ISYS3401
Information Technology Evaluation

Week 3 Lecture

Dr Vincent Pang

Vincent.Pang@sydney.edu.au

Agenda

- Recap
- Planning
- Usability Testing:
  - Formative assessment
  - Summative assessment
- Understand Your Users
  - Performance
  - Satisfaction
- Evaluation Method
  - Traditional (Moderated) Usability Tests
  - Online (Unmoderated) Usability Tests
  - Online Surveys

- Class Activities:
  - T-test
  - Chi-Square

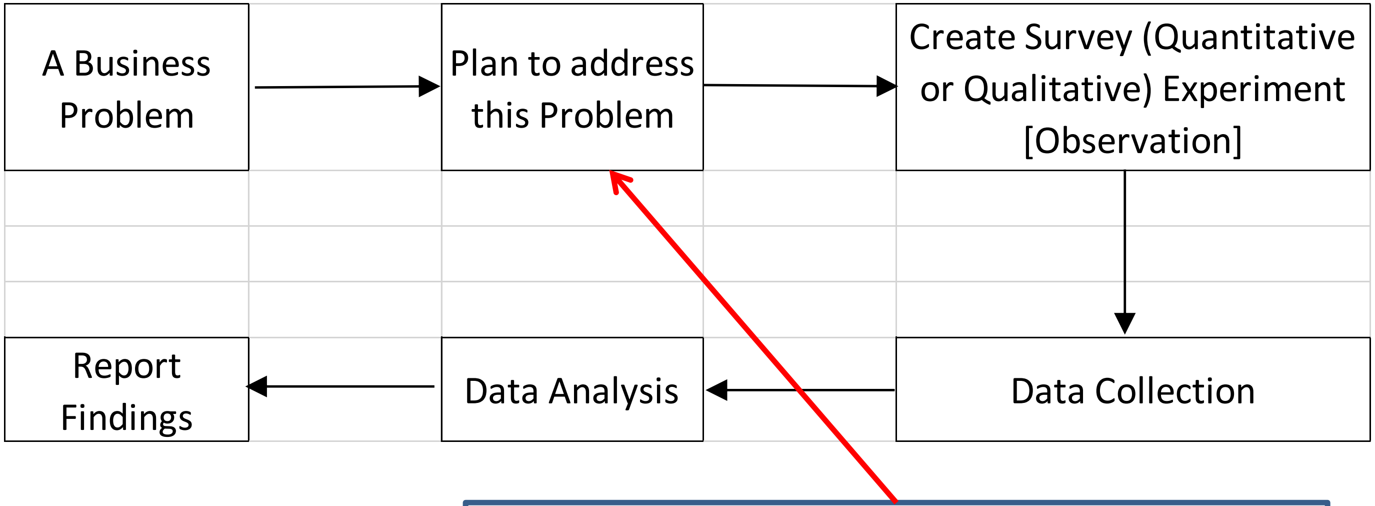
Recap from Last Week

A quick recap of what we have done last week:

- ***Descriptive statistics*** – *mean*, median, *standard deviation*, and the concept of *confidence intervals*
- Relationships Between Variables
- Look at Data Types
- *Class Exercise (in Excel):*
  - A bit of Data Cleaning
  - Do they make any sense, i.e. looking for outliers?
  - Look for data types, e.g. Nominal data type – Male Versus Female
- Normal Distribution
- Bionomial Distribution

Recap from Last Week Class Activities

In Week 1, you were asked by Mr Apple to conduct UX Experience (for $10 million).



A Business Problem

Plan to address this Problem

Create Survey (Quantitative or Qualitative) Experiment [Observation]

Data Collection

Data Analysis

Report Findings

Today, we look at the Planning stage: you plan what you are going to do with you study!

Reference

Measuring the User Experience: Collecting, Analyzing, and Presenting Usability Metrics, by William Albert, Thomas Tullis, ***Chapter 3***

Planning (1)

1. Need to understand the goals of the study.
   - Are you benchmarking the UX for an existing product, or are you trying to see if a new piece of functionality will improve UX.
2. Funding of the study (if any).
   - For example, who fund the study?
   - How much is the funding, i.e. you budget?
   - What is the key the objective of the funding?
3. Need to understand the goals of the users.
   - Are the users looking to simply complete a task and then stop using the product or will they use the product many times on a daily basis?
4. Need to know the users.
   - Who are the users?
   - Where are come from?
   - Which population these users can draw from?
   - Can you access to these users?
   - Will different users give different findings?
   - How many participants are enough to get reliable feedback?

Planning (2)

5. Any cultural and ethical issues.
    - Will these issues impact on your research
6. Need to know the collecting metrics
    - How are you going to collect the data?
    - What is the best tool to collect the data?
    - How much data collected is enough?
7. Time to deliver.
    - When you have to deliver?
    - Is this a short term or longitudinal study?
8. Need to know the best tool to analyse data.
    - Which is the best tool to analysis the data?
    - Who will analyse the data?

➢ By answering these questions, you will be well prepared to carry out any UX study involving metrics.

USABILITY TESTING

The first stage of planning is to understand how the data will ultimately be used within the product development life cycle:

1. **Formative assessment** is done on an ongoing basis (e.g. informal observation and mid-term quiz)
2. **Summative assessment** is done at the end of some significant period of time (e.g. final exam).

Differences: Formative assessment focuses on identifying ways of making improvements, whereas **summative assessment** focuses on evaluating products against a set of criteria.

Formative Assessment

Formative usability is always done before the design has been finalised.

- The goal in UX is to make improvements in the design prior to release.
- This means identifying or diagnosing the problems, making and implementing recommendations, and then evaluating again.
- A UX professional evaluates a product or design periodically while it is being created, identifies shortcomings, makes recommendations, and then repeats the process, until, ideally, the product comes out as close to perfect as possible.

Questions on Formative Assessment

- What are the most significant usability issues preventing users from accomplishing their goals or resulting in inefficiencies?
- What aspects of the product work well for the users?
- What do users find frustrating?
- What are the most common errors or mistakes users are making?
- Are improvements being made from one design iteration to the next?
- What usability issues can you expect to remain after the product is launched?

Summative Assessment

- The goal of summative usability is to evaluate how well a product or piece of functionality meets its objectives.
- Summative assessment can also be used to compare two or more products.
- Usually there are some follow-up activities after summative assessment include securing funding to enhance functionality on a product, or to address some outstanding usability issues, or even benchmarking changes to the user experience against which senior managers will be evaluated.

Questions on Summative Assessment

- Did we meet the usability goals of the project?
- What is the overall usability of our product?
- How does our product compare against the competition?
- Have we made improvements from one product release to the next?

Understand Your Users

The second stage of planning a usability study is you need to understand the users and what they are trying to accomplish.

- Are users required to use the product every day as part of their job?
- Are they likely to use the product only once or just a few times?
- Are they using it frequently as a source of entertainment?
- Does the user simply want to complete a task or is its efficiency the primary driver?
- Do users care at all about the design aesthetics of the product?

All these questions lead to measure two main aspects of the user experience: **performance** and **satisfaction**.

Performance

Performance is what the user actually does in interacting with the product.

- It includes measuring the degree to which users can accomplish a task or set of tasks successfully.
- Measure the performance of these tasks such as:
  - time taken to perform each task or
  - the amount of effort to perform each (such as number of mouse clicks or amount of cognitive effort),
  - the number of errors committed, or
  - the amount of time it takes to become proficient in performing the tasks (learnability).

(More in Chapter 4)

Satisfaction

Satisfaction is what the user says or thinks about his interaction with the product.

- Is it easy to use?
- Is it better than what you expect?
- Is the product visually appealing?
- Is the product trustworthy or untrustworthy?

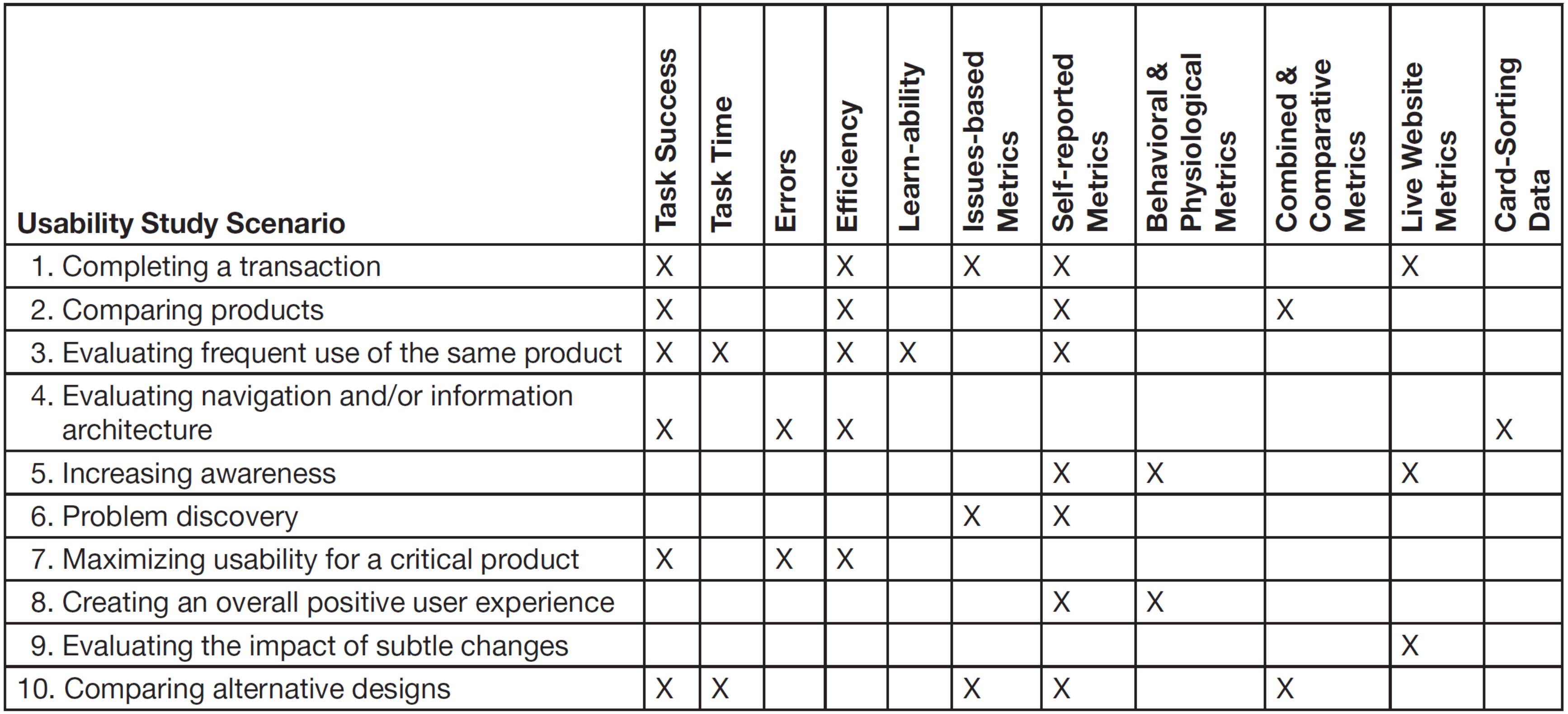
User satisfaction is important when the users have some choice in their usage of product such as websites, software applications, and consumer products.

(More in Chapter 6)

Performance and Satisfaction

- Performance and Satisfaction do not always Correlate
  - In some studies, users gave poor satisfaction ratings to an application that worked perfectly, and vice versa

Choosing the Right Metrics: Ten Types of Usability Studies [pp. 45-52]



Usability Study Scenario	Task Success	Task Time	Errors	Efficiency	Learn-ability	Issues-based Metrics	Self-reported Metrics	Behavioral & Physiological Metrics	Combined & Comparative Metrics	Live Website Metrics	Card-Sorting Data
1. Completing a transaction	X			X		X	X			X	
2. Comparing products	X			X			X		X		
3. Evaluating frequent use of the same product	X	X		X	X		X				
4. Evaluating navigation and/or information architecture	X		X	X							X
5. Increasing awareness							X	X		X	
6. Problem discovery						X	X				
7. Maximizing usability for a critical product	X		X	X							
8. Creating an overall positive user experience							X	X			
9. Evaluating the impact of subtle changes										X	
10. Comparing alternative designs	X	X				X	X		X		

Evaluation Method

Choosing an evaluation method is based on how many participants are needed, predict on how many participants will participate, and selection of metrics:

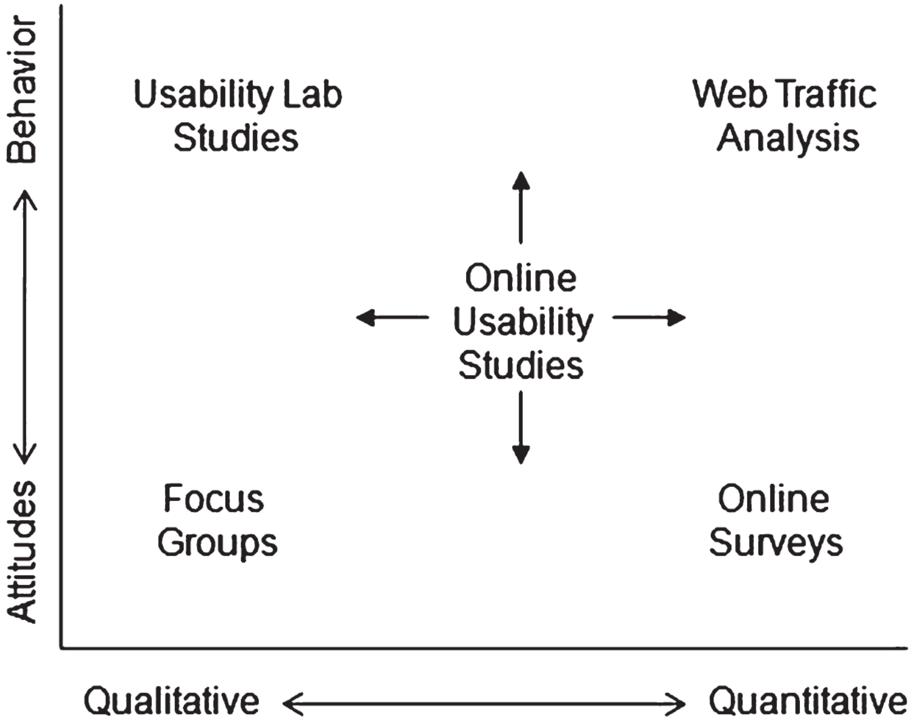
1. Traditional (Moderated) Usability Tests
2. Online (Unmoderated) Usability Tests
3. Online Surveys

Traditional (Moderated) Usability Tests

- The most common usability method is a lab test that utilises a small number of participants (5 to 10), but some are more (15-20)
  - A one-on-one session between a moderator (usability specialist) and a test participant.
  - The moderator asks questions of the participants and gives them a set of tasks to perform on the product in question.
  - The test participant is likely to be thinking aloud as she performs the various tasks.
  - The moderator records the participant’s behaviour and responses to questions.
  - Lab tests are used most often in formative studies where the goal is to make iterative design improvements.
- Self-reported metrics can be collected by having participants answer questions regarding each task, or at the conclusion of the study.
  - Warning – it can be to overgeneralised the results to a larger population without an adequate sample size.

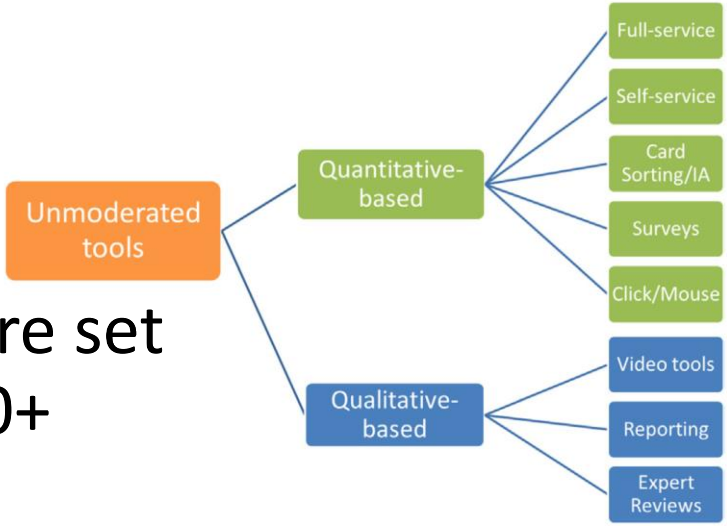
Online (Unmoderated) Usability Tests (1)

- Collect plenty of usability data in a short amount of time from users in dispersed geographically
- Online questions include background or screener questions, tasks, and follow-up question, data collected automatically.
- A wide range of data could be collected including many performance metrics and self-reported metrics.



Online (Unmoderated) Usability Tests (2)

- The use of different tools to collect quantitative or qualitative data from the participants:
  - Quantitative-based tools are set up to collect data from 100+ participants
  - Qualitative-based online tools are designed to collect data from a small number of participants to gain greater **insight** of the issues or “problems” with the product


Unmoderated tools
Quantitative-based
Full-service
Self-service
Card Sorting/IA
Surveys
Click/Mouse
Qualitative-based
Video tools
Reporting
Expert Reviews

Online Surveys

- Today, many online survey tools allow you to include images, which will allow you to collect feedback on visual appeal, page layout, perceived ease of use, and likelihood to use, to name just a few metrics.
- The strength of online surveys is it is a quick and easy way to compare different types of visual designs, measure satisfaction with different web pages, and even preferences for various types of navigation schemes.
- An online survey suits when you do not require participants to interact with the product directly.
- The main drawback of online surveys is that the data received from each participant are somewhat limited, but that may be offset by the larger number of participants.

Significance Testing (1)

(Hypothesis Testing)

1. Formulate a null hypothesis (H_0) that there is no difference (or no effect)
2. Collect data to test the hypothesis
3. Calculate the probability (P) of these or more extreme data occurring if the null hypothesis is true
4. If **P is large**, the null hypothesis cannot be rejected. This does not necessarily mean that the null hypothesis is true. The result is said **to be not statistically significant.**
5. If **P is small**, we **reject the null hypothesis**, and conclude that there is an effect which is said **to be statistically significant**.

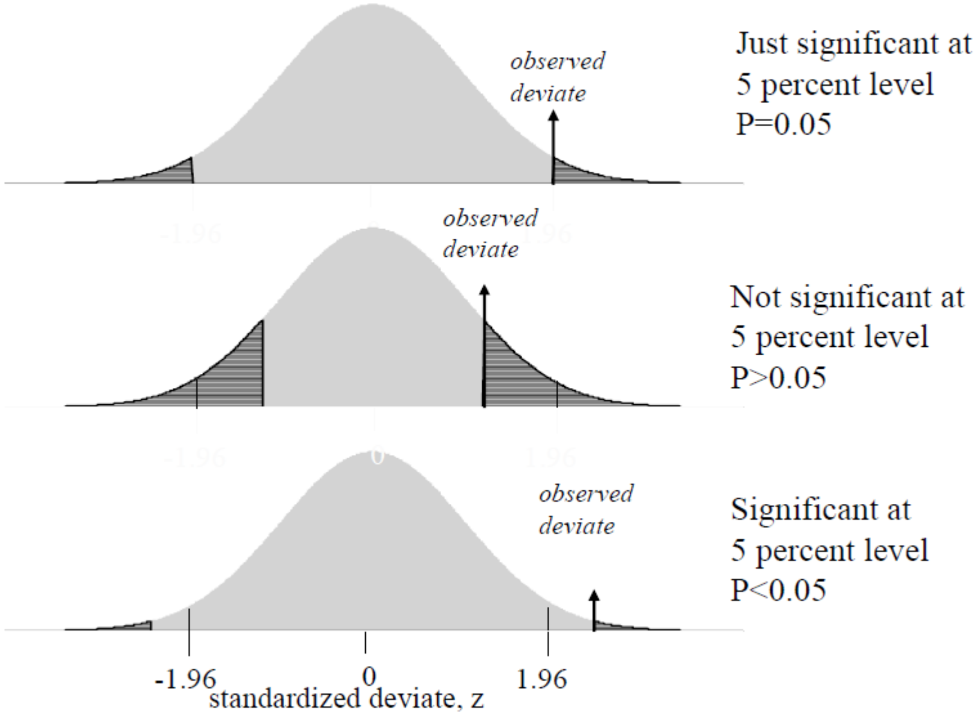
Significance Testing (2)

- *If the probability P is small, then either:*
  1. The null hypothesis is true and a rare event has occurred by chance

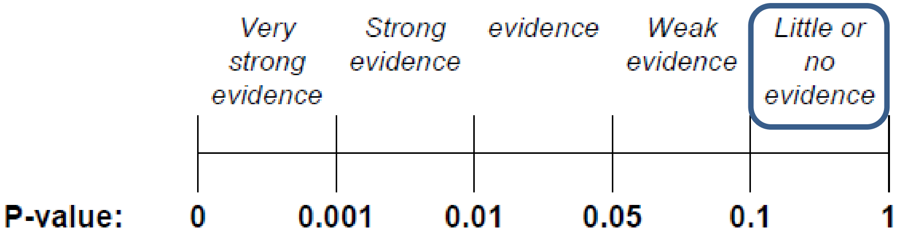
  OR

  2. The null hypothesis is false and can be rejected, and hence there is a difference (or an effect)

- *Explanation 2 is usually preferred, and the null hypothesis is rejected in favour of the alternative hypothesis that there is an effect, but there is a small risk (α) that this explanation is incorrect.*



The strength of the empirical evidence for rejecting the null hypothesis can be regarded as:



Anonymous quotation:

> *“Statistics is the only profession which demands the right to make mistakes 5% of the time.”*

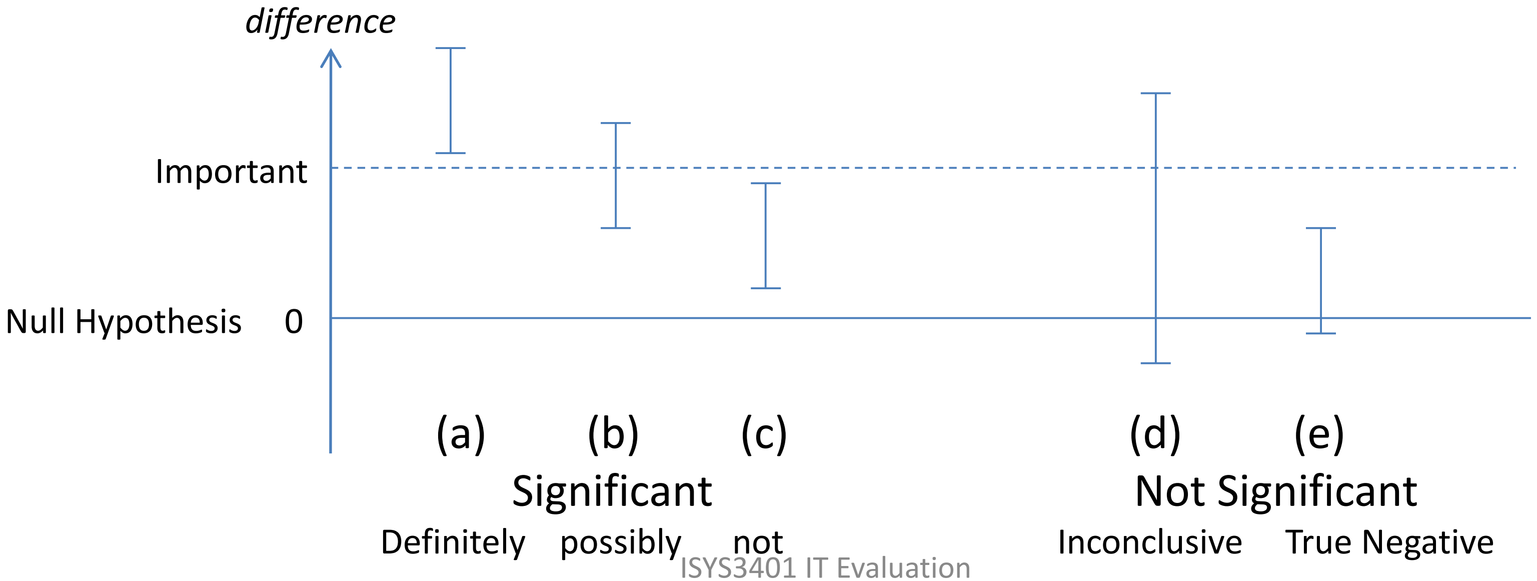
Statistical Testing

- Null Hypothesis (H0): there is no difference, or no effect.

- Question: *Is there evidence of a different (or an effect)?*
  - Is the difference **statistically significant**?
    - Are the data inconsistent with the null hypothesis?
  - Is the difference **statistically non-significant**?
    - Are the data consistent with the null hypothesis?

Statistically Significant versus Practically Important

- Null Hypothesis (H_0):
  - *Under the assumption that the null hypothesis is true:*
    - Are the data surprising? Are data that have a low probability of occurring by chance if the null hypothesis is true
    - Are the data what we might expect by chance? The P-value gives the probability of the observed data, or data more extreme from the null hypothesis
- Confidence Interval:
  - is centred around the observed sample estimate (mean); and
  - consists of those values of an effect with which the observed data are consistent, at a given level of confidence (usually 95%)



Testing Methods (σ unknown)

For Evaluation Studies

	One Sample	
	Continuous variable	**Binomial variable**
One sample	Estimate ***s*** from the sample and use the student's ***t***	Normal approximation to Binomial (equivalent to χ^2 test)

	Two Samples (to be compared)	
	Continuous variable	**Binomial variable**
Paired samples	Normally distributed (approximately) Paired ***t*** test	McNemar's test
Independent samples	Normally distributed (approximately) 2-sample ***t*** test	2 samples χ^2 test (2 x 2 table) Comparison of 2 proportions

Class Activities

- t-test
- χ2 test

[To be updated after Class Activities]

[McNemar’s Test]

This Week Tutorial

- t-test
- χ2 test
- McNemar’s Test