UNIT - IV

Empirical Research Methods in HCI

Objective

- The key concerns in user evaluation.
- Data collection procedure.
- Data analysis techniques.

Empirical Research

- Empirical research is broadly defined as the observation-based investigation seeking to discover and interpret facts, theories, or laws.
- Collection and analysis of end user data for determining usability of an interactive system is an "observation-based investigation", hence it qualifies as empirical research

Themes of Empirical Research

Generally speaking, empirical research is based on three themes

- Answer and raise questions about a new or existing UI design or interaction method
- Observe and measure
- User studies

Research Question

- It is very important in an empirical research to formulate "appropriate" research questions
- For example, consider some questions about a system
  - Is it viable?
  - Is it as good as or better than current practice?
  - Which of several design alternatives is best?
  - What are its performance limits and capabilities?
  - What are its strengths and weaknesses?
  - How much practice is required to become proficient?

Testable Research Question

- Preceding questions, while unquestionably relevant, are not testable
- We have to come-up with testable questions in empirical research
- **Example**: Suppose you have designed a new text entry technique for mobile phones. You think the design is good. In fact, you feel your method is better than the most widely used current technique, multi-tap. You decide to undertake some empirical research to evaluate your invention and to compare it with multi-tap? What are your research questions?

Weak question

Q. Is the new technique better than multi-tap?

- Better

Q. Is the new technique faster than multi-tap?

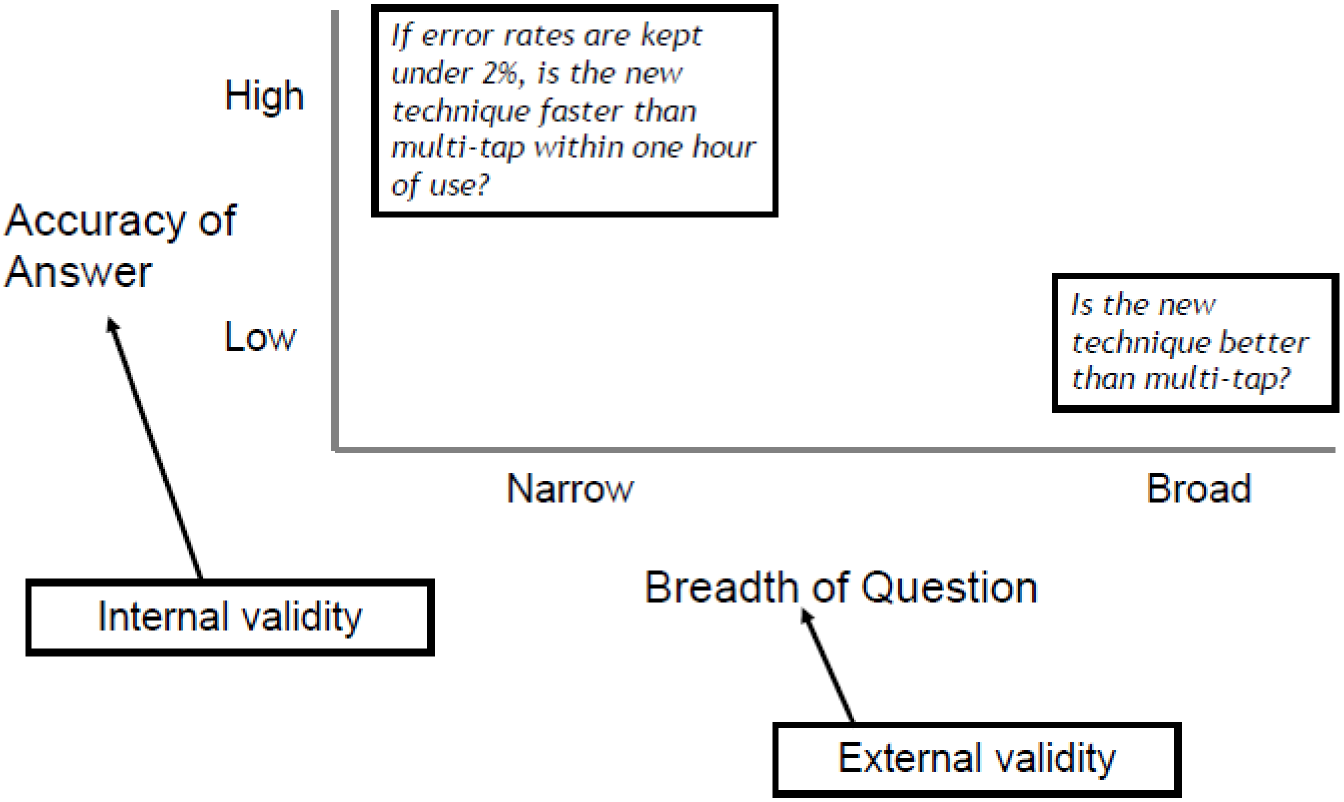
- Better still

Q. Is the new technique faster than multi-tap within one hour of use?

- Even better

Q. If error rates are kept under 2%, is the new technique faster than multi-tap within one hour of use?

- The questions are testable (we can actually conduct experiments to test the answer to the questions)
- We can ask very specific questions (the last one) or relatively broad questions (the first one).
- For very specific questions, the accuracy of answers is high whereas for broader questions, the breadth or generalizability is high.


If error rates are kept under 2%, is the new technique faster than multi-tap within one hour of use?
High
Accuracy of Answer
Low
Is the new technique better than multi-tap?
Narrow
Broad
Breadth of Question
Internal validity
External validity

Internal and External Validity

- The extent to which the effects observed are due to the test conditions is called internal validity of the research question.
- The extent to which results are generalizable to other people and other situations is known as the external validity of the research question.

Key Differences between Internal and External Validity

The points presented to you describe the differences between internal and external validity:

1. The extent to which the experiment is free from errors and any difference in measurement is due to independent variable and nothing else is known as an independent variable. The extent to which the research results can be inferred to the world at large is known as a dependent variable.
2. Internal Validity is nothing but the measure of the accuracy of the experiment. On the contrary, external validity examines whether the cause and effect connection between the dependent and independent variable found in the experiment can be generalized or not.

3. Internal validity is concerned with control of extraneous variable, whereas external validity stresses on the applicability of the outcome to the practical situations.
4. Internal validity ascertains the strength of the research methods and design. Conversely, external validity examines the generality of the research outcomes to the real world.
5. Internal Validity determines the extent to which the conclusion is warranted. As against this, external validity ascertains the extent to which the study is warranted to generalize the result to another context.
6. Internal Validity either addresses or eliminates alternative explanation for the result. In contrast, external validity is used to generalize the outcome.

More Examples on Validity

- Suppose you wish to compare two input devices for remote pointing (e.g., at a projection screen)
- External validity is improved if the test environment mimics expected usage.
  –The test environment should use a projection screen, position participants at a significant distance from screen, have participants stand and include an audience.
- Note that creating the test environment mimicking the real usage scenario is not easy.
- Instead you can go for controlled experiments where you can ask the user to sit in front of a computer in a laboratory and use the pointing devices to operate an application on the screen
  -The above setting can answer research questions with high internal validity but cannot help in determining if the answers are applicable in real world.
- Consider another scenario where you wish to compare two text entry techniques for mobile devices.
- To improve external validity, the test procedure should require participants to enter representative samples of text (e.g., phrases containing letters, numbers, punctuation, etc.) and correct mistakes
  –This may require compromising on internal validity.

Trade-off

There is tension between internal and external validity.
–The more the test environment and experimental procedures are “relaxed” (to mimic real-world situations), the more the experiment is susceptible to uncontrolled sources of variation, such as pondering, distractions, or secondary tasks.

Resolving the Trade-off

- Internal and external validity are increased by posing multiple narrow (testable) questions that cover the range of outcomes influencing the broader (un-testable) questions.

–E.g., a technique that is faster, is more accurate, takes fewer steps, is easy to learn, and is easy to remember, is generally better.

- The “good news” is that there is usually a positive correlation between the testable and un-testable questions.

–For example, participants generally find a UI better if it is faster, more accurate, takes fewer steps, etc.
The “good news”, in fact, is not so good after all as it raises more confusions.

Implication

- The “good news” actually implies we do not need empirical research!!
- We just do a user study and ask participants which technique they preferred.
  –Because of the “positive correlation”, we need not take the pain in collecting and analyzing data. However, this is not true.
- If participants are asked which technique they prefer (a broad question), they’ll probably give an answer… even if they really have no particular preference!
- There are many reasons, such as how recently they were tested on a technique, personal interaction with the experimenter, etc.
- Therefore, such preferences need not be indicative of the system performance.
  –We need to scientifically ascertain the validity of the preferences expressed by the participants, which requires formulation of testable questions.

- Also, with broader questions, we may not get idea about the feasibility or usefulness of the system
  –It is not enough to know if a system is better than another system only but we also need to know ”how much better” (for example, it may not be feasible economically to develop a system that is only 5% better than the current system).

- Seeking feedback from users on broader questions is not very helpful from another perspective
  –It does not help to identify the strengths, weaknesses, limits, capabilities of the design, thereby making it difficult to identify opportunities for improvements.

- Such concerns can be addressed only with the raising of testable research questions.
- An important point to note is, in order to test the validity of research questions through observations, we need measurements.

Observe and Measure

In empirical research, observation is the most fundamental thing to do. Observational (empirical) data can be gathered in two ways:

- Manual: in this case, a human observer manually records all the relevant observational data.
- Automatic: The observation can also be recorded automatically, through the use of computers, software, sensor, camera and so on.
- A measurement is, simply put, a recorded observation.
- There are broadly four scales of measurements that are used (nominal, ordinal, interval and ratio).

Scales of Measurements

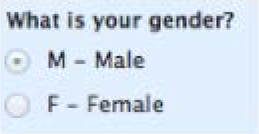
Each scale of measurement satisfies one or more of the following properties of measurement.

- **Identity**. Each value on the measurement scale has a unique meaning.
- **Magnitude**. Values on the measurement scale have an ordered relationship to one another. That is, some values are larger and some are smaller.
- **Equal intervals**. Scale units along the scale are equal to one another. This means, for example, that the difference between 1 and 2 would be equal to the difference between 19 and 20.
- **A minimum value of zero**. The scale has a true zero point, below which no values exist.

Nominal Scale of Measurement

The nominal scale of measurement only satisfies the identity property of measurement. Values assigned to variables represent a descriptive category, but have no inherent numerical value with respect to magnitude.

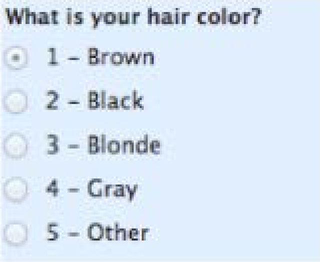
Gender is an example of a variable that is measured on a nominal scale. Individuals may be classified as "male" or "female", but neither value represents more or less "gender" than the other. Religion and political affiliation are other examples of variables that are normally measured on a nominal scale.



What is your gender?

● M – Male

○ F – Female



What is your hair color?

● 1 – Brown

○ 2 – Black

○ 3 – Blonde

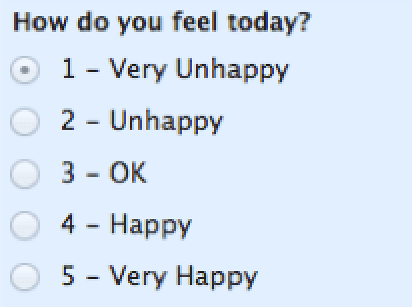
○ 4 – Gray

○ 5 – Other

Ordinal Scale of Measurement

The ordinal scale has the property of both identity and magnitude. Each value on the ordinal scale has a unique meaning, and it has an ordered relationship to every other value on the scale.

Take a look at the example below. In each case, we know that a #4 is better than a #3 or #2, but we don't know–and cannot quantify–how *much* better it is. For example, is the difference between "OK" and "Unhappy" the same as the difference between "Very Happy" and "Happy?" We can't say.



How do you feel today?

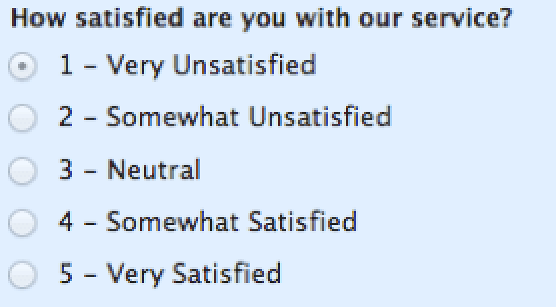
● 1 – Very Unhappy

○ 2 – Unhappy

○ 3 – OK

○ 4 – Happy

○ 5 – Very Happy



How satisfied are you with our service?

● 1 – Very Unsatisfied

○ 2 – Somewhat Unsatisfied

○ 3 – Neutral

○ 4 – Somewhat Satisfied

○ 5 – Very Satisfied

Interval Scale of Measurement

The interval scale of measurement has the properties of identity, magnitude, and equal intervals.

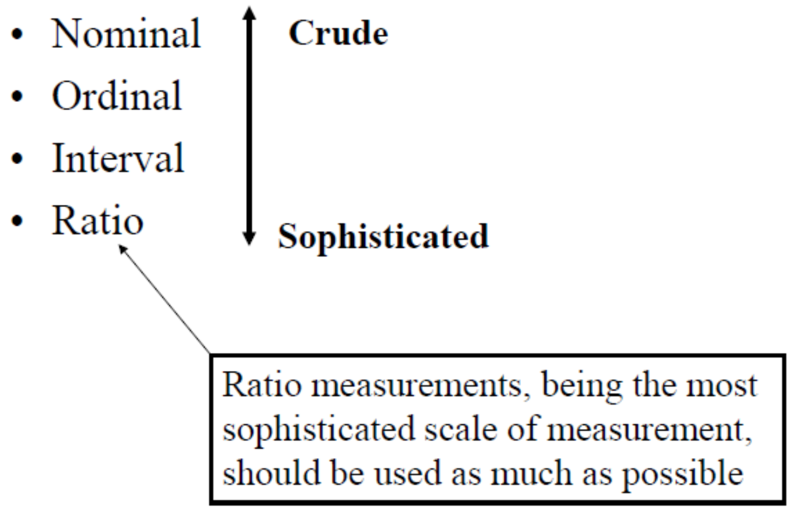
A perfect example of an interval scale is the Fahrenheit scale to measure temperature. The scale is made up of equal temperature units, so that the difference between 40 and 50 degrees Fahrenheit is equal to the difference between 50 and 60 degrees Fahrenheit.

With an interval scale, you know not only whether different values are bigger or smaller, you also know *how much* bigger or smaller they are. For example, suppose it is 60 degrees Fahrenheit on Monday and 70 degrees on Tuesday. You know not only that it was hotter on Tuesday, you also know that it was 10 degrees hotter.

Ratio Scale of Measurement

The ratio scale of measurement satisfies all four of the properties of measurement: identity, magnitude, equal intervals, and a minimum value of zero.

The weight of an object would be an example of a ratio scale. Each value on the weight scale has a unique meaning, weights can be rank ordered, units along the weight scale are equal to one another, and the scale has a minimum value of zero.
Weight scales have a minimum value of zero because objects at rest can be weightless, but they cannot have negative weight.



Ratio measurements, being the most sophisticated scale of measurement, should be used as much as possible

Ratio Measurements

- Ratio scales are the most preferred scale of measurement. This is because ratio scales make it convenient to compare or summarize observations.
- If you are conducting an empirical research, you should strive to report "counts" as ratios wherever possible.
- For example, assume you have observed that "a 10-word phrase was entered by a participant in an empirical study in 30 seconds". What should you measure?
  –If you measure the "time to enter text" (i.e., t = 30 seconds) as an indicator of system performance, it is a bad measurement
  –However, if you go for a ratio measurement (e.g., entry rate = 10/0.5 = 20 wpm), that is much better and gives a general indication of performance
- Let us consider another example. Suppose in an empirical study, you observed that a participant committed two errors while entering a 50 character phrase
  –If you measure the "number of errors committed" (i.e., n = 2) as an indicator of system performance, it is a bad measurement
  –However, if you go for a ratio measurement (e.g., error rate = 2/50 = 0.04 = 4%), that is much better and is a more general performance indicator

User Study

- A user study, in the context of HCI, is a scientific way of collecting and analyzing observational data from end users on an interactive system
- Collection of data involve experiments and design of experiments.

Experiment Design

- Experiment design is a general term referring to the organization of variables, procedures, etc., in an experiment
- The process of designing an experiment is the process of deciding on which variables to use, what procedure to use, how many participant to use, how to solicit them etc.

Terminology

- Participant
- Independent variable (test conditions)
- Dependent variable
- Control variable
- Random variable
- Confounding variable
- Within subjects vs. between subjects
- Counterbalancing and Latin square

Participant

The people participating in an experiment are referred to as participants

- When referring specifically to the experiment, use the term participants (e.g., "all participants exhibited a high error rate…")
- General comments on the problem or conclusions drawn from the results may use other terms (e.g., "these results suggest that users are less likely to…"

Independent Variable

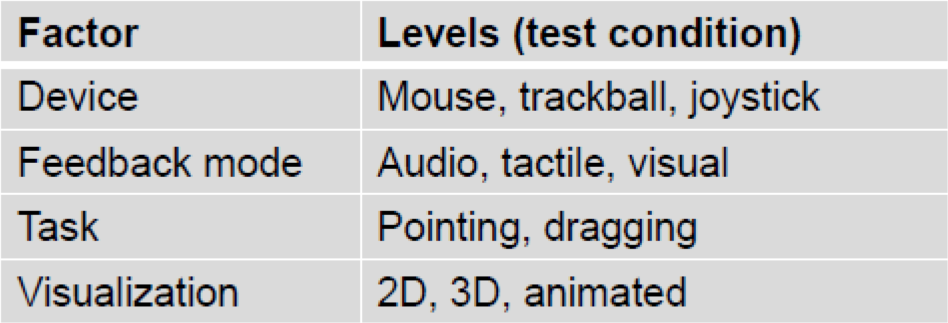
An independent variable is the manipulated variable in an experiment. It's called "manipulated" because it's the one you can change. In other words, you can decide ahead of time to increase it or decrease it. In an experiment you should only have one manipulated variable at a time.

An independent variable is a variable that is selected or controlled through the design of the experiment –Examples include device, feedback mode, button layout, visual layout, gender, age, expertise, etc.

The terms independent variable and factor are synonymous

Test Conditions

- The levels, values, or settings for an independent variable are the test conditions
- Provide names for both an independent variable (factor) and the test conditions (levels) for the controlled variable (see examples below)



Factor	Levels (test condition)
Device	Mouse, trackball, joystick
Feedback mode	Audio, tactile, visual
Task	Pointing, dragging
Visualization	2D, 3D, animated

Dependent Variable

The dependent variable (DV) is just like the name sounds; it *depends* upon some factor that you, the researcher, controls. For example:

- How well you perform in a race depends on your training.
- How much you weight depends on your diet.
- How much you earn depends upon the number of hours you work.

- ➢ "(Independent variable) causes a change in (Dependent Variable) and it isn't possible that (Dependent Variable) could cause a change in (Independent Variable)."
- ➢ A variable representing the measurements or observations on a independent variable
- ➢ It is required to provide a name for both the dependent variable and its unit

–Examples: Task completion time (ms), speed (word per minute, selections per minute, etc), error rate (%), throughput (bits/s)

Control Variable

An experiment has several types of variables, including a control variable (sometimes called a controlled variable). Variables are just values that can change; a good experiment only has two changing variables: the independent variable and dependent variable. Let's say you are testing to see how the amount of light received affects plant growth:

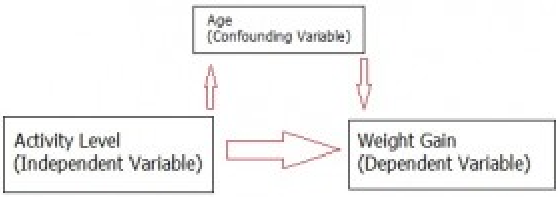
- The independent variable, in this case the amount of light, is changed by you, the researcher.
- As you change the independent variable, you watch what happens to the dependent variable. In this case you see how much the plants grow.
- A control variable is another factor in an experiment; it must be held constant. In the plant growth experiment, this may be factors like water and fertilizer levels.

Random Variable

- ➢ Instead of controlling all circumstances or factors, some might be allowed to vary randomly
- ➢ Such circumstances are random variables

Confounding Variable

A confounding variable is an "extra" variable that you didn't account for. They can ruin an experiment and give you useless results. They can suggest there is correlation when in fact there isn't. They can even introduce bias. That's why it's important to know what one is, and how to avoid getting them into your experiment in the first place.



In an experiment, the independent variable typically has an effect on your dependent variable. For example, if you are researching whether lack of exercise leads to weight gain, lack of exercise is your independent variable and weight gain is your dependent variable. Confounding variables are any other variable that also has an effect on your dependent variable.

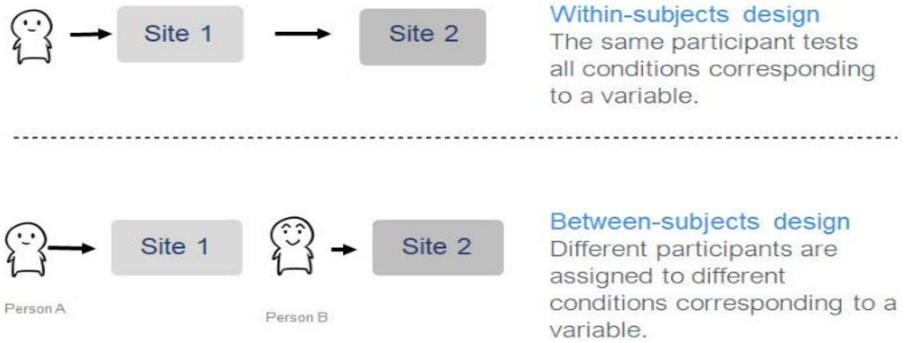
- Any variable that varies systematically with an independent variable is a confounding variable

– For example, if three devices are always administered in the same order, participant performance might improve due to practice; i.e., from the 1st to the 2nd to the 3rd condition; thus "practice" is a confounding variable (because it varies systematically with "device")

Within Subjects, Between Subjects

When you want to compare several user interfaces in a single study, there are two ways of assigning your test participants to these multiple conditions:

- **Between-subjects** (or **between-groups**) study design: different people test each condition, so that each person is only exposed to a single user interface.
- **Within-subjects (or repeated-measures)** study design: the same person tests all the conditions (i.e., all the user interfaces).

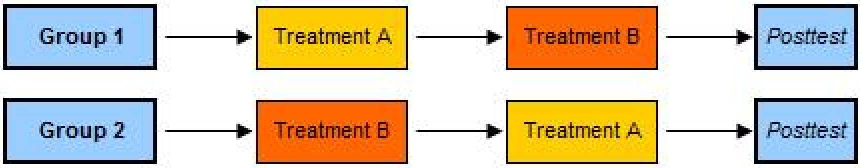

Site 1
Site 2
Within-subjects design
The same participant tests all conditions corresponding to a variable.
Site 1
Site 2
Person A
Person B
Between-subjects design
Different participants are assigned to different conditions corresponding to a variable.

- A relevant question is, which of the two approaches (within subject and between subject) should be chosen in designing an experiment
- Answer: It depends!
  –Sometimes a factor must be between subjects (e.g., gender, age)
  –Sometimes a factor must be within subjects (e.g., session, block)
  –Sometimes there is a choice. In this case there is a trade-off
- The advantage of within subject design is, the variance due to participants' pre-dispositions should be the same across test conditions
- Between subjects design, on the other hand, has the advantage of avoiding interference effects (e.g., the practice effect while typing on two different layouts of keyboards)

Counterbalancing

- Counterbalancing is a procedure that allows a researcher to control the effects of nuisance variables in designs where the same participants are repeatedly subjected to conditions, treatments, or stimuli (e.g., within-subjects or repeated-measures designs). Counterbalancing refers to the systematic variation of the order of conditions in a study, which enhances the study's interval validity.
- For repeated measures designs, participants' performance may tend to improve with practice as they progress from one level to the next. Thus, participants may perform better on the second level simply because they benefited from practice on the first (this is undesirable)
- To compensate, the order of presenting conditions is counterbalanced
- The simplest type of counterbalanced measures design is used when there are two possible conditions, A and B. As with the standard repeated measures design, the researchers want to test every subject for both conditions. They divide the subjects into two groups and one group is treated

with condition A, followed by condition B, and the other is tested with condition B followed by condition A.



Latin Square

- Participants are divided into groups, and a different order of administration is used for each group
- The order is best governed by a Latin Square (The defining characteristic of a Latin Square is that each condition occurs only once in each row and column)

- Example: suppose we want to administer 4 levels (denoted by A, B, C and D) of a factor to 4 participants (represented by P1, P2, P3 and P4)
  –We can construct a 4×4 Latin square arrangement to depict the order of administering the levels to each participant

P1	A	B	C	D
P2	B	C	D	A
P3	C	D	A	B
P4	D	A	B	C

- In a balanced Latin Square, each condition both precedes and follows each other condition an equal number of times
  –We can construct a balanced 4×4 Latin square arrangement for the previous example

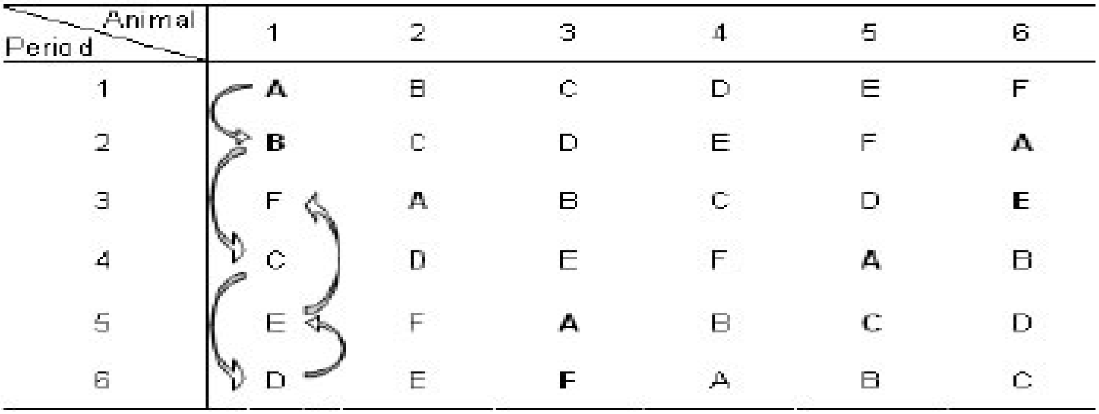
P1	A	B	C	D
P2	B	D	A	C
P3	D	C	B	A
P4	C	A	D	B

A systemic method for balanced Latin square designs

The systemic method balances the residual effects when a treatment is an even number. A balanced 6 × 6 Latin square design using this method is illustrated in Figure 2. In this example, treatments A to F are ordinarily assigned in the first row (animal). Treatments for the fi rst column (period) are assigned as an order of:

1, 2, n, 3, n-1, 4, ···, and (n + 2)/2,

And treatments for the second to the last row are assigned using the same sequence as in the first row, but starting with the treatment assigned to the first column (i.e., BCDEFA, FABCDE, CDEFAB, EFABCD, and DEFABC in this example). All treatments are preceded and followed by all other treatments exactly once. The balance of this design is maintained by a randomization of the treatments in the first period before assigning treatments in the other periods or a post randomization by column.



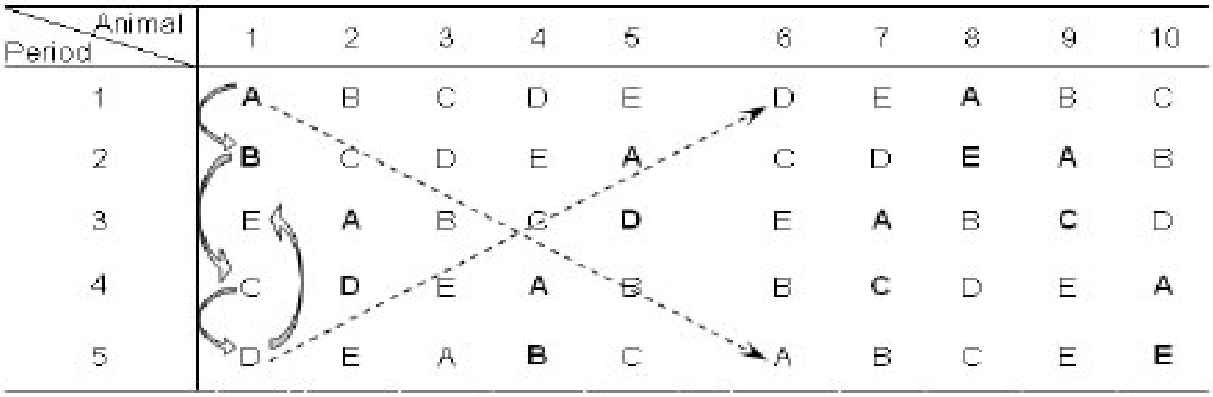
Animal / Period	1	2	3	4	5	6
1	A	B	C	D	E	F
2	B	C	D	E	F	A
3	F	A	B	C	D	E
4	C	D	E	F	A	B
5	E	F	A	B	C	D
6	D	E	F	A	B	C

Figure 2. A systematic method to balance the first order residual effect in a Latin square design with an even number of rows (periods) and columns (animals). Each capital letter represents a treatment. Treatments for the first column are assigned as: 1, 2, n, 3, n-1, 4, ···, and n/2 + 1 based on the treatment sequence in the first row, and treatments for the rest columns are assigned as the same sequence of the first row. Each treatment is immediately followed by every other treatment only once (adapted from Williams, 1949).

For Latin square designs with an odd number of treatments, the first order residual effects can be balanced only if the square is replicated an even number of times. In Figure 3, a replicated 5 × 5 Latin square design is illustrated. Treatments A to E are ordinarily assigned in the first row. Based on this treatment sequence, treatments for the first column are assigned as an order of:

1, 2, n, 3, n-1, 4, ···, (n + 1)/2, and (n + 3)/2,

And treatments for the second to the last row are assigned using the same sequence as in the first row, but starting with the treatment assigned to the first column (i.e., BCDEA, EABCD, CDEAB, and, DEABC in this example). Treatments for the first column of the second square are assigned in a reverse order of the first column of the first square, and treatments for the remaining columns are assigned based on the treatment sequence in the first row of the first square starting with the treatment assigned to the first column. All treatments are preceded and followed by all other treatments by row exactly twice.



Animal Period	1	2	3	4	5	6	7	8	9	10
1	A	B	C	D	E	D	E	A	B	C
2	B	C	D	E	A	C	D	E	A	B
3	E	A	B	C	D	E	A	B	C	D
4	C	D	E	A	B	B	C	D	E	A
5	D	E	A	B	C	A	B	C	E	E

Figure 3. A systematic method to balance the first order residual effect in a Latin square design with an odd number of rows (periods) and columns (animals) and an even number of squares. Each capital letter represents a treatment. Treatments for the first column are assigned as: 1, 2, n, 3, n-1, 4, ···, (n + 1)/2, and (n + 3)/2 based on the treatment sequence in the first row of the first square, and treatments for the rest columns in the first square are assigned as the same sequence of the first row. Treatments for the first column of the second square are assigned in a reverse order of the first column of the first square, and treatments for the rest columns are assigned based on the treatment sequence in the first row of the first square. Each treatment is immediately followed by every other treatment equally twice (adapted from Williams, 1949).

Expressing Experiment Design

- Consider the statement "3 x 2 repeated-measures design".
  –It refers to an experiment with two factors, having three levels on the first, and two levels on the second. There are six test conditions in total. Both factors are repeated measures, meaning all participants were tested on all test conditions
- Any type of experiment is expressed similarly.

Analysis of Empirical Data
Answering Empirical Questions:

- Suppose, we want to determine if the text entry speed of a text input system we proposed is more than an existing system
- We know how to design an experiment and observe and measure
- We conduct a user study and measure the performance on each test condition (our system and the existing system) over a group of participants
- For each test condition we compute the mean score (text entry speed) over the group of participants.
- We are faced with three questions

Q1. Is there a difference?

- This is obvious as we are most likely to see some differences. However, can we conclude anything from this difference? This brings us to the second question

Q2. Is the difference large or small?

- This is more difficult to answer. If we observe a difference of, say, 30%, we can definitely say the difference is large. However, we can't say anything definite about, say, a 5% difference. Clearly, the difference figure itself can't help us to draw any definite conclusion. This brings us to the third question

Q3. Is the difference significant or is it due to chance?

- Even if the observed difference is "small", it can still lead us to conclude about our design if we can determine the nature of the difference. If the difference is found to be "significant" (not occurred by chance), then we can say something about our design

- It is important to note that the term "significance" is a statistical term
- The test of (statistical) significance is an important aspect of empirical data analysis
- We can use statistical techniques for the purpose
  –The basic technique is ANOVA or Analysis Of Variance

ANOVA

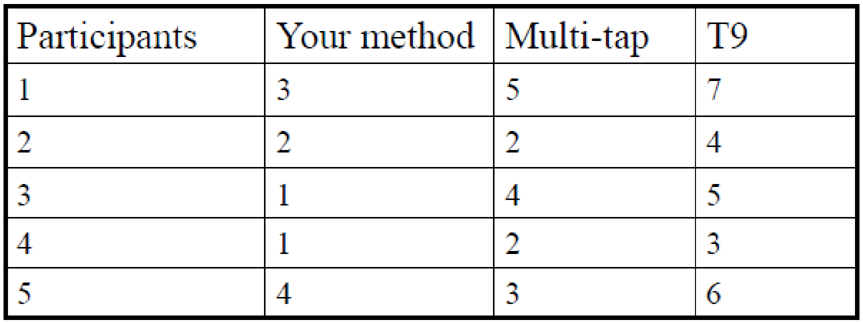
- Let us go through the procedure for one-way ANOVA
  –That means, one independent variable
- Multi-way ANOVA computations are very cumbersome to do manually
  –Better to do with statistical packages

Explanation of one-way ANOVA Illustrated

Suppose you have designed a new text entry technique for mobile phones. You think the design is good. In fact, you feel your method is better than the most widely used current techniques, multi-tap and T9. You decide to undertake some empirical research to evaluate your invention and to compare it with the current techniques? Suppose "better" is defined in terms of error rate

Data

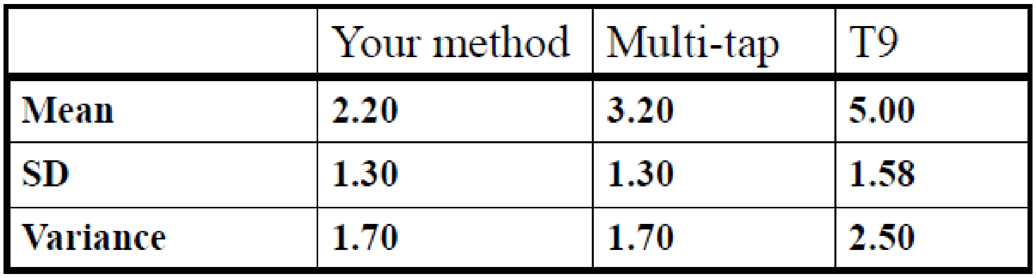
In order to ascertain the validity of your claim, you conducted experiments and collected the following data (error rate of participants under different test conditions)



Participants	Your method	Multi-tap	T9
1	3	5	7
2	2	2	4
3	1	4	5
4	1	2	3
5	4	3	6

ANOVA Steps -1

Calculate means, standard deviations (SD) and variances for each test condition (over all participants)



	Your method	Multi-tap	T9
Mean	**2.20**	**3.20**	**5.00**
SD	**1.30**	**1.30**	**1.58**
Variance	**1.70**	**1.70**	**2.50**

Calculate means:
Your Method= (3+2+1+1+4)/5 = 11/5= 2.20
Multi-tap= (5+2+4+2+3)/5= 16/5=3.20
T9 = (7+4+5+3+6)/5= 25/5=5.00

Calculate standard deviations (SD):

The formula for **Sample Standard Deviation**:

$$s = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \overline{x})^2}$$

According to formula now calculate the standard deviation and variances:

$\overline{x}$ = Mean
N= no. of participant
N-1=4

Your Method:

x	3	2	1	1	4
$\overline{x}$	**0.64**	**0.04**	**1.44**	**1.44**	**3.24**

$= (3-2.20)^2 + (2-2.20)^2 + (1-2.20)^2 + (1-2.20)^2 + (4-2.20)^2$
= 0.64+0.04+1.44+1.44+3.24
= 6.8
6.8/N-1 = 6.8/4 = 1.70 (variance)
√1.70 = 1.30 (standard deviation)

For rest (multi-tap and T9) calculate same as above.

Also calculate “grands” – values involving all irrespective of groups

- Grand mean (mean of means) = 3.467 = (2.20+3.20+5.00)/3
- Grand SD (w.r.t. grand mean) =1.8647
- Grand variance (w.r.t. grand mean) = 3.6442

Grand standard deviation and variances w.r.t.
For your method
$= (3-3.467)^2 + (2-3.467)^2 + (1-3.467)^2 + (1-3.467)^2 + (4-3.467)^2$
= 0.218 + 2.152+ 6.086 + 6.086 + 0.284

= 14.826 = (14.826/4= 3.7065)
Variance= 3.7065
Standard deviation= √3.7065 = 1.925

Calculate as above for multi tap and T9:
For multi tap method
Variance=1.789
Standard deviation=1.3375

For T9 method
Variance=5.4373
Standard deviation=2.3318

Grand Standard deviation= 1.925+1.3375+2.3318/3 = 1.8647
Grand Variance= 3.7065+1.789+5.4373/3 = 3.6442

ANOVA Steps -2
Calculate "total sum of squares (SS_T)"

$$\begin{aligned} \mathrm{SS\_T} &= \sum (x\_i - mean\_grand)^2 \\ &= 43.74 \end{aligned}$$

x_i is the error rate value of the i-th participant (among all)
Calculation:
Row1 in data table:
= (3-3.467)2 + (5-3.467)2 + (7-3.467)2
=0.218+2.350+12.482
= 15.05
Row2 in data table:
= (2-3.467)2 + (2-3.467)2 + (4-3.467)2
= 4.58
Row3 in data table:
= (1-3.467)2 + (4-3.467)2 + (5-3.467)2
= 8.72
Row4 in data table:
= (1-3.467)2 + (2-3.467)2 + (3-3.467)2
= 8.456
Row5 in data table:
= (4-3.467)2 + (3-3.467)2 + (6-3.467)2
= 6.912
Row1+ Row2+ Row3+ Row4+ Row5= 43.74

- An associated concept is the degrees of freedom (DoF), which is the number of observations that are free to vary
- DoF can be calculated simply as the (number of things used to calculate –1)
  –For SS_T calculation, DoF = N-1

ANOVA Steps -3

Next calculate the "model sum of square (SS_M)"

- Calculate (mean_group_i-mean_grand) for the i-th group
- Square the above
- Multiply by n_i, the number of participants in the i-th group
- Sum for all groups

In the example,

$$\text{SS\_M} = 5(2.200-3.467)^2 + 5(3.200-3.467)^2 + 5(5.000-3.467)^2$$
$$= 20.135$$

DoF = number of group means –1
= 3 -1 = 2 (in our example)

ANOVA Steps -4

Calculate the "residual sum of square (SS_R)" and the corresponding DoF

$$\text{SS\_R} = \text{SS\_T} - \text{SS\_M}$$
$$\text{DoF (SS\_R)} = \text{DoF (SS\_T)} - \text{DoF (SS\_M)}$$

Thus, in the example,

$$\text{SS\_R} = 43.74 - 20.14 = 23.60$$
$$\text{DoF (SS\_R)} = 14 - 2 = 12$$

ANOVA Steps -5

Calculate two "average sum of squares" or "mean squares (MS)"

- Model MS (MS_M) = SS_M/DoF(SS_M)
  = 20.135/2 = 10.067 (for our example)
- Residue MS (MS_R) = SS_R/DOF(SS_R)
  = 23.60/12 = 1.967 (for our example)

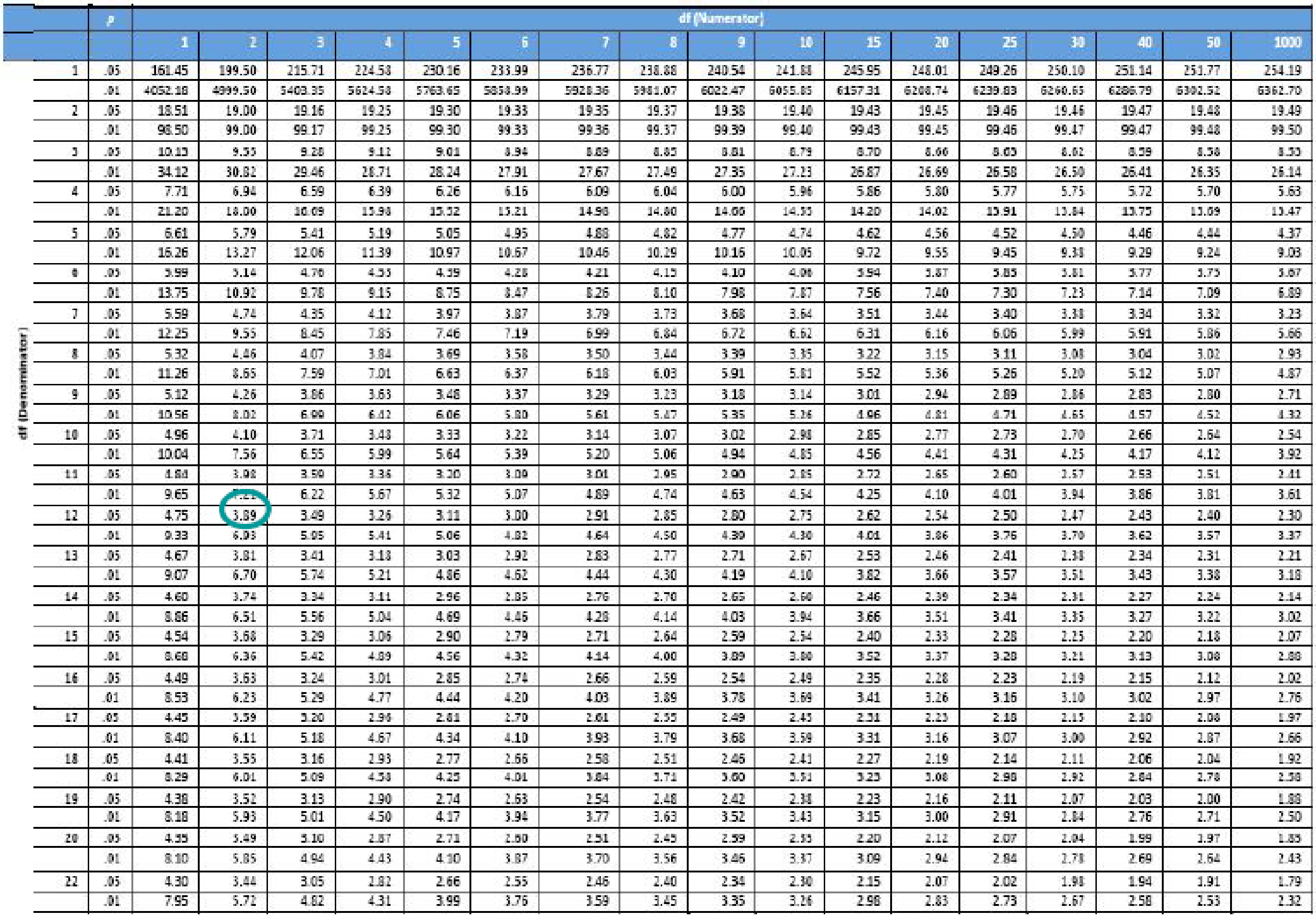
ANOVA Steps -6

- Calculate the "F-ratio" (simply divide MS_M by MS_R)
  –F = 10.067/1.967 = 5.12 (for our example)

- DoF associated with F-ratio are the DoFs used to calculate the two mean squares [that is DoF(SS_M) and DoF(SS_R)]
  –In our case, these are 2, 12
- Hence, in our case, the F-ratio would be written as F(2, 12) = 5.12

- Look up the critical value of F
  –The critical values for different "significance levels"/thresholds (α) are available in a tabular form
  –The critical values signifies the value of F that we would expect to get by chance for α% of tests

- Example
  –To find critical value of F(2, 12) from the table for α=.05, look at 2ndcolumn, 12throw for .05
  –Which is 3.89
  –That means, 3.89 is the F-value we would expect to get by chance for 5% of the tests.



df (Denominator)	p	df (Numerator)																
		1	2	3	4	5	6	7	8	9	10	15	20	25	30	40	50	1000
1	.05	161.45	199.50	215.71	224.58	230.16	233.99	236.77	238.88	240.54	241.88	245.95	248.01	249.26	250.10	251.14	251.77	254.19
	.01	4052.18	4999.50	5403.35	5624.58	5763.65	5858.99	5928.36	5981.07	6022.47	6055.85	6157.31	6208.74	6239.83	6260.65	6286.79	6302.52	6362.70
2	.05	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40	19.43	19.45	19.46	19.46	19.47	19.48	19.49
	.01	98.50	99.00	99.17	99.25	99.30	99.33	99.36	99.37	99.39	99.40	99.43	99.45	99.46	99.47	99.47	99.48	99.50
3	.05	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.70	8.66	8.63	8.62	8.59	8.58	8.53
	.01	34.12	30.82	29.46	28.71	28.24	27.91	27.67	27.49	27.35	27.23	26.87	26.69	26.58	26.50	26.41	26.35	26.14
4	.05	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.86	5.80	5.77	5.75	5.72	5.70	5.63
	.01	21.20	18.00	16.69	15.98	15.52	15.21	14.98	14.80	14.66	14.55	14.20	14.02	13.91	13.84	13.75	13.69	13.47
5	.05	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.62	4.56	4.52	4.50	4.46	4.44	4.37
	.01	16.26	13.27	12.06	11.39	10.97	10.67	10.46	10.29	10.16	10.05	9.72	9.55	9.45	9.38	9.29	9.24	9.03
6	.05	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	3.94	3.87	3.83	3.81	3.77	3.75	3.67
	.01	13.75	10.92	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87	7.56	7.40	7.30	7.23	7.14	7.09	6.89
7	.05	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.51	3.44	3.40	3.38	3.34	3.32	3.23
	.01	12.25	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72	6.62	6.31	6.16	6.06	5.99	5.91	5.86	5.66
8	.05	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.22	3.15	3.11	3.08	3.04	3.02	2.93
	.01	11.26	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91	5.81	5.52	5.36	5.26	5.20	5.12	5.07	4.87
9	.05	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.01	2.94	2.89	2.86	2.83	2.80	2.71
	.01	10.56	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35	5.26	4.96	4.81	4.71	4.65	4.57	4.52	4.32
10	.05	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.85	2.77	2.73	2.70	2.66	2.64	2.54
	.01	10.04	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94	4.85	4.56	4.41	4.31	4.25	4.17	4.12	3.92
11	.05	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.72	2.65	2.60	2.57	2.53	2.51	2.41
	.01	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63	4.54	4.25	4.10	4.01	3.94	3.86	3.81	3.61
12	.05	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.62	2.54	2.50	2.47	2.43	2.40	2.30
	.01	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39	4.30	4.01	3.86	3.76	3.70	3.62	3.57	3.37
13	.05	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.53	2.46	2.41	2.38	2.34	2.31	2.21
	.01	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19	4.10	3.82	3.66	3.57	3.51	3.43	3.38	3.18
14	.05	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.46	2.39	2.34	2.31	2.27	2.24	2.14
	.01	8.86	6.51	5.56	5.04	4.69	4.46	4.28	4.14	4.03	3.94	3.66	3.51	3.41	3.35	3.27	3.22	3.02
15	.05	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.40	2.33	2.28	2.25	2.20	2.18	2.07
	.01	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89	3.80	3.52	3.37	3.28	3.21	3.13	3.08	2.88
16	.05	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.35	2.28	2.23	2.19	2.15	2.12	2.02
	.01	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78	3.69	3.41	3.26	3.16	3.10	3.02	2.97	2.76
17	.05	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45	2.31	2.23	2.18	2.15	2.10	2.08	1.97
	.01	8.40	6.11	5.18	4.67	4.34	4.10	3.93	3.79	3.68	3.59	3.31	3.16	3.07	3.00	2.92	2.87	2.66
18	.05	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.27	2.19	2.14	2.11	2.06	2.04	1.92
	.01	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60	3.51	3.23	3.08	2.98	2.92	2.84	2.78	2.58
19	.05	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38	2.23	2.16	2.11	2.07	2.03	2.00	1.88
	.01	8.18	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52	3.43	3.15	3.00	2.91	2.84	2.76	2.71	2.50
20	.05	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.20	2.12	2.07	2.04	1.99	1.97	1.85
	.01	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46	3.37	3.09	2.94	2.84	2.78	2.69	2.64	2.43
22	.05	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30	2.15	2.07	2.02	1.98	1.94	1.91	1.79
	.01	7.95	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35	3.26	2.98	2.83	2.73	2.67	2.58	2.53	2.32

Implication

- Thus, we get critical value = 3.89 for F(2,12), α=.05
- Note that F(2, 12)=5.12 > the critical value
  –Implies that the effect of test conditions has a significant effect on the outcome w.r.t. α

Reporting F-Statistic

- You can report the result as "my method has a significant effect on reducing user errors [F(2,12)=5.12, p<.05] compared to the other methods."

- If it is found that the effect is not significant, it is reported as “my method has no significant effect on reducing user errors [F(1,9)=0.634, ns] compared to the other methods.”

A Note of Caution

- ANOVA requires that
  - Data should have normally distributed sampling distribution and from a normally distributed population
  - Variances in each experimental condition are fairly similar
  - Observations should be independent
  - Dependent variables should be measured on at least an interval scale

- The first two may be ignored if group sizes are equal
  –Otherwise, ALL conditions MUST have to be met

Empirical Research Case Study

Objective

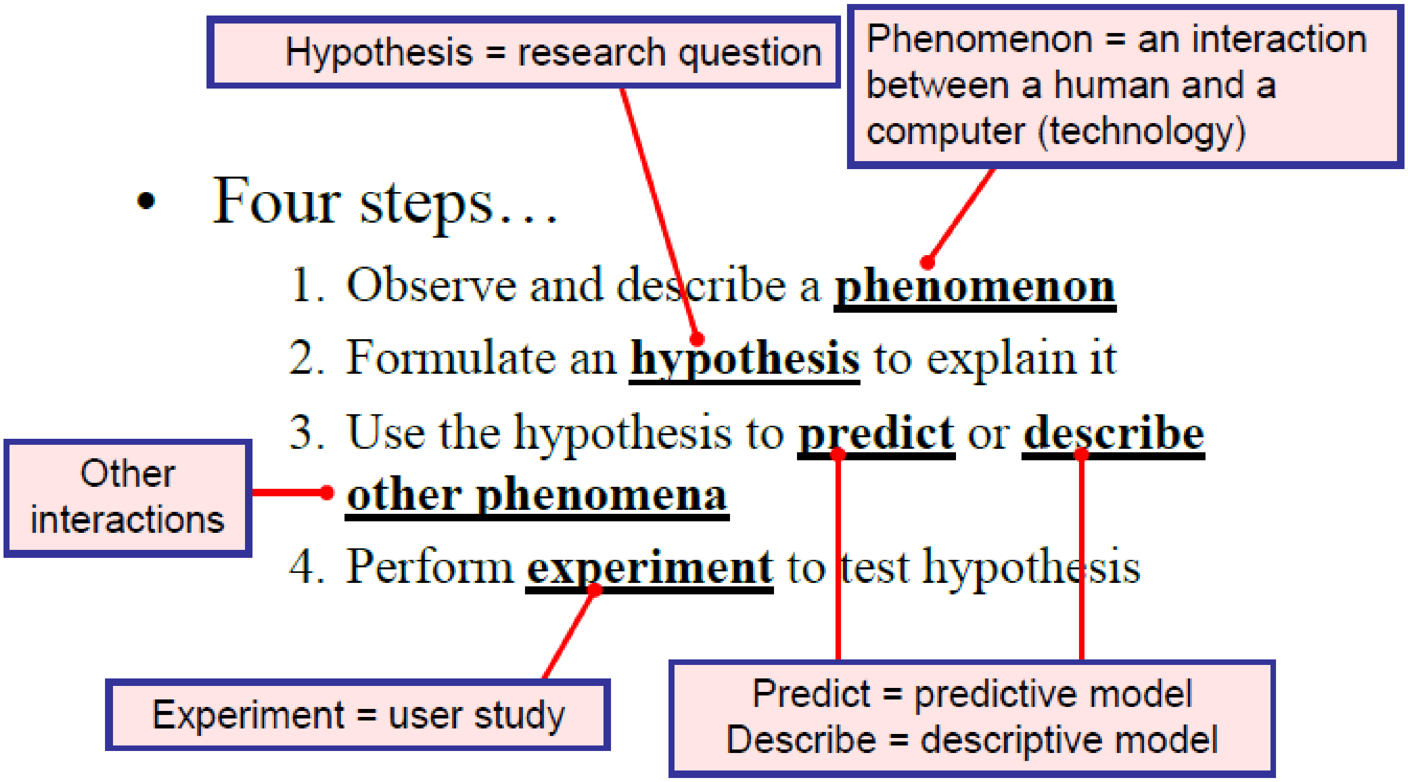
- In this module, we have so far learned about the basics of empirical research
- We got introduced to several concepts such as testable question formulation, experiment design, data collection and statistical analysis of data
- In this lecture, we shall illustrate these concepts further with a case study

Case Study

Suppose, you want to study the application of eye tracking technology to text entry (i.e., typing through eye gaze). You initiate an empirical inquiry to explore the performance limits and capabilities of various feedback modalities for keys in on-screen keyboards used with eye typing

–Suppose four feedback modalities are considered by you: Audio only[A], click+visual[C], speech+visual[S], visual only[V]

Steps in Empirical Research(Classical View)



- Four steps…
  1. Observe and describe a **phenomenon**
  2. Formulate an **hypothesis** to explain it
  3. Use the hypothesis to **predict** or **describe** **other phenomena**
  4. Perform **experiment** to test hypothesis

Hypothesis = research question

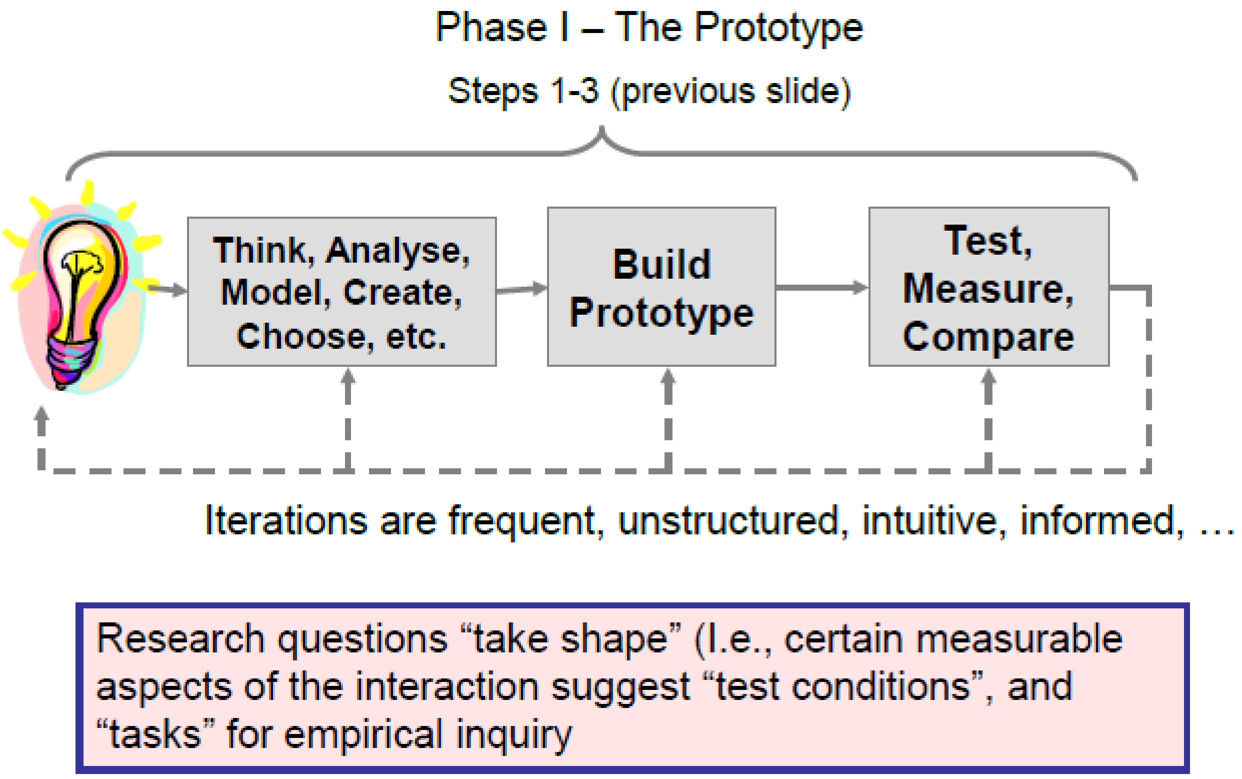
Phenomenon = an interaction between a human and a computer (technology)

Other interactions

Experiment = user study

Predict = predictive model
Describe = descriptive model

Steps in Empirical Research(Practical View)



Phase I – The Prototype

Steps 1-3 (previous slide)

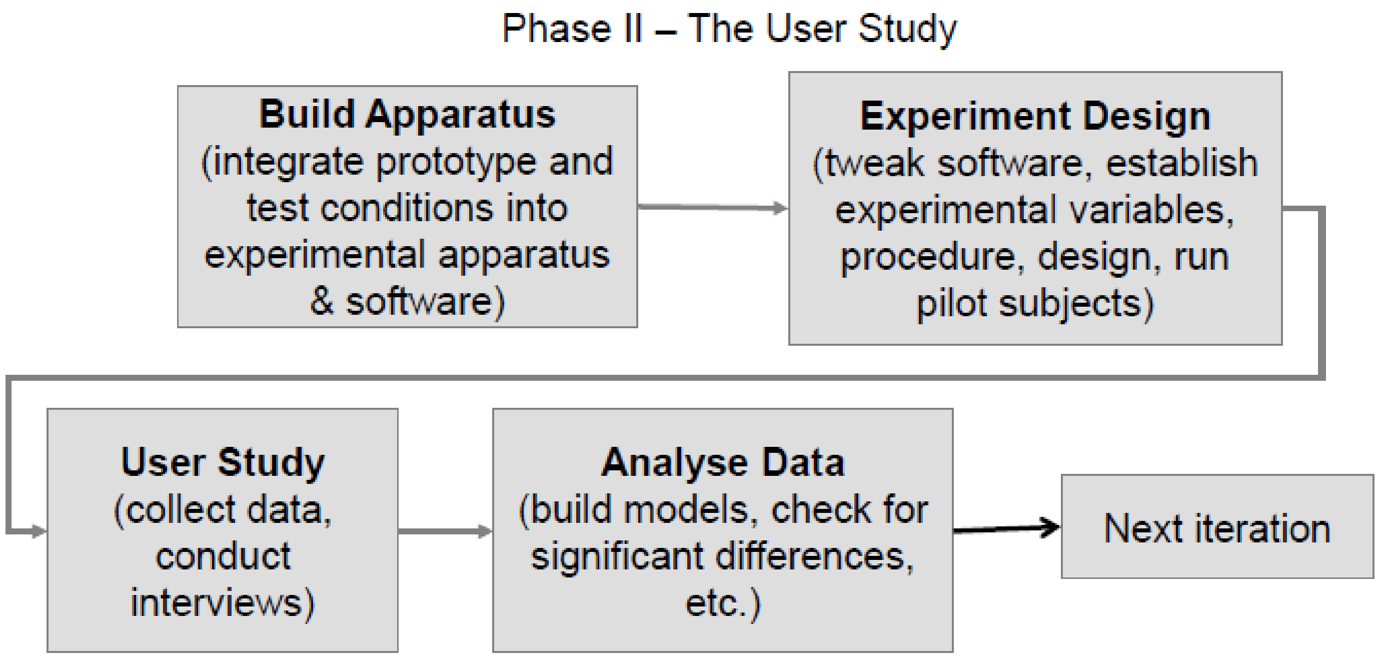
Think, Analyse, Model, Create, Choose, etc.

Build Prototype

Test, Measure, Compare

Iterations are frequent, unstructured, intuitive, informed, …

Research questions "take shape" (I.e., certain measurable aspects of the interaction suggest "test conditions", and "tasks" for empirical inquiry



Phase II – The User Study

Build Apparatus (integrate prototype and test conditions into experimental apparatus & software)

Experiment Design (tweak software, establish experimental variables, procedure, design, run pilot subjects)

User Study (collect data, conduct interviews)

Analyse Data (build models, check for significant differences, etc.)

Next iteration

The User Study

- Describe the participants employed for your study
  –Thirteen, all volunteers, recruited from university campus, age, gender, computer experience, eye tracking/typing experience
- Apparatus
  –Describe hardware and software, etc.

- Experiment design
  –You decided to have a 4 x 4 repeated measures design
  –There are two independent variables (factors) with four levels each
- Feedback modality (with the levels A, C, S, V)
- The participants were asked to enter blocks of text at a time and four such blocks were there for each participant. So, "block" is a factor with four levels 1, 2, 3, 4

- Experiment design
  –You have identified dependent variables (measures)
- Speed of text entry (in "words per minutes")
- Accuracy of text entry (in "percentage of characters in error")
- Key selection activity (in "keystrokes per character")
- Also… responses to "broad" questions
  –Order of conditions
- Feedback modality order differed for each participant (using a Latin square method)

- Procedure for data collection
  –You first explained to the participants the general objectives of the experiment
  –Then the eye tracking apparatus was calibrated
  –The participants were put through some practice trials for familiarization
  –Afterwards, you began data collection

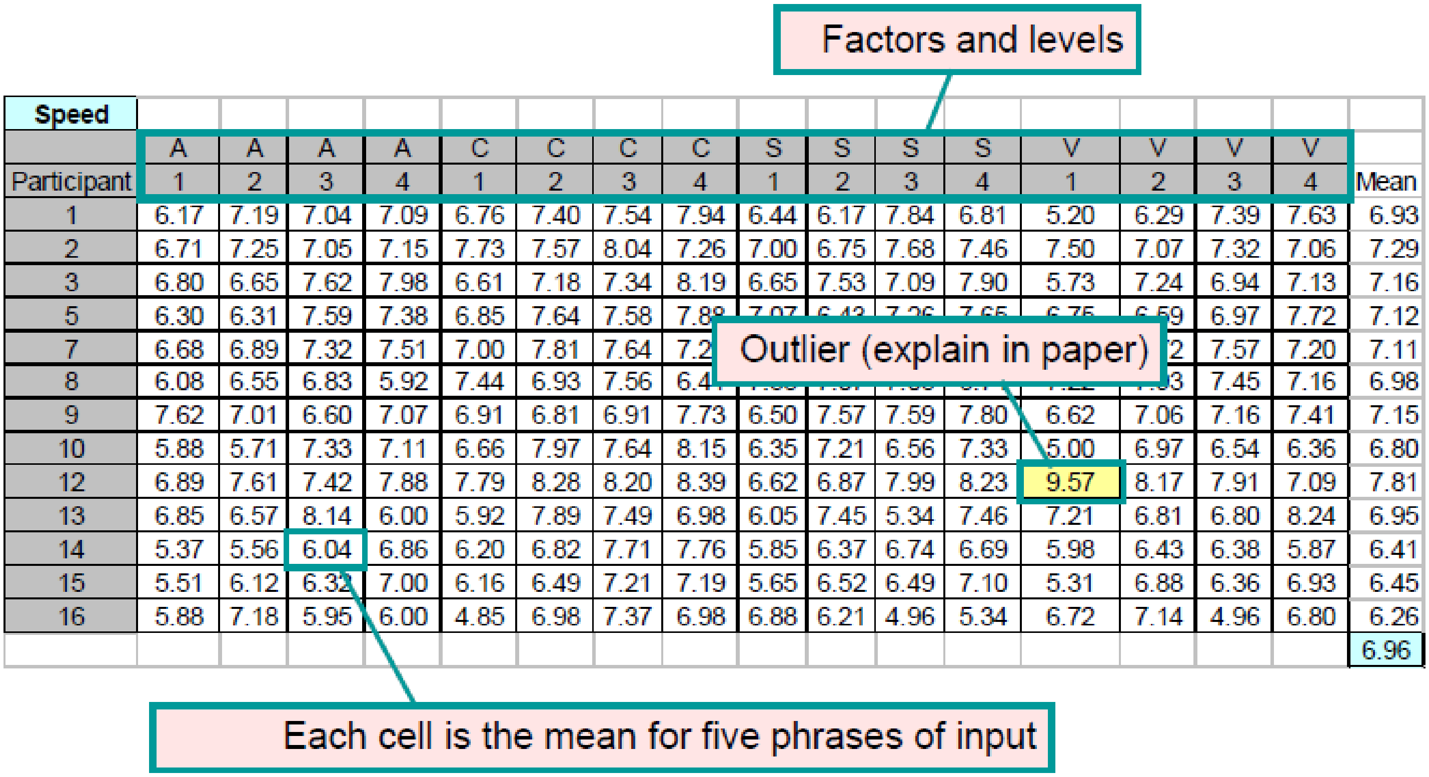
- Procedure for data collection
  –Phrases of text presented to the participants by experimental software
  –Participants instructed to enter phrases "as quickly and accurately as possible"
  –Five phrases were entered by the participants per block
  –Total number of phrases entered in experiment = 13 x 4 x 4 x 5 = 1040

Experiment Replication

- The description of the experimental methodology (i.e., participants, participant selection, apparatus, design, procedure) must be sufficient to allow the experiment to be replicated by other researchers
  –This is necessary to allow the possibility for the results to be verified or refuted
  –An experiment that cannot be replicated is useless

Data Tables

- Example data on text entry speed, recorded in the study
  –Create a table to arrange data
  –From the table, calculate other quantities such as grand mean = 6.96 wpm
  –The table also allows you to make salient observations (for example, 4thblock speed for best condition was…)



Factors and levels

Speed																	
	A	A	A	A	C	C	C	C	S	S	S	S	V	V	V	V	
Participant	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	Mean
1	6.17	7.19	7.04	7.09	6.76	7.40	7.54	7.94	6.44	6.17	7.84	6.81	5.20	6.29	7.39	7.63	6.93
2	6.71	7.25	7.05	7.15	7.73	7.57	8.04	7.26	7.00	6.75	7.68	7.46	7.50	7.07	7.32	7.06	7.29
3	6.80	6.65	7.62	7.98	6.61	7.18	7.34	8.19	6.65	7.53	7.09	7.90	5.73	7.24	6.94	7.13	7.16
5	6.30	6.31	7.59	7.38	6.85	7.64	7.58	7.88	7.07	6.43	7.26	7.65	6.75	6.59	6.97	7.72	7.12
7	6.68	6.89	7.32	7.51	7.00	7.81	7.64	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	7.57	7.20	7.11
8	6.08	6.55	6.83	5.92	7.44	6.93	7.56	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	7.45	7.16	6.98
9	7.62	7.01	6.60	7.07	6.91	6.81	6.91	7.73	6.50	7.57	7.59	7.80	6.62	7.06	7.16	7.41	7.15
10	5.88	5.71	7.33	7.11	6.66	7.97	7.64	8.15	6.35	7.21	6.56	7.33	5.00	6.97	6.54	6.36	6.80
12	6.89	7.61	7.42	7.88	7.79	8.28	8.20	8.39	6.62	6.87	7.99	8.23	9.57	8.17	7.91	7.09	7.81
13	6.85	6.57	8.14	6.00	5.92	7.89	7.49	6.98	6.05	7.45	5.34	7.46	7.21	6.81	6.80	8.24	6.95
14	5.37	5.56	6.04	6.86	6.20	6.82	7.71	7.76	5.85	6.37	6.74	6.69	5.98	6.43	6.38	5.87	6.41
15	5.51	6.12	6.32	7.00	6.16	6.49	7.21	7.19	5.65	6.52	6.49	7.10	5.31	6.88	6.36	6.93	6.45
16	5.88	7.18	5.95	6.00	4.85	6.98	7.37	6.98	6.88	6.21	4.96	5.34	6.72	7.14	4.96	6.80	6.26
																	6.96

Outlier (explain in paper)

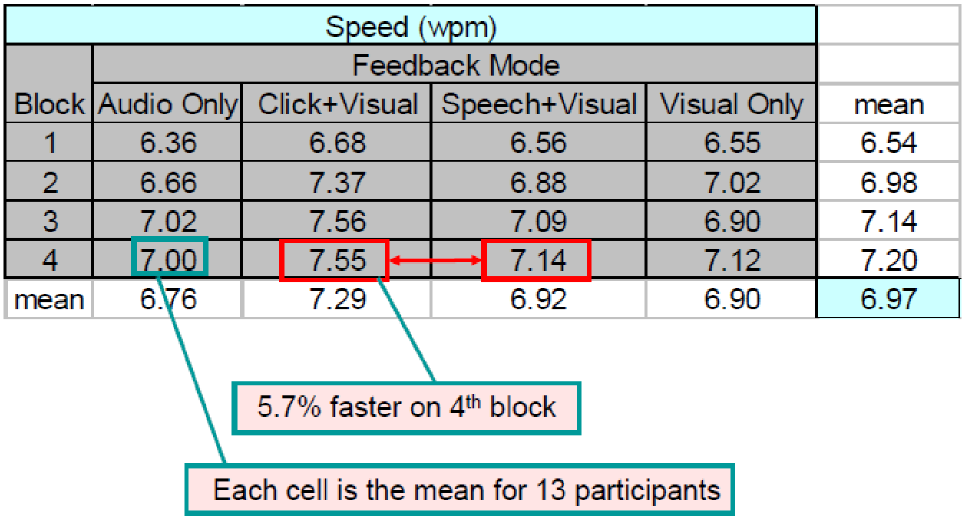
Each cell is the mean for five phrases of input

Statistical Analysis of Data

- The data recorded in the table are analyzed statistically to identify (statistically) significant effects
- For example, you may have the following findings
  –Main effect for Feedback mode significant [F(3,36)=8.77, p<.0005]
  –Feedback mode by block interaction not significant [F(9,108)=0.767, ns]

Data Tables

- Apart from the main tables, other tables are also created, which helps in making more useful observations
- The next example of a summary table created from the data on text entry speed



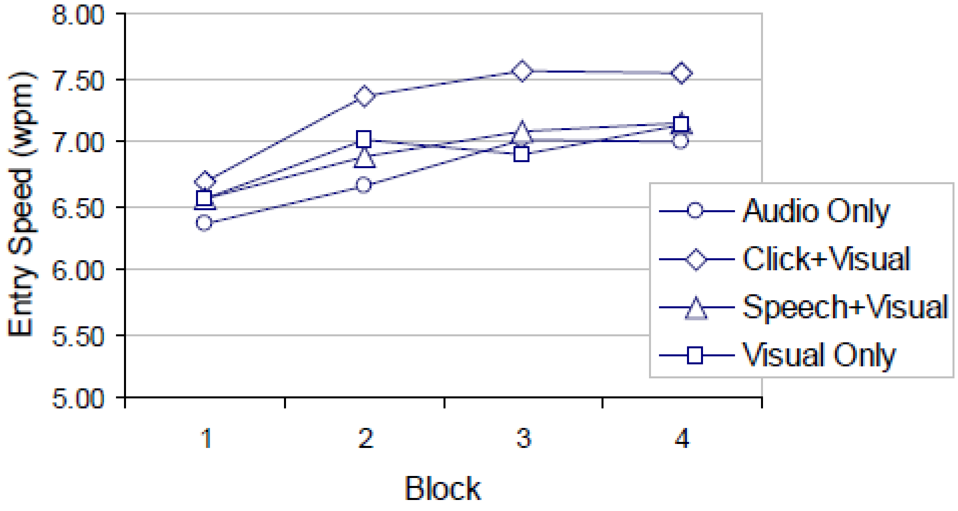
Speed (wpm)					
	Feedback Mode				
Block	Audio Only	Click+Visual	Speech+Visual	Visual Only	mean
1	6.36	6.68	6.56	6.55	6.54
2	6.66	7.37	6.88	7.02	6.98
3	7.02	7.56	7.09	6.90	7.14
4	7.00	7.55	7.14	7.12	7.20
mean	6.76	7.29	6.92	6.90	6.97

5.7% faster on 4th block

Each cell is the mean for 13 participants

Charts/Graphs

Also create graphs/charts to visualize findings



The Broad Questions

- Along with the data analysis, an empirical study typically collects direct feedback from participants on "broad" questions
  –For example, participants can be asked about their preferences, satisfaction levels or even their suggestions for improvements

- In the study, you asked the participants to rank (between 1 to 4) the feedback mode based on personal preference
- You obtained the following results
  –Six of thirteen participants gave a 1st place ranking to the fastest feedback modality

- The results obtained is not strong enough to come to any conclusions
  –A reason may be that the differences just weren't large enough for participants to really tell the difference in overall performance

- However, you have also made another observation, namely ten of the thirteen participants gave a 1st or 2nd place ranking to the fastest feedback modality
  –This can be treated as a strong indication that better performance yields a better preference rating

What's Missing?

- The case study just described show that the user study involves collection and analysis of usage data as well as participants' feedback
- However, that's not all (it misses an important aspect of empirical research)
  –There is no theoretical account of the phenomena

- There is no delineation, description, categorization of the known and observed behaviors (…that can form such a theoretical account)
- It is not sufficient to simply observe and conclude, it is also necessary to theorize about the observations (e.g., why the text entry speed is the least in a particular feedback mode)

- The direct conclusions from observations help us decide an interaction method; a theory about observed behavior can help us do much more
  –Such theories can eliminate need for further investigations as well as can suggest ways for improvement

Empirical Research in HCI

- Such theories, if found, are another motivation for conducting empirical research in HCI (in fact, many models in HCI have been derived empirically)
- Case Study: The Case for a Model
- Is there a "model of interaction" suggested by the observations in the case study?
- Perhaps. Here's one possibility
  –All gaze point changes were logged as "events". What was the total number of such events? Are there categories of such events?
- The identification, labeling, and tabulation of such could form the basis of a model of interaction for eye typing