Evaluation and CSCW

Margaret-Anne Storey

mstorey@uvic.ca

CSCW 2018

User testing

- Aim: improve products
- Few participants
- Results inform design
- Not perfectly replicable
- Controlled conditions
- Procedure planned
- Results reported to developers

User testing versus research

- User testing is not necessarily “research”
- User testing may be done to inform a product team or company about the software
- Research is expanding what we know in some way that can be **generalized** beyond a single product and small population
- In user testing, we may focus on usability rather than usefulness

What is research?

- A process of steps used to collect and analyze information in order to increase our **understanding** of a topic or issue
- It involves three steps:
  - Pose a question
  - Collect data to answer the question
  - Present an answer to the question

Why is research important in CSCW?

- Adds to our knowledge of what is good practice in CSCW and what is not good practice
  - May lead to improved tools or processes
- Different ways of adding to our knowledge:
  - ***Address gaps*** in knowledge
  - ***Replicate knowledge*** by testing old knowledge with new participants or new research sites
  - ***Expand knowledge*** by extending research to new ideas or practices
  - ***Broaden our perspectives*** – e.g. add voices of individuals to the body of knowledge
  - ***Inform practice*** by developing new ideas

Process of research

Identifying a research **problem**

Reviewing the **literature**

Specifying a **purpose** for the research

- Provide a purpose statement
- Then refine it to research **questions or predictions**

Collecting data

Analyzing and **interpreting** the data

Reporting and **evaluating** research

From the blog posts...

- Importance of asking the right and well formed research question (discussed in several blog posts, e.g., [gturney])

Skills required to conduct research

- Solving puzzles
- Attention span
- Using libraries
- Writing and editing

Research methods

- **Tools:** instruments, techniques and procedures used by a science to gather information
- Different methods tell us different things
- Each has its **advantages** but also its **limitations**
  - E.g. What about questionnaires?
- We can and should use **multiple methods**

Controlled experiments

- Permits **precise measurement** of the effects from manipulating some presumed causes
- But we have to greatly narrow the scope of the problem – **artificial setting** and conditions due to the methods used

Thinking critically about research

- Ask: is the study **valid**?
- Do the results, as presented, identify the **strengths** and **weaknesses** of the research strategies used?
- Are the results **consistent** not just with similar studies, but from studies using other strategies?
- Do the results **converge** across different kinds of studies?

Practical issues

- Difficult to have expertise and equipment/resources to conduct multiple strategies…

[lindenqu]

FROM THE BLOG POSTS...

"Although software development and social sciences may seem like wildly different topics, these two articles made me realize they are more similar than I expected."

McGrath paper: Methodology Matters….

Research process

Involves three sets of things:

- Some **content** that is of interest
- Some **ideas** that give meaning to that content and
- Some **techniques** or procedures for studying the content and ideas

More formally…

- **Substantive domain**: from where we draw contents to study
- **Conceptual domain**: from where we draw ideas that will give meaning to our results
- **Methodological domain**: from where we draw techniques that may be useful in the research

An aside: Methods and more methods…

- Overloaded term…
- Research method – also referred to as research approach, or mode of treatment
- Modes of treatment will involve some **measurement methods** which may also be referred to as techniques and also involve techniques for **manipulating some feature** in a research situation as well as techniques for **controlling the impact** of various extraneous features in the situation

Research strategy

- In CSCW, always involves: somebody doing something, in some situation
- ***Who, what, where***
  - **Actors:** human systems – individuals, groups, organizations, communities
  - **Behavior:** all aspects of the states and actions of those human systems
  - **Context:** temporal, locational and situational features in which the human system is embedded

All methods are inherently *flawed*

But each method is flawed differently!

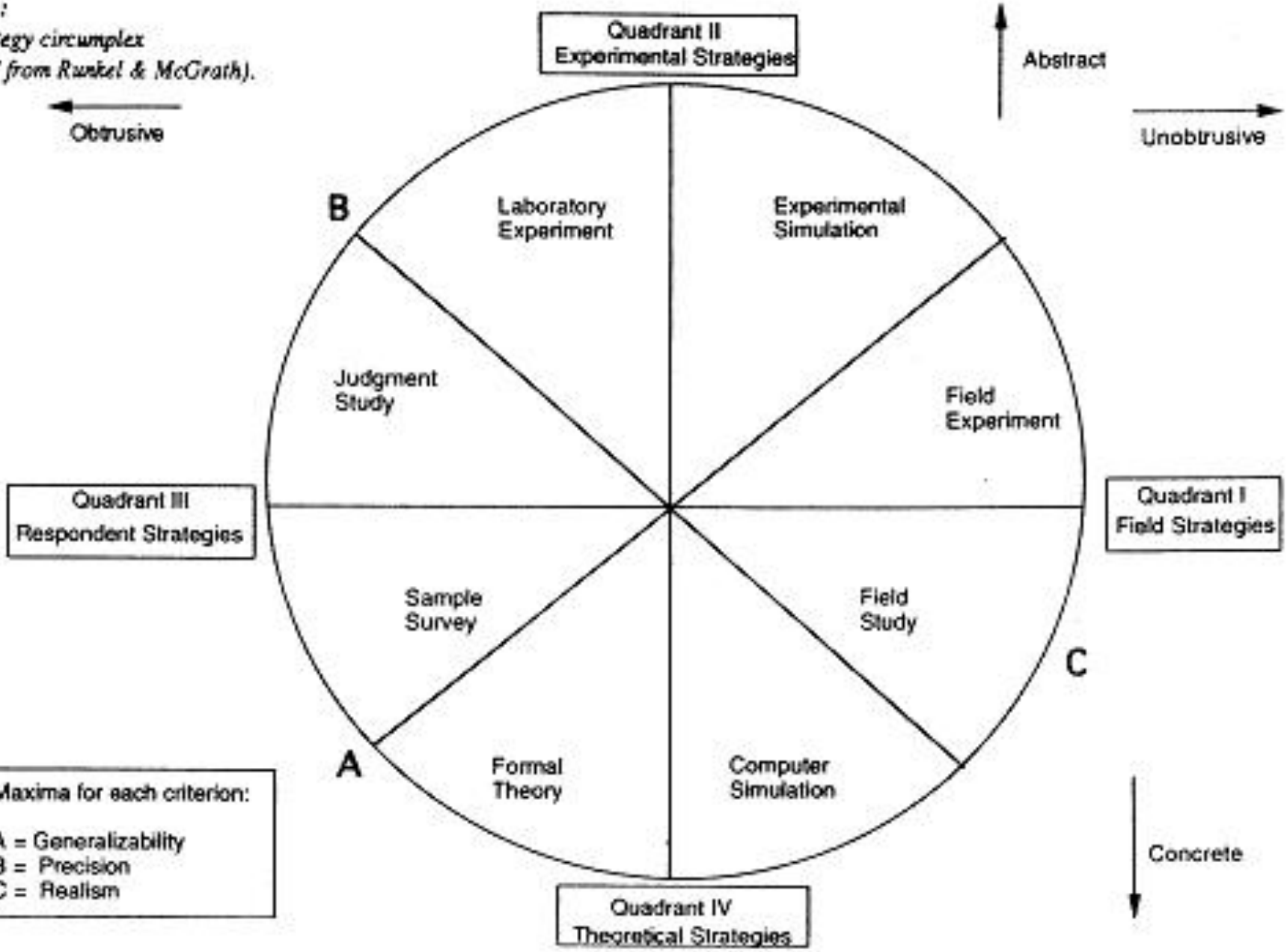
- Questionnaires?
- Observations?
- Controlled experiments?

Desirable features of research evidence: Choosing a setting

- **Generalizability** of the evidence over the populations of actors
- **Precision of measurement** of the behaviours being studied
- **Realism** of the situation or context where the evidence is gathered

Although goal is to maximize the above three things – we cannot!

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

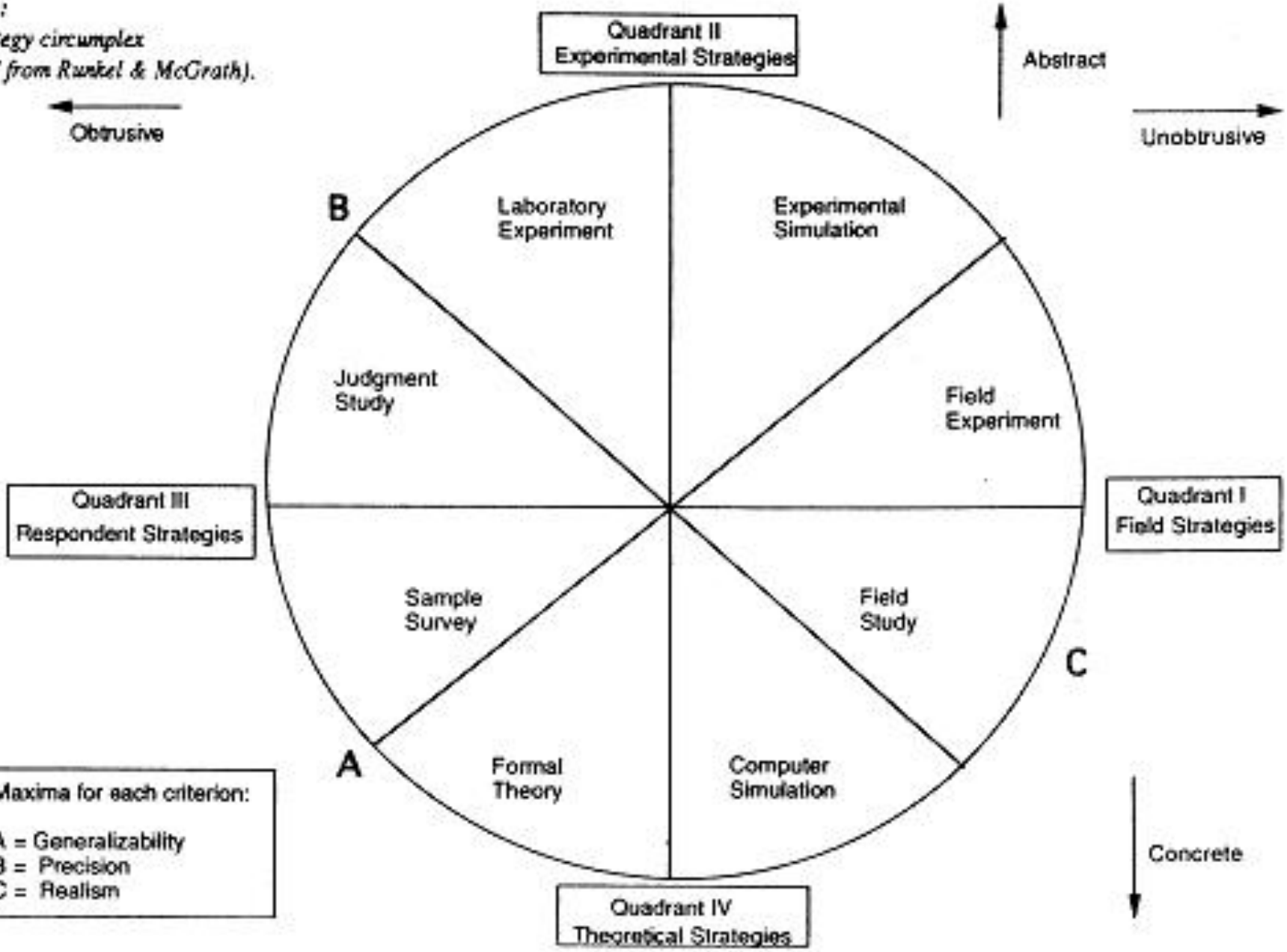


Quadrant 1: Field Strategies

Observations made are done in natural settings, and systems are disturbed as little as possible

- **Field studies:**
  - Ethnography
  - Case studies of organizations
- **Field experiment**
  - Some compromise is made – some of the naturalness is given up in favour of increasing precision of the measurements done
  - In a field experiment some variable may be manipulated (e.g. tool used or process) – more obtrusive than a field study but still in natural setting

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

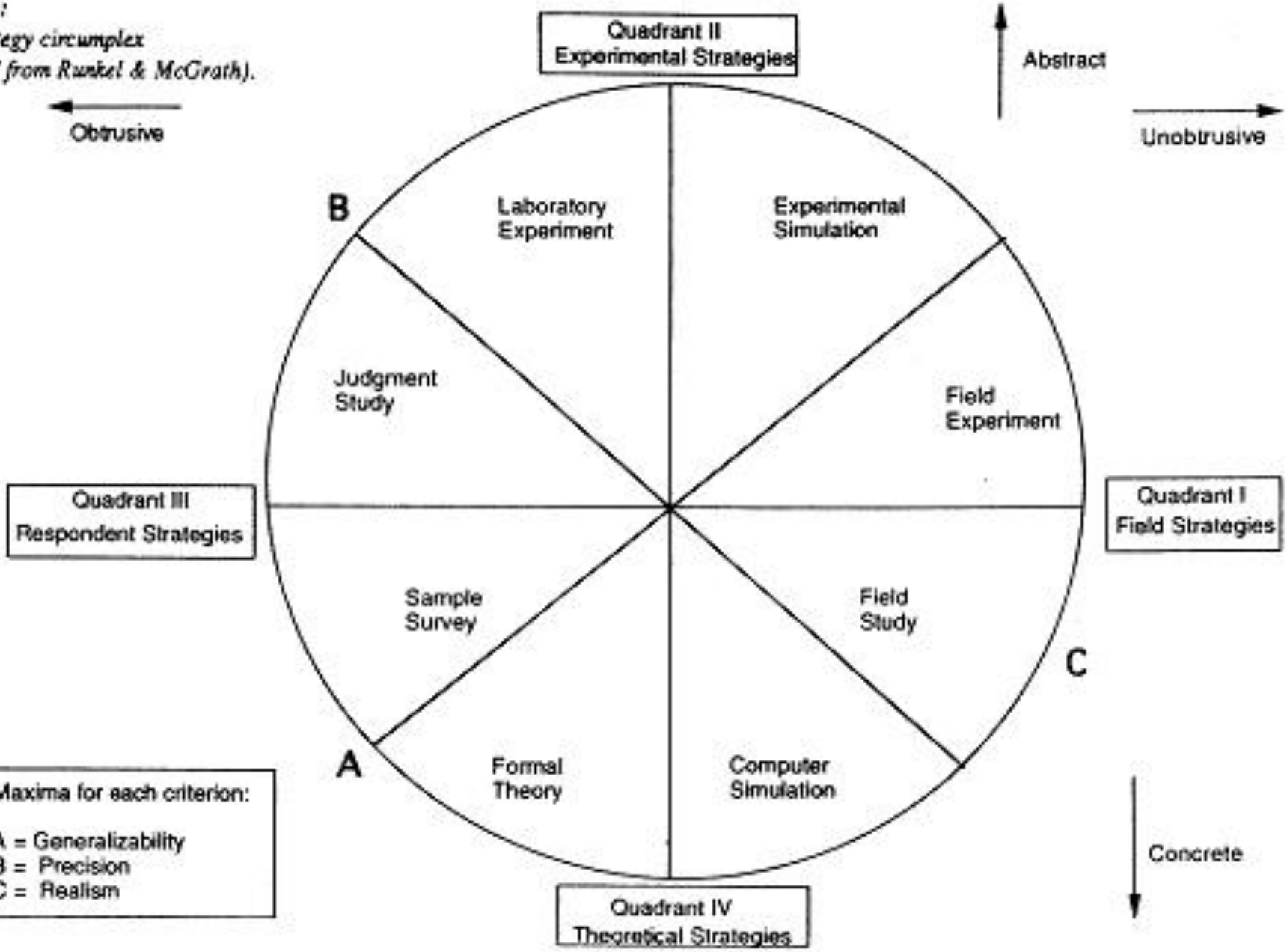


Quadrant II: Experimental Strategies

Concocted rather than natural settings

- **Laboratory experiment:**
  - Investigator creates the setting, defines the rules for its operation and then induces actors to enter this system
  - Increased precision of measurement
  - Increased obtrusiveness, unrealistic setting and reduced generalizability
- **Experimental simulation:**
  - Similar to a laboratory experiment, experimenter has control over the setting and conditions – but made to feel more like the real setting
  - E.g. flight simulators

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

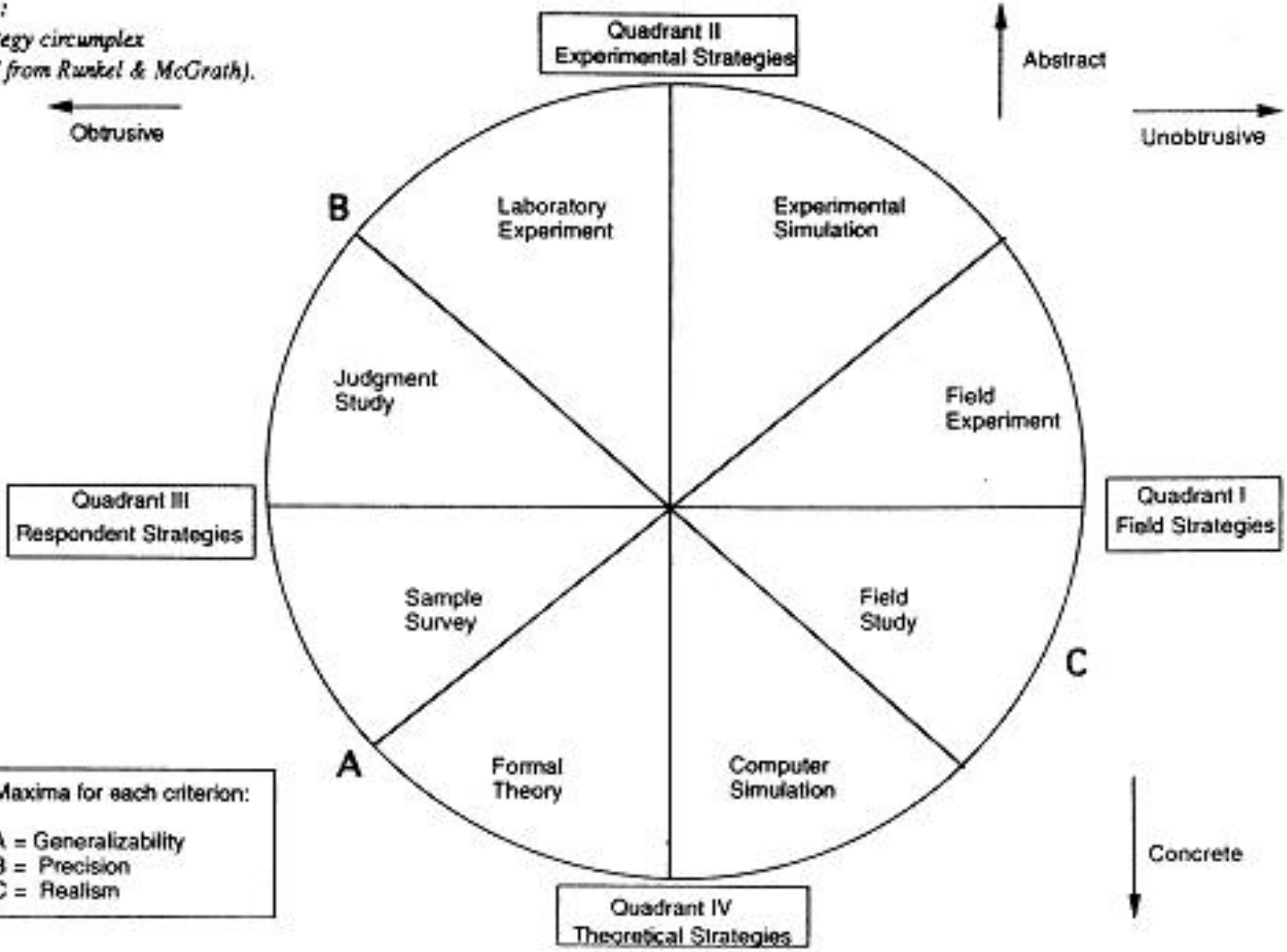


Quadrant III: Respondent strategies

Systematic gathering of participant responses to questions/stimuli where the setting is irrelevant

- **Sample survey:**
  - Collects evidence across a distribution of some variables or relationships among them, within a specific population
  - Careful sampling must be done to maximize generalizability
  - Little opportunity for much precision of measurements
- **Judgment study:**
  - For obtaining information about properties of a certain set of stimulus materials (focus of the study)
  - Usually done with actors of convenience
  - More precise measurements, but low generalizability, setting is not realistic

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



Quadrant IV: Theoretical strategies

- **Formal theory:**
  - Theories based on previous empirical evidence or other theories
  - Does not involve the gathering of empirical observations – researcher focuses on formulating general relations among a number of variables of interest
  - These relations (hypotheses, propositions) are intended to hold over a broad range of populations – generalizability is high but precision/realism are low
- **Computer simulation:**
  - Similar to experimental simulation as setting is contrived
  - Computer simulation is a closed system that models the operation of the concrete system but without participants
  - Behaviour must also be modeled, so all behavioural parameters must be known in advance – based on previous empirical evidence

Group discussion (see handout)

- From blog: “Runkel and McGrath’s strategy circumplex clearly attempts to split research strategies, but can we argue that a computer simulation is more concrete than a field study after reading the previous paper? What if our claims are incorrectly justified? I mean, in general computer sims have less human error than field studies.” [spencervatrtwatts]

- Discussion: do the dimensions abstract/concrete and obtrusive/unobtrusive make sense?

Comparison techniques

Critical to every empirical study, comparisons are at the heart of the research – depend on elements, relations and context

Three basic forms of comparison techniques:

- Baserates
- Correlational questions
- Difference (or comparison) questions

Baserates

- How **often** does Y occur? (at what rate, what proportion of the time)
- Often done as a precursor to more complex questions…
- Need to know what the **rate** of something is in the general case

Correlational questions

- Is there a **systematic or covariation** in the values of 2 or more properties (or features) of some system?
  - Positive correlation: As X increases (decreases), so does Y
  - Negative correlation: As X increases (decreases), then Y decreases (increases)
  - Zero or low correlation: No observable connection between X and Y
  - Non-linear correlations: e.g. X and Y may covary, then flatten, then covary again… need more powerful statistical tools to study non-linear correlations
- May look at more than two variables
- Note correlations do not necessarily indicate causal relationships

correlation versus causation [leonli] video:
https://www.youtube.com/watch?v=8B271L3NtAw

Science media hype [mlruss]
https://www.nature.com/news/study-points-to-press-releases-as-sources-of-hype-1.16551

Difference (comparison) questions

- Is Y **present** (**absent**) when X is present or high (absent or low)?
  - E.g. Do software engineers collaborate more effectively when they have had face-to-face meetings?
- Need to look for **interaction effects** of other variables – not always easy to hold other factors constant

Dealing with other factors!

- **Randomization** – 2 aspects
  - Sampling: how we select actors from a given population
  - How we allocate cases to conditions
- Note: you do not select a random sample, you select a sample using a *random procedure*!
- *Sample size* is critical – the larger the sample, the more likely it is you have a random sample (but be careful with this too!)

- Even doing all of the above won't lead to logical conclusions – just increases the likelihood or **probability** that X causes Y (could be other factors that were not evenly distributed)
- Need to reduce the scope to improve the power of the randomization -- realism is removed as we selected the participants, created the tasks and created the conditions

Validity

- Internal validity
- Construct validity
- External validity
- Threats to validity

Internal Validity

- What can we **conclude** from the study?
- Could it have been due to **chance** (statistical conclusion validity)?
- Some **other variables** may have been covarying with X (e.g. age and money) that we did not measure/control
- Have you considered all plausible **rival hypotheses**?

Construct validity

- How well defined are the **theoretical ideas** in your study?
- Do the **methods** you select match the problem?
- Are you really **measuring** what you are trying to measure?

External validity

- Will the findings hold under replication, that is how **generalizable** are they? What are the limits of how they hold?
- External validity can not be determined from one study – need follow-up/multiple studies

Threats to validity

- In all cases, we need to think what are the threats to validity…
  - What other **hypotheses** could explain the results?
  - **Mono-method** bias?
  - Did you **measure** what you thought you measured? Did your participants understand the vocabulary terms the way you did?
  - **Interaction effects**?
- Perhaps an experiment is better to be described as a pre or quasi experiment – stating the limitations

Classes of measures in social psychology

- Techniques for measuring the **presence or absence of specific features** in the human systems under study
  - Each has strengths and weaknesses

- For each case in our study, we need a record of what they did with information about the **context of the collected data**
  - This is needed so that later on the researcher can score it, aggregate it with other data, compare etc.

Who makes the record?

- Actor
- Investigator
- Third party

- **When** is the record made?
- Are the **participants aware** the data is being collected?

Classification of measurement types

- **Self reports:**
  - Made by participants with their knowledge
  - E.g. interviews, questionnaires, rating scales
  - Versatile, low costs, low “dross” rates
  - Potentially reactive
  - May be inaccurate
- **Trace measures:**
  - Made by participants, but often unknowingly
  - Typically not reactive, unobtrusive
  - Not so versatile, can’t always get them, often not closely linked to the things you study, costly

Classification of measurement types (2)

- **Observations:**
  - Made by researcher, participant usually knows but not always (hidden observer), only can view overt behaviour
  - **Reactivity** is very high, may have **observer errors,** costly
  - Advantages?
- **Archival records:**
  - Made by **3rd party** without research in mind
  - E.g. newspaper, birth records
  - Not so costly, but **difficult to cross validate**
  - Maybe reactive if records were to be made public
  - Ethical concerns?

Manipulating variables

- Options:
  - Can select **cases** with desired variable values
    - Does not lead to a true experiment as can't randomize allocation to conditions
  - **Direct intervention:**
    - May not always be possible, or could be difficult
    - Participants aware -> reactivity
    - But can do random assignment to conditions
  - Try to **induce desired values**:
    - Often involves deception of some form, ethics
    - Participants may guess!

- Summary:
  - Not one **right** or **best** way to measure – exclusive use of one technique can compromise the results

Easterbrook, Singer, Storey and Damian: Selecting a method…

Asking vague questions!

- Jane: "Is a fisheye view file navigator more efficient than the traditional view for file navigation?"
- Joe: "How widely used are UML diagrams used as collaborative shared artifacts during design?"

What is wrong with these questions?

What kind of research question are you asking?

Exploratory questions:

- Existence question
- Description and classification question
- Descriptive-comparative questions

Base-rate questions:

- Frequency and distribution questions
- Descriptive-process questions

Relationship questions:

- Correlation questions
- Causality questions
- Causality-comparative questions
- Causality-comparative interaction questions

Design questions

Philosophical perspectives (1)

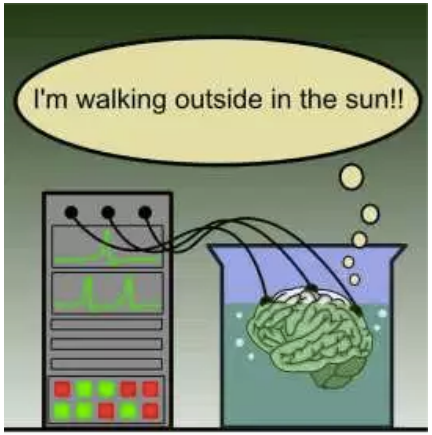
Positivitism

- states that all knowledge must be based on logical inference from a set of basic observable facts (post-positivist variation)
- Methods: controlled experiments also case studies and surveys

[superpenshine]

From the blog

- Brain in the vat thought experiment – our observations can’t be trusted



[spencervatrtwatts]

Jaden Smith ✓
@officialjaden

Follow

How Can Mirrors Be Real If Our Eyes Aren't Real

6:23 PM - 1 May 2013

43,174 Retweets **33,519** Likes

3.8K 43K 34K

Liv tbh @Oliviaaa02 · 1 May 2013
Replying to @officialjaden
@officialjaden stop trying to be deep Jaden. Just stop.
1

Brian J. Hunt @BrianJHunt · 4 Aug 2014
Replying to @officialjaden
"@officialjaden: How Can Mirrors Be Real If Our Eyes Aren't Real"
@BrianShortall you got an answer for this?
1

Brian Shortall ✓ @BrianShortall · 4 Aug 2014
@BrianJHunt @officialjaden passagesmalibu.com/home.html?kmas...
1

Brian J. Hunt @BrianJHunt · 4 Aug 2014
@BrianShortall @officialjaden How can we walk through a passage if our legs are our eyes?
2

Philosophical perspectives (2)

Constructivism

- Argues that scientific knowledge can not be separated from its human context
- Meanings of theoretical terms are socially constructed: theories emerge, not verified
- Methods: ethnographies, and exploratory case studies and survey research

Philosophical perspectives (3)

Critical theorists

- Choose what research to undertake based on whom it helps. They prefer participatory approaches in which the groups they are trying to help are engaged in the research, including helping to set its goals.
- Methods: Case studies used to draw attention to areas of research; action research (advocacy role)

Philosophical perspectives (4)

Pragmatists

- acknowledge that knowledge is judged by how useful it is for solving practical problems, emphasize the importance of consensus, choose research methods they feel will work (more of an engineering approach)
- Methods: mixed methods

Which methods go well together [timchancscw]

Potential biases in the papers?

- This could mean the author of the article (McGrath) is inclined towards pragmatism [AlisonG]
- “Although they [Easterbrook et al.] try to remain unbiased in their descriptions of these philosophies I got the impression that pragmatism was their preferred choice, mostly due to use of more favourable language when being discussed and the fact that it is the only philosophy entirely missing from the “empirical validity” section where each has its inherent flaws analyzed.” [jongrandfield]

WHAT ARE YOU?

[leonli] A paper on successful engineering / education collaborations.. and how to consider different backgrounds...
https://onlinelibrary.wiley.com/doi/full/10.1002/j.2168-9830.2008.tb00962.x

Theory building…

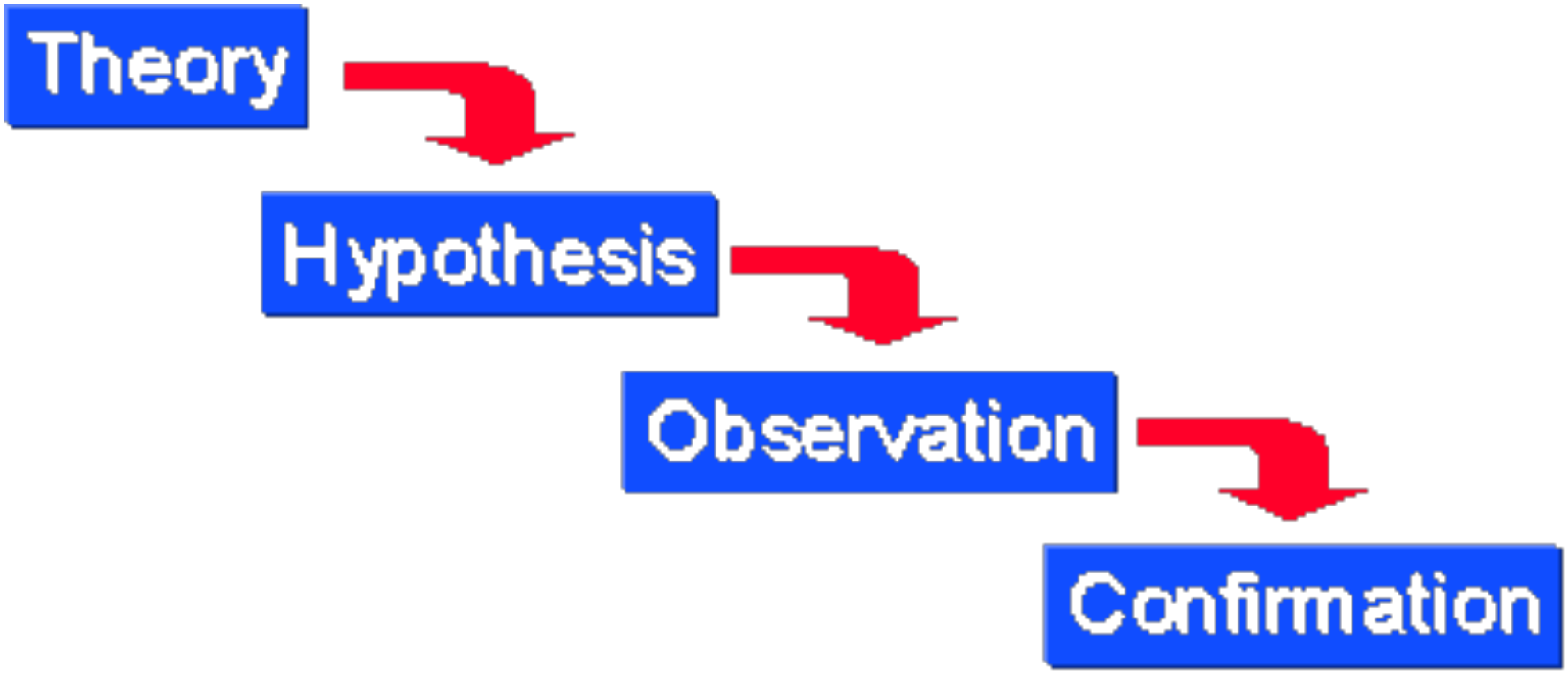
Joe's theory – describes how UML diagrams are a stylized form of external memory used in a collaborative group

- His theory says what they are used for (meetings, shared understanding..)
- His theory must define meaning of the terms such as "diagram", "discussion"
- Should explain why the diagrams are used in some settings and not in others
- Why some things are included in the diagrams, and other things are not
- His theory should be predictive of how a team may use UML based on certain factors

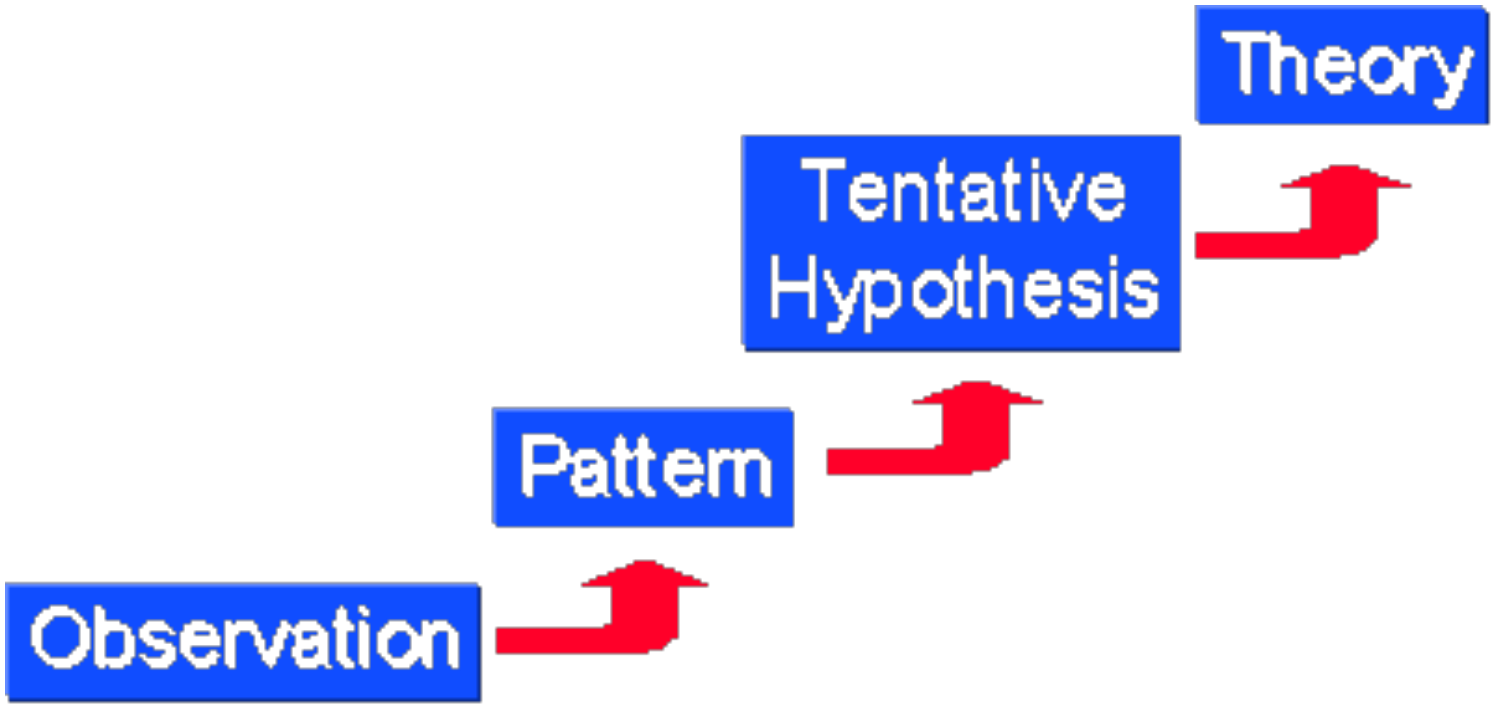
The role of theory in your studies….

- In a **quantitative study**, the **theory** is used as **a lens to guide which variables** should be measured or isolated
- But in a **qualitative study**, the **theory** is used to help **label and categorize** (code) the data
- But a theory may not be available at the outset, may be an **emerging theory**
- Theories also play a role in connecting research to the relevant **literature**

Deductive reasoning



Inductive reasoning



Controlled Experiment

- Investigation of a testable and clear hypothesis where one or more independent variables are manipulated to measure their effect on one or more dependent variables
- Each combination of values of the independent variables is a treatment
- We measure the effects of treatments on subjects
- Control is important
- Risks of using a theory?

Case Studies

- Term often misused to describe a worked example
- Yin: *"an empirical inquiry that investigates a contemporary phenomenon within its real-life context, especially when the boundaries between phenomenon and context are not clearly evident"*
- Exploratory case studies: derive new theories
- Confirmatory case studies: test existing theories
- Need to have a study proposition in advance, that guides the selection of cases and types of data to collect

Case Studies (2)

- Types of case studies:
  - Critical case (for testing a particular aspect of a theory)
  - Extreme or unique case
  - Typical case
  - Literal replications (to show same results for confirmatory case studies)
  - Theoretical replications (to show contrasting results)
- What is your unit of analysis (determines what data you collect)
- Benefits/risks?
- Which philosophical stance do case studies apply to?

From the blog...

- [jianwuuvic] "This article examines five common misunderstandings about case-study research: (a) theoretical knowledge is more valuable than practical knowledge; (b) one cannot generalize from a single case, therefore, the single-case study cannot contribute to scientific development; (c) the case study is most useful for generating hypotheses, whereas other methods are more suitable for hypotheses testing and theory building; (d) the case study contains a bias toward verification; and (e) it is often difficult to summarize specific case studies. This article explains and corrects these misunderstandings one by one and concludes with the Kuhnian insight that **a scientific discipline without a large number of thoroughly executed case studies is a discipline without systematic production of exemplars, and a discipline without exemplars is an ineffective one. Social science may be strengthened by the execution of a greater number of good case studies**."
- http://journals.sagepub.com/doi/pdf/10.1177/1077800405284363

Survey Research

- Usually questionnaires, but could be interviews or data logging techniques
- Need a representative sample from a population so we can generalize
- Need a clear research question
- Need to control for sampling bias, low response rates increase the risk of bias

Ethnography

- Field observation
- Study a community of people to understand how members make sense of social interactions
- Result is a rich description of how the community's culture
- May involve participant observation
- Ethnographic research takes an explicit constructivist stance, create theories
- Challenge: how to collect so much data, what to collect and how to analyze it

From the blog posts...

- “ethnographies also help the community make sense of their **social and culture setting**.” [superpenshine]

- “Ethnography and participant observation are often used interchangeably. But we prefer ethnography because participant observation seems to imply just observation. An ethnographer or a participant observer immerses him or herself in a group, observing behavior, listening to what is said, and asking questions. **Ethnography is a study in which participant observation is the prevalent method, but that also has a specific focus on the culture of the group being studied.** There are two types of ethnography: overt ethnography and covert ethnography.” [jianwuuvic]

Action Research

- Simultaneously study a problem and try to solve it iteratively (change the world)
- Need a project owner
- Research should be authentic
- Are there authentic knowledge outcomes for the participants?
- Most closely associated with critical theorists
- Risks?

From the blog...

https://www.emeraldinsight.com/doi/full/10.1108/09593849910267206
"Questions have been raised such as why is it important for action research to declare the intent of the study? What bias on roles might a researcher's philosophical stance have? Can an iteration be made if there was no reflective learning from the last step? Why is it important that action research has an intended change?" [jianwuuvic]

Mixed Methods

- Sequential explanatory strategy: quantitative data followed by qualitative data (latter helps explain the former)
- Sequential exploratory strategy: qualitative followed by quantitative (for testing emerging theory, explain early qualitative findings)
- Concurrent triangulation strategy: different methods used concurrently, improve validity

Mixed methods can fit with any of the philosophical stances. Usually associated with a pragmatist stance.

Threats to Validity: Positivist stance

- Construct validity
- Internal validity
- External validity
- Reliability: would the study yield the same results if done by different researchers?

“Validity”: Constructive stance

- Triangulation
- Member checking
- Rich, thick descriptions
- Clarify bias (report researcher bias)
- Report discrepant information
- Prolonged contact with participants
- Peer debriefing (plan ahead for this!)
- External auditor (also need to plan)

[mlruss]

Grounded theory...

- Connection between grounded theory and training process in machine learning! Overtraining similar to too narrow a focus in grounded theory


[Andreas]

Verification versus Validation

- https://www.unf.edu/~cwinton/html/cop4300/s09/class.notes/VerifyValidate.pdf
- Simply put, **verification** is the task of determining if the implementation of a model has been done correctly. Beyond program debugging, this means that verification data needs to be generated at various points in the model for comparison with expected values.
- **Validation** is the task of determining if the model constructed accurately represents the underlying real system being modeled. For any simulation model that is to be used in actual application it is very important to validate the model insofar as practicable, since real decisions are going to be made based on the simulation outcomes. ... Because a simulation model provides a surface "realism", it is possible to be fooled by the realistic appearance of the simulation. The best defense against this kind of mistake is to employ multiple means of comparing model performance against real data (if available), including statistical testing.

Pragmatic issues

- Access to field sites?
- Experience?
- Time, resources?
- Access to subjects?

Many challenges – but we must do our best!

[kunye]

Useful link for authoring a research paper: https://www3.nd.edu/~pkamat/pdf/researchpaper.pdf

Additional references

- Research Design, by John W. Cresswell
- Methods: Doing Social Research, By Winston Jackson
- Educational Research, by John W. Cresswell
- Case study tutorial (highly recommended): http://www.cs.toronto.edu/~sme/case-studies/index.html
- Activity theory: http://en.wikipedia.org/wiki/Activity_theory
- Case study research, books by Robert Yin!