	○	○	○	○	Very High

Sickness and Cybersickness

- A popular questionnaire used to assess sickness using Virtual Reality devices is the SSQ (**Simulator Sickness Questionnaire**)
- The SSQ is a 27-item scale correspondent to a list of 27 symptoms which are commonly experienced by users of virtual reality systems

Sickness and Cybersickness



Pre-exposure Simulator Sickness Questionnaire

SYMPTOM CHECKLIST (Pre-exposure)

Pre-exposure instructions: please fill in this questionnaire. Circle below if any of the symptoms apply to you now. You will be asked to fill this again after the experiment

一般不適	1.	General discomfort	None	Slight	Moderate	Severe
疲 倦	2.	Fatigue	None	Slight	Moderate	Severe
沉 悶	3.	Boredom	None	Slight	Moderate	Severe
想 睡	4.	Drowsiness	None	Slight	Moderate	Severe
頭 痛	5.	Headache	None	Slight	Moderate	Severe
眼 痛	6.	Eyestrain	None	Slight	Moderate	Severe
很難集中視力	7.	Difficulty focusing	None	Slight	Moderate	Severe
口水分秘增加	8.	Salivation increase	None	Slight	Moderate	Severe
口水分秘減少		Salivation decrease	None	Slight	Moderate	Severe
出 汗	9.	Sweating	None	Slight	Moderate	Severe
作 嘔	10.	Nausea	None	Slight	Moderate	Severe
很難集中精神	11.	Difficulty concentrating	None	Slight	Moderate	Severe

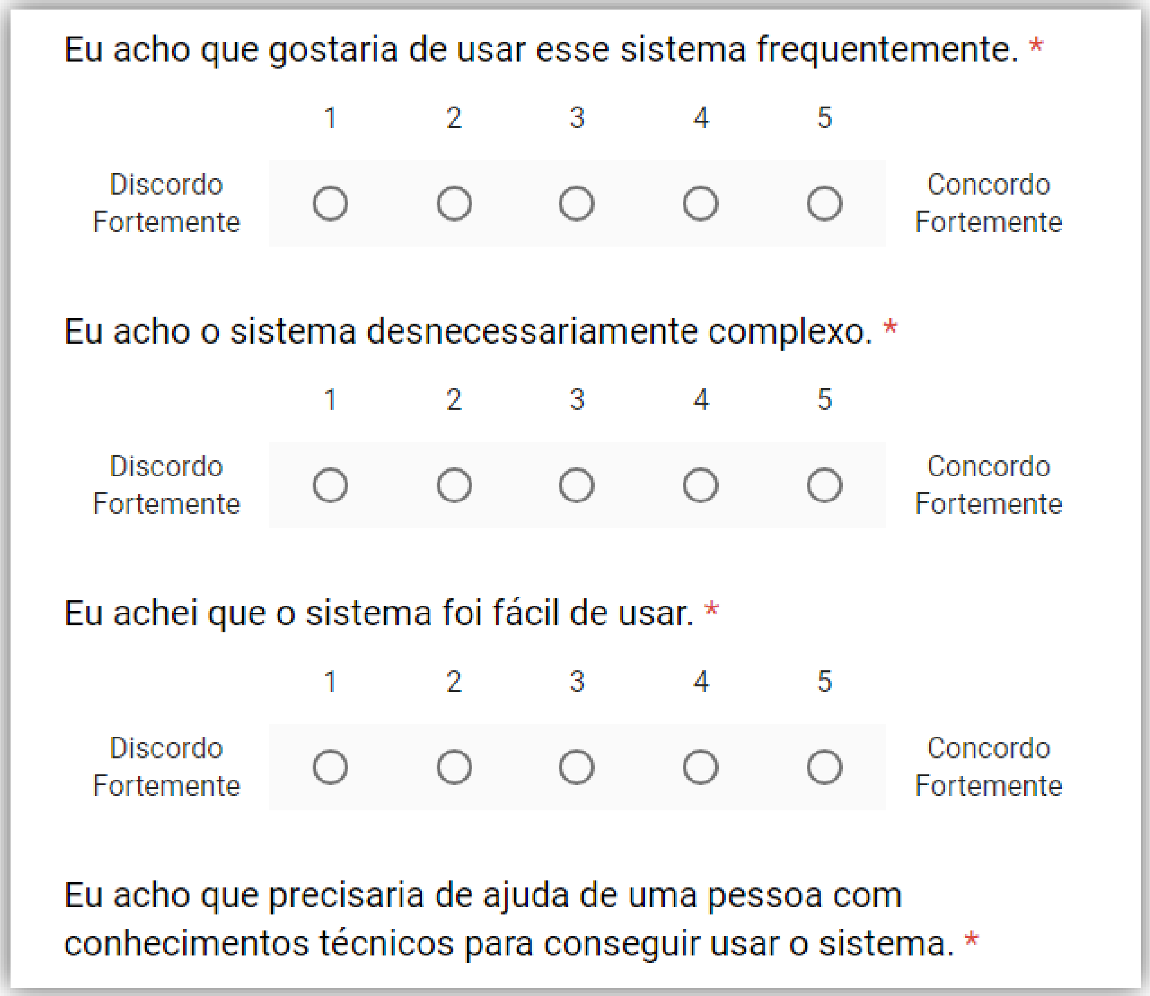
Usability

- Standardized usability questionnaires are questionnaires designed for the assessment of perceived usability
- Standardized questionnaires are available for assessment of a product at the end of a study (**post-study**) and after each task in a study (**post-task**)

Usability

- Examples of post-study questionnaires:
  - SUS - System Usability Scale (10-item) *is one of the most used
  - QUIS - User Interface Satisfaction (6 to 27 items)
  - SUMI - Software Usability Measurement Inventory (50-item)
  - PSSUQ - Post-Study Usability Questionnaire (13 to 19 items)
  - UMUX - Usability Metric for User Experience (4-item)
  - UMUX-LITE (2-item)

Usability



Eu acho que gostaria de usar esse sistema frequentemente. *

	1	2	3	4	5	
Discordo Fortemente	○	○	○	○	○	Concordo Fortemente

Eu acho o sistema desnecessariamente complexo. *

	1	2	3	4	5	
Discordo Fortemente	○	○	○	○	○	Concordo Fortemente

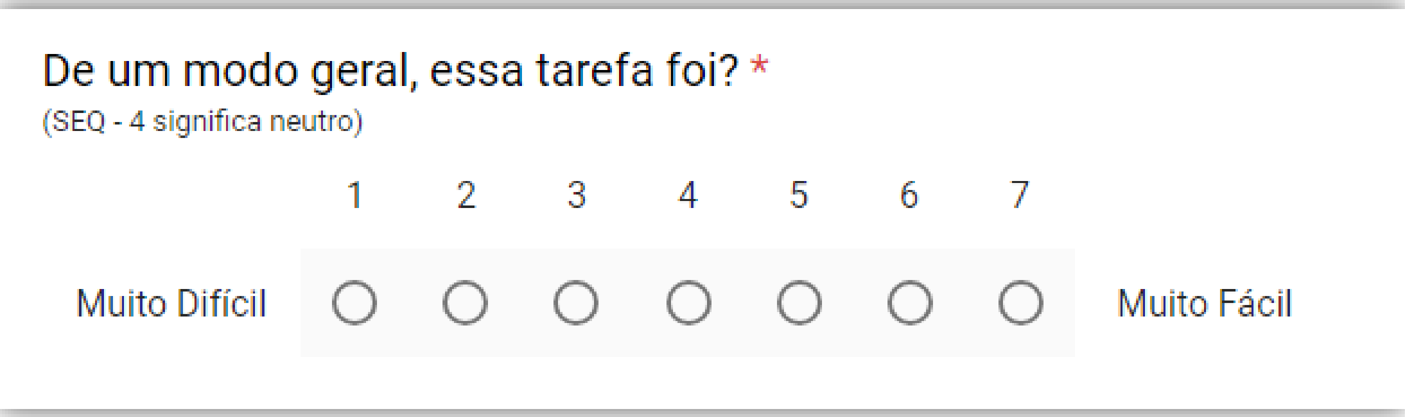
Eu achei que o sistema foi fácil de usar. *

	1	2	3	4	5	
Discordo Fortemente	○	○	○	○	○	Concordo Fortemente

Eu acho que precisaria de ajuda de uma pessoa com conhecimentos técnicos para conseguir usar o sistema. *

Usability

- Examples of post-task questionnaires:
  - ASQ - After Scenario Questionnaire (3-item)
  - SEQ - Single Ease Question (1-item)
  - SMEQ - Subjective Mental Effort Questionnaire (1-item)
  - UME - Usability Magnitude Estimation (1-item)



De um modo geral, essa tarefa foi? *

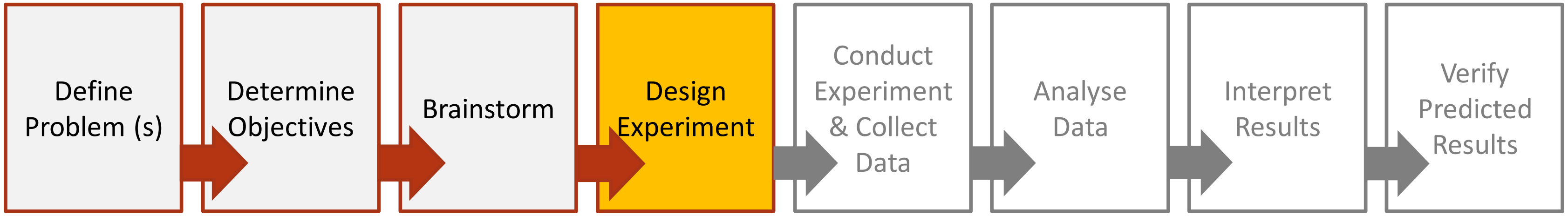
(SEQ - 4 significa neutro)

	1	2	3	4	5	6	7	
Muito Difícil	○	○	○	○	○	○	○	Muito Fácil

Questionnaires

https://goo.gl/KpjdmY

Experiment Design Process



- H: I predict that learning will increase according the level of immersion caused by the technique

-- What more can we assess in this setup? --

-- Which scales or questionnaires can we use? --

PARTE II

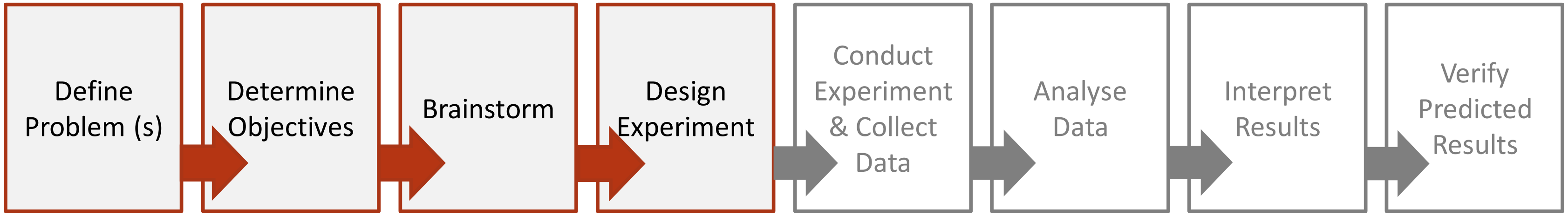
Design and Application of Experiments and User Studies

Victor Adriel de Jesus Oliveira
INF - UFRGS

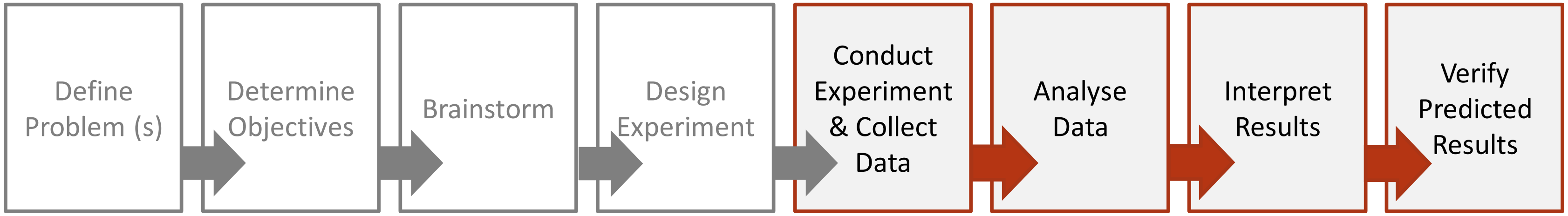
.Inf INSTITUTO DE INFORMÁTICA UFRGS

GRAPHICS VISUALIZATION INTERACTION | LAB

Experiment Design Process



Experiment Design Process



Define Problem (s)

Determine Objectives

Brainstorm

Design Experiment

Conduct Experiment & Collect Data

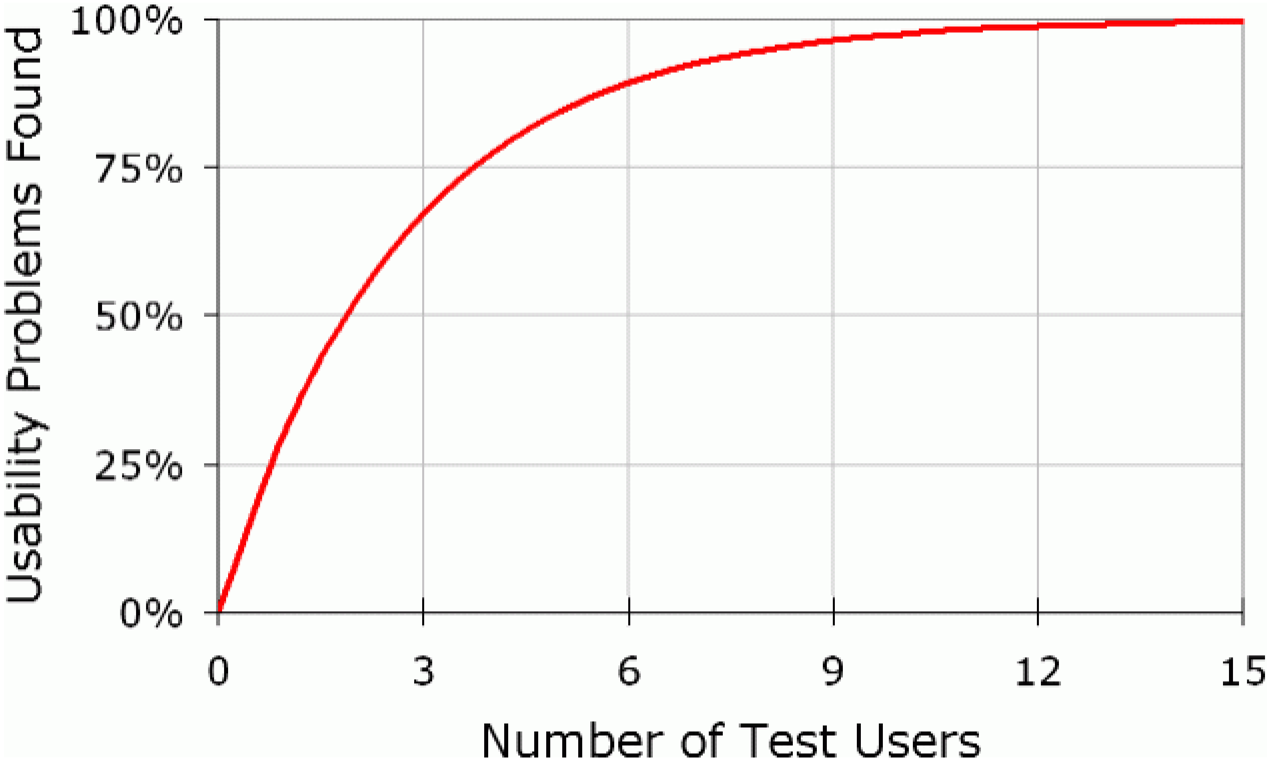
Analyse Data

Interpret Results

Verify Predicted Results

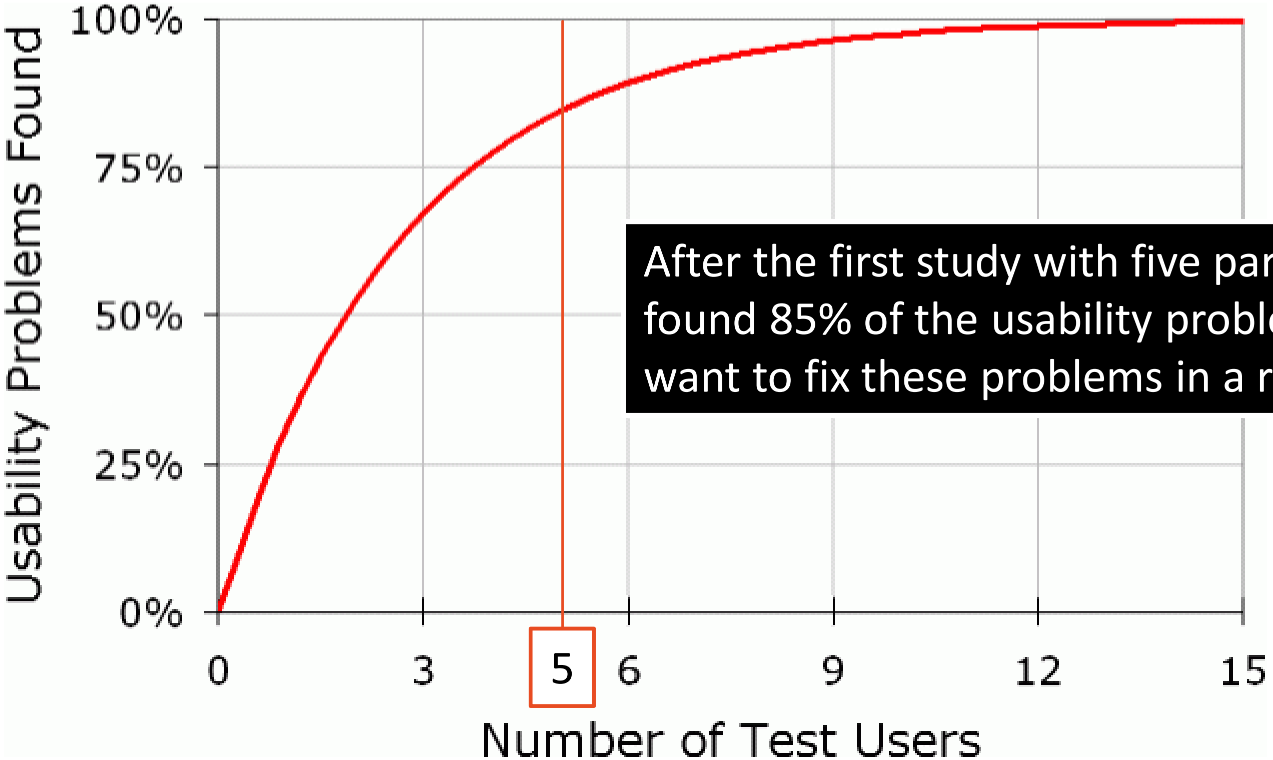
Application of User Studies

Recruiting Users



$N\,(1-(1-L\,)^{n}\,)$

Recruiting Users



After the first study with five participants has found 85% of the usability problems, you will want to fix these problems in a redesign

$N(1-(1-L)^{n})$

Recruiting Users

- Statistical Analysis
  - Sample size >= 30
  - Sample size should be at least of 5 participants per variable
  - Sample size should be of 10 participants per variable (10:1)
  - Sample size = $ available / $ per sample
  - Power increases with sample size

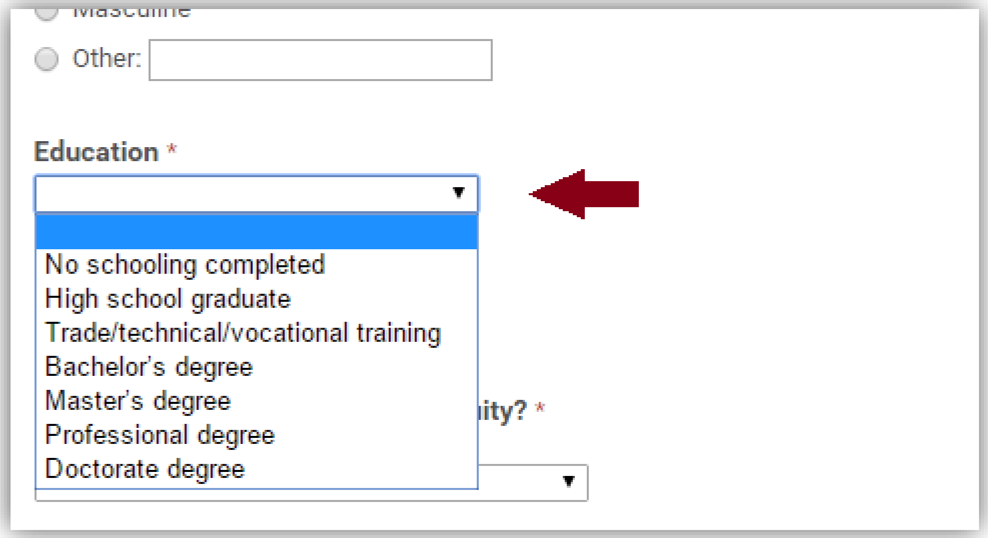
Power corresponds to the chance that you reject H0 when H0 is false (i.e. you [correctly] conclude that there is a treatment effect when there really is a treatment effect)

Recruiting Users

- A Screener survey has questions to determine who is the right fit for a particular study
  - Keep your screener short
  - Use simple sentence structure
  - Start broad and narrow down to your audience
  - Eliminate leading questions
  - Present questions with multiple answer options, as opposed to binary (yes/no) questions (and include “Other”/“none of the above” questions)
  - Place *necessary* demographic questions

Demographics

- Education
  - The user should chose the highest degree or level of school the subject has completed



○ Masculine

○ Other:

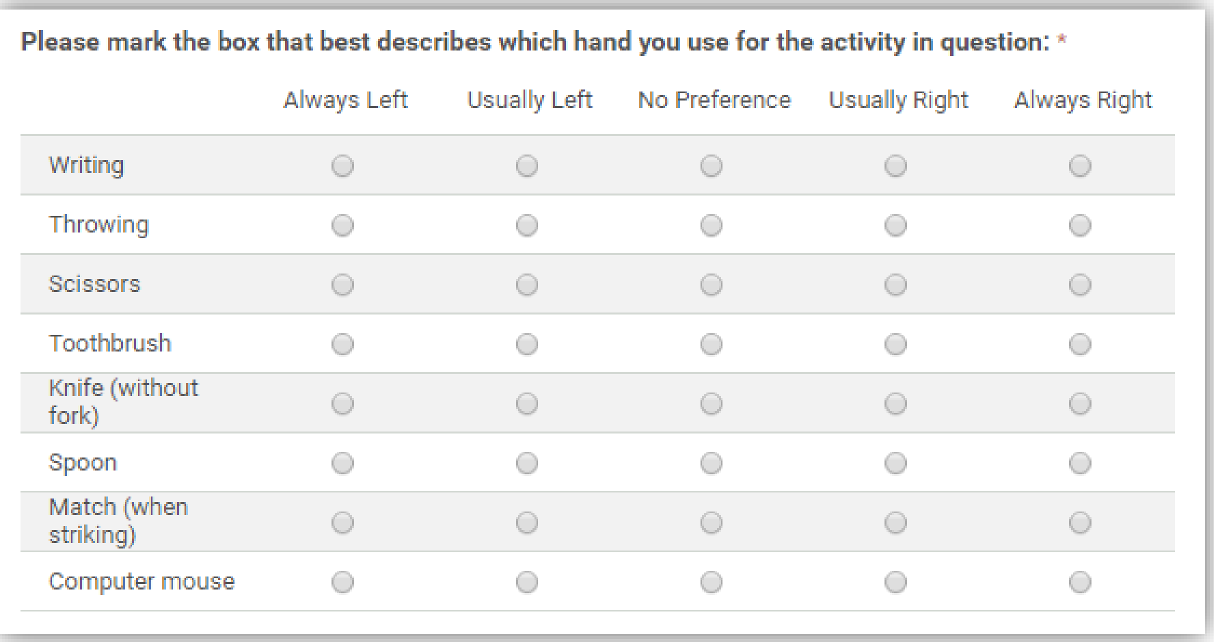
Education *

No schooling completed
High school graduate
Trade/technical/vocational training
Bachelor's degree
Master's degree
Professional degree
Doctorate degree

ity? *

Demographics

- Handedness
  - There are standard questionnaires for assessing handedness, such as the Edinburgh Inventory



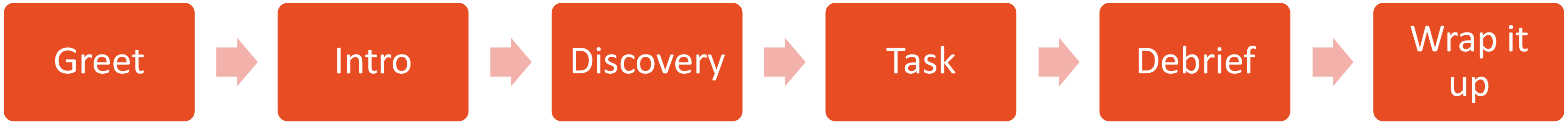
Please mark the box that best describes which hand you use for the activity in question: *

	Always Left	Usually Left	No Preference	Usually Right	Always Right
Writing	○	○	○	○	○
Throwing	○	○	○	○	○
Scissors	○	○	○	○	○
Toothbrush	○	○	○	○	○
Knife (without fork)	○	○	○	○	○
Spoon	○	○	○	○	○
Match (when striking)	○	○	○	○	○
Computer mouse	○	○	○	○	○

Legal and Ethical Issues

- The purpose of the research can not precede the rights and interests of each research subject
- Risks must be predicted, evaluated and managed
- The research should be based on thorough knowledge of the scientific literature
- The privacy of the research subject and the confidentiality of your personal information should be protected
- Subjects must give informed consent

General Protocol


Greet → Intro → Discovery → Task → Debrief → Wrap it up

- Keep the place clean and organized
- Maintain a standard for every subject
- Make the user comfortable
- Design task scenarios
- Do not lead the user`s answers

Analysis and Report of Results

Tools

- R, Excel, Python, BioEstat, SPSS…

Data in R

User ID	Cond	Value
1	1	75.0
1	2	42.0
1	3	80.3

Data in Excel

User ID	Cond1	Cond2	Cond3
1	75.0	42.0	80.3

Tools

- R, Excel, Python, BioEstat, SPSS…

Data in R

User ID	Cond	Value
1	1	75.0
1	2	42.0
1	3	80.3

Data in Excel

User ID	Cond1	Cond2	Cond3
1	75.0	42.0	80.3
2	90.6	88.4	95.0
3	53.6	45.9	60.0
4	89.0	60.0	88.5
5	60.0	55.0	75.9

Tables

- Reporting data

User ID	Cond1	Cond2	Cond3
1	75.0	42.0	80.3
2	90.6	88.4	95.0
3	53.6	45.9	60.0
4	89.0	60.0	88.5
5	60.0	55.0	75.9

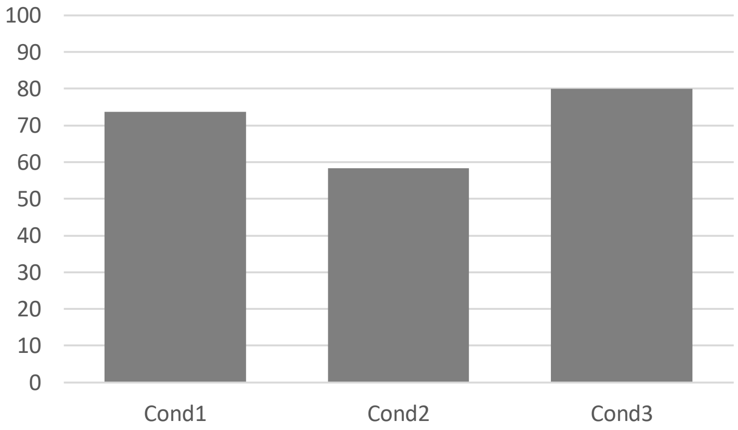
Good Table

User ID	Cond1	Cond2	Cond3
1	75	42.0	80.3
2	90.65	88.43	95.0
3	53.665	45.9	60
4	89	60.0	88.5
5	60.0	55.000	75.9

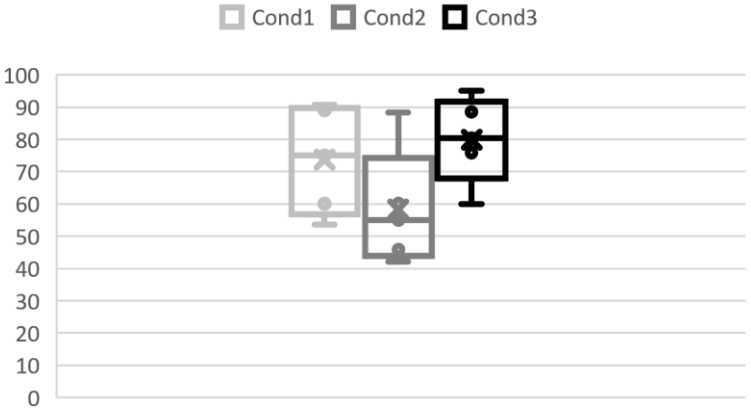
Bad Table

Charts

- Visualizing data



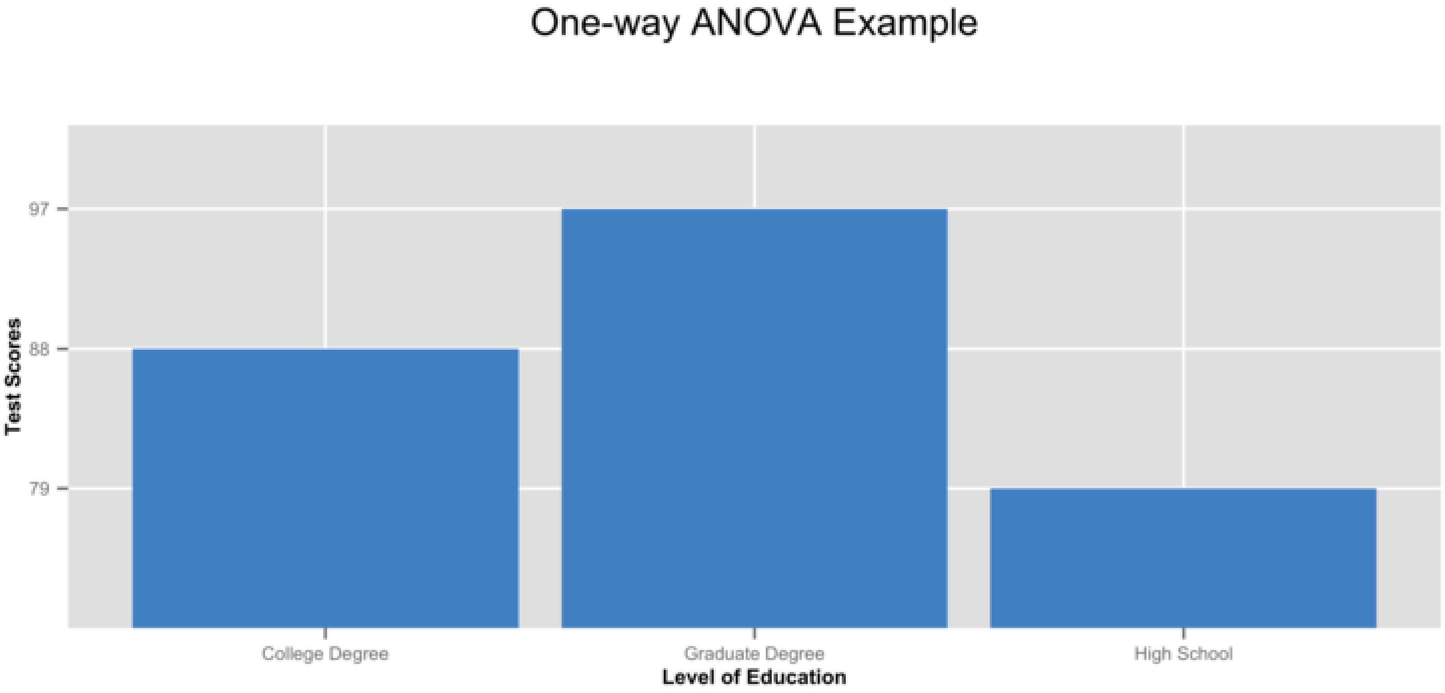
Less informative



More informative

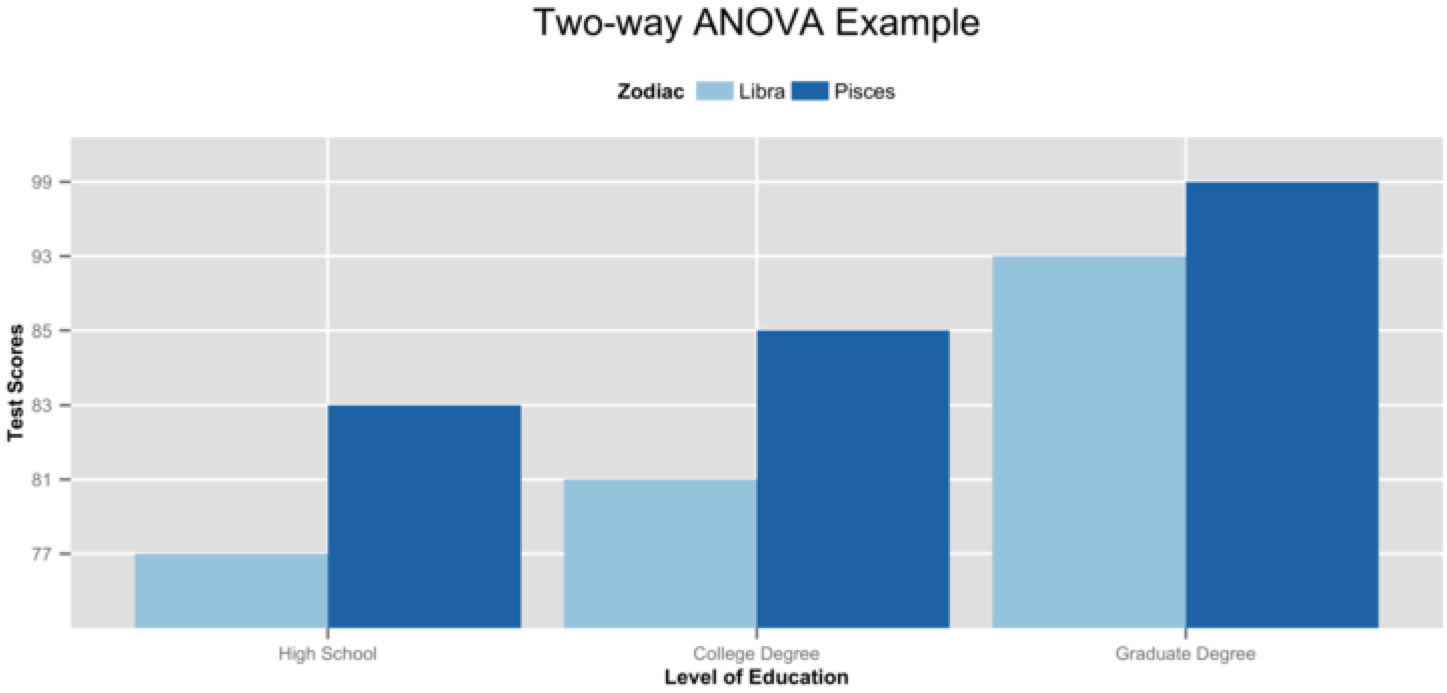
Test Choice

- "Analysis of Variance" (ANOVA)



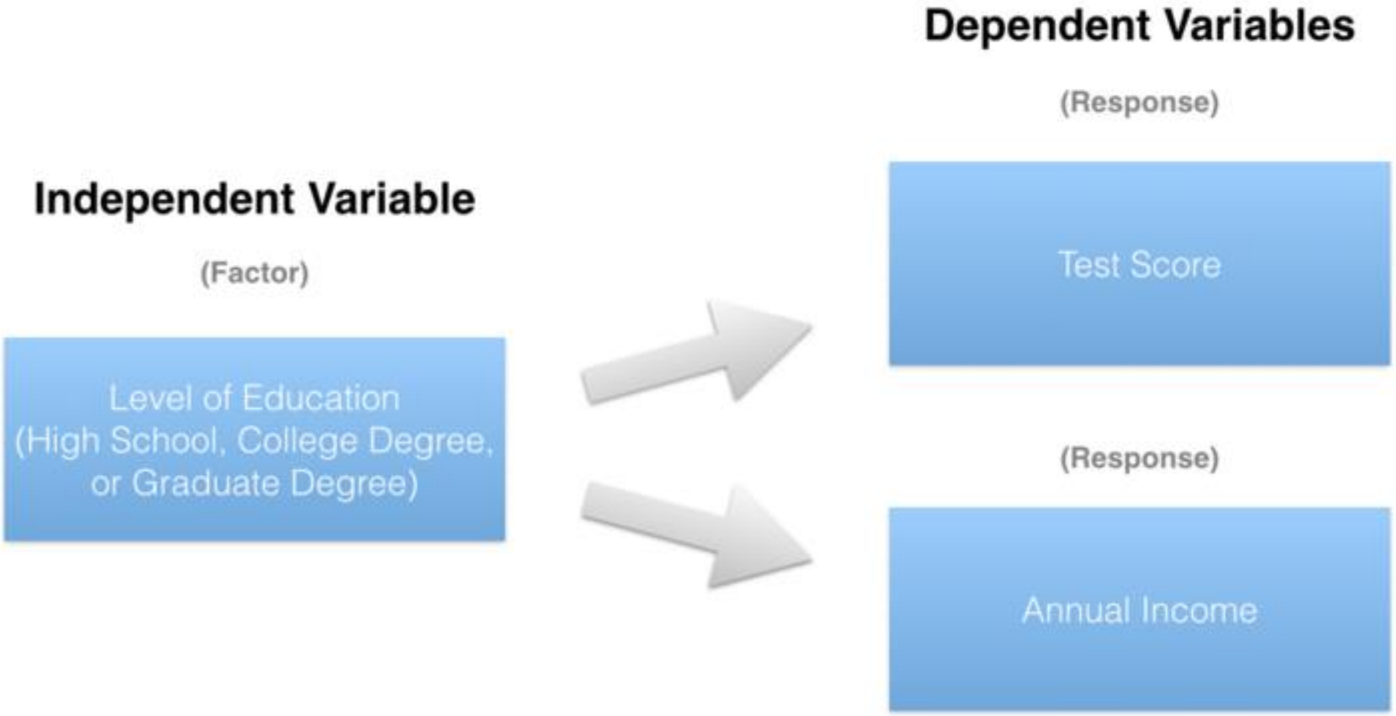
Test Choice

- "Analysis of Variance" (ANOVA)



Test Choice

- MANOVA: "Multivariate Analysis of Variance“
  - One-way



Dependent Variables

(Response)

Test Score

Independent Variable

(Factor)

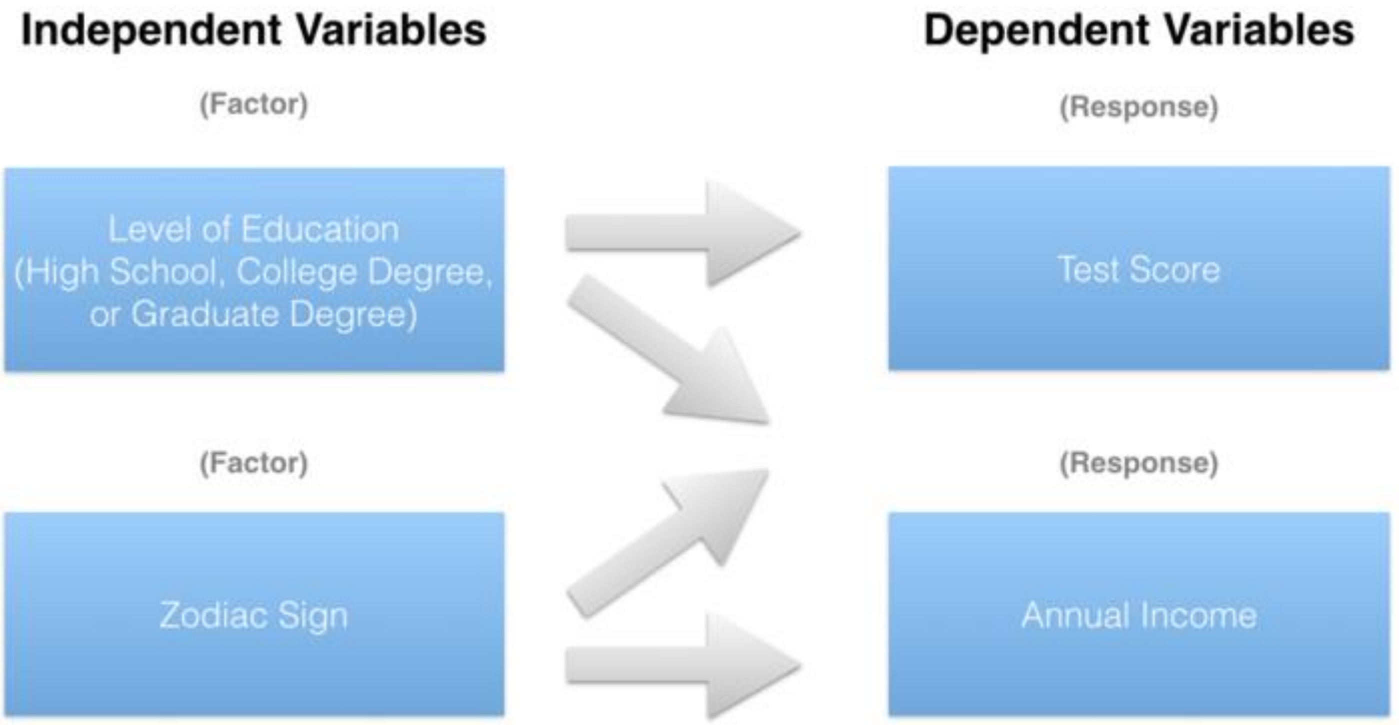
Level of Education (High School, College Degree, or Graduate Degree)

(Response)

Annual Income

Test Choice

- MANOVA: "Multivariate Analysis of Variance“
  - Two-way



Test Choice

- Test for normality (Kolmogorov-Smirnov (K-S), Shapiro-Wilk, etc.)

Test Choice

- Test for normality (Kolmogorov-Smirnov (K-S), Shapiro-Wilk, etc.)
- If (NORMAL DISTRIBUTION)
  - Analysis of Variance: ANOVA/MANOVA
  - Post-hoc (means): Tukey's HSD, Student's t-test...

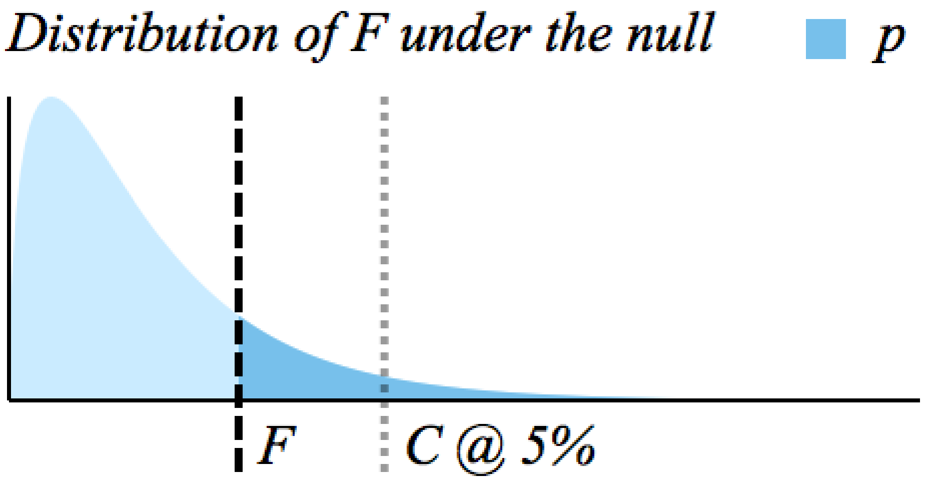
Test Choice

- Test for normality (Kolmogorov-Smirnov (K-S), Shapiro-Wilk, etc.)
- If (NORMAL DISTRIBUTION)
  - Analysis of Variance: ANOVA/MANOVA
  - Post-hoc (means): Tukey's HSD, Student's t-test...
- Else // proceed with non-parametric tests

Test Choice

- Test for normality (Kolmogorov-Smirnov (K-S), Shapiro-Wilk, etc.)
- If (NORMAL DISTRIBUTION)
  - Analysis of Variance: ANOVA/MANOVA
  - Post-hoc (means): Tukey's HSD, Student's t-test...
- Else // proceed with non-parametric tests
  - If (WITHIN SUBJECTS)
    - Analysis of Variance: Friedman test
    - Post-hoc (means): Wilcoxon analyses
  - If (BETWEEN SUBJECTS)
    - Analysis of Variance: Kruskal-Wallis
    - Post-hoc (means): Dunn analyses

Report Results



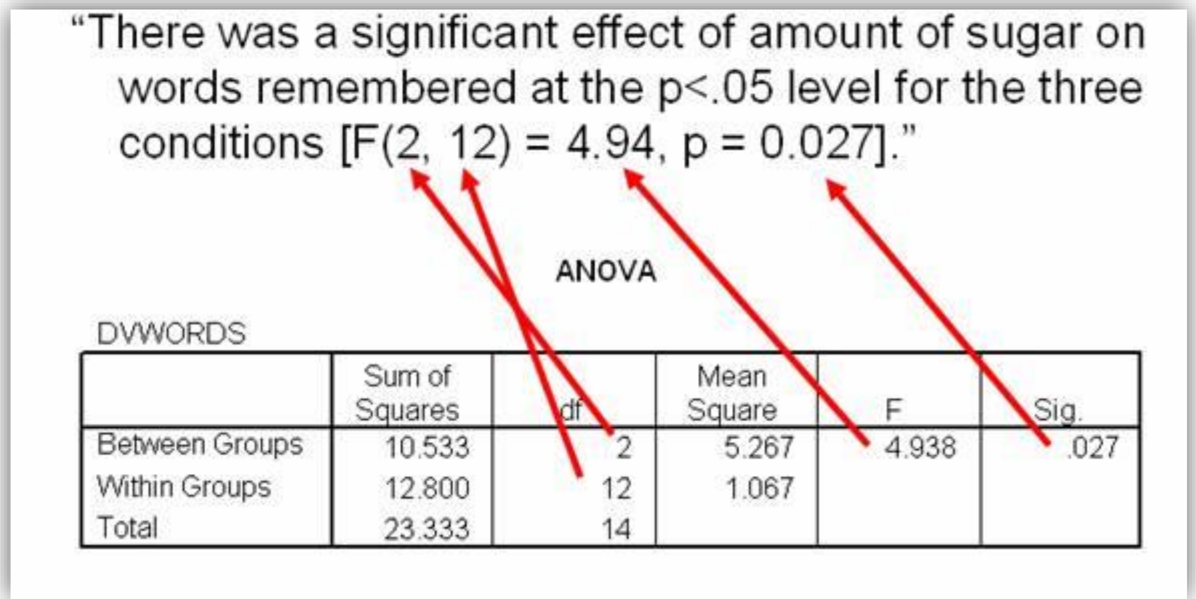
- F distribution: the distribution of F statistics that we'd see if the null hypothesis were true
- F statistic here would result in a failure to reject the null hypothesis because it is less than C, that is, its p value is greater than .05

Report Results

- APA Style
  - “There was a significant (not a significant) effect of IV ______________ on DV ________________ at the p<.05 level for the three conditions [F(*degrees of freedom*) = *F-value*, p = *p-value*].

Report Results

- APA Style
  - “There was a significant (not a significant) effect of IV ______________ on DV ________________ at the p<.05 level for the three conditions [F(*degrees of freedom*) = *F-value*, p = *p-value*].



“There was a significant effect of amount of sugar on words remembered at the p<.05 level for the three conditions [F(2, 12) = 4.94, p = 0.027].”

ANOVA

DVWORDS

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	10.533	2	5.267	4.938	.027
Within Groups	12.800	12	1.067		
Total	23.333	14			

Report Results

- Multiple Factor
  - There was a significant main effect for treatment, F(1, 145) = 5.43, p < .01, and a significant interaction, F(2, 145) = 3.13, p < .05.
- Mean and Standard Deviation
  - (M = 12.4, SD = 2.26)
- Correlations
  - The two variables were strongly correlated, r(55) = .49, p < .01.

Examples

Referências

- https://www.nngroup.com/articles/why-you-only-need-to-test-with-5-users/
- https://www.moresteam.com/toolbox/design-of-experiments.cfm#purposeExperimentation
- https://en.wikipedia.org/wiki/Design_of_experiments
- https://www.isixsigma.com/tools-templates/design-of-experiments-doe/design-experiments-%E2%90%93-primer/
- https://www.passeidireto.com/arquivo/23301968/psicometria-hutz-bandeira-trentini
- http://www.statsmakemecry.com/smmctheblog/stats-soup-anova-ancova-manova-mancova

Design and Application of Experiments and User Studies

Victor Adriel de Jesus Oliveira

vajoliveira@inf.ufrgs.br

Skype: victor.adriel