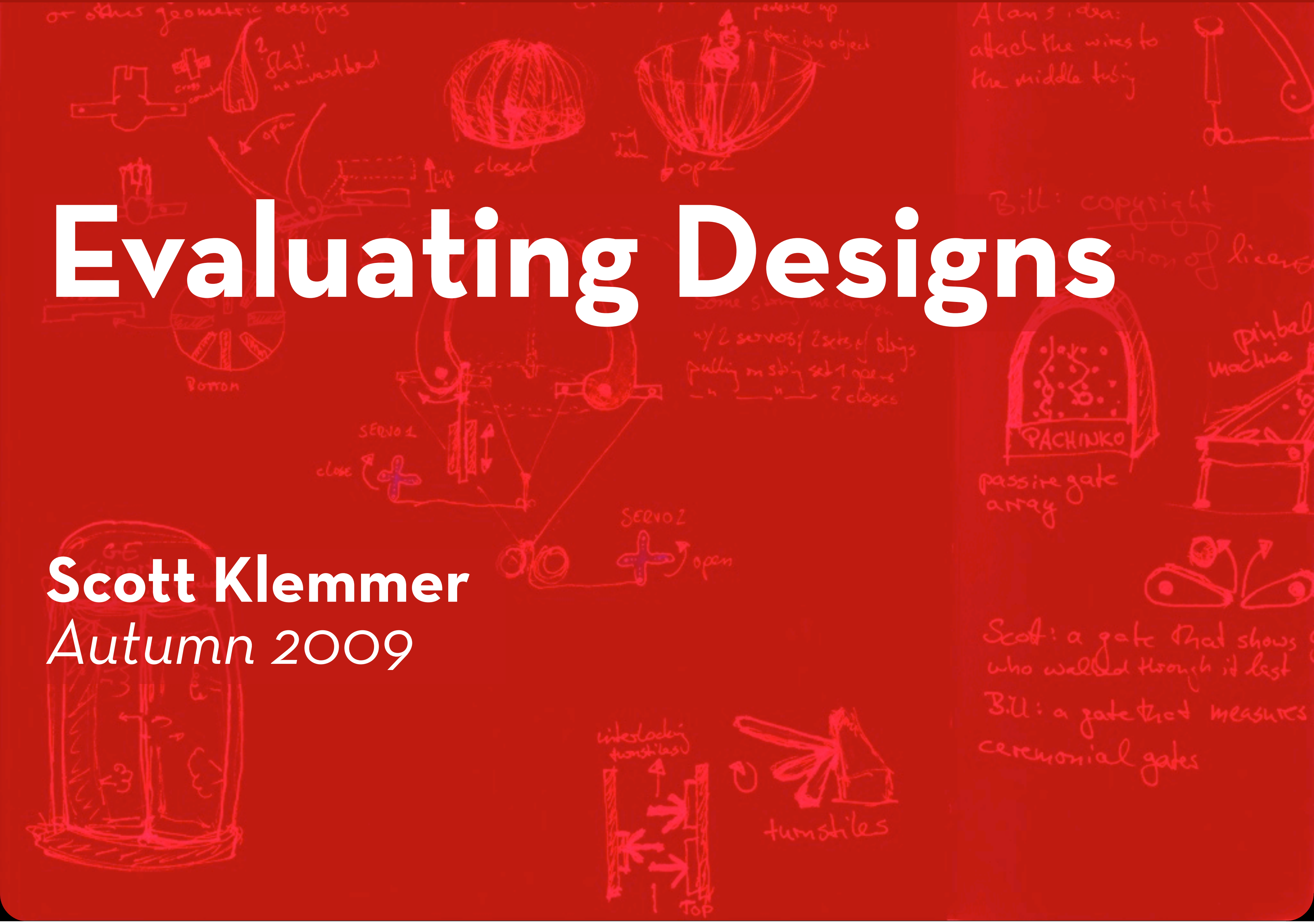
STANFORD HCI GROUP HTTP://CS147.stanford.edu


or other geometric designs
flat
open
closed
lift
Alan's idea: attach the wires to the middle tubing
Bill: copyright
pinball machine
PACHINKO
passive gate array
BOTTOM
w/ 2 servos, 2 sets of strings
pulling on string set 1 opens
2 closes
SERVO 1
close
SERVO 2
open
Scott: a gate that shows who walked through it last
Bill: a gate that measures
ceremonial gates
interlocking turnstiles
turnstiles
TOP

Evaluating Designs

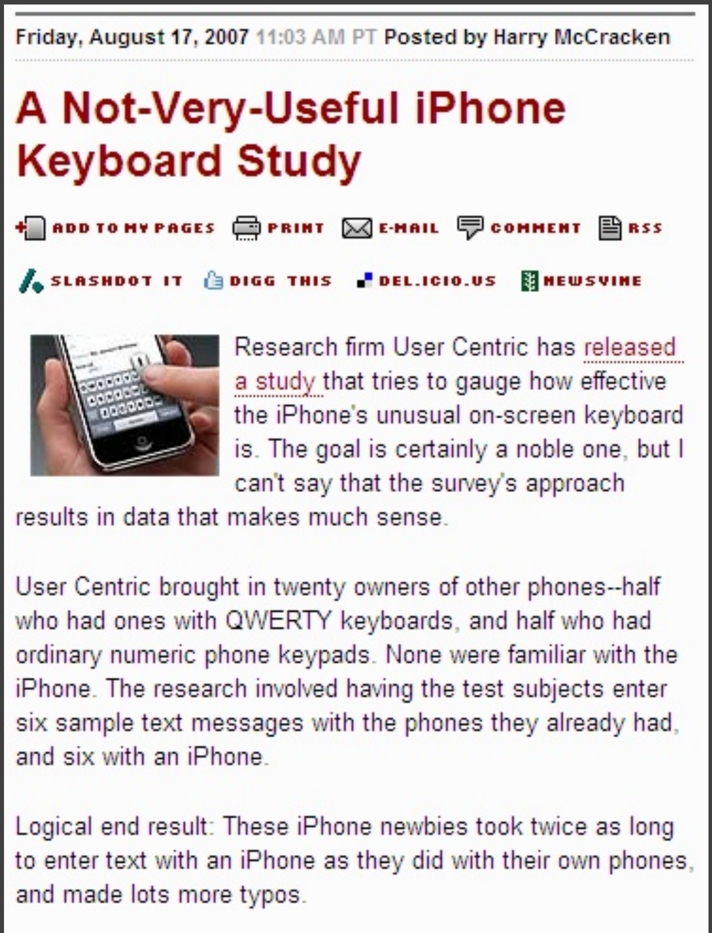
Scott Klemmer

Autumn 2009

How can we measure success?

How do we know?

Poor tests yield poor results



Friday, August 17, 2007 11:03 AM PT Posted by Harry McCracken

A Not-Very-Useful iPhone Keyboard Study

Research firm User Centric has released a study that tries to gauge how effective the iPhone's unusual on-screen keyboard is. The goal is certainly a noble one, but I can't say that the survey's approach results in data that makes much sense.

User Centric brought in twenty owners of other phones--half who had ones with QWERTY keyboards, and half who had ordinary numeric phone keypads. None were familiar with the iPhone. The research involved having the test subjects enter six sample text messages with the phones they already had, and six with an iPhone.

Logical end result: These iPhone newbies took twice as long to enter text with an iPhone as they did with their own phones, and made lots more typos.

Issues

- user sample
- statistical significance
- "newbie" effect / learning effects

Source: PC World

If you read a bit more carefully into the study, you'll notice that the study is about initial adoption of the iPhone keyboard compared to users' current phones. Also, it isn't a survey, it was a study with one on one interviews where users typed and were timed.

The multitap (Non-QWERTY) users did the same or better with the iPhone than their current method, which suggests that multitappers may have an easier time adopting the iPhone's keyboard than QWERTY users. Which to me is interesting.

The study does not at any time attempt to say that QWERTY users will be twice as slow on the iPhone for as long as they use the iPhone, but it does say they may have more difficulty than multitap users initially. Which to me is interesting.

It would be interesting to see ia study some expert iPhone texters and have them switch to a QWERTY phone to see if there is a similar difference in typing efficiency.

Why do User Testing?

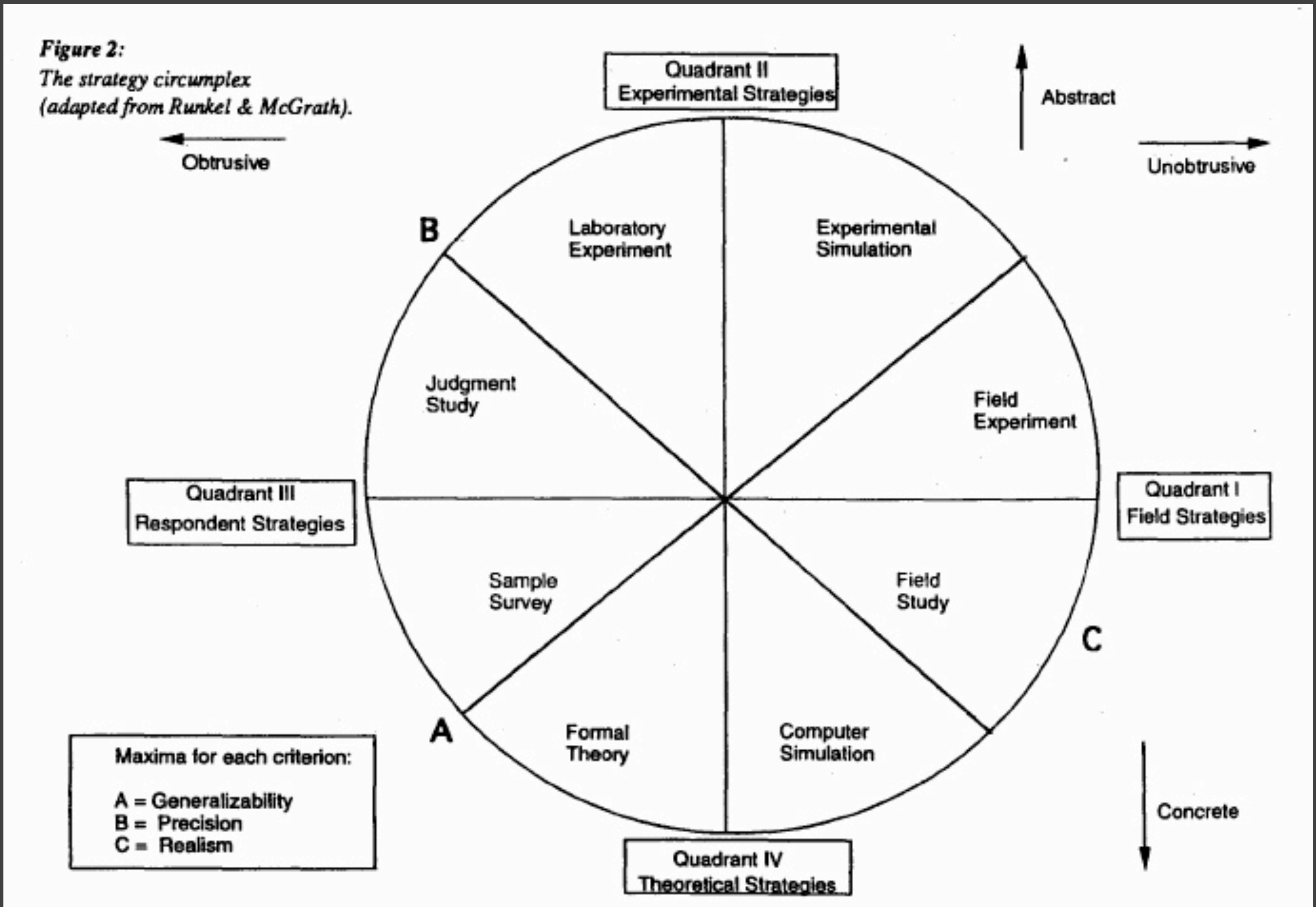
- Can’t tell how good UI is until?
  - people use it!
- Other methods are based on evaluators who
  - may know too much
  - may not know enough (about tasks, etc.)
- Hard to predict what real users will do

Different claims, different methods

- This idea/system/method
- is innovative
  - analysis of prior work/competitors
  - design alternatives & rationale
- may solve a known problem
  - analysis of the problem, its context
  - formative technique, e.g., concept validation, case study, or (gulp) think-aloud usability study
- is better than another idea/system/method
  - summative empirical or analytic technique, e.g., controlled lab experiment or quasi-experimental field study

If you don’t like the method, don’t make the claim

Taxonomy of Methods



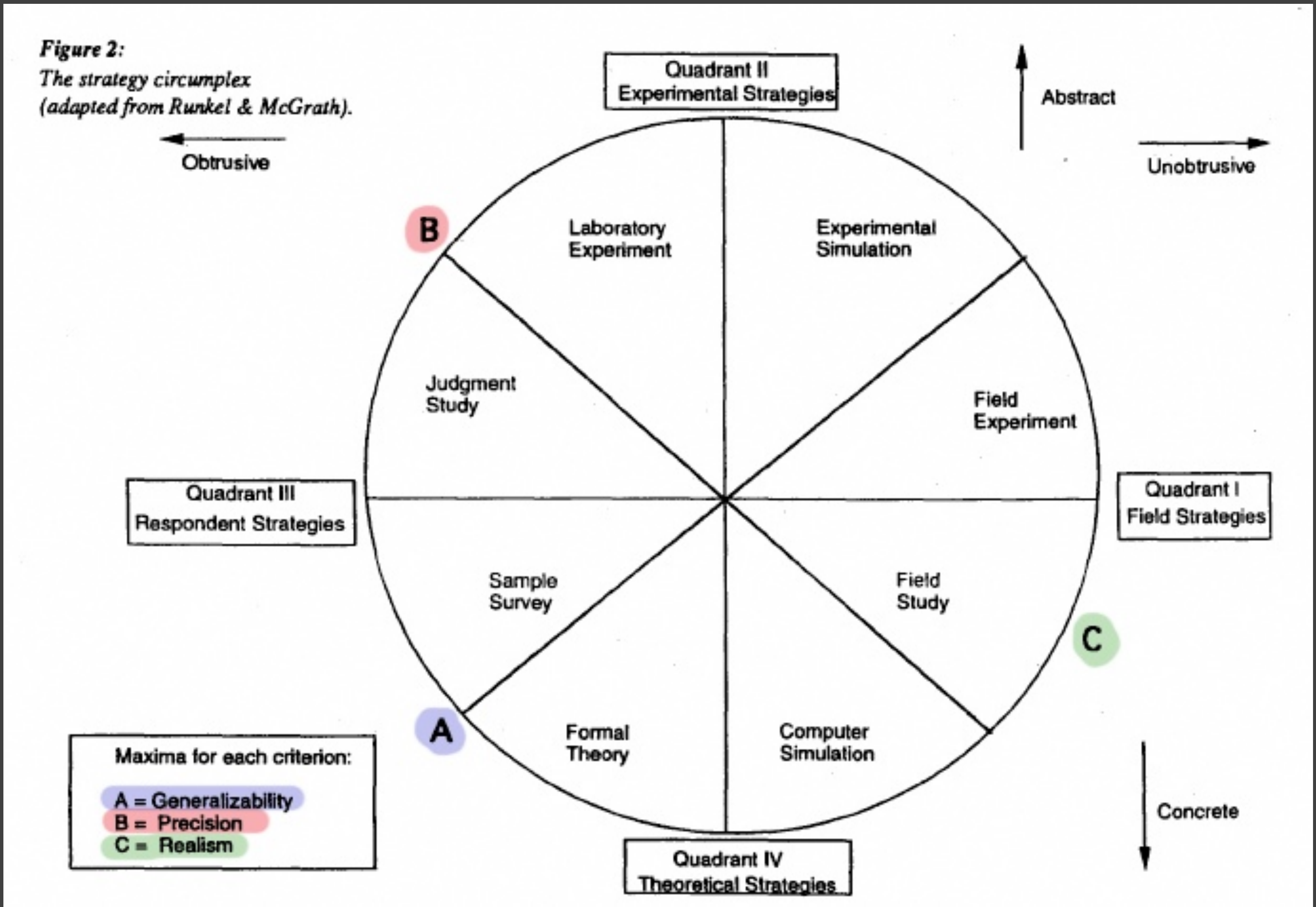
McGrath et al. 1994

Interview transcripts – kind of like a judgment study? Yeah. The stimulus is the software, and the conditions under which you are measuring the interactions are controlled for (i.e. the interviewer is the same, etc.) But it is a pretty non-experimental version...

Abstract->concrete: concrete meaning that nothing is glossed over? Abstract meaning that things have been simplified?

Quadrant II: experimental situation is concocted, doesn't already exist (vs. quadrant 1 where it is natural)

Taxonomy of Methods



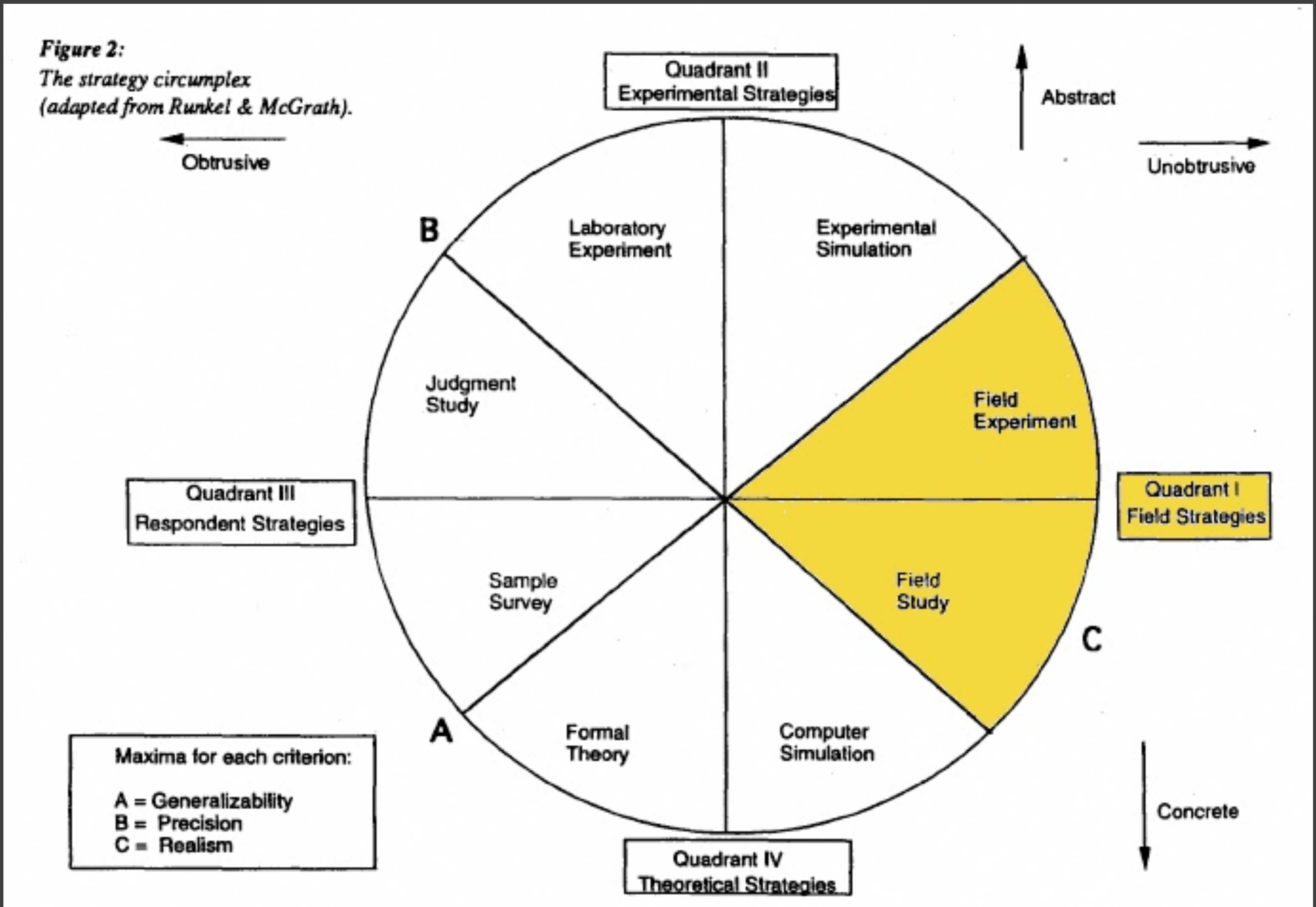
McGrath et al. 1994

Interview transcripts – kind of like a judgment study? Yeah. The stimulus is the software, and the conditions under which you are measuring the interactions are controlled for (i.e. the interviewer is the same, etc.) But it is a pretty non-experimental version...

Abstract->concrete: concrete meaning that nothing is glossed over? Abstract meaning that things have been simplified?

Quadrant II: experimental situation is concocted, doesn't already exist (vs. quadrant 1 where it is natural)

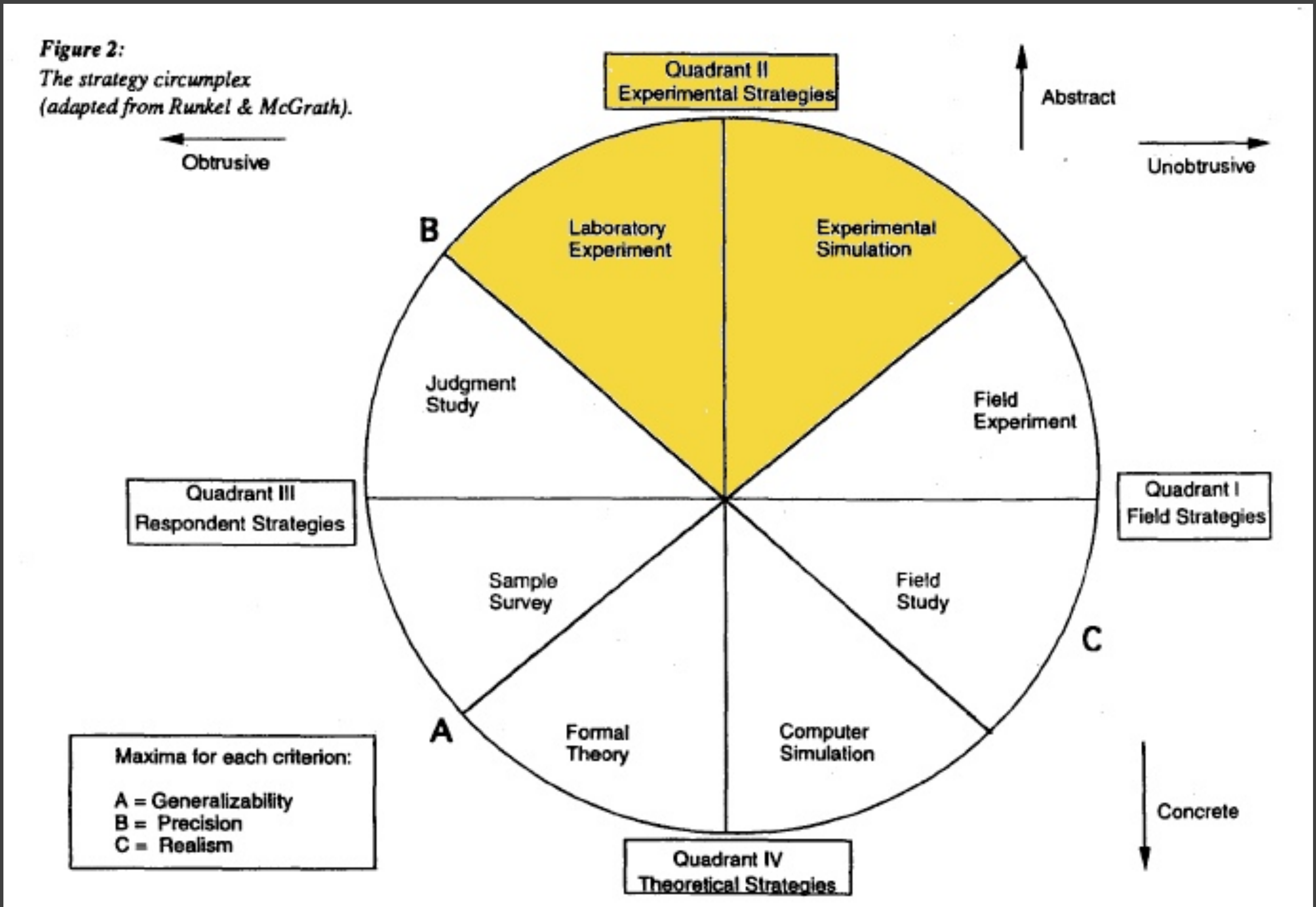
Taxonomy of Methods



McGrath et al. 1994

Count = number of “insights” –coded for by analyzing videotapes and coding them (using experts)
Value = total value of all insights, added together.

Taxonomy of Methods



McGrath et al. 1994

Count = number of “insights” –coded for by analyzing videotapes and coding them (using experts)
Value = total value of all insights, added together.

Taxonomy of Methods

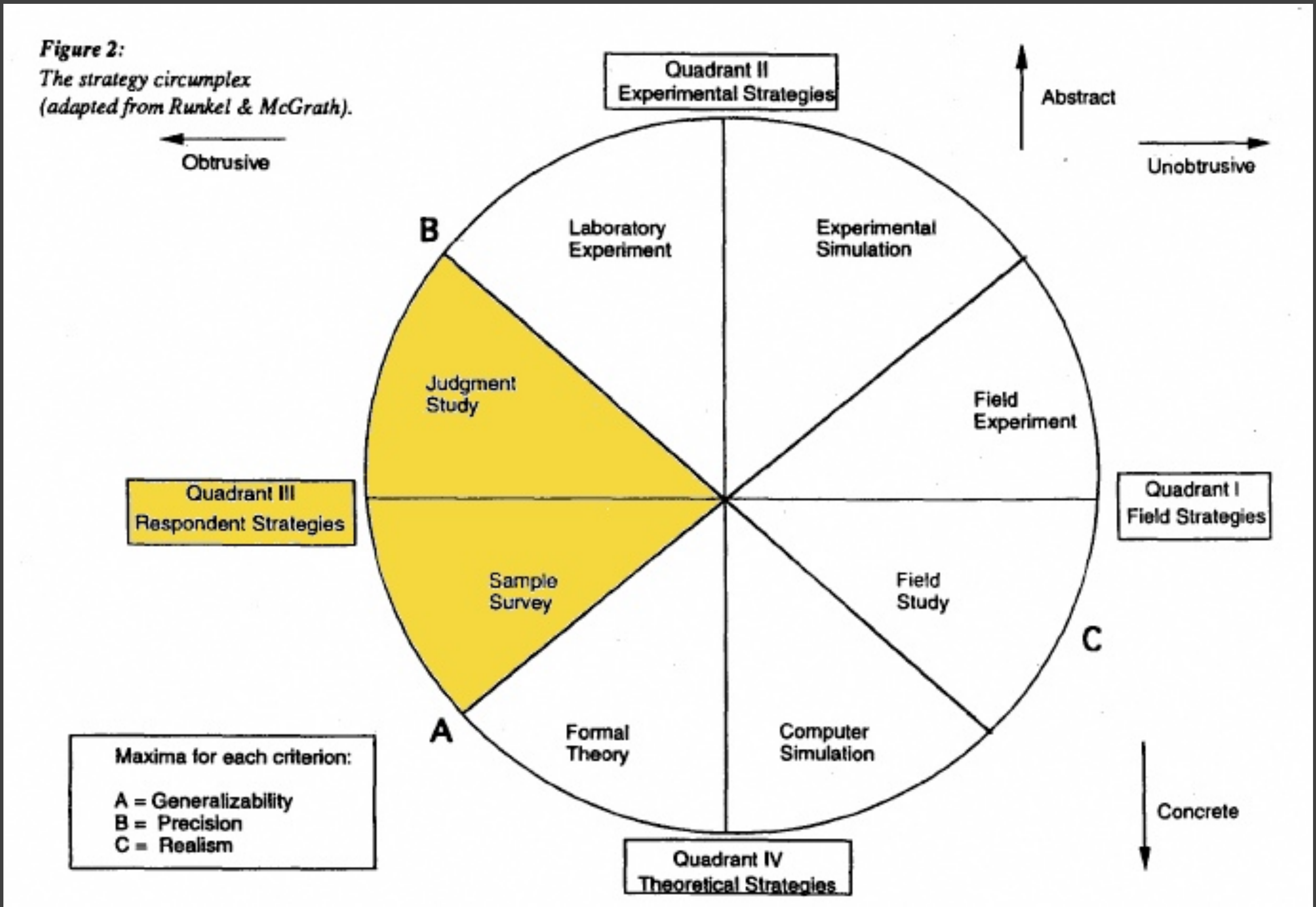
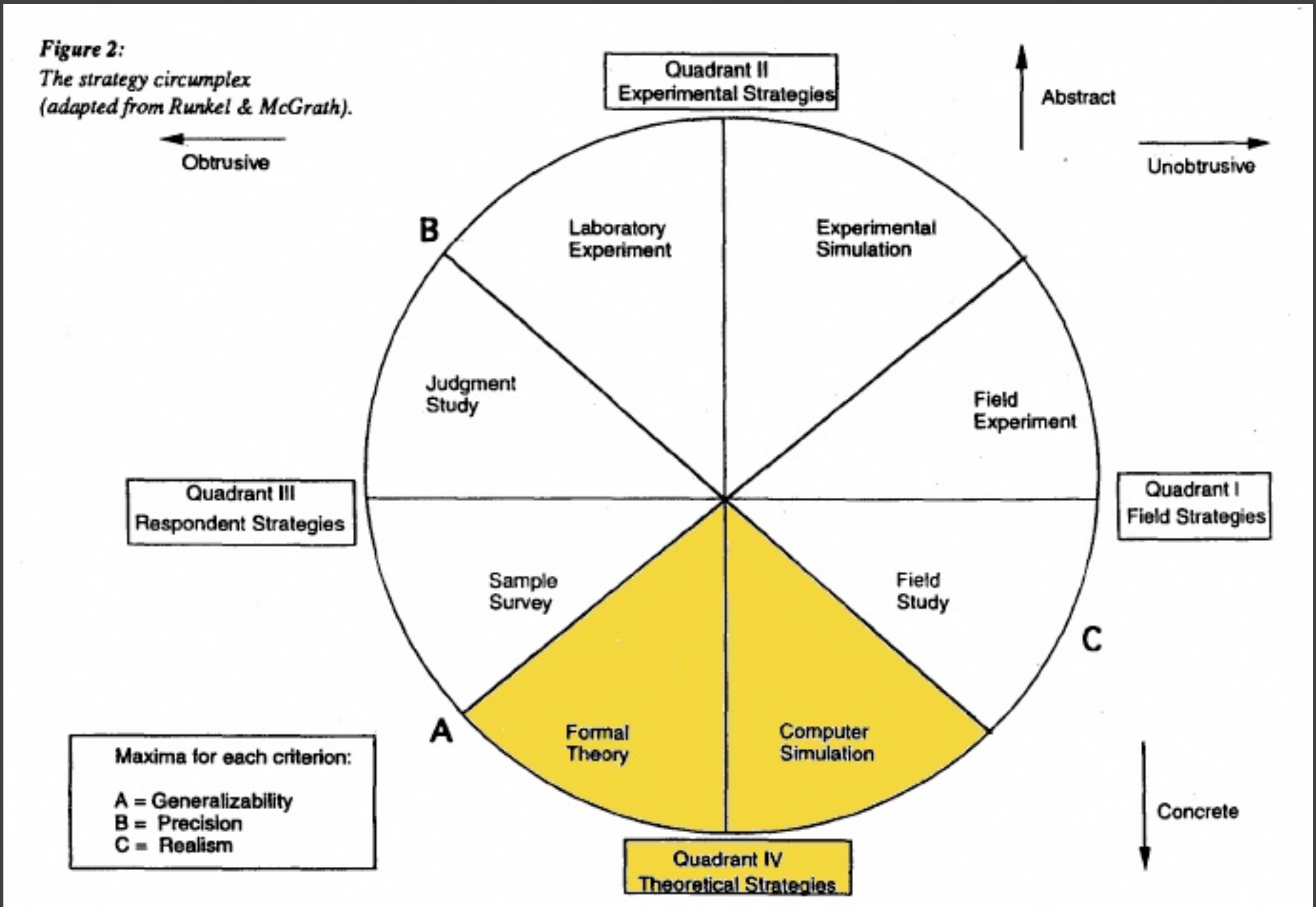


Figure 2:
The strategy circumplex
(adapted from Runkel & McGrath).

McGrath et al. 1994

Count = number of “insights” –coded for by analyzing videotapes and coding them (using experts)
Value = total value of all insights, added together.

Taxonomy of Methods



McGrath et al. 1994

Count = number of “insights” –coded for by analyzing videotapes and coding them (using experts)
Value = total value of all insights, added together.

Empirical Questions

- Baserates: How often does Y occur?
  - Requires measuring Y.

- Correlations: Do X and Y co-vary?
  - Requires measuring X and Y.

- Causes: Does X cause Y?
  - Requires measuring X and Y, and manipulating X.
  - Also requires somehow accounting for the effects of other independent variables (confounds)!

What will you measure?

- Time on Task -- How long does it take people to complete basic tasks? (For example, find something to buy, create a new account, and order the item.)
- Accuracy -- How many mistakes did people make? (And were they fatal or recoverable with the right information?)
- Recall -- How much does the person remember afterwards or after periods of non-use?
- Emotional Response -- How does the person feel about the tasks completed? (Confident? Stressed? Would the user recommend this system to a friend?)

Source: Wikipedia

Two kinds of variables

- Response variables (a.k.a. *dependent* variable(s))
  - Outcomes of experiment
- Factors (a.k.a. *independent* variables))
  - Variables we manipulate in each condition

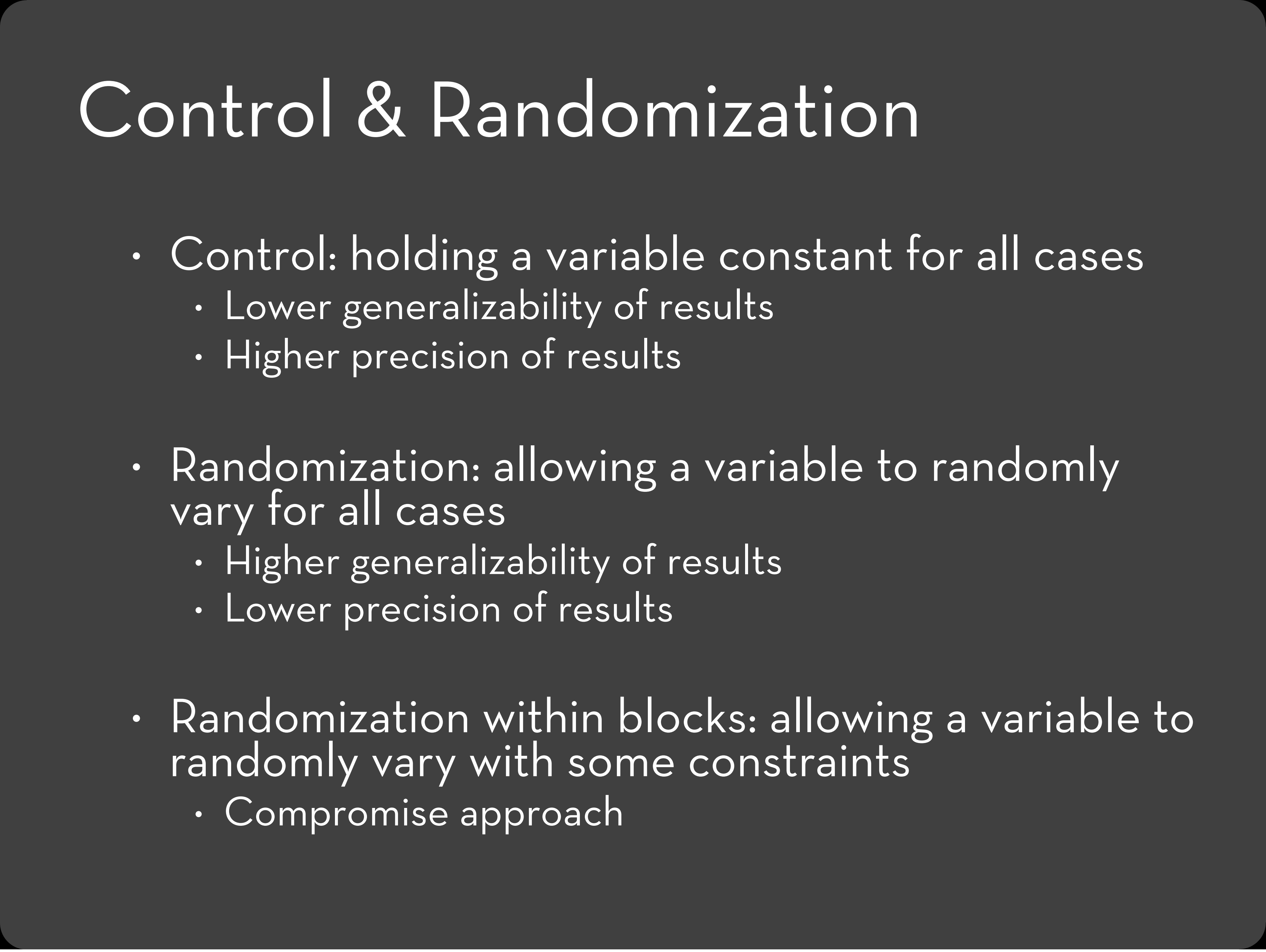
Goals

- Internal validity
  - Manipulation of independent variable is cause of change in dependent variable
    - Requires removing effects of confounding factors
    - Requires choosing a large enough sample size, so the result couldn't have happened by chance alone.
- External validity
  - Results generalize to real world situations
  - Requires that the experiment be replicable
  - No study "has" external validity by itself!

Confounding variables are those that change with the independent variable and could be cause of effect ☺

There's a tradeoff between internal validity and external validity – the

Example of trade-off – my phd work – strongly controlled experiments (e.g., no rotation) – how does this relate to what people do in the real world



Control & Randomization

- Control: holding a variable constant for all cases
  - Lower generalizability of results
  - Higher precision of results
- Randomization: allowing a variable to randomly vary for all cases
  - Higher generalizability of results
  - Lower precision of results
- Randomization within blocks: allowing a variable to randomly vary with some constraints
  - Compromise approach

Confounding variables are those that change with the independent variable and could be cause of effect ☺

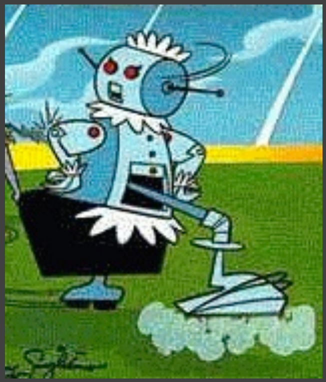
There's a tradeoff between internal validity and external validity – the

Example of trade-off – my phd work – strongly controlled experiments (e.g., no rotation) – how does this relate to what people do in the real world

Between subjects design

- Wilma and Betty use one interface

Dino and Fred use the other

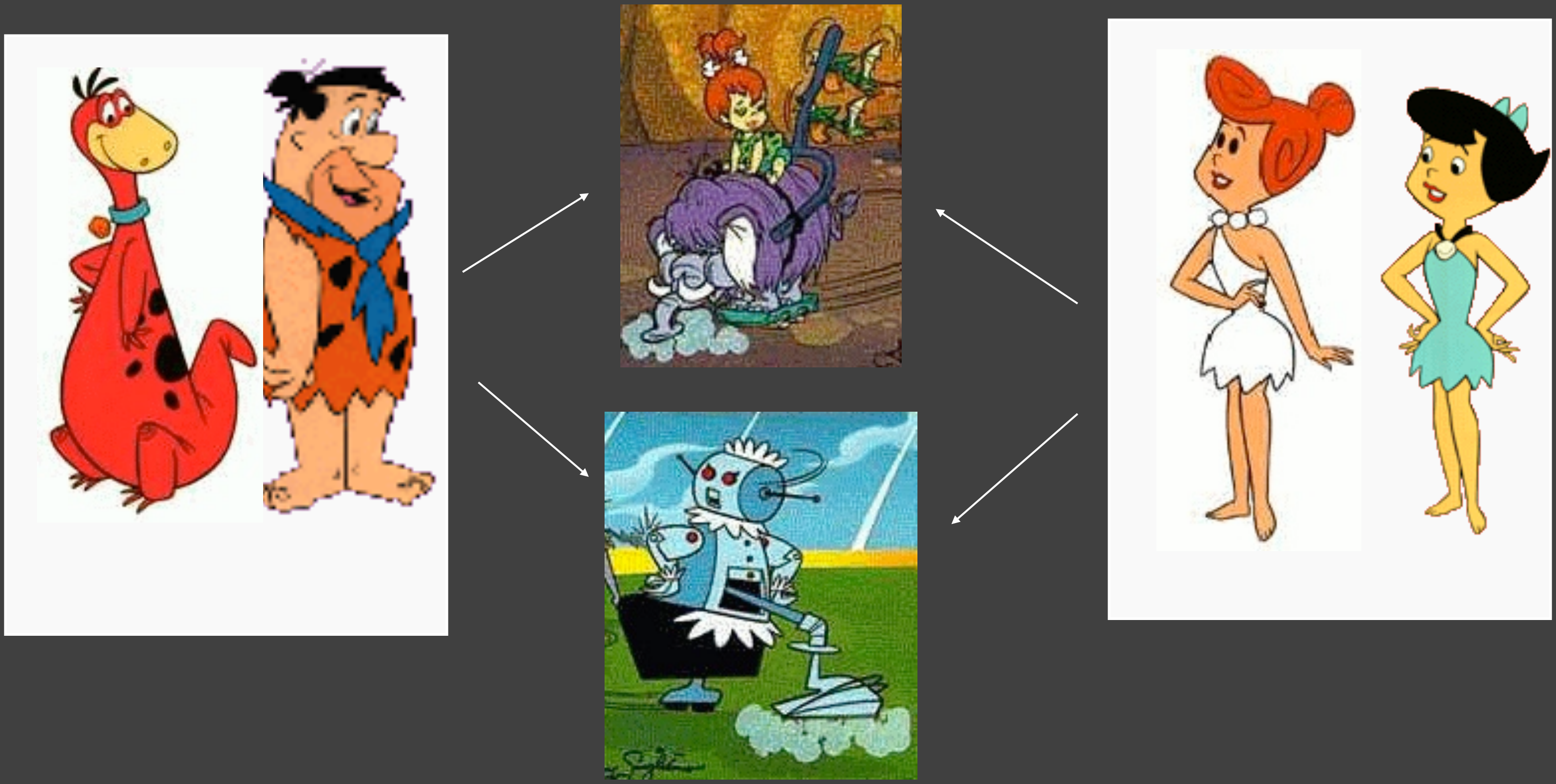


Between subjects design

Subjects wearing 7-mm spikes	Time (in seconds)	Subjects wearing 13–mm spikes	Time (in seconds)
Mike	11.7	Don	15.7
Bob	18.2	Hector	13.4
Homer	12.2	Ron	18.0
George	15.4	Tom	12.8
Harry	15.8	Steve	13.6
Gordon	13.2	Dale	19.0
John	13.7	Pete	16.2
Bill	19.1	Juan	11.9
Randy	12.9	Dan	14.6
Tim	16.0	Paul	18.0

Within subjects design

- Everyone uses both interfaces



Within subjects design

Subjects on manual typewriter	Typing speed (wpm)	Subjects on electric typewriter	Typing speed (wpm)
Mike	35	Mike	40
Bob	42	Bob	45
Homer	32	Homer	35
George	25	George	30
Harry	30	Harry	32
Gordon	30	Gordon	35
John	30	John	40
Bill	36	Bill	37
Randy	36	Randy	42
Tim	30	Tim	34

Ordering effects

- Ordering of conditions is a variable that can confound the results
  - Randomization
  - Counterbalancing
  - Latin square (partial counterbalancing)

Between vs. within subjects

- Within subjects
    - All participants try all conditions
        - + Can isolate effect of individual differences
        - + Requires fewer participants
        - - Ordering and fatigue effects

- Between subjects
    - Each participant tries one condition
        - + No ordering effects, less fatigue.
        - - Cannot isolate effects due to individual differences.
        - - Need more participants

Choosing Participants

- Representative of target users
  - job-specific vocab / knowledge
  - tasks
- Approximate if needed
  - system intended for doctors
    - get medical students
  - system intended for engineers
    - get engineering students
- Use incentives to get participants

What should you keep in mind?

- You are testing the site not the users.
- Rely more on what you learn about performance than preference.
- Make use of what you learn.
- Try to find the best solution given the reality of your many users.

- Follow University Human Subject Guidelines

Source: Usability.gov

Ethical Considerations

- Sometimes tests can be distressing
  - users have left in tears
- You have a responsibility to alleviate
  - make voluntary with informed consent
  - avoid pressure to participate
  - let them know they can stop at any time
  - stress that you are testing the system, not them
  - make collected data as anonymous as possible