Qualitative vs. Quantitative

- Qualitative: Develop understanding of human experience
- Quantitative: Objectively measure human performance

- Less about more vs. more about less

When are each appropriate?

Quantitative Evaluation

- Gather (performance) measurements

- Methods
  - User events collection
    - *Mouse clicks, keys pressed,…*
    - *Data collected during system use*
      - Google, Amazon
  - Controlled experiments
    - *Set forth a testable hypothesis*
    - *Manipulate one or more* independent *variable*
    - *Observe effect on one or more* dependent *variable*
    - *Can be reproduced by others*

Controlled experiment

- State a lucid, testable hypothesis
- Identify independent and dependent variables
- Design the experimental protocol
- Choose the user population
- Apply for human subjects protocol review
- Run a couple of pilots
- Run the experiment
- Run statistical analysis
- Draw conclusions

Question Experiment

- Is it reliable?
  - Does the experiment take into account variations between subjects?
    - *Need for testing a sample of subjects*

- Is it valid?
  - Does the experiment reflects target use?
    - *Were users typical?*
    - *Were tasks typical?*
    - *Was the setting realistic?*
    - *Was the experience biased?*

Are results significant?

- Statistical significance
  - Comparing to the null hypothesis: “There is no effect”
  - Type I errors are the most disruptive

Researcher’s Decision	Actual Situation: Null Hypothesis is True	Actual Situation: Null Hypothesis is False
Fail to reject the null hypothesis	Correct decision	**Type II error**
Reject the null hypothesis	**Type I error**	Correct decision

- Design significance?
  - 3.00s versus 3.05s?

Are results significant?

- Statistical significance
  - Comparing to the null hypothesis: “There is no effect”
  - Type I errors are the most disruptive

Researcher’s Decision	Actual Situation	
	NO effect	Effect
NO effect	Correct decision	**Type II error**
Effect	**Type I error**	Correct decision

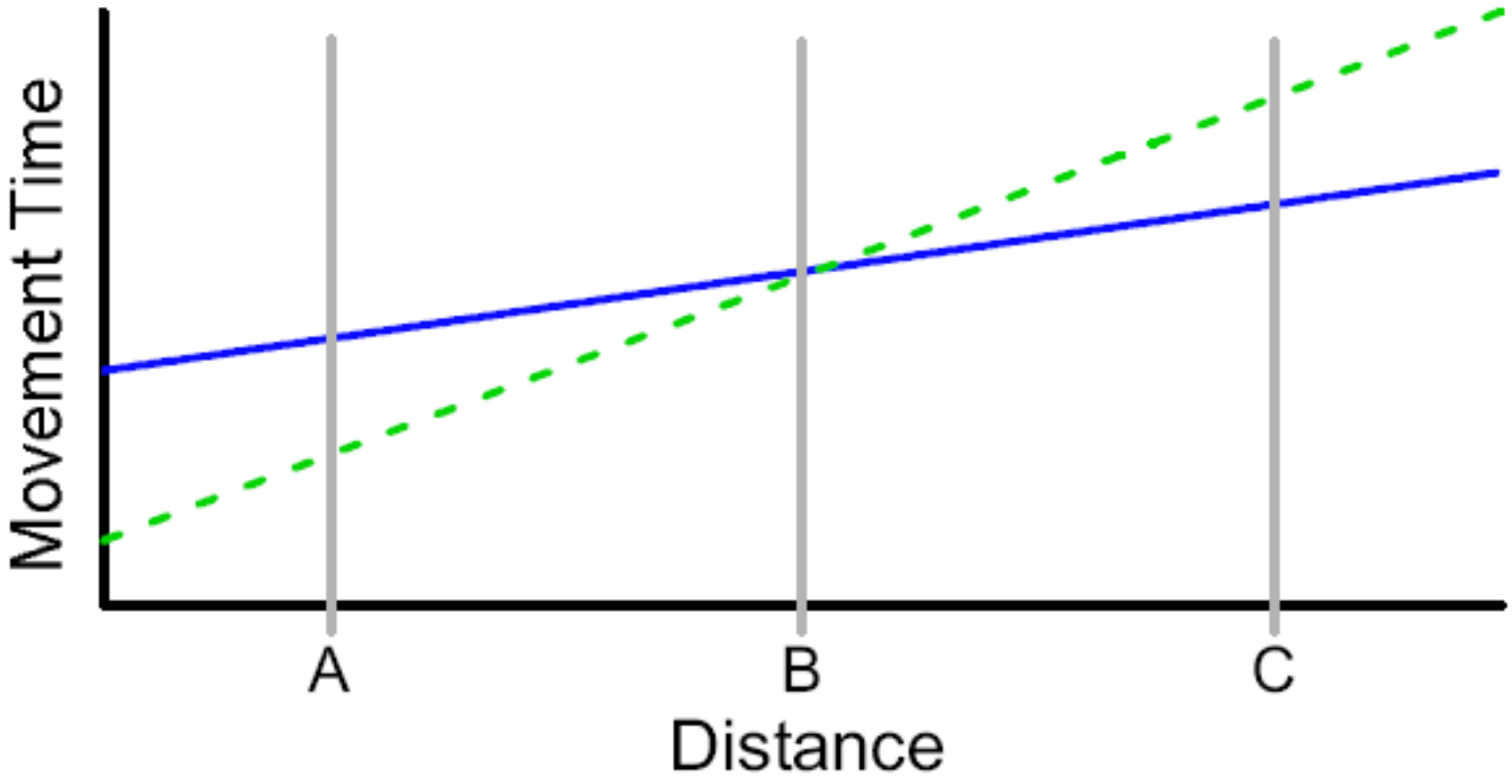
- Design significance?
  - 3.00s versus 3.05s?

Running example

- Compare Scrolling Techniques [Hinckley et al. ’02]
  - ScrollPoint
  - Standard Wheel
  - Accelerated Wheel (2 methods)

State a lucid, testable hypothesis

"With a proper acceleration function, a scroll-wheel based system can be faster than a ScrollPoint."



Choose the variables

- Manipulate one or more *independent* variable
  - Method
  - Device type…
- Observe effect on one or more *dependent* variable
  - Time to completion
  - Accuracy
  - Error rate…

- Running example
  - Independent variable: method
  - Dependent variable: speed, error rate, user satisfaction…

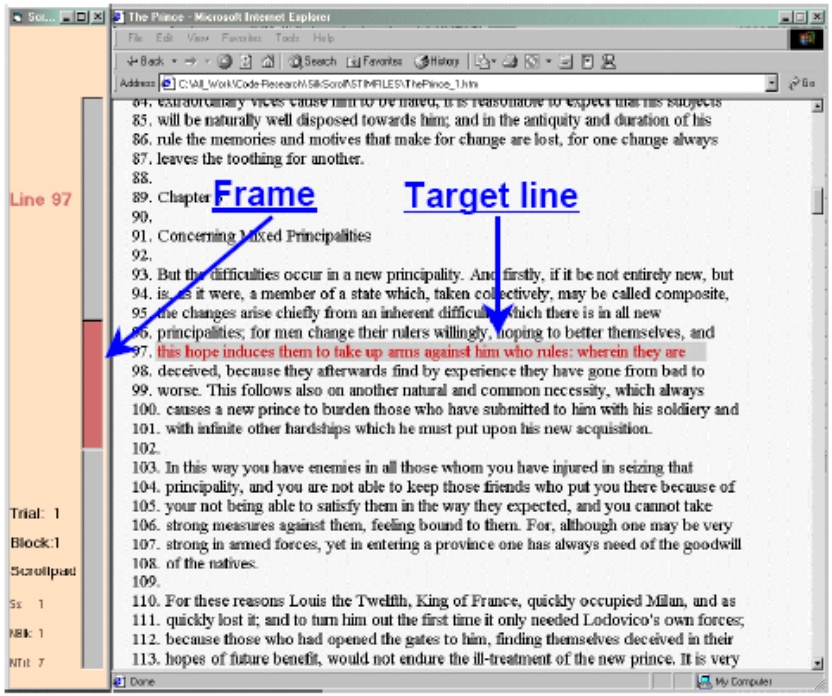
Design the experimental protocol

- Between or within subjects?
  - Between subjects: each subject runs one condition
    - *Need more subjects*
    - *Difference between subjects might introduce a bias*
  - Within subjects: each subject runs several conditions
    - *Need fewer subjects but possible problem with learning effects*
  - Very important for the statistical analysis phase

- Which task?
  - Must reflect the hypothesis
  - Must avoid bias
    - *Instructions, ordering…*
    - *In doubt, always favor the null hypothesis*

Design the experimental protocol

- Running Example:
  - Navigating in a document
    - *Using a simplified navigation task*
  - Use Fitts' law as the experimental framework



Chose the user population

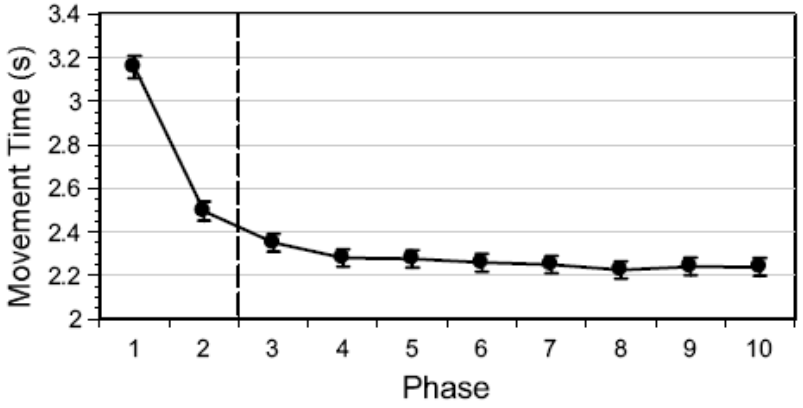
- Pick a well balanced sample
  - Novices, experts, average
  - Age group
  - Sex…
- Population group may be one of the independent variable

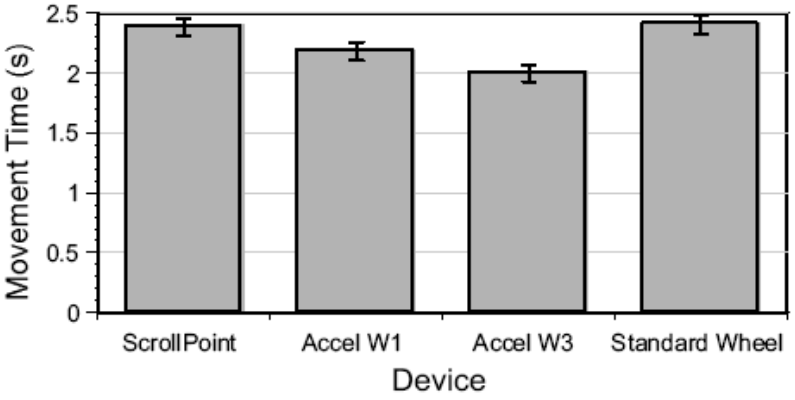
- Running example
  - Used a wide range of age

Run the experiment

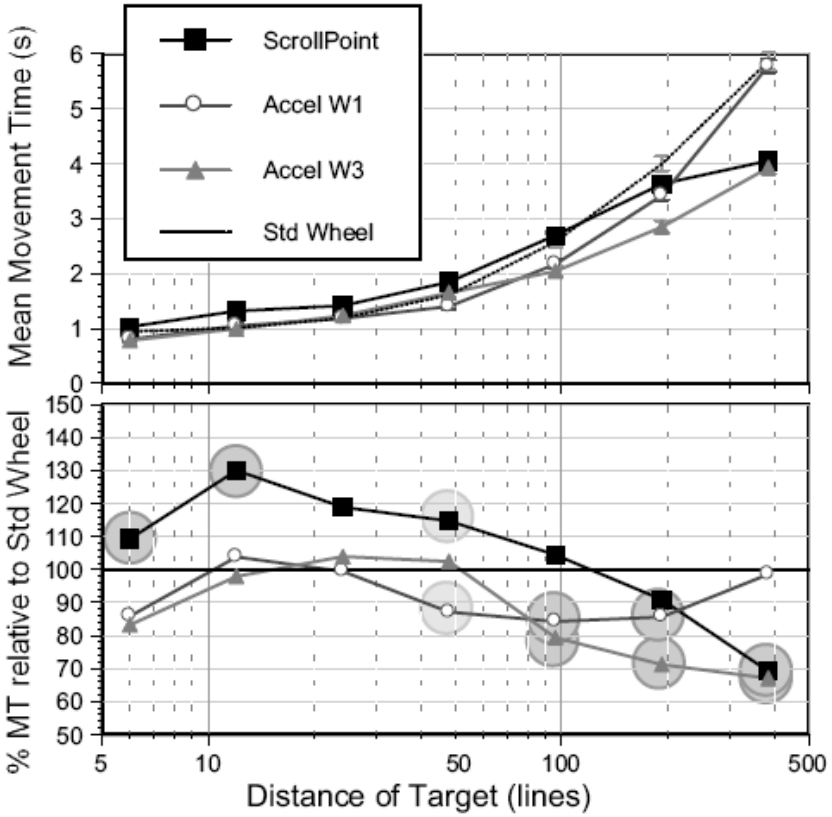
- Always run pilots first!
  - There are always unexpected problem!
  - When the experiment has started you cannot pick and choose
- Use a check-list so that all subjects follow the same steps
- Don’t forget the consent form!
- Don’t forget to debrief each subject

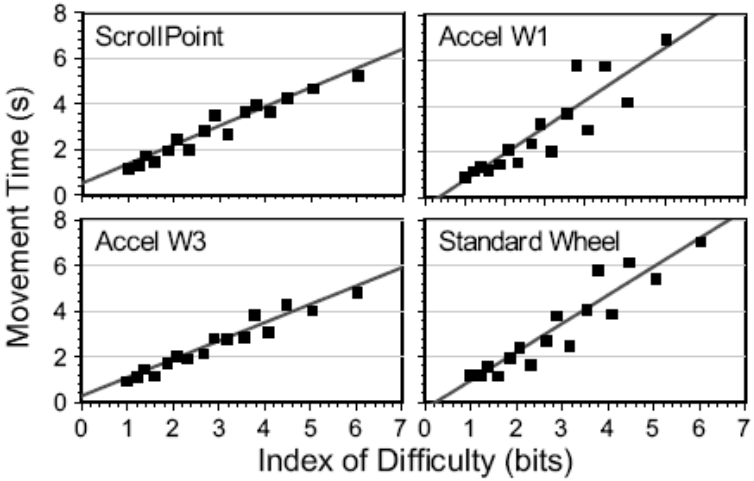
Running example result I





Running example result II





Run statistical analysis

- Properties of our population
  - Mean, variance…
- How different data sets relate to each other
  - Are we sampling from similar of different distributions?
- Probability that our claims are correct
  - Statistical significance:

    "The hypothesis that technique X is faster is accepted ($p < .05$)" means that there is a higher than 95% chance the hypothesis is true
  - Typical level are .05 and .01 level

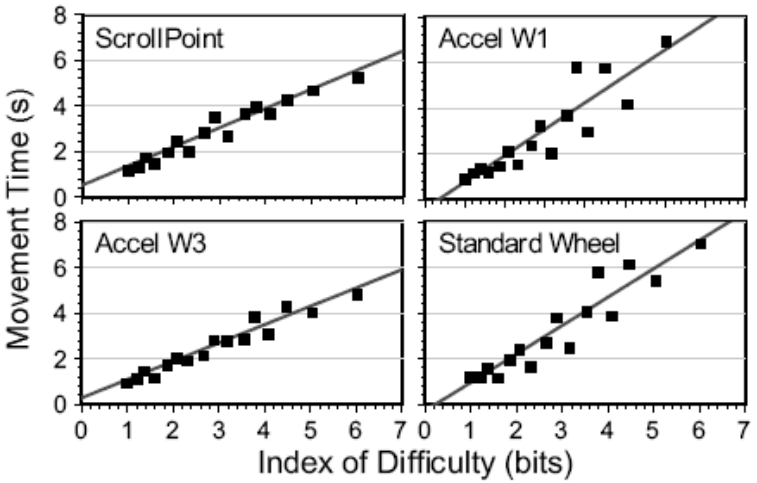
Statistical tools I

- T-test
  - Compare the mean of 2 populations
    - *Null hypothesis: no difference between means*
    - *Can only examine a single independent variable*
  - Assumptions
    - *Samples are normally distributed*
      - Very robust in practice
    - *Population variances are equal*
      - Reasonably robust for differing variances
    - *Individual observations in samples are independent*
      - Very important

Statistical tools II

- Correlation
  - Measure the extent to which 2 concepts are related
  - Caveats
    - *Correlation does not imply cause and effect (hidden variable)*
      - Ice cream consumption and drowning
      - Third variable problem
      - Directionality problem
    - *Need a large enough group*
- Regression
  - Calculate the “best fit”

	R	R^2	Slope	Intercept (s)	IP (bps)
ScrollPoint	0.97	0.94	0.84	0.42	1.19
Accel W1	0.90	0.81	1.16	-0.51	0.86
Accel W3	0.97	0.95	0.80	0.18	1.25
Wheel Std	0.94	0.88	1.25	-0.42	0.80



Statistical tool III

- ANOVA
  - Single factor analysis of variance
    - *Compare three or more means*
  - Analysis of variance
    - *Compare relationship between many factors*
      - Beginners type at the same speed on all keyboards,
      - Touch-typist type fastest on the qwerty

- Running example
  - Accept the hypothesis

- Your protocol influences the kind of test you can use
  - If in doubt, consult with a statistician before starting the experiment!

Reporting Results

- “This analysis revealed a significant main effect for Device, $F(2,15)=15.2$, $p<0.001$.”
- “As one would expect, movement times increased as either W decreased or D increased (i.e., as the task got more difficult: for W, $F(2,25)=801$, $p<0.001$; and for D, $F(3,54)=1429$, $p<0.001$).”