

I USED TO THINK
CORRELATION IMPLIED
CAUSATION.



THEN I TOOK A
STATISTICS CLASS.
NOW I DON'T.



SOUNDS LIKE THE
CLASS HELPED.

WELL, MAYBE.



Evaluating Designs

SCOTT KLEMMER

FALL 2011

cs147.stanford.edu

What should you do...
If one of your teammates
isn't pulling their weight

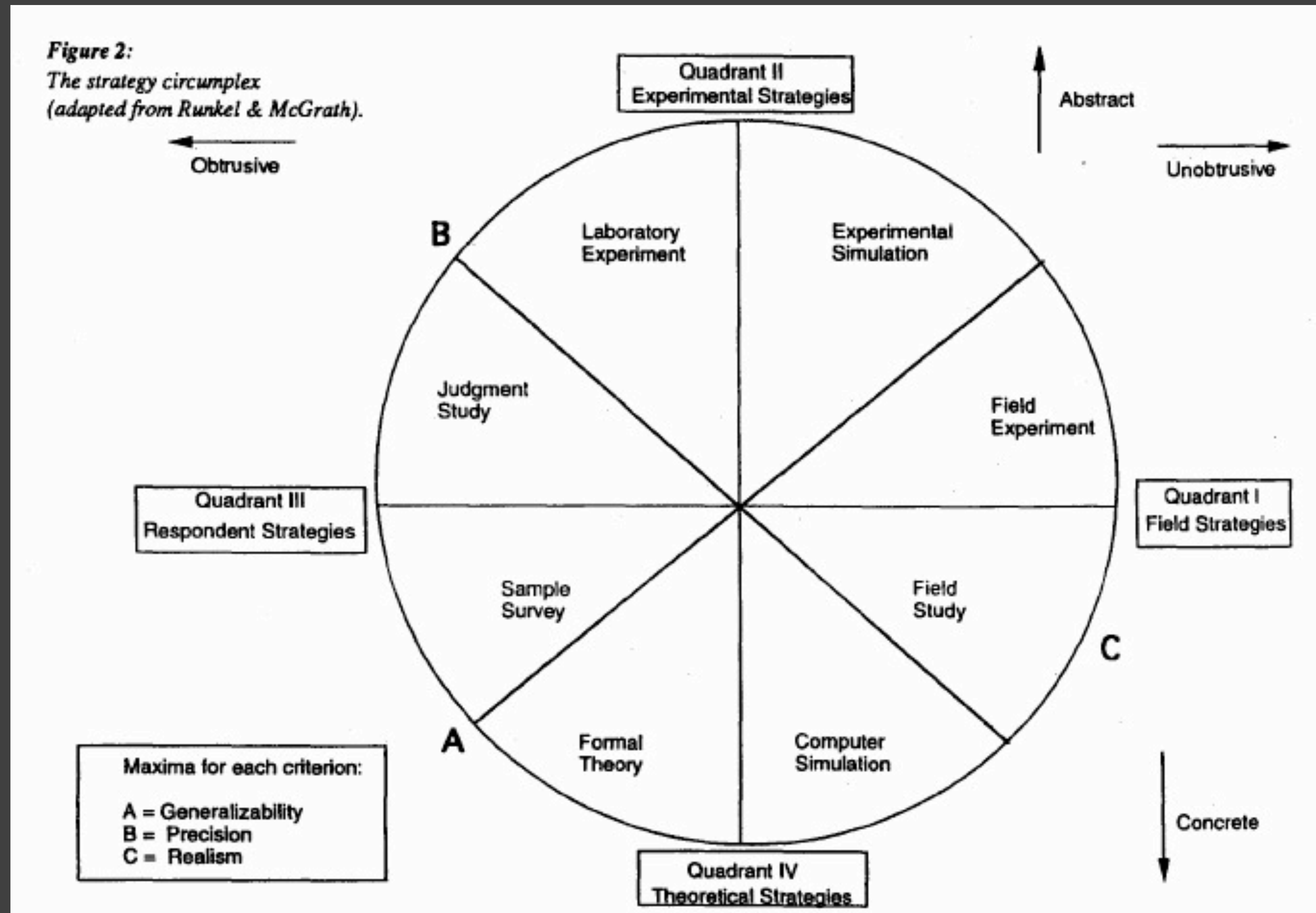
How can we measure success?

How do we know?

Why Evaluate Designs with People?

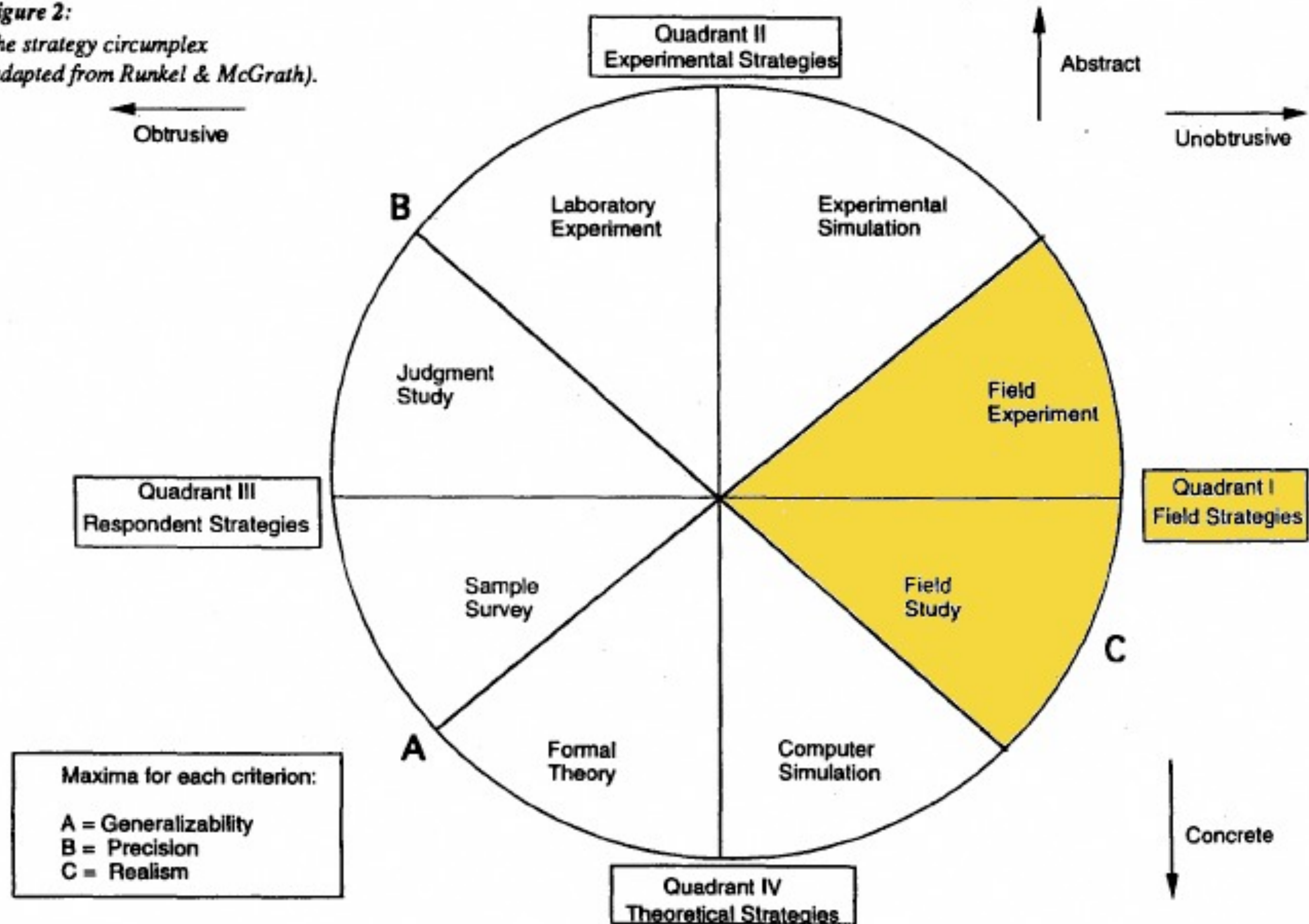
What are some claims you might make about an interface? How would you evaluate them?

Different Methods Achieve Different Goals



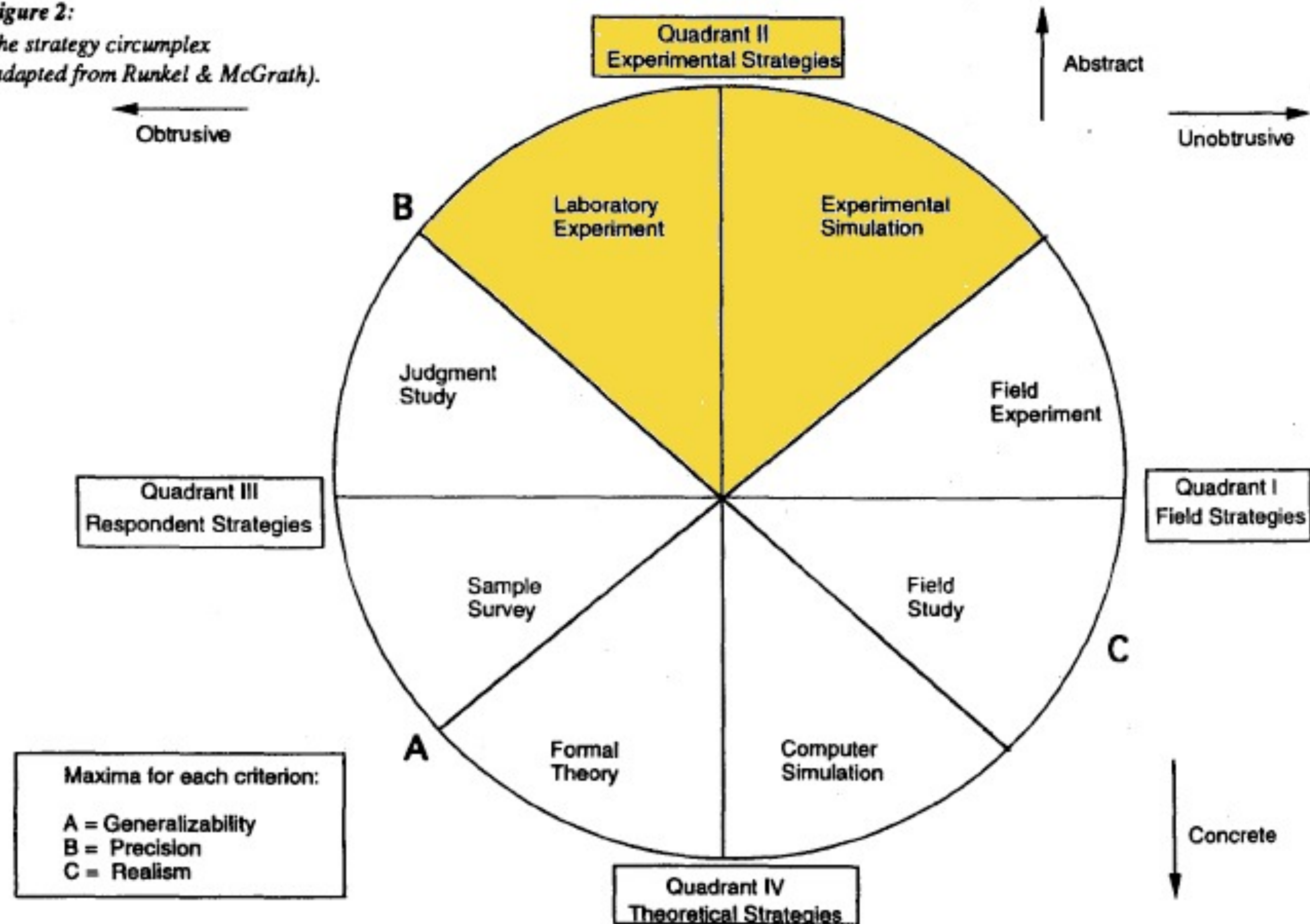
Taxonomy of Methods

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



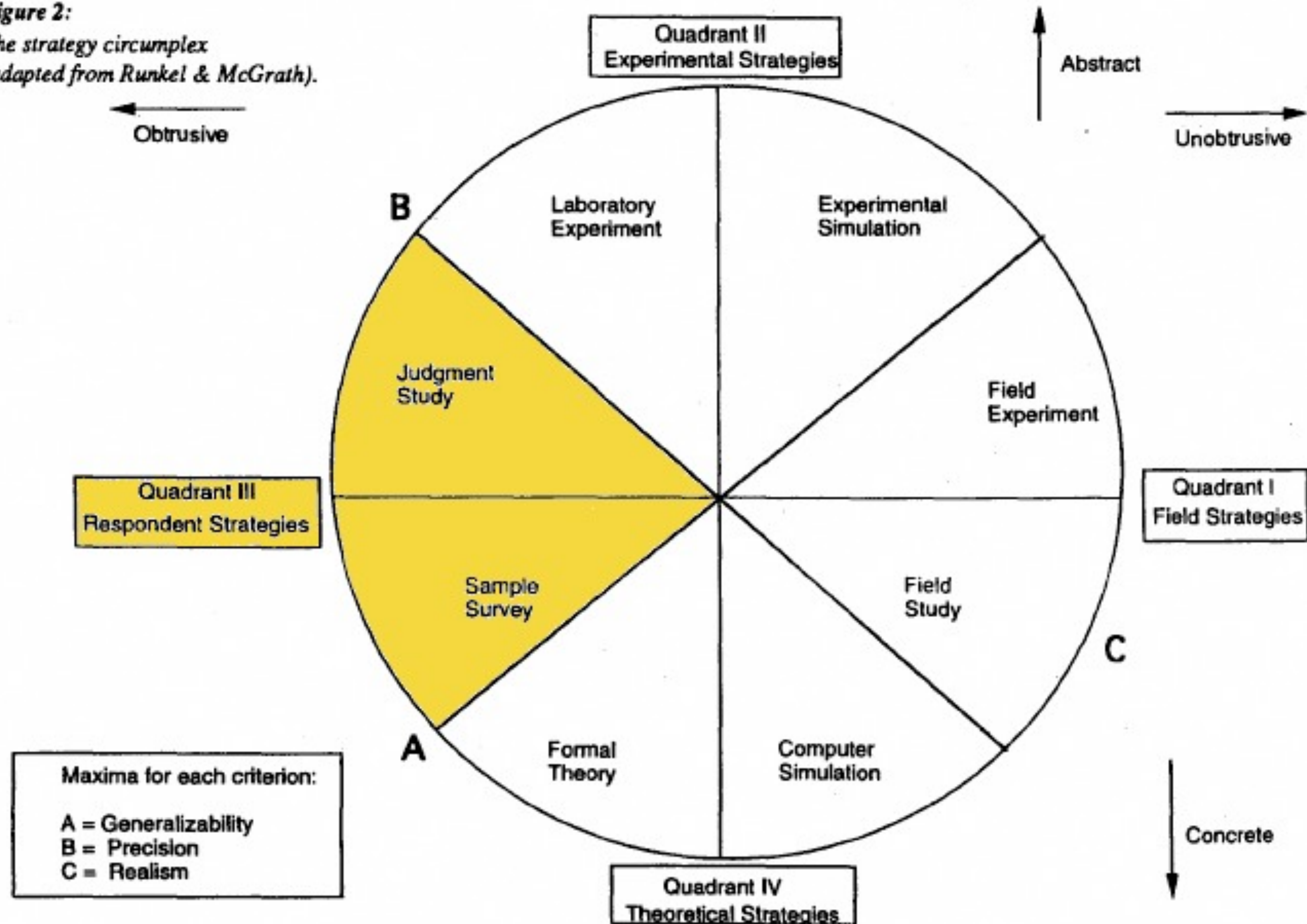
Taxonomy of Methods

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



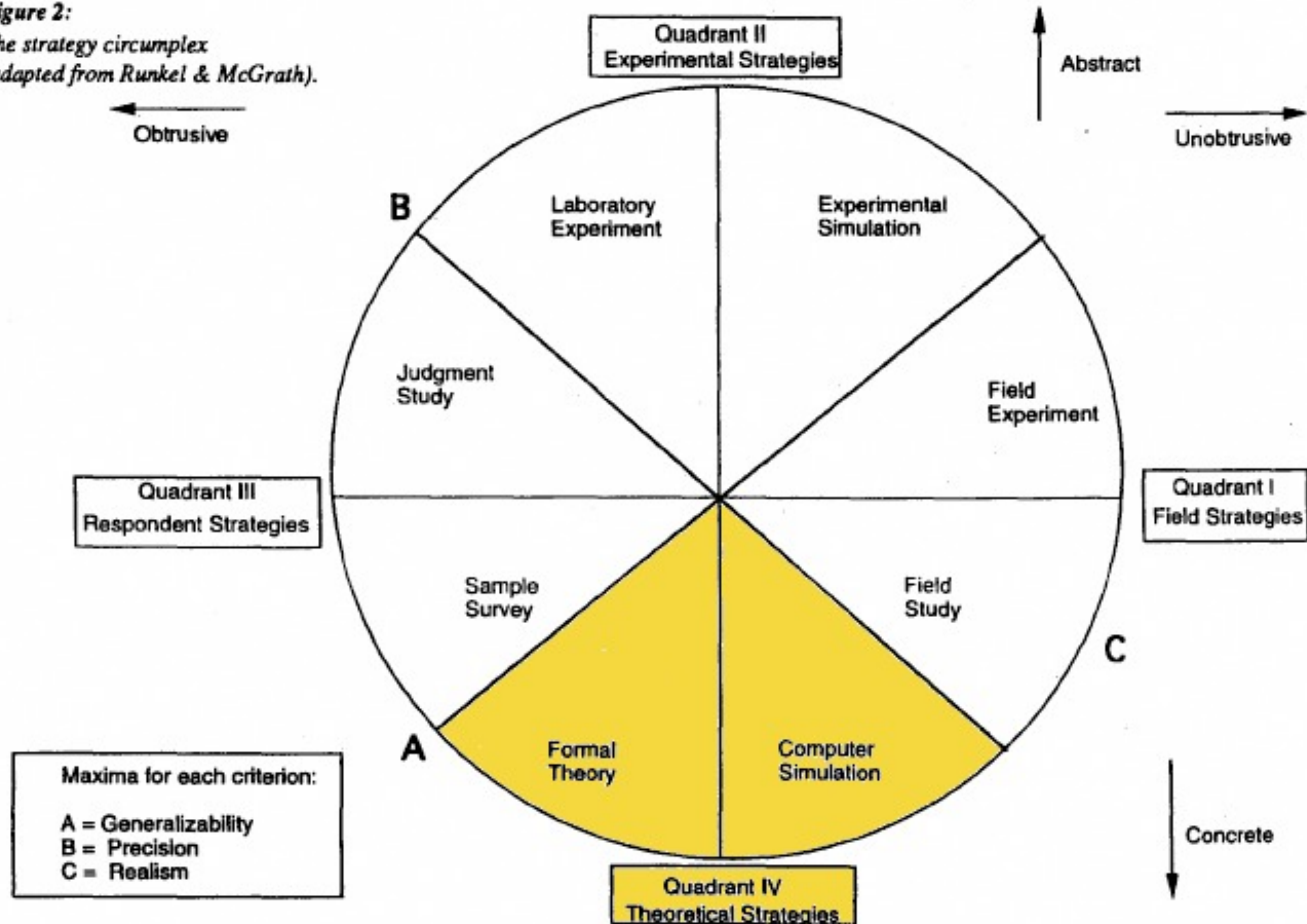
Taxonomy of Methods

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



Taxonomy of Methods

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



MEASURES

Dependent Variables

MANIPULATIONS

Independent Variables

PRECISION

Internal Validity

GENERALIZABILITY

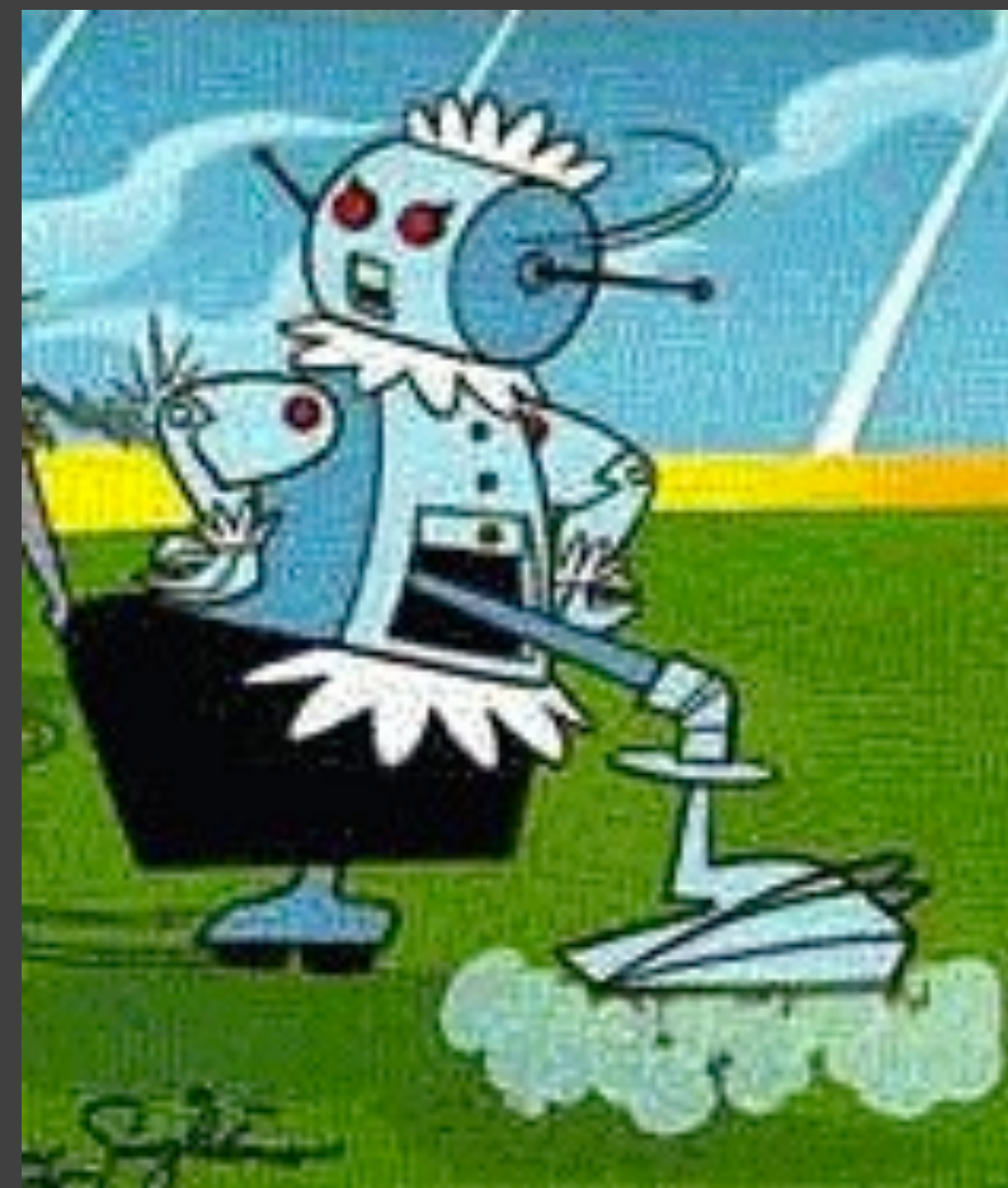
External Validity

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

Should every participant
use every alternative?

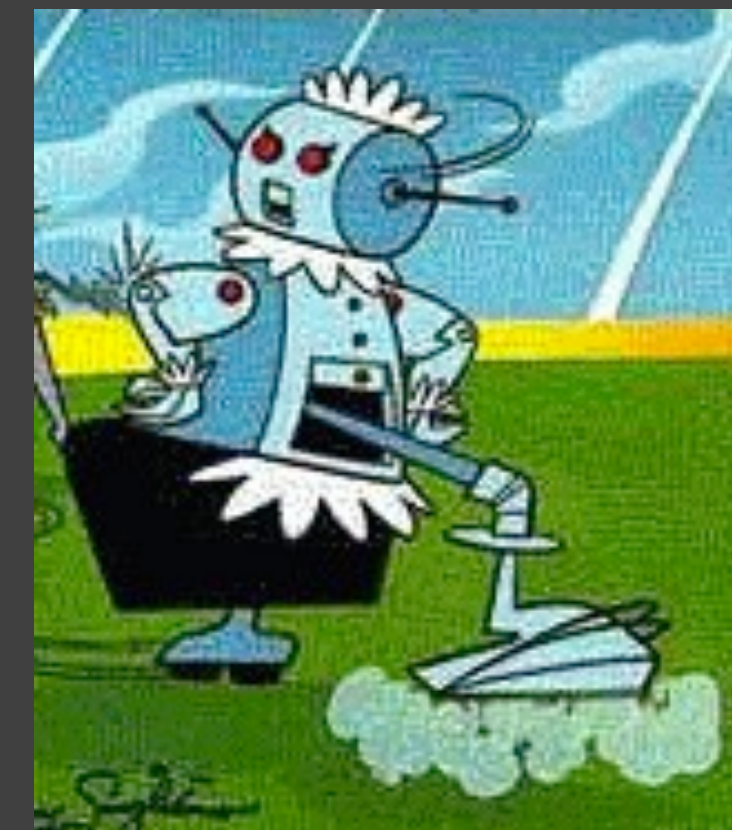
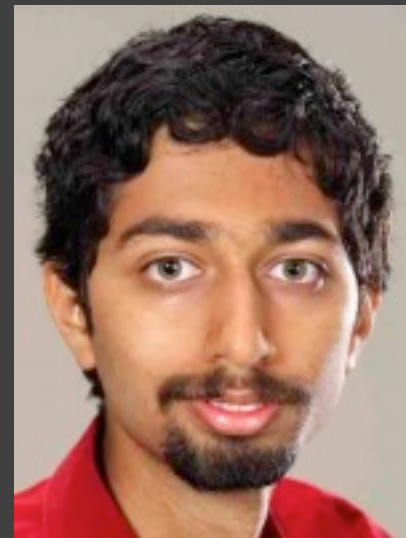
Which vacuum cleaner is more effective?



Between subjects design

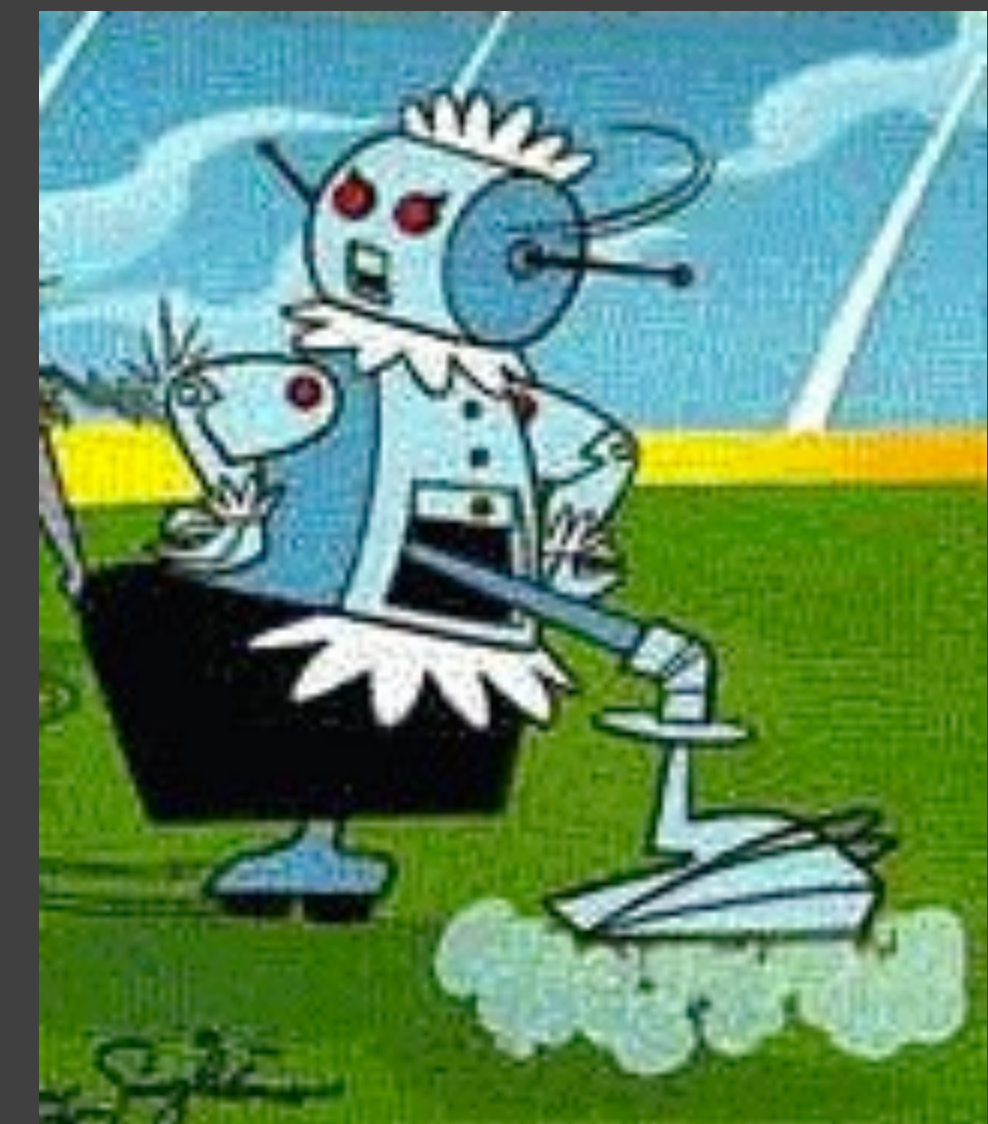
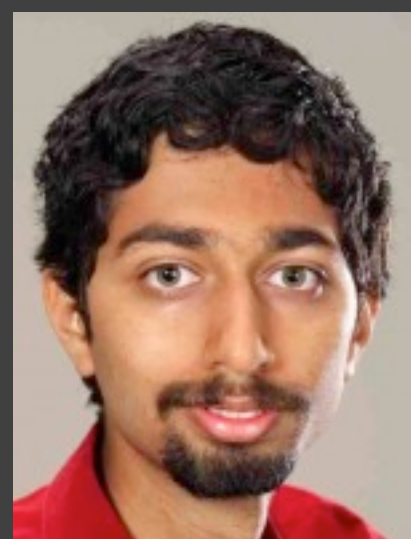
- Half the participants use one interface

The other half use the other



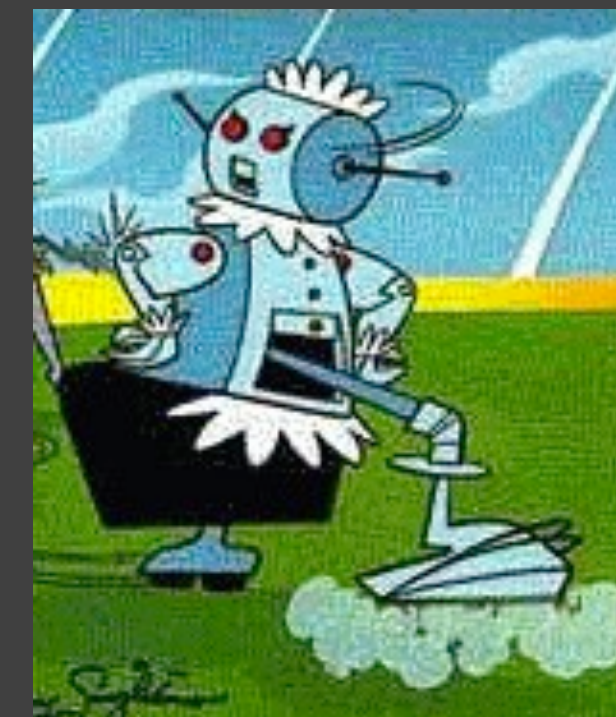
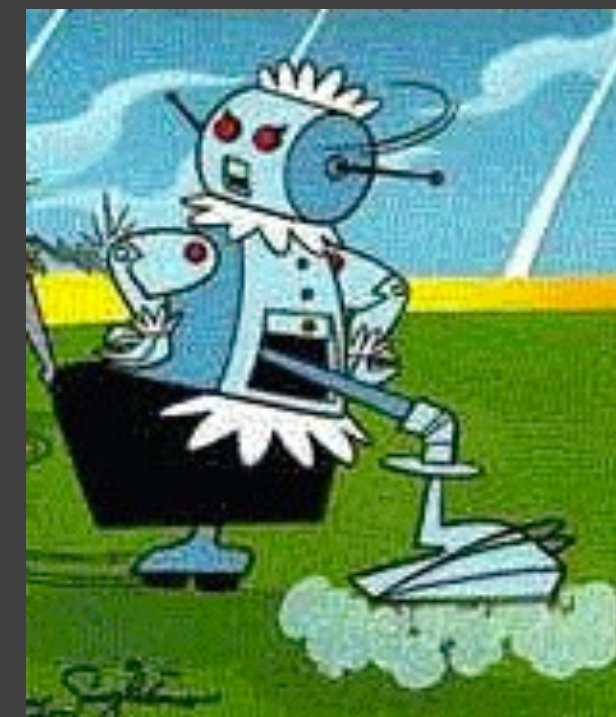
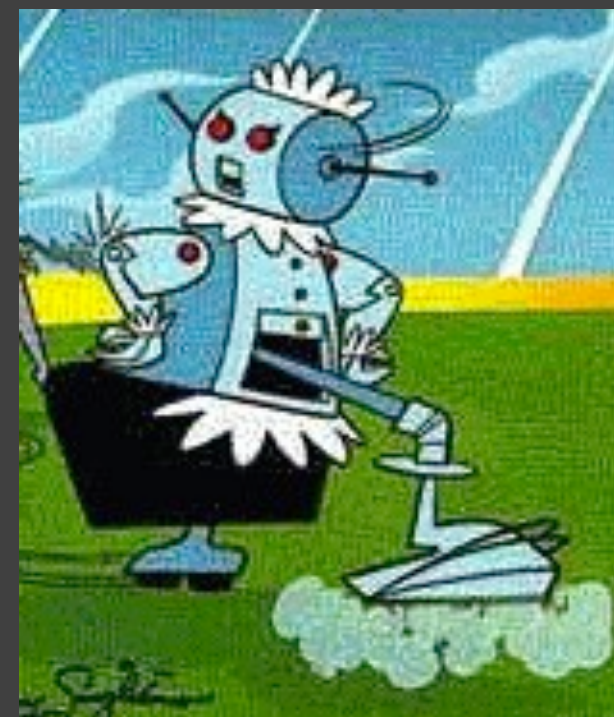
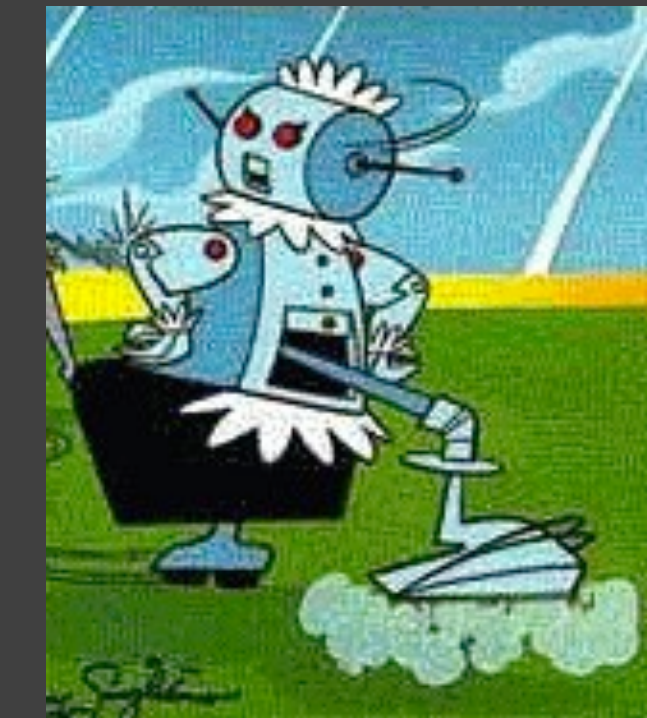
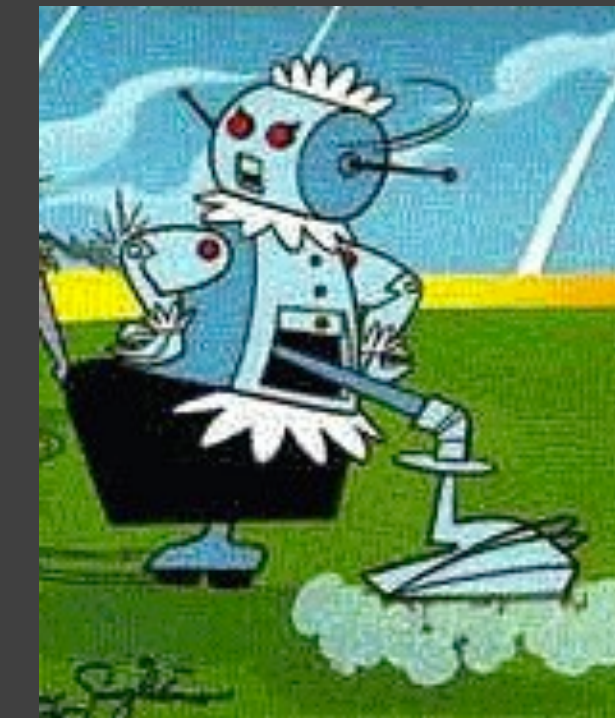
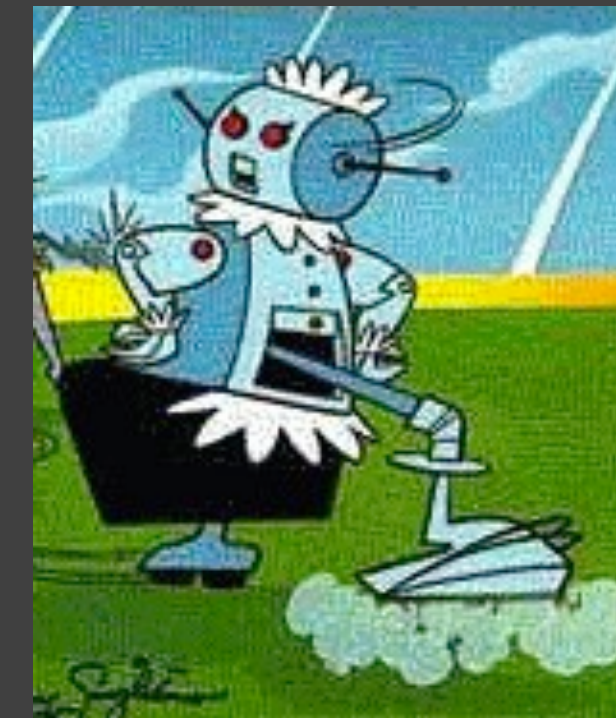
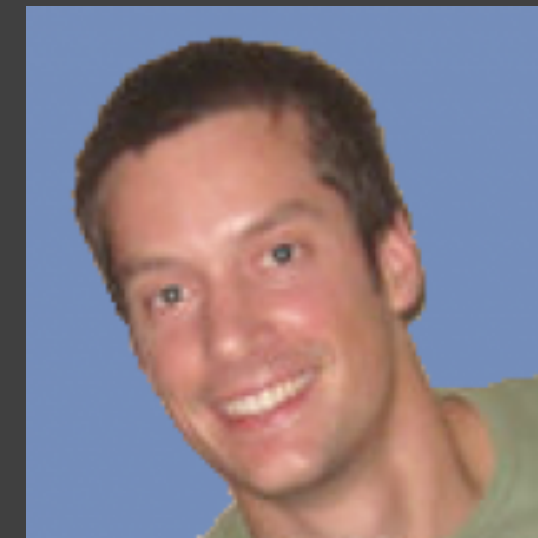
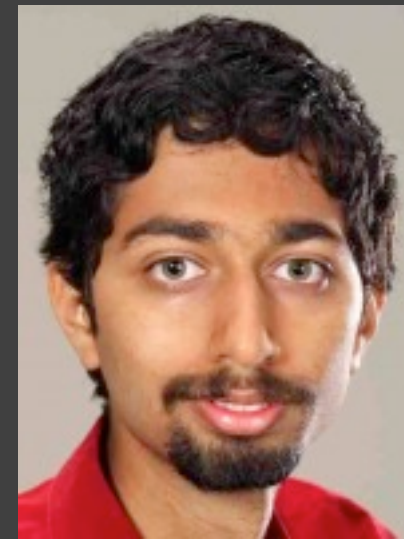
Within subjects design

- Everyone uses both interfaces



How Can We Address
Ordering Effects?

How Can We Address Ordering Effects?



How about individual
differences?

What about for Three
or More Alternatives?

Latin Square

1	2	3
2	3	1
3	1	2

Should every participant
use every alternative?

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

The importance of random assignment

Counterbalanced Assignment

- Say you worry that running speed will affect input speed (that fast runners are physically fast people, and they'll do better)
- You can assign participants so that running speed is balanced across conditions
- Do so with care: form matched pairs using a pre-test (work this out on the board)

A danger: regression

- Let's find heady coins
- First, let's flip all the coins (our pre-test)
- If they land heads more than half, we'll call them heady
- Now let's feed them a snack
- After a snack, do heady coins beat taily coins?
- Similarly, some of our runners will have had a relatively good run, and others a relatively bad one. They'll exhibit regression to the mean if re-tested.
- If the pre-test is used to counterbalance, and assignment is random, then the error goes away

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



A real potential pitfall

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

A Not-Very-Useful iPhone Keyboard Study

ADD TO MY PAGES PRINT E-MAIL COMMENT RSS
SLASHDOT IT DIGG THIS DEL.ICIO.US NEWSVINE



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


The Hawthorne Effect







Choosing Participants

Selecting Tasks

A Not-Very-Useful iPhone Keyboard Study

 [ADD TO MY PAGES](#)  [PRINT](#)  [E-MAIL](#)  [COMMENT](#)  [RSS](#)

 [SLASHDOT IT](#)  [DIGG THIS](#)  [DELICIOUS.US](#)  [NEWSVINE](#)



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.

A better version: actual users

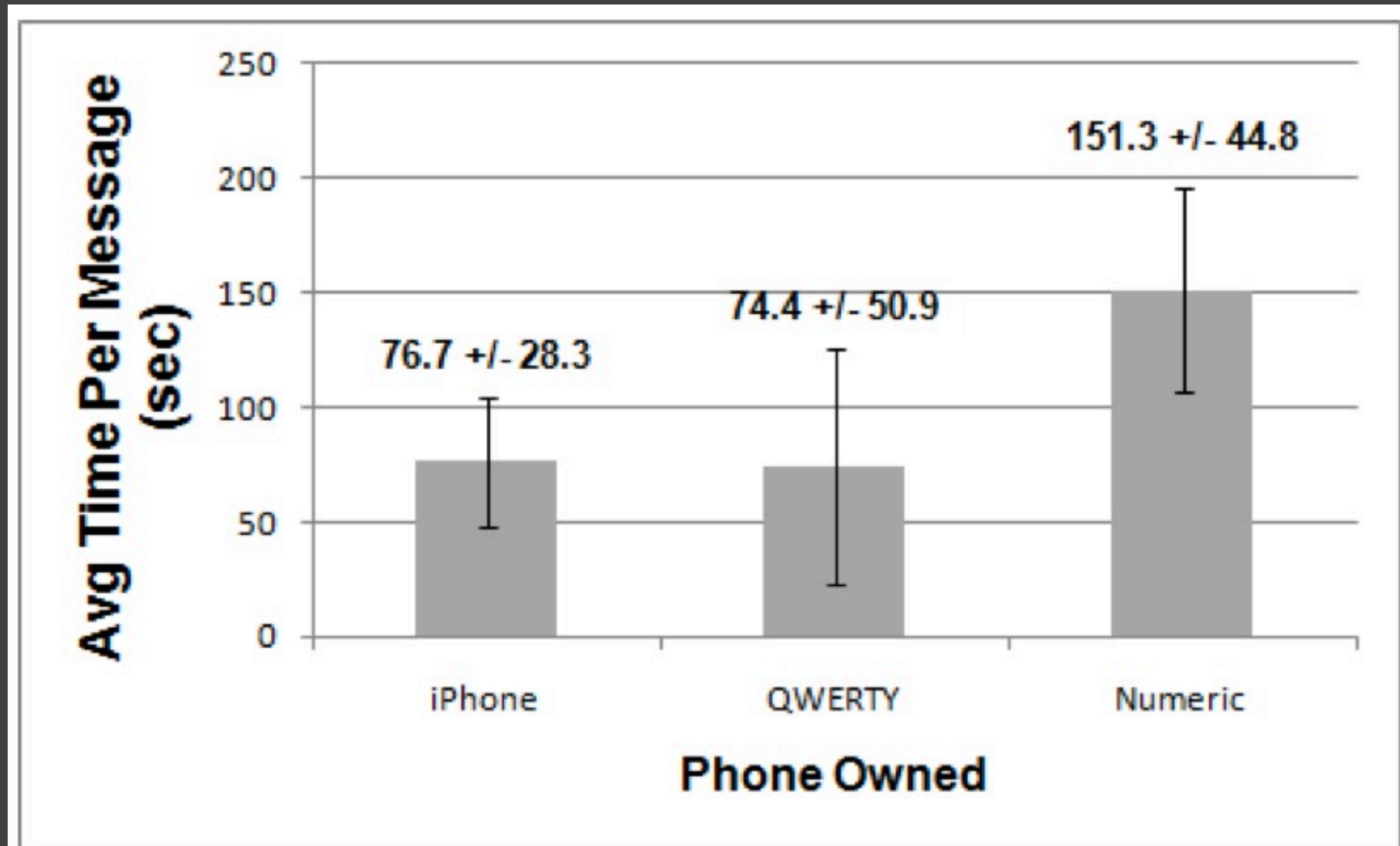


Figure 1. Average time to type a message on phones owned by the participants ($M \pm SD$).

iPhone users are almost as fast, but make more errors

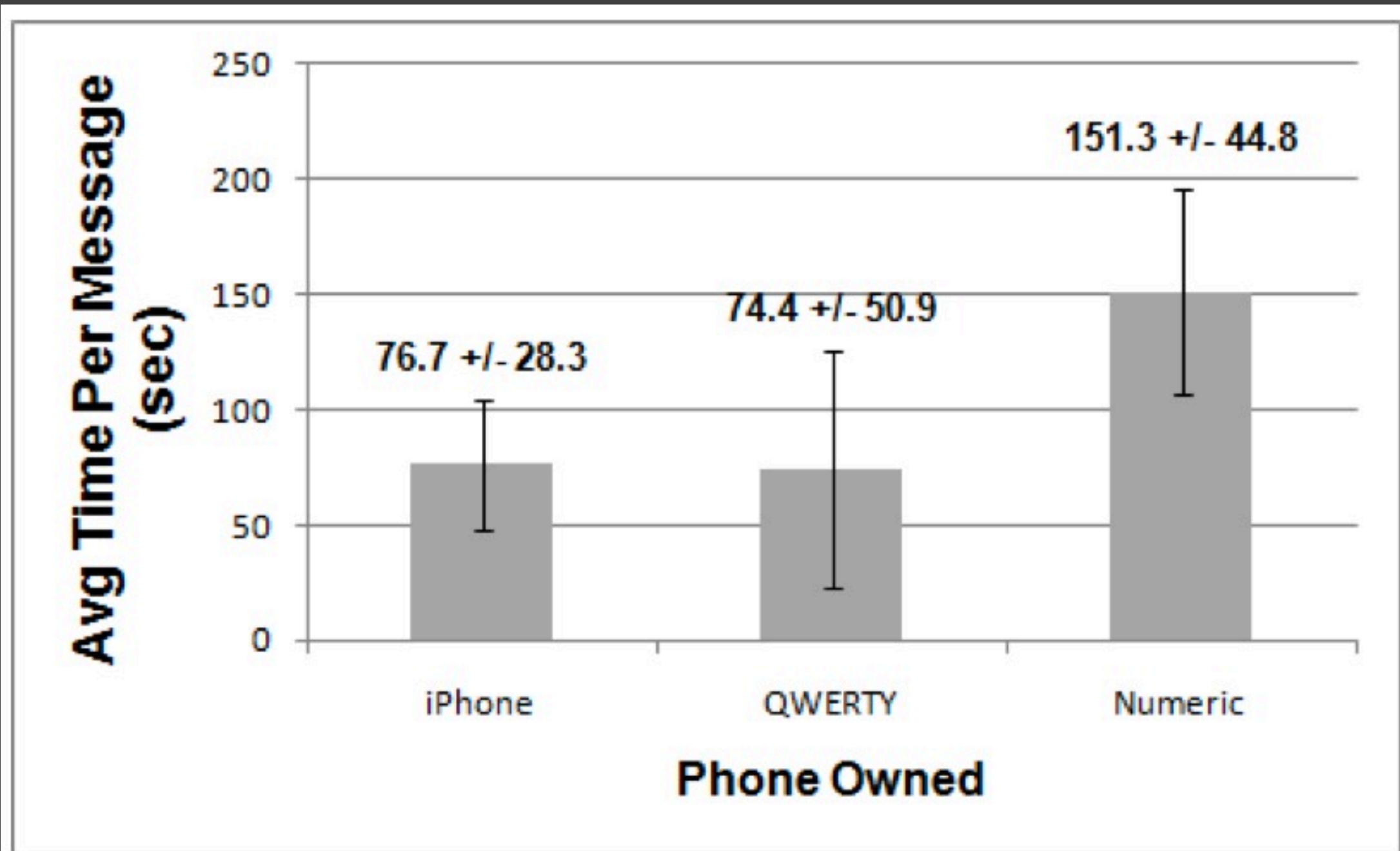


Figure 1. Average time to type a message on phones owned by the participants ($M \pm SD$).

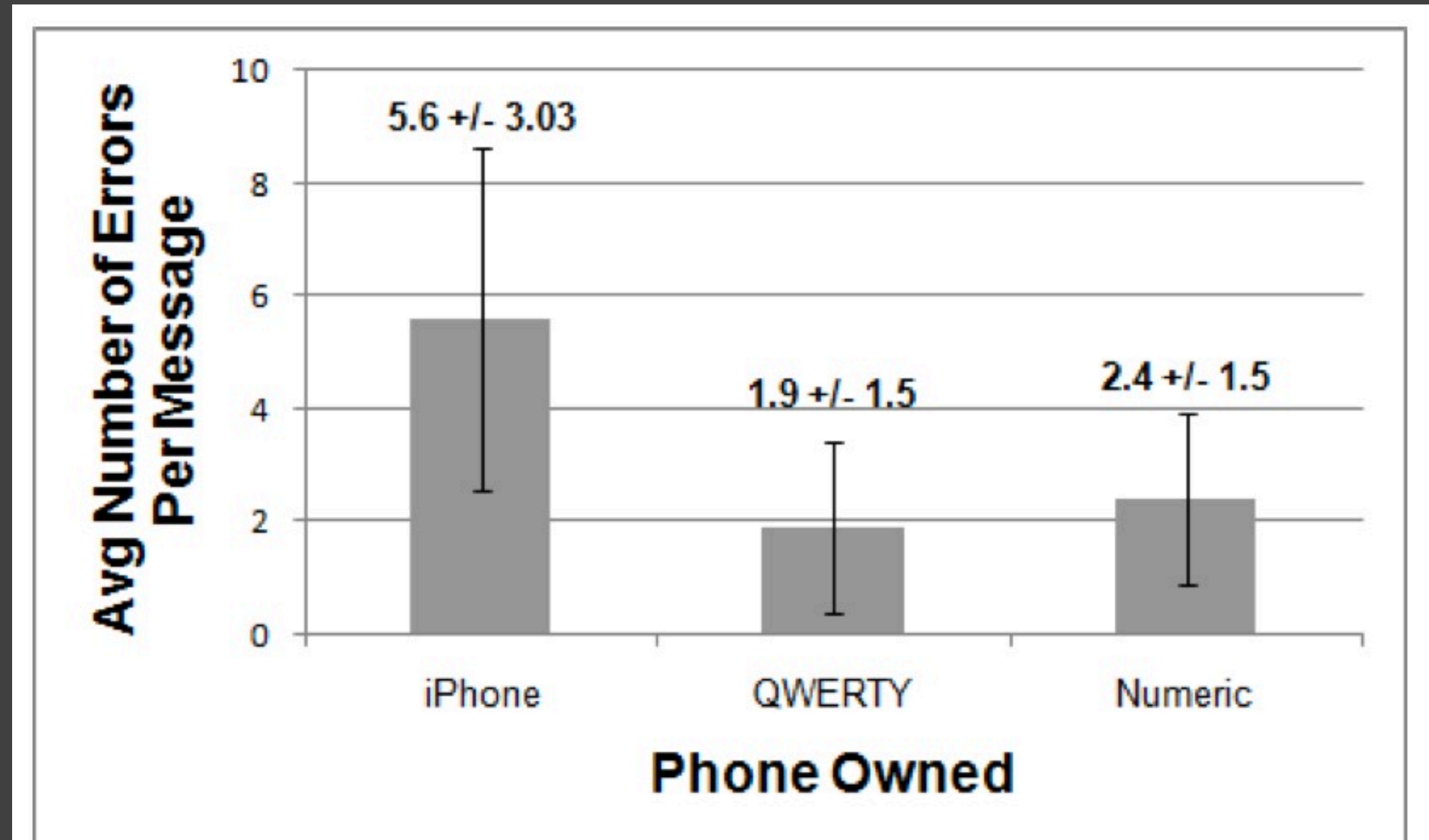


Figure 2. Average number of total errors per message made by participants using their own phones ($M \pm SD$).