Experimental Validity

Roy A. Maxion

Dependable Systems Laboratory
Computer Science Department
Carnegie Mellon University
Pittsburgh, PA 15213
Email: maxion@cs.cmu.edu

03 July 2009

IFIP 10.4 Workshop on Experimentation
Obidos, Portugal

Claim

- Poor experimental methods (and/or inadequate reporting of them) impede scientific progress.
- Example: keystroke dynamics – unresolved after 30 years.
- Reasons: studies are …
    - invalid
    - uncontrolled
    - unreliable
    - unrepeatable / unreproducible
    - inadequately reported
        - no method section where (a) readers can find relevant details; (b) authors are prompted by section structure to include the essentials of the experiment
- We will focus on a few factors relevant to the validity of experiments in keystroke dynamics.

Experimental validity

- Valid – Well grounded; justifiable; logically correct.
- Experimental validity refers to the manner in which variables influence the results of the research.
- If a study is valid, then it truly represents what it was intended to represent.
- Validity is broken down into two general types:
  - Internal validity – concerned with ruling out rival explanations for the phenomenon under study
  - External validity – concerned with being able to generalize the results beyond the confines of the study

Consequences of not being valid

- Can’t predict accurately
- Results don’t generalize
- Experiments can’t be repeated, replicated, reproduced
- Previous work can’t serve as a foundation for future work – everyone has to start over

Outline

- Example - invalidities & keystroke dynamics
- Definition of keystroke dynamics
- Typical keystroke experiment
- What is it good for?
- Forensics – a new twist
- Taxonomy of typing behaviors
- What’s the state of the art?
- Current accomplishments, rationales
- Why keystrokes as investigative framework?
- Challenges for the future

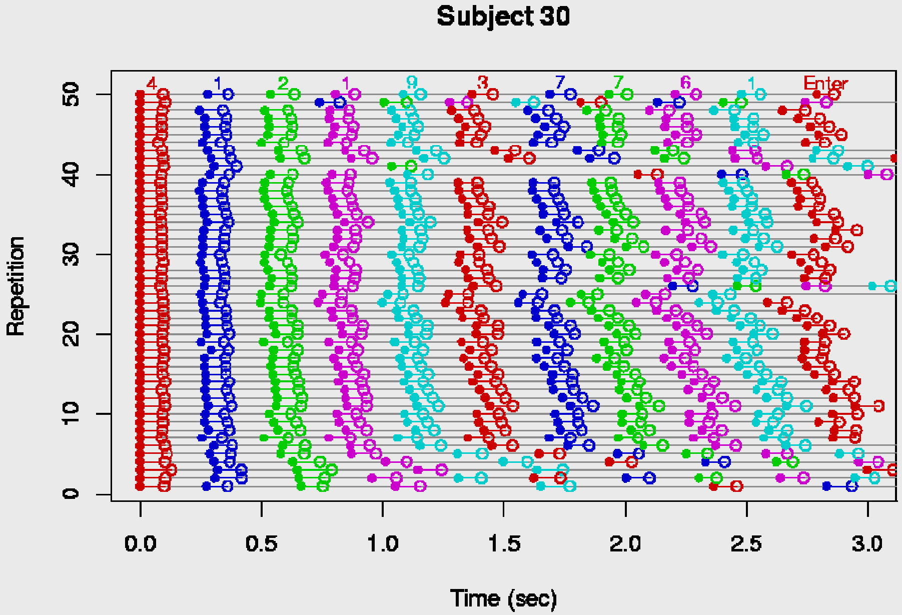
What is keystroke dynamics?

- Keystroke dynamics is the term given to the procedure of measuring and assessing a user's typing style, the characteristics of which appear to be unique to one’s physiology, behavior, & habits.
  - Like digital fingerprints in cyberspace
- The technique is based on (1) the timing latencies between keystrokes, (2) the time that a key is held down, and (3) other typing features (e.g., errors).
- These measures are compared to a user profile; a match or a non-match can be used to decide whether or not the claimed user is authenticated, or whether or not the user is the true author of a typed sequence or document.

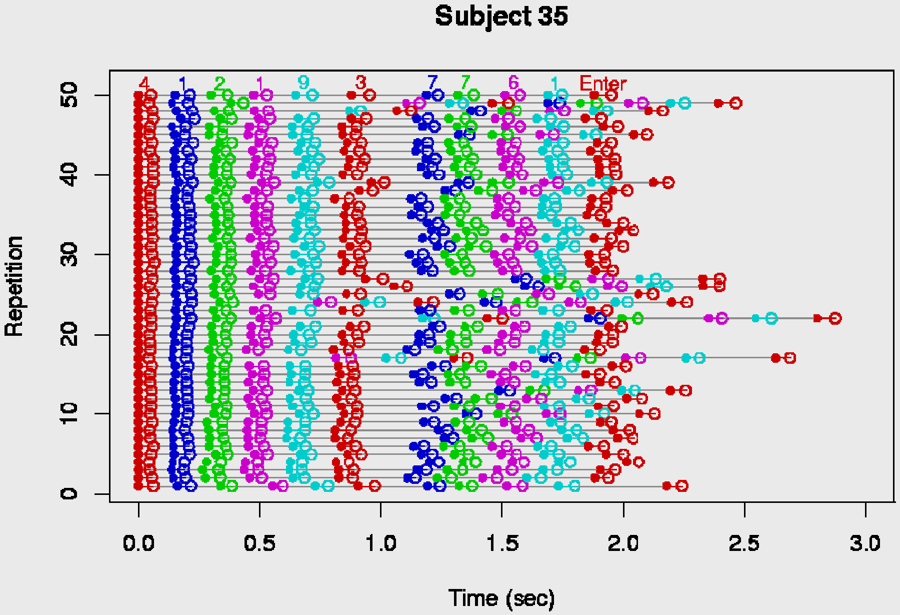
A typical keystroke experiment

- How a typical keystroke experiment is done:
  - N participants type a string repeatedly
  - Other participants act as impostors, typing the same string, but fewer times
  - A typing profile is constructed for each of the N participants
  - Participants are matched against their own profile for judging false-alarm (false-reject) rates
  - Impostors are run against participant profiles for judging miss (false-accept) rates

Compare: differences between two typists



Times between keystrokes are spread out and reasonably even.



Times between keystrokes are tighter and more consistent.

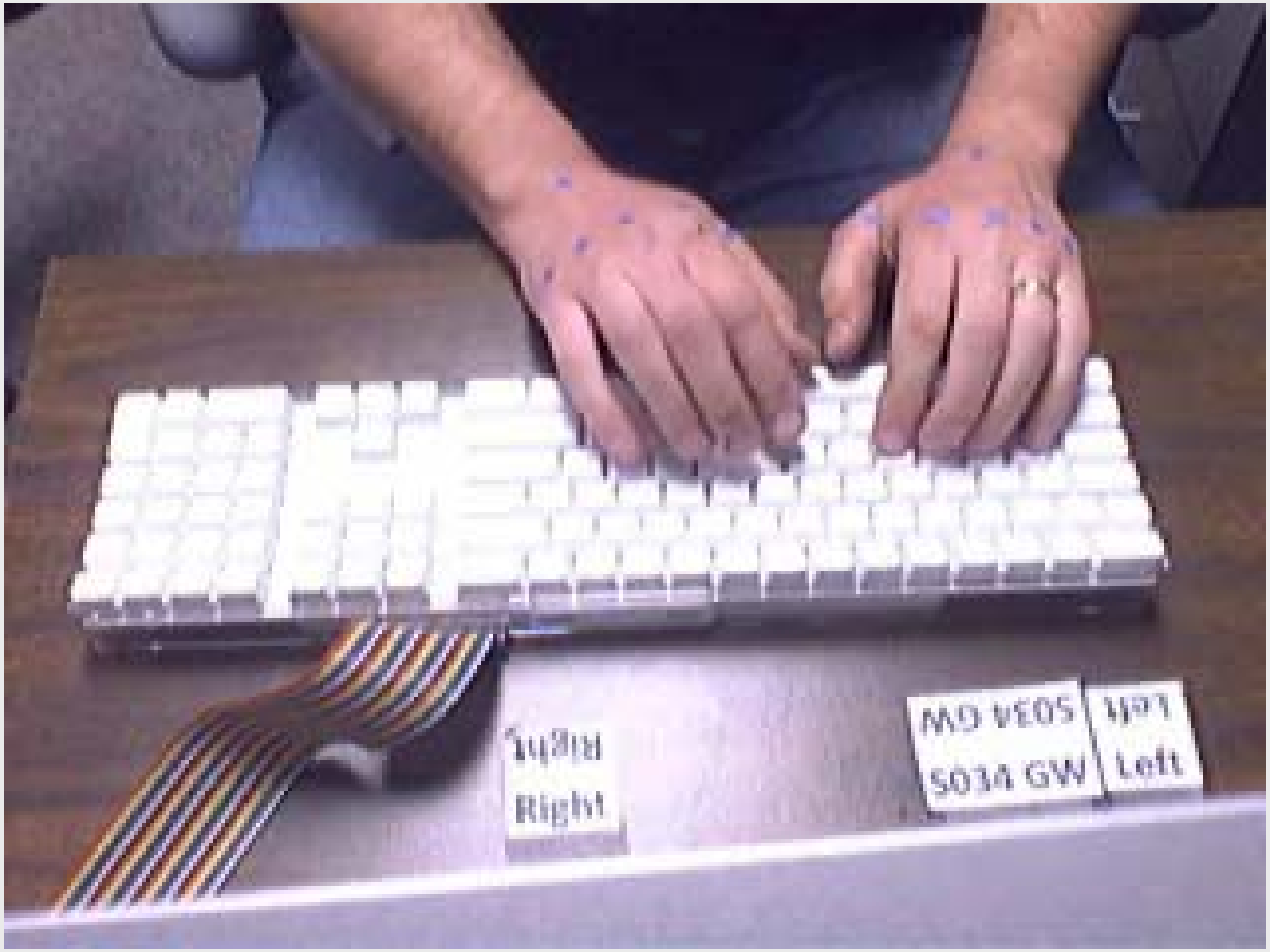
Two different users typed the passcode 412 193 7761 50 times each. Their typing patterns are remarkably different and unique.

Closed circles on the timeline indicate key-down events; open circles represent key-up events.

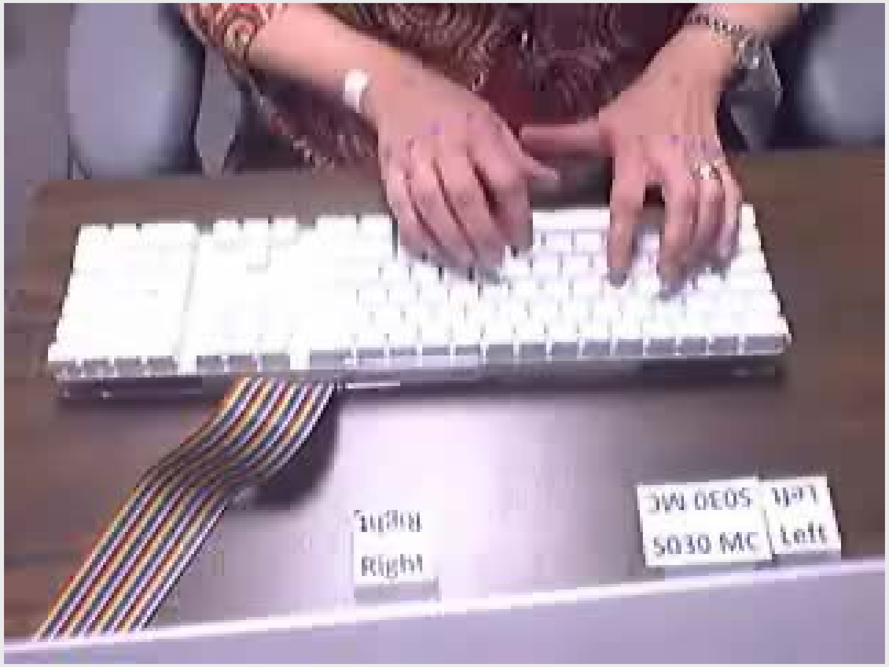
Many things can affect typing rhythm

- Gender
- Handedness
- Touch typist vs. hunt-and-peck
- Neural conditions
- Injury
- Native language
- Canonical typing errors
- Stress
- Keyboard defects (sticky or broken keys)
- Posture
- Hand geometry
- Typing behaviors (taxonomy of 8)

Blue dots trace hand points while typing

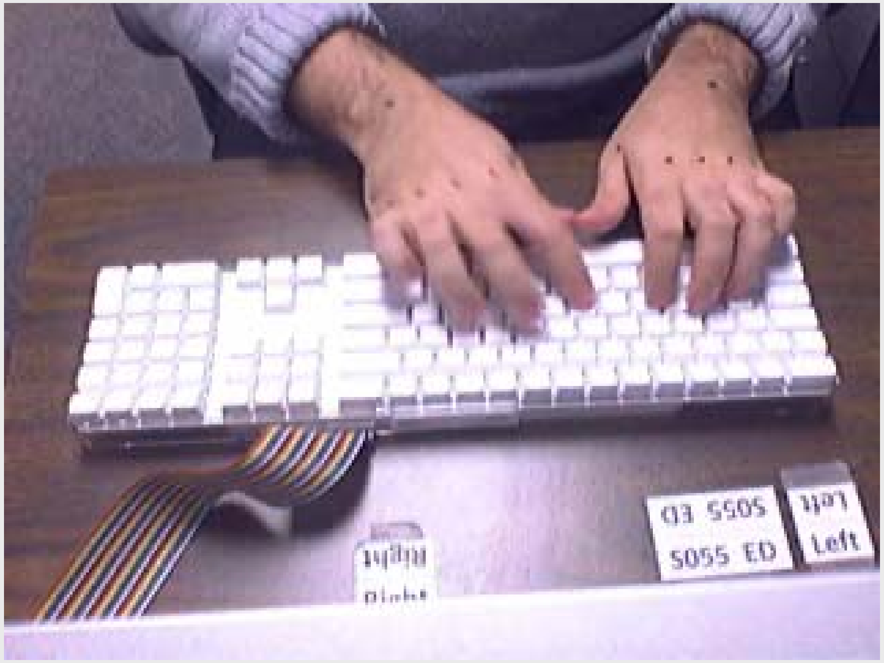


Isolated thumb



- The subject’s left thumb protrudes continuously while typing.
- Due to the tension in her thumb, the left hand is more open than the right hand.

Flexed (vs. extended) finger joints



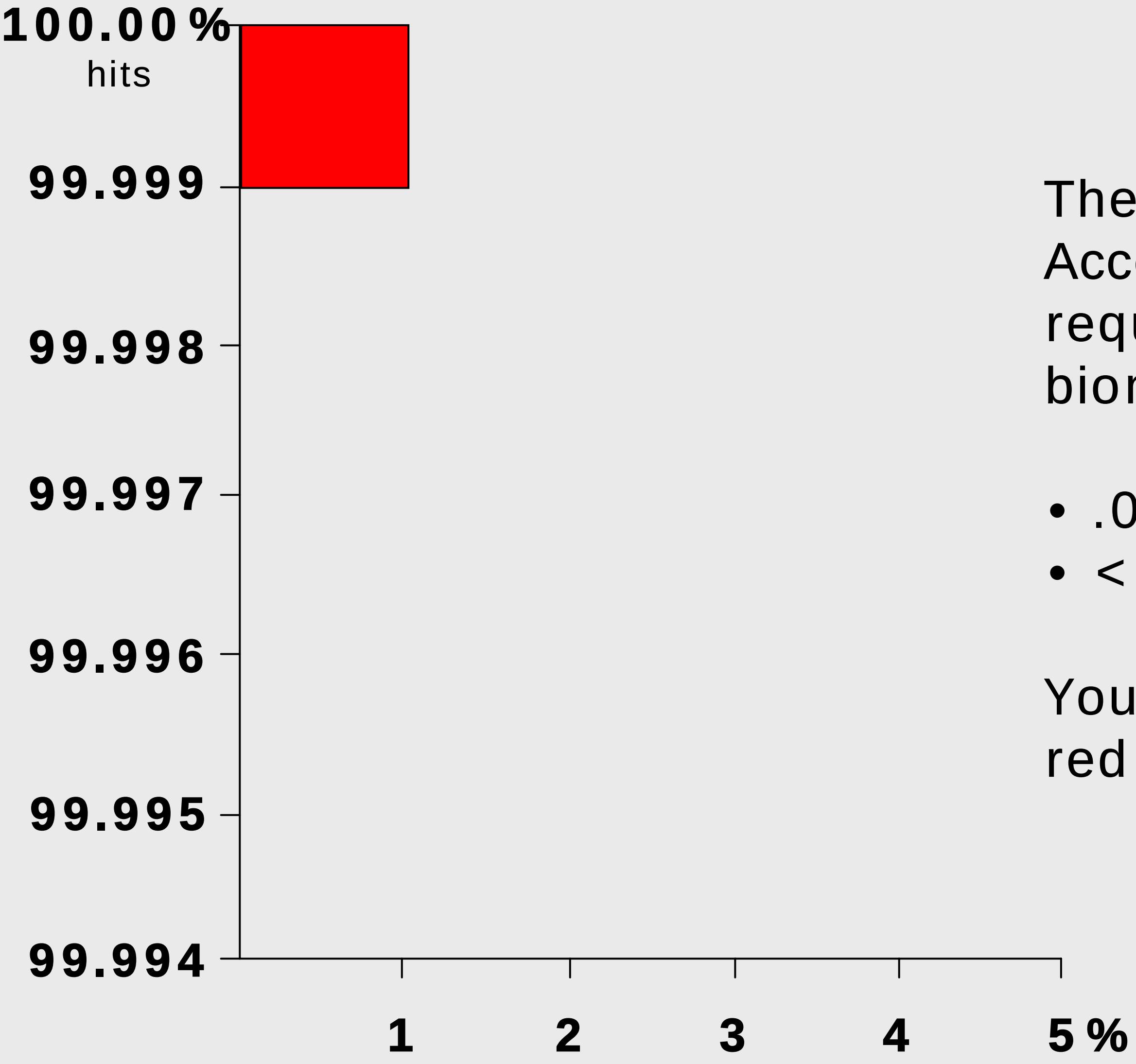
- The subject’s PIP/DIP joints are curved >25°on both hands.
- The pinky fingers (5th digit) are curled up so tightly that they are not used for typing.

What is keystroke dynamics good for?

- Two-factor authentication
  - What if you only had to have one password for all of your accounts?
- Continuous re-authentication
  - Checks to see that it’s still you, as you type.
- Questioned-document analysis / forensics
  - Did you write that email? Who issued that command? Who wrote or edited that document?
- Insider detection
  - Did someone else enter commands at your keyboard while you were out for coffee?
- Forensics
  - Ruling out (in) a class of suspects in network cyber-crime

No special equipment needed; just a keyboard.

How good do you have to be?

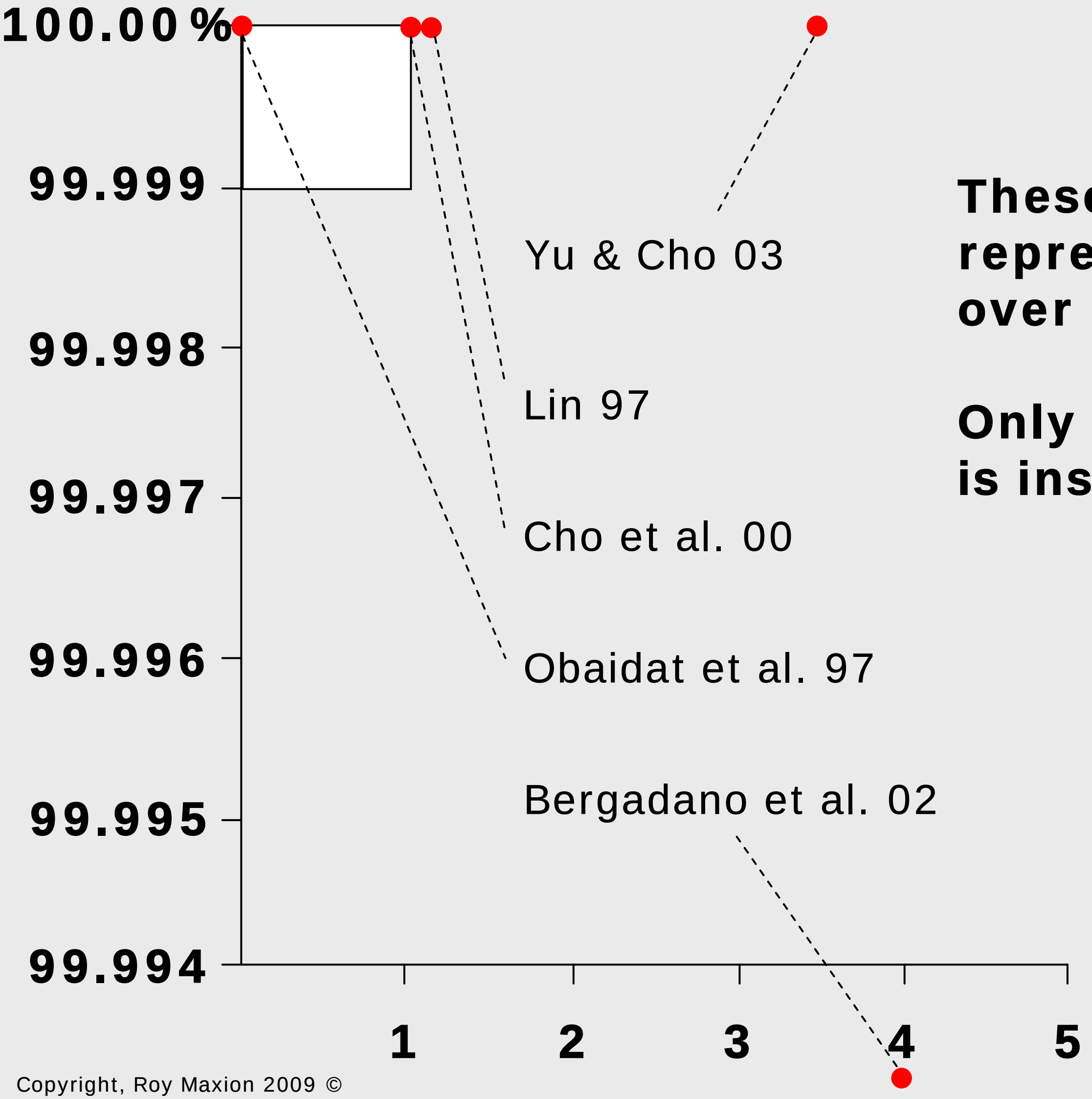


The European Standard for Access Control (EN 50133-1) requires a commercial biometric system to have a

- .001% miss rate; and
- < 1% false alarm rate.

You need to be in the little red square.

Best results



These citations and x,y points represent the best five results over the last dozen years.

Only one of them, Obaidat 97, is inside the square.

Best results, perhaps … but …

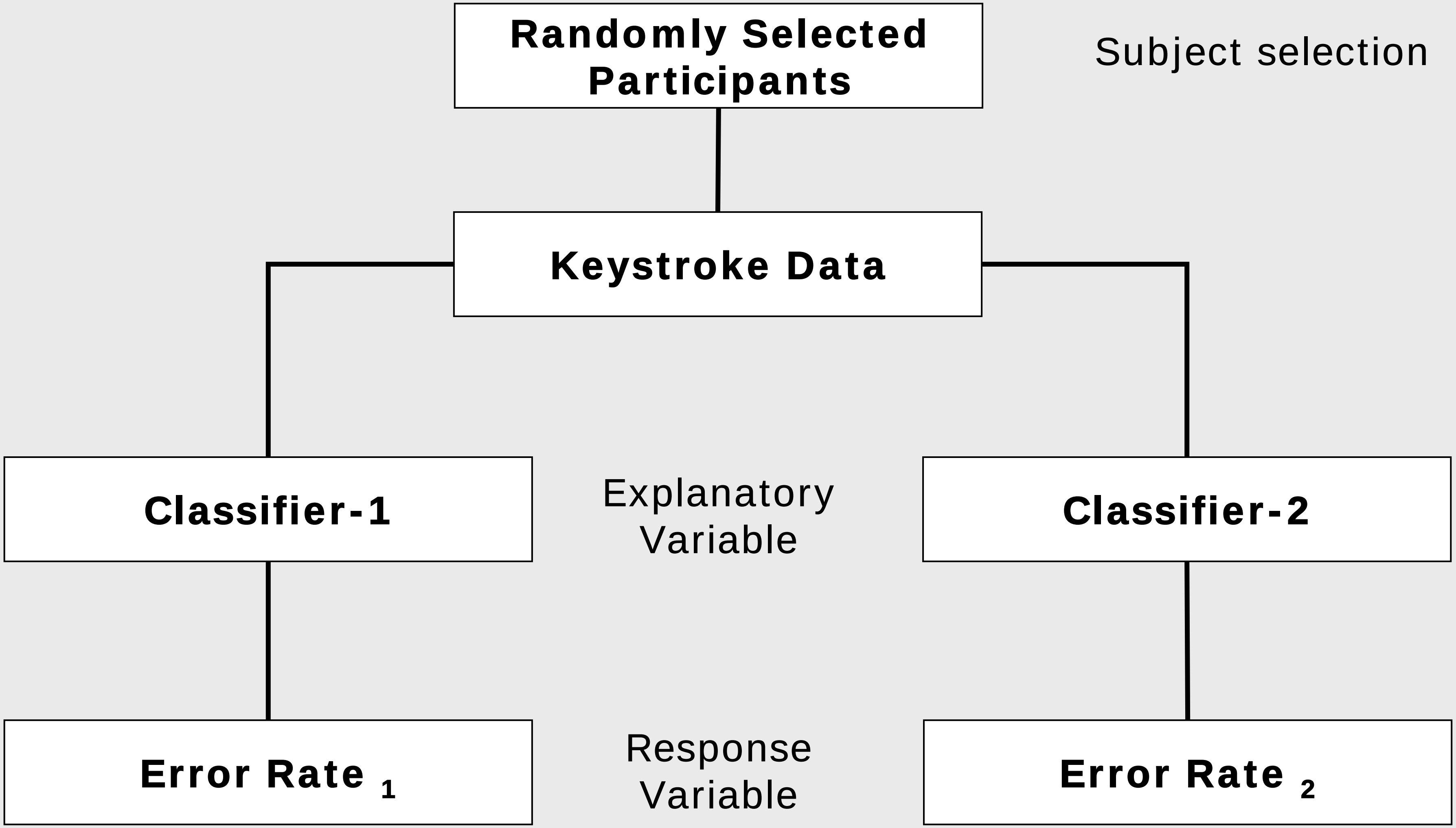
- In general, every study contained confounds – factors other than typing behavior that could explain the results. Previous work is not reproducible, due to:
- Apparatus factors
  - PS-2 vs USB keyboards, clock resolution, CPU load, network path
- Instrumentation factors
  - MS vs QPC timer, timestamp assignment mechanism, logger
- Task-structure factors
  - How many repetitions, how frequent, over what time period
- Stimulus factors
  - Self-selected and different passwords vs assigned passwords
- Practice factors
  - Typing a string/password 100 times vs 300 times (vs 14,000)
- Analysis factors
  - Feature extraction, feature transformation, training procedure, training repetitions, testing procedure, outlier handling, test-data selection, parameter tuning, cost/loss function

What is an experiment?

- Experiment: A procedure in which an intervention is deliberately introduced to observe its effects.

- There are several types of experiment:
  - True experiment: random assignment to the treatment or alternative condition.
  - Quasi-experiment: not assigned randomly.
  - Natural experiment: Not really an experiment; the cause usually cannot be manipulated, e.g., in a study contrasting a naturally occurring event such as before and after an earthquake.
  - Correlational / observational experiment: a study that simply observes the size and direction of a relationship among variables.

True (randomized) experiment



Subject selection

Explanatory Variable

Response Variable

Some hallmarks of a good experiment

- Valid
- Reliable
- Repeatable
- Reproducible
- Properly reported

Hallmarks of a good experiment (1)

- Valid
  - Internal - An experiment is internally valid if there are no alternative explanations for the outcome, other than the one posited for the experiment.
    - Example – distinguishing users by mouse movements … but letting users choose their own web content means that the content could have explained the outcome, not the user mousing style
  - External - An experiment is externally valid if the conclusions drawn from the experiment can be extended beyond the bounds of the experiment.
    - Example – college students would pay extra to make purchases from web site with a strong privacy policy … but college students are not representative of the general population

Hallmarks of a good experiment (2)

- Reliable/repeatable
  - Repeatability refers to the variation in measurements taken by a single person or instrument … on the same item … and under the same conditions; we seek high agreement from one measured instance to another.
  - A measurement is said to be repeatable when this variation is smaller than some agreed limit.

Hallmarks of a good experiment (3)

- Reproducible
  - Reproducibility relates to the agreement of experimental results with independent researchers using similar but physically different test apparatus, and different laboratory locations, but trying to achieve the same outcome as was published in an article.
  - Replication or reproduction allows an assessment of the control on the operating conditions, i.e., the ability to reset the conditions to some desired value. Ultimately, replication estimates our control over the procedure used.
  - Measurements give the same results each time they are taken, irrespective of who does the measuring.

Hallmarks of a good experiment (4)

- Properly reported
  - All methodological details are provided (preferably together, in one place), enabling readers to reproduce the experiment (even if only mentally), and to obtain the same results.

Note on internal vs. external validity

- This may be seen as a trade-off.
- One kind of validity may be increased at the expense of the other.
- Example is keystrokes.
- The more realistic (externally valid) the experiment is, the less internally valid it is.

Eh?

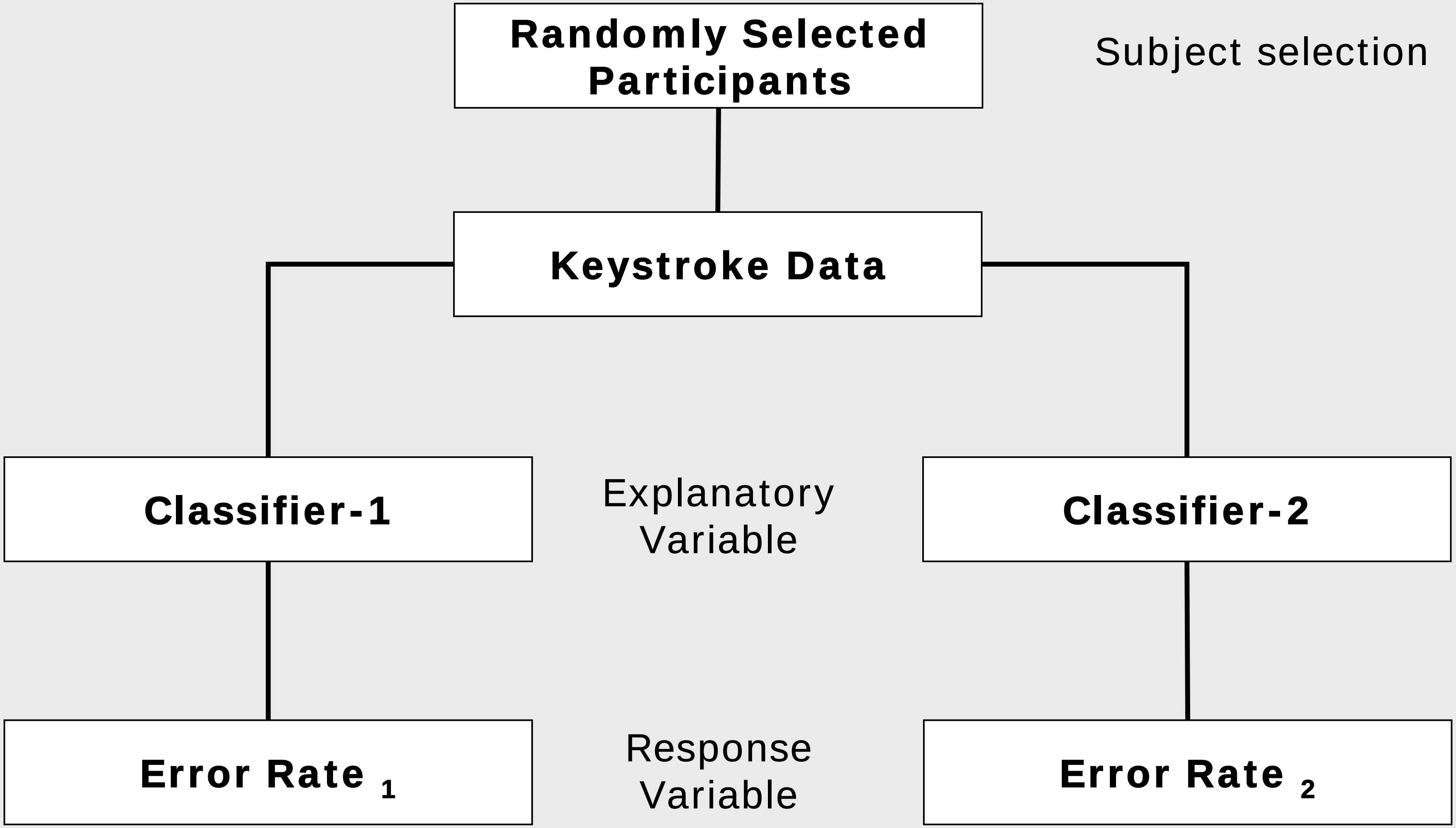
- Why generalize from experiments?
  - We can’t test every situation, so we need to extend (infer) from a sample to a broader population.
  - Side note: if you happen to conduct an experiment that doesn’t generalize, not all is lost; just don’t claim (or imply) that it generalizes.

- Why replicate or reproduce experiments?
  - To check results about which you are skeptical.
  - To establish a starting point for extending previous work.

When an experiment is not valid, neither generalization nor reproduction is possible.

- Let’s look at some examples of how a simple experiment can be rendered invalid.

True (randomized) experiment



Randomly Selected Participants

Subject selection

Keystroke Data

Classifier-1

Explanatory Variable

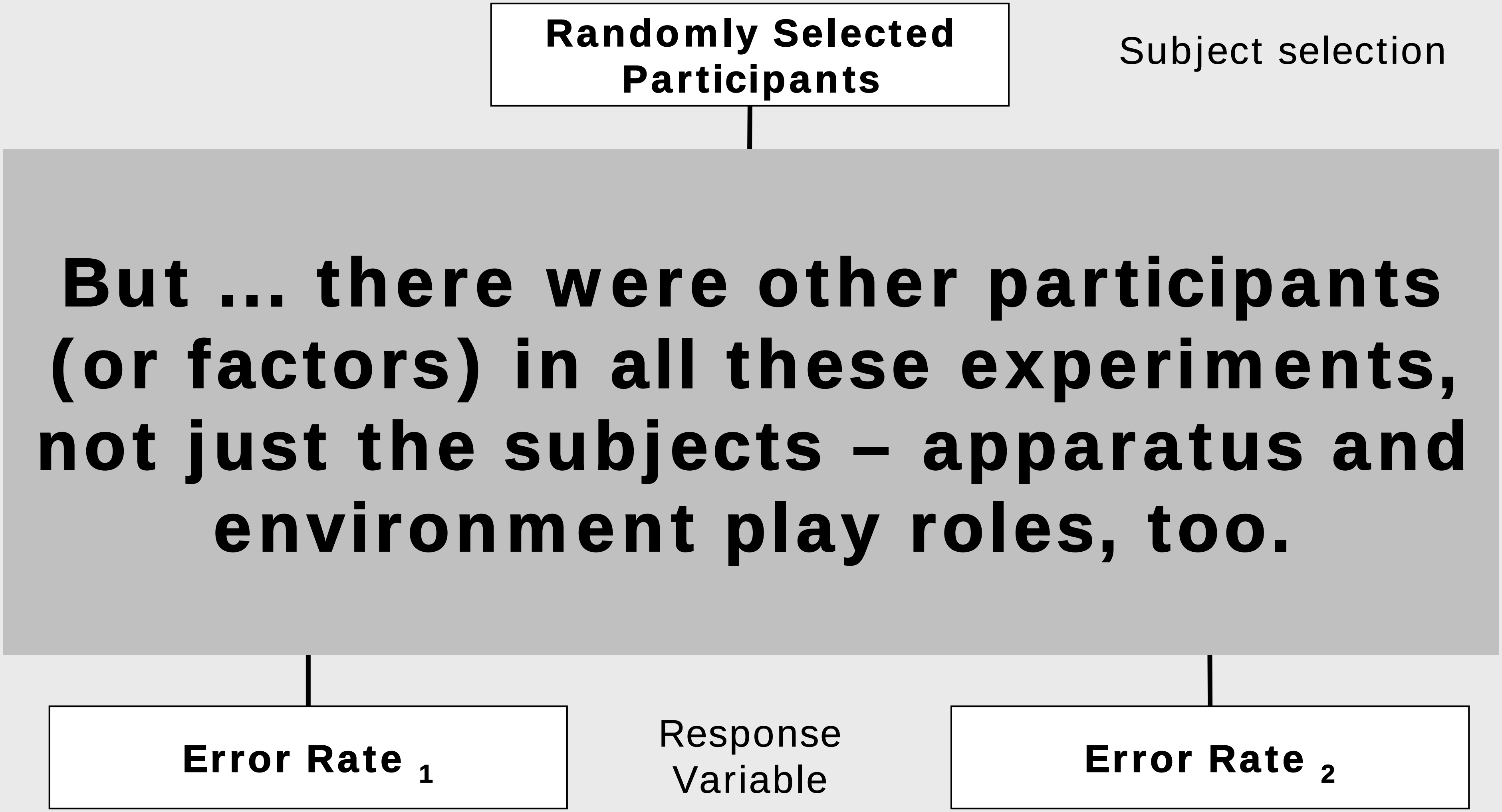
Classifier-2

Error Rate $_1$

Response Variable

Error Rate $_2$

True (randomized) experiment



Randomly Selected Participants

Subject selection

But … there were other participants (or factors) in all these experiments, not just the subjects – apparatus and environment play roles, too.

Error Rate $_1$

Response Variable

Error Rate $_2$

We will examine some of these factors

- Apparatus - An appliance designed for a specific operation
- Instrumentation - A measuring device for determining the present value of a quantity under observation; often software-based
- Materials - Physical and stimulus items that might inject differences or manipulations into an experiment
- Subjects - The people or objects of study
- Instructions to subjects - Details of what the subjects should do
- Design - Description of the variables to be manipulated/measured, how the exptl & control groups were constituted, how subjects/objects were assigned to groups
- Procedure - Chronological account of the experiment, start to finish
- Analysis - How the data were treated

Parts of an experimental paper

- Title
- Author(s)
- Abstract
- Introduction
- Problem being solved
- Background and related work
- Approach
- **Method**
  - Apparatus & instrumentation
  - Materials
  - Subjects
  - Instructions to subjects
  - Design
  - Procedure
- Data
- Analysis
- Results
- Discussion
- Limitations
- Conclusion
- Future
- Acknowledgements
- References
- Appendices
- Endnotes and footnotes

Apparatus & instrumentation

- Hardware
  - Computer, keyboard – must be uniformly used
  - Timer – must be calibrated; it’s not accurate just because you say so
- Software
  - Presenter – reproducible?
  - Logger – run under uniform conditions?

Calibration procedure

Keystroke timing accuracy was calibrated by pulsing the keyboard matrix with a known signal; we used a Hewlett Packard model 33120A, 15 MHz function and arbitrary waveform generator.

We used a square wave whose characteristics were: frequency of one Hertz, amplitude of 3.8 volts peak-to-peak, duty cycle of 50%, DC offset of 2 volts, and rise time of 20 nanoseconds.

The keyboard matrix was triggered by the square wave via a simple TTL logic tri-state output latch, with the “enable” input tied to the clock line (the output of the function generator).

Three thousand keystroke events (one key-press and one key-release per event) were triggered. 81.3% had zero error, and 18.7% had an error of 200 microseconds (or 0.2 milliseconds).

At worst, timing is accurate to a precision of 200 microseconds.

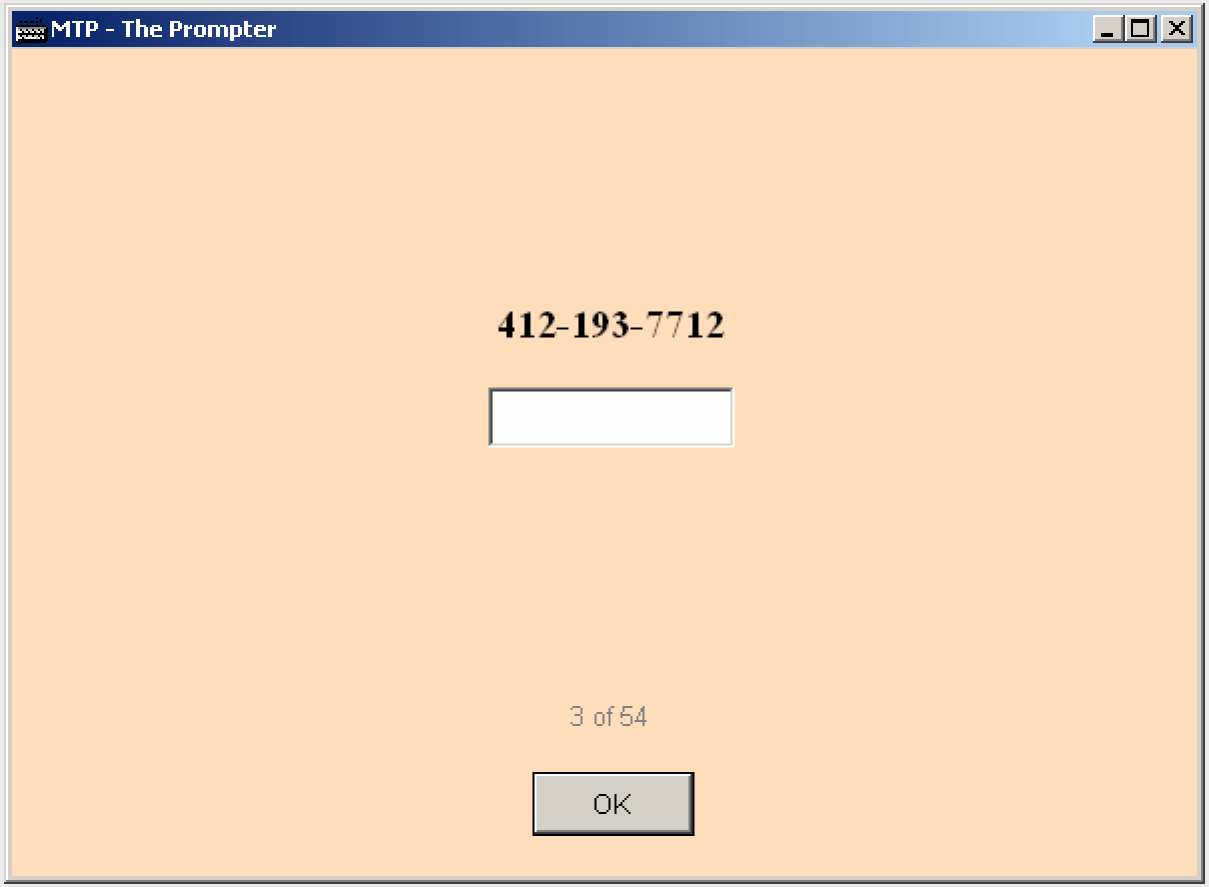
Presentation and logging software

- Two components: prompter and logger

- Prompter
  - Takes script-driven text configuration file
  - Presents instructions to user
  - Prompts text to be typed
  - Checks for typing errors

- Logger
  - Timestamps every event
  - Writes log in XML
  - Includes session information
  - Includes copy of configuration script

Example presentation screen



MTP - The Prompter

412-193-7712

3 of 54

Presentation and logging software

- Two components: prompter and logger

- Prompter
  - Takes text configuration file
  - Presents instructions to user
  - Prompts text to be typed
  - Checks for typing errors

- Logger
  - Timestamps every event
  - Writes log in XML
  - Includes session information
  - Includes copy of configuration script

XML log snippet (“4” make & break)

```
- <event type="keystroke">
  <timestamp source="ticks">20.8910</timestamp>
  <timestamp source="qpc">20.8995</timestamp>
  <virtual_key>VK_NumPad4</virtual_key>
  <key>4</key>
  <key_event>make</key_event>
  </event>
- <event type="keystroke">
  <timestamp source="ticks">20.9530</timestamp>
  <timestamp source="qpc">20.9598</timestamp>
  <virtual_key>VK_NumPad4</virtual_key>
  <key>4</key>
  <key_event>break</key_event>
  </event>
```

Apparatus & instrumentation - validity

- Hardware
  - Computer, keybd –uniform across experiments
  - Timer – calibrated; procedure described
- Software
  - Presenter – script driven; facilitates reproducibility
  - Logger – not run in the presence of other loads

- If one uses diverse hardware, or logs under different loads at different times, validity is threatened.

Materials – stimulus items

- We will stick to stimulus items – what things are people asked to type?
  - Passwords
  - Passphrases
  - Free text

- Issues
  - Choose your own password, or be assigned one?
  - Everyone types the same material, or different?
  - Materials chosen on principle, or ill-considered?

Materials

- Choosing own password
  - Subjects have different motives, such as easy or hard to type
- Same pwd or different across subjects?
  - If different, then results could be attributed to people having typed different items, not to style
- Stimuli chosen on principled basis
  - One-finger number – rationale for each digit
  - 5 passphrases – all 31 characters, "pre-practiced"
  - Free text – make pictures as uniform as possible

Materials - validity

- Making stimuli realistic and uniform across subjects reduces threats to validity.

- If different subjects use different materials, then these differences could account for the classification/discrimination results, not the subjects' typing rhythms.

- A demographic questionnaire can be used for identifying subjects' handedness, gender, etc. ... to be recorded for ground-truth purposes.

Subjects

- Sampling the population – is the sample representative of the population?
- Is the population the one to which you want to generalize?
- There are many sampling techniques; random is best.
- We took a “convenience sample,” which is often regarded with skepticism, but in our case it seems reasonable.
- We did not draw the sample from outside the Big-and-Tall Shop, so it was not likely to be biased with lots of large-handed people.
- We photographed people’s hands so we could assess the effect of hand geometry on results.

Instructions to subjects

- Participants must be told exactly what to do. If they're not told, they will infer the goals of the task (perhaps incorrectly), and they will respond accordingly.
- Getting instructions right is similar to getting a questionnaire right (like a review form).
- There is a risk that each subject will respond in a different way than you expected … but you will never know.
- Example:
  - Type naturally; this is neither a speed nor an accuracy test.
  - Explain that this experiment involves no trickery or deceit … unlike typical psychology experiments, for example.
- Pilot-test all the instructions, several times; use verbal protocols
- Query subjects after the task to probe for hidden motives or "perceived" instructions.

- Validity is reasonably assured, but this can only be judged by readers if the instructions are provided in the paper.

Design

- Our measured variables are simple times: key-down and key-up.
- We have no manipulations, so there are no control and treatment groups.
- But we have issues of sessions, repetitions, and return visits over time:
    - Ask for 400 repetitions of a password … in one sitting???
    - Have subjects type 50 repetitions per sitting for 8 sittings.
    - This raises issues of validity, because this is not a natural way to learn a new password (usually).
- A discrimination task could be closed-world or open-world, which introduces other validity issues – how real are either of these two paradigms?
- Similarly, we could do anomaly detection or multi-class classification; how appropriate are these?
- Reader cannot assess these issues unless written.

Procedure

- The experimental procedure comprises all the activities of a subject’s experience, from the time s/he walks into the lab until the time s/he walks out.

- Set up data-collection environment
- Explain experiment to subject
- Consent form / IRB
- Give instructions
- Check for uniform lighting, noise, environmental conditions
- Run the experiment
- Debrief the subject (& administer demographic survey)
- Ensure integrity of data, and archive
- A checklist with operationally-defined procedures leads to identical conditions and reproducible processes.

- Validity must be assessed with regard to experimental goals.

Analysis

- There are many issues here
  - Outlier handing – how done, if at all? Can make a huge difference in outcome.
  - Dropped subjects – this is ok, but only for clear and principled reasons, not because the drops make the results come out better.
  - How training and testing data were drawn across the session boundaries can reflect the extent of practice that subjects had at any point in the experiment.
    - If training data are drawn from early in the data, and testing from late in the data, is this valid? Or vice versa?
  - How are decision threshold determined? What is the tuning procedure for the classification algorithms (e.g., number of hidden layers in a neural net)?
- Validity of these elements of the experiment must be judged in the context of the experimental goals.

- But they cannot be judged if they do not appear.

What cannot be done w/ invalid expt.

- You can’t reconcile results across studies if the clocks are different.
  - Or if the network paths are different, or if the system loads are different, or the keyboards are different.
- Can’t make claims about having the best algorithm if all the algorithms are run under different conditions
  - And not on the same data.
- Can’t claim generality with a biased sample of an unrepresentative population.
- Can’t advance the field – remember … 30 years of invalid experiments, not well reported.

Consequences of invalidity

- Can’t predict accurately (isn’t science about prediction?).
- Results don’t generalize.
- Experiments can’t be repeated, replicated, reproduced.
- Previous work can’t serve as a foundation for future work – everyone must start over.

Summary

- Factors affecting experimental validity can be subtle; we need to watch for them.

- We are a dependability community.

- Our goal should include being dependable in all respects, including the correct and thorough conduct and reporting of experiments.

- The least we could do is ensure that experimental work is valid.

- And that authors tell the story so that readers (and reviewers) don’t have to struggle to understand it.

Aphorismus …

The unity of all science consists alone in its method, not in its material.

Karl Pearson
The Grammar of Science
Meridian Books: NY. 1911
(First published in 1892)