Manuel Giuliani
Praminda Caleb-Solly

Week 4

Human-Robot Interaction

User Studies

UWE Bristol | University of the West of England

Lectures overview

W.	Lecture
1	Introduction to HRI
2	Human factors and context
3	Design for HRI systems
4	User studies
5	Social signal processing
6	Natural language processing

W.	Lecture
7	Speech synthesis
8	Human-aware motion planning
9	Symbolic reasoning for HRI
10	Architecture for HRI
11	Statistics for HRI user studies
12	Exam revision

UWE Bristol | University of the West of England

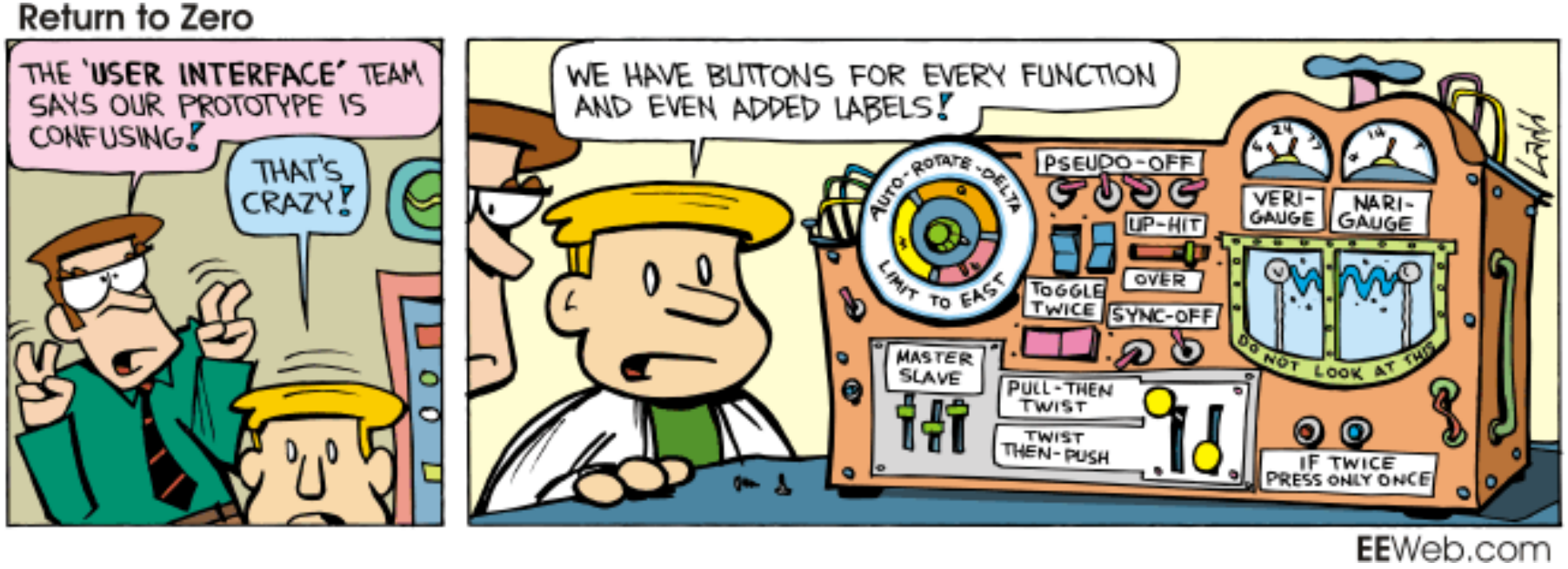
Learning outcomes

Being able to explain the concepts validity and reliability.

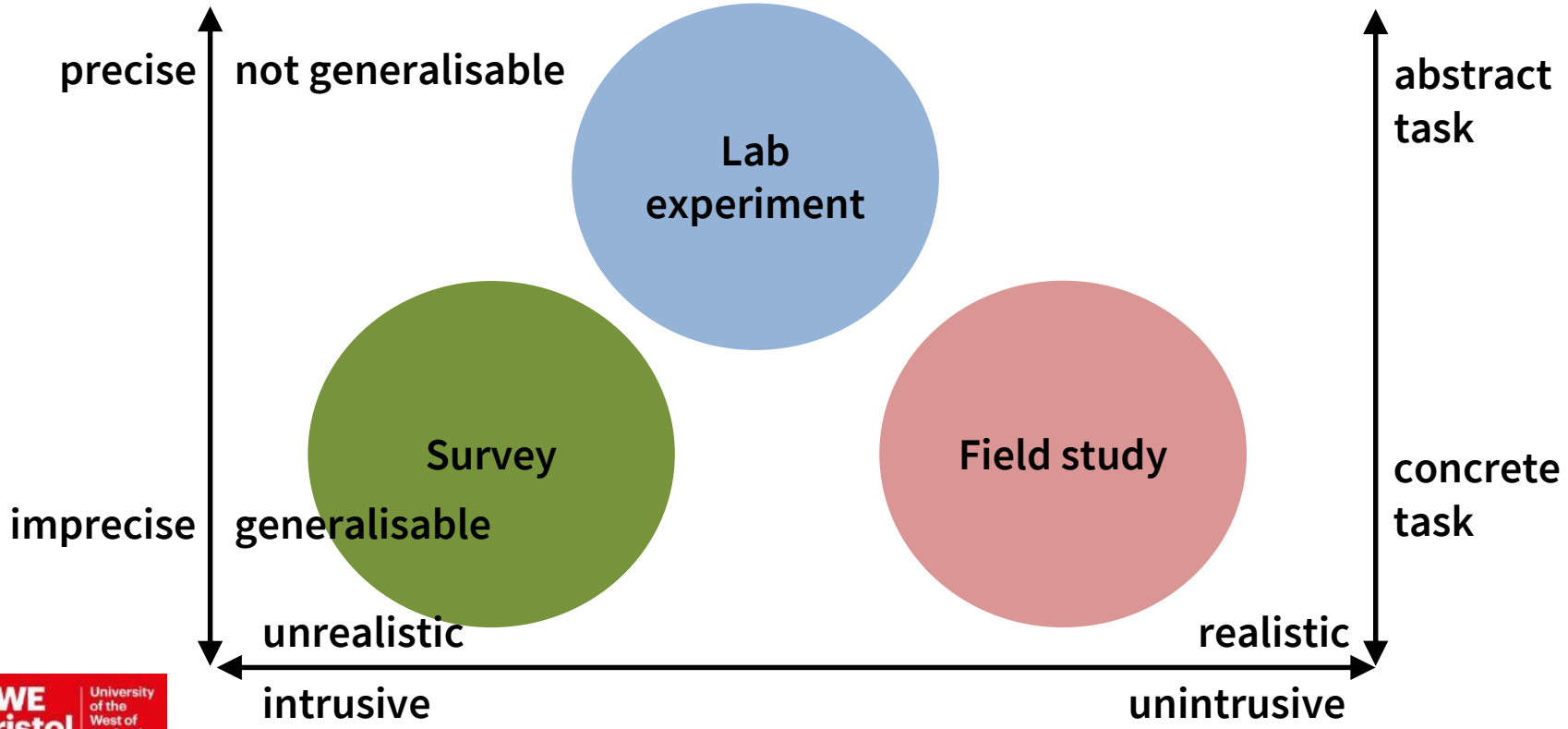
Knowing the parts of a controlled experiment.

UWE Bristol | University of the West of England

User testing



User study methods



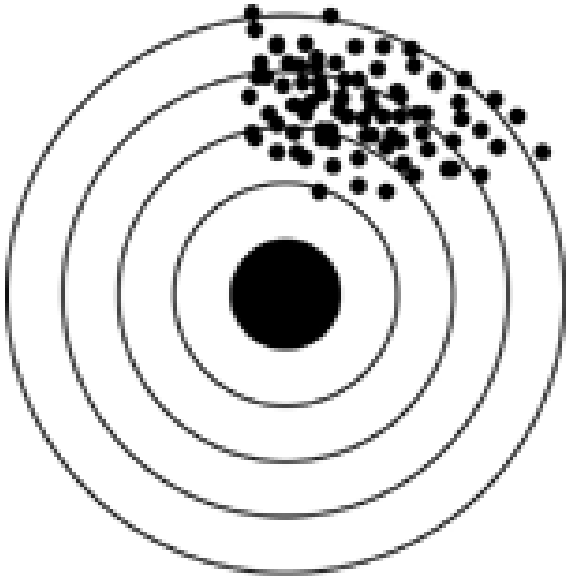
UWE Bristol University of the West of England

User study methods

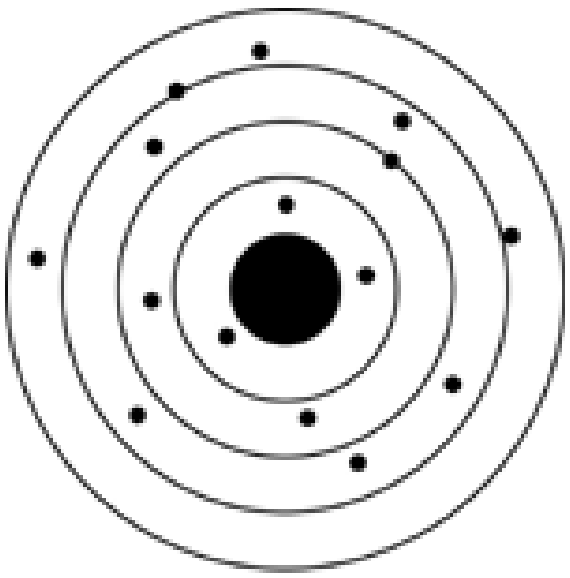
Methods for testing with users can be compared on several dimensions. The horizontal dimension distinguishes whether a test is intrusive or unobtrusive. That means whether the user has to put the full attention focus on the test and whether the user is aware that he is being watched. Also on the horizontal dimension you will find the realism of the test. For example, field studies take part directly in the context in which the user interacts with the robot, that is in a real environment. Laboratory experiments, as the name implies, take place in a laboratory, i.e. in an unrealistic environment.

The vertical dimension distinguishes whether a test is precise or imprecise. The more one controls the test, e.g. eliminating disturbing environmental influences, the more precise it becomes. However, the more precise a test is, the more abstract the tasks are the user is supposed to work on in the test. The experiments of Paul Fitts, with which he proved the Fitts' Law named after him, are an example of this kind of precise experiment with an abstract task (https://en.wikipedia.org/wiki/Fitts's_law). Lastly, the generalization of the test can be applied to the vertical dimension. Surveys are particularly well-suited for exploring generalizable results, as it can reach a large, representative number of users.

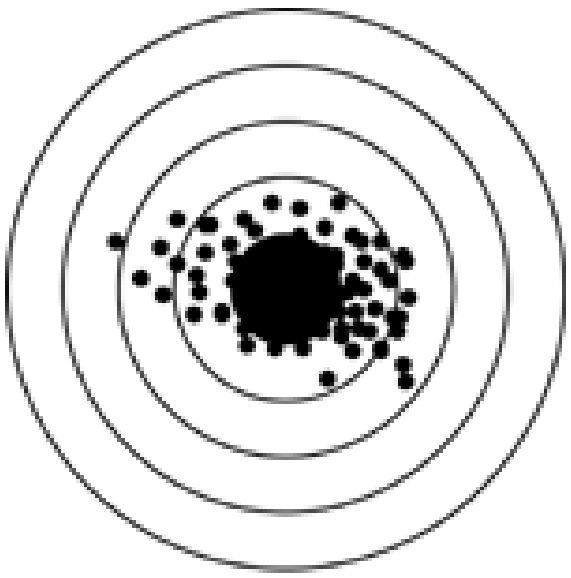
Validity and reliability



Reliable but Not Valid



Valid but Not Reliable



Valid and Reliable

Validity and reliability

When designing user tests, one must always be concerned with the validity and reliability of the test.

Validity is used to indicate whether the test actually answers the question that one wants to research or is relevant to the question. To design a valid test, you need a good basic knowledge of the test method you use as well as common sense.

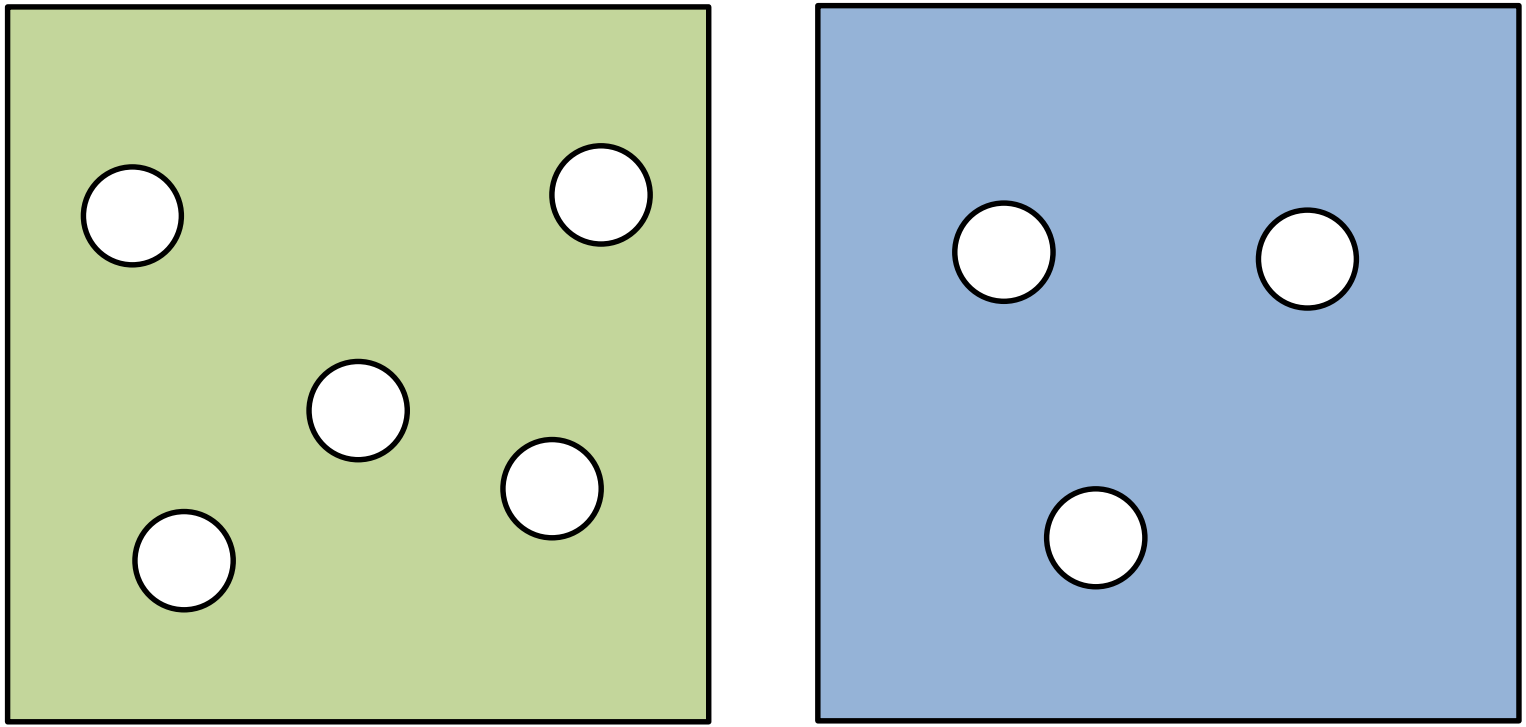
The **reliability** of a test describes how reliably the results of the test are reproducible and generalizable. The reliability of a test can be increased by repetition of the test and statistical evaluation.

Validity and reliability

- **Internal validity**
  - Are the experimental results influenced only by the experimental variables?
- **External validity**
  - Can the results be generalized beyond the experiment?
- **Reliability**
  - If I repeat the experiment, will I get the same results?

UWE Bristol | University of the West of England

Example: Balls in boxes



UWE Bristol University of the West of England

Example: Balls in boxes

- Test question: Which box has more balls?
- **Reliability**
  - Counting the balls is only reliable if the boxes contain a few balls
  - Repeated counting increases reliability
- **Internal validity**
  - You could weigh the boxes instead of counting the balls
  - What if the balls are different in weight?
  - What if the boxes are different in weight?
- **External validity**
  - Are our results valid for all green and blue boxes in the world?

UWE Bristol | University of the West of England

Influences on internal validity

- Order effects
- Selection effects
- Experimenter bias

UWE Bristol | University of the West of England

Influences on internal validity

- Order effects
  - Study participants get better because they learn how the robot behaves
  - Subjects get worse because they get tired

  → Randomizing the order of experiment tasks is important
- Selection effects
  - Do not use already existing groups of study participants (e.g. all students of one class, who all have the same knowledge)
  - Assign participants randomly to experimental conditions

UWE Bristol | University of the West of England

Influences on internal validity

- Experimenter bias
  - Depending on the hypothesis, experimenter has a preference for one of the experimental conditions
  - → Give instructions on paper or video
  - → Provide the same instructions for all experiment conditions
  - → Double-blind experiment: Experimenter and participant both do not know which experimental condition the participant is experiencing

Influences on external validity

- Participants
- Environmental influences
- Instructions
- Tasks

UWE Bristol | University of the West of England

Influences on external validity

- Participants
  - Select participants from target group
  - Separate existing subgroups
- Environmental influences
  - Produce the most realistic environment possible in laboratory studies
  - Eliminate disturbing environmental influences that are not part of the experiment
- Instructions
  - Imitate human-human interaction
  - Example: Robot bartender behaves similar to a human bartender
- Tasks
  - The task of the user should be based on previous context analysis

Influences on reliability

- Uncontrolled variations
  - Pre-experience of the users
  - Inaccurate task description
  - Measuring errors
- Eliminate uncontrolled variables as much as possible
  - Query the prior experience of the users
  - Precise measuring
- Repetition
  - Many participants, many experiment runs
  - Statistical evaluation of measurements

Learning outcomes

Being able to explain the concepts validity and reliability.

Knowing the parts of a controlled experiment.

UWE Bristol | University of the West of England

Controlled experiments

- **Hypothesis**
- Independent variables
- Dependent variables
- Experiment design

Controlled experiments

- Hypothesis
  - Hypothesis must be testable (= quantifiable and measurable)
  - Hypotheses involve assumptions by the experiment designer

UWE Bristol | University of the West of England

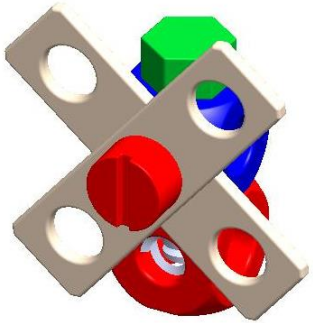
UWE Bristol | University of the West of England

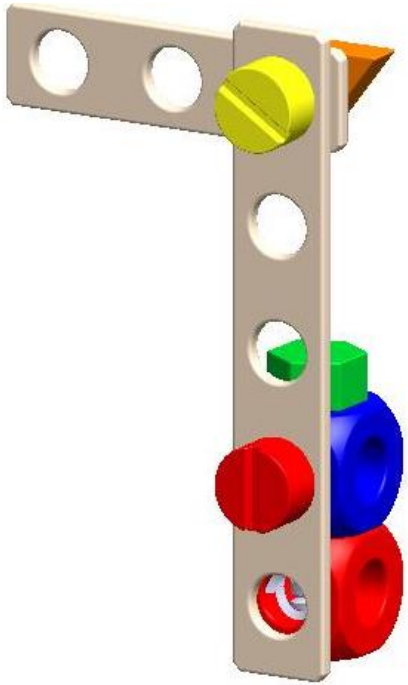
Experiment description

- Participant and robot build two objects according to a blueprint
- The blueprint of the participant contains errors
- Robot detects errors and explains what participant did wrong

UWE Bristol | University of the West of England

Objects to build





Hypotheses

- When the robot adapts its language to the task context, it is perceived as a better dialogue partner
- Adaptive speech output can cause user confusion if it is not implemented properly

Controlled experiments

- Hypothesis
- **Independent variables**
- Dependent variables
- Experiment design

UWE Bristol | University of the West of England

Controlled experiments

- Hypothesis
  - Hypothesis must be testable (= quantifiable and measurable)
  - Hypotheses involve assumptions by the experiment designer
- Independent variables
  - Test variables that are changed between experiment conditions (for example, two different versions of a robot behaviour)

UWE Bristol | University of the West of England

Independent variables

- Robot has two ways to refer to objects
- Constant: Robot always uses the same expression (e.g., "I'll give you a green cube")
- Adaptive: robot includes context (e.g., "I give you this green cube")

UWE Bristol | University of the West of England

Constant language generation

Adaptive language generation

Controlled experiments

- Hypothesis
- Independent variables
- **Dependent variables**
- Experiment design

Controlled experiments

- Hypothesis
  - Hypothesis must be testable (= quantifiable and measurable)
  - Hypotheses involve assumptions by the experiment designer
- Independent variables
  - Test variables that are changed between experiment conditions (for example, two different versions of a robot behaviour)
- Dependent variables
  - Variables that are measured to test the hypothesis
  - Objective variables: time, counting events, ...
  - Subjective variables: questionnaires, satisfaction, ...

UWE Bristol | University of the West of England

Dependent variables

- Subjective: questionnaire with 4 categories
  - Intelligence of the robot
  - Difficulty of the task
  - Feelings of the users
  - Quality of conversation
- Objective:
  - Counted events in experiment videos
  - Log data
  - E.g., user and robot turns, dialogue length, number of correctly built objects

UWE Bristol | University of the West of England

Controlled experiments

- Hypothesis
- Independent variables
- Dependent variables
- **Experiment design**

Between subjects vs within subjects design

	Between Subjects	Within Subjects
Condition A		
Condition B		

UWE Bristol | University of the West of England

Between subject vs within subject design

- Between Subjects
  - Users are split into multiple groups
  - One group per experiment condition
  - Comparison of results between groups
  - Eliminates order effects
- Within Subjects
  - Each user sees all experimental conditions
  - Results are compared individually for each user
  - Eliminates variation due to differences between users

- For between subjects experiments, more participants are needed because user differences are larger than order effects

- External validity is higher in between subjects experiment, since in reality users will only use one of the two robot variants

Counterbalancing

- For within subjects experiments
- Assignment of subjects in a Latin square
- Elimination of order effects
- Well suited for small numbers of users
- Guarantees that every condition of the experiment comes in every position equally often

EC1	EC2
A	B
B	A

EC1	EC2	EC3
A	C	B
B	A	C
C	B	A

EC1	EC2	EC3	EC4
A	D	C	B
B	A	D	C
C	B	A	D
D	C	B	A

UWE Bristol University of the West of England

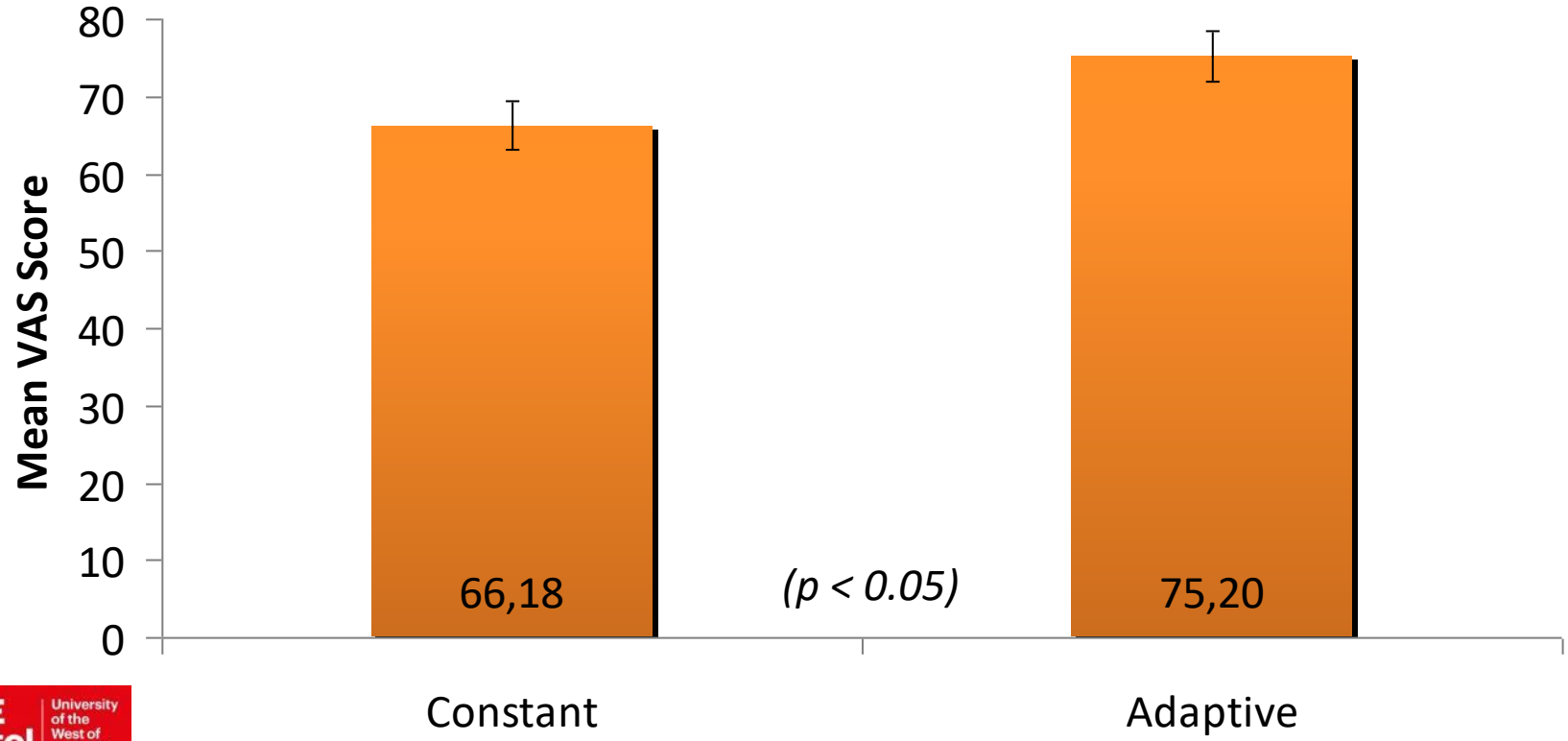
Results JAST Experiment

- Between subjects experiment
- Participants heard either Constant or Adaptive Language Generation
  - 41 participants
  - Average age 24.5 years (19 - 42)
  - 33 male, 9 female
- Participants were alternately assigned to conditions (constant, adaptive)

UWE Bristol | University of the West of England

Results

Significant difference for subjective dependent variable "quality of conversation"



UWE Bristol University of the West of England

Self check

- What are the internal and external validity of a user study?
- Name the components of a controlled experiment?