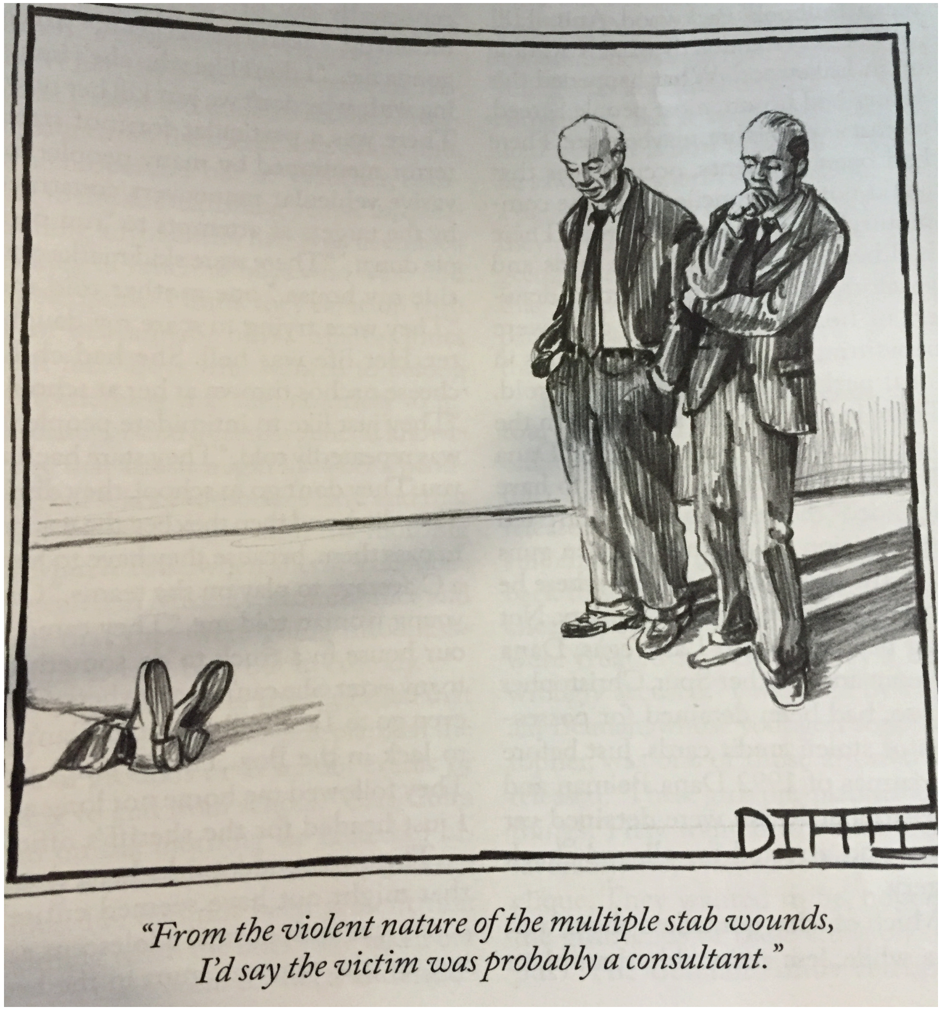
Evaluation, Measurement, Research and Being the 'Expert'

Brooks Applegate

January 25, 2017

Eval Cafe



"From the violent nature of the multiple stab wounds, I'd say the victim was probably a consultant."

The New Yorker Reader (2015)

How can we prevent this from happening?

SO WHERE DO I START?

Oh, yes with measurement, of course

Why Do We Measure?

Answer Questions and Solve Problems

Who Measures?

- Pretty simple -

Scientists & Non-scientists

BUT ….

Is measurement in the social sciences different from that in the natural sciences?

Is “science” in the social sciences different from that in the natural sciences?

A Step Back: Science is ……

- A collection of methods for describing, explaining, predicting, controlling …..
- Theory building, testing, rebuilding
- Possibly probabilistic
- System that utilizes the language/logic of inference
  - A method for reasoning: premise leads to a conclusion which results in an Inference
    - Deduction
    - Induction

Deduction

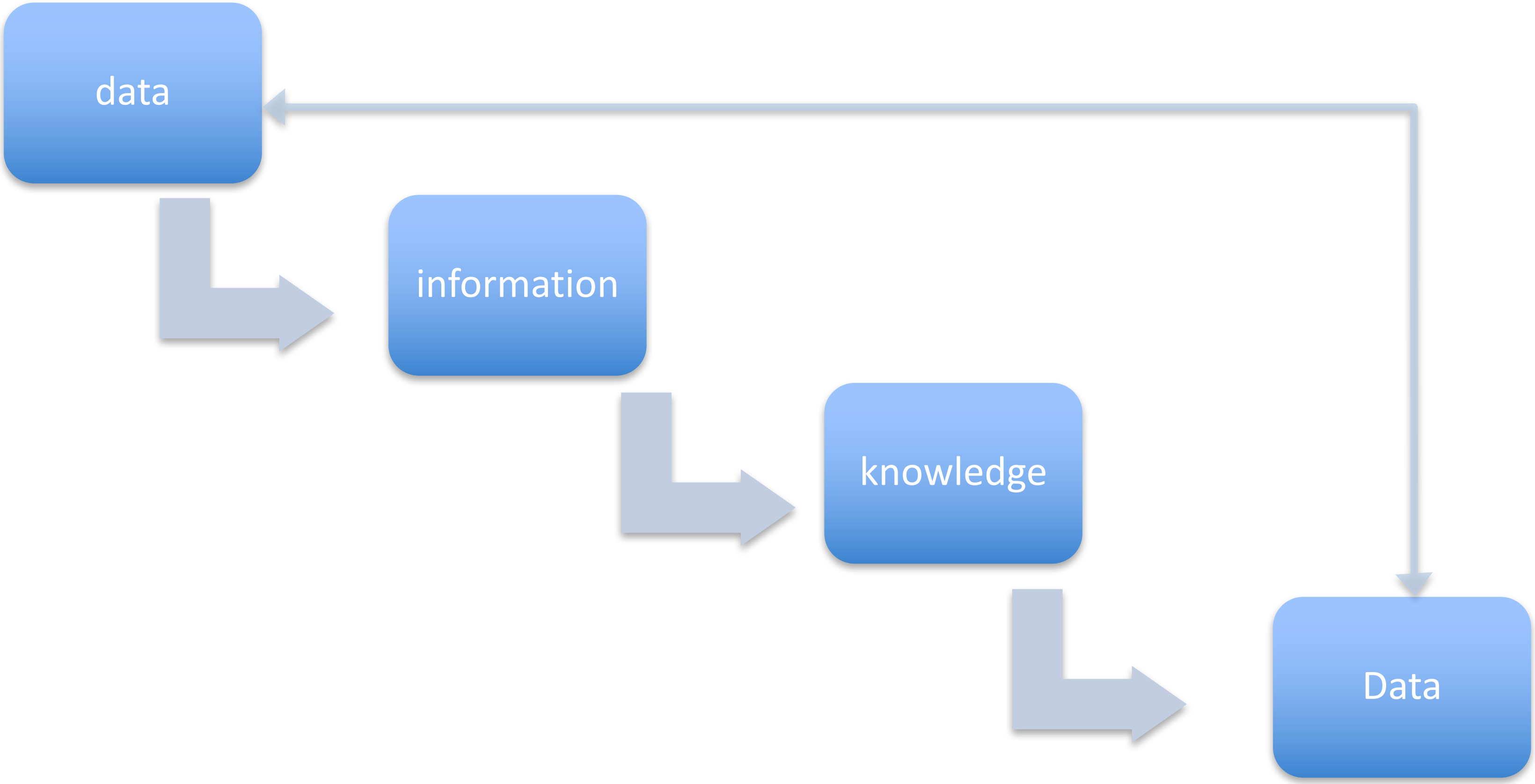
- Requires the existence of an appropriate relation between premise and conclusion
- Generalizing from many to one
- "proof" or "prove"

Induction

- Requires the existence of an appropriate relation between premise and conclusion
- Generalizing from few to many
  - Ah! So what do you know about sampling?

The overwhelming preponderance of (our) *knowledge* is based on inductive inference

Science Does This By Accumulation



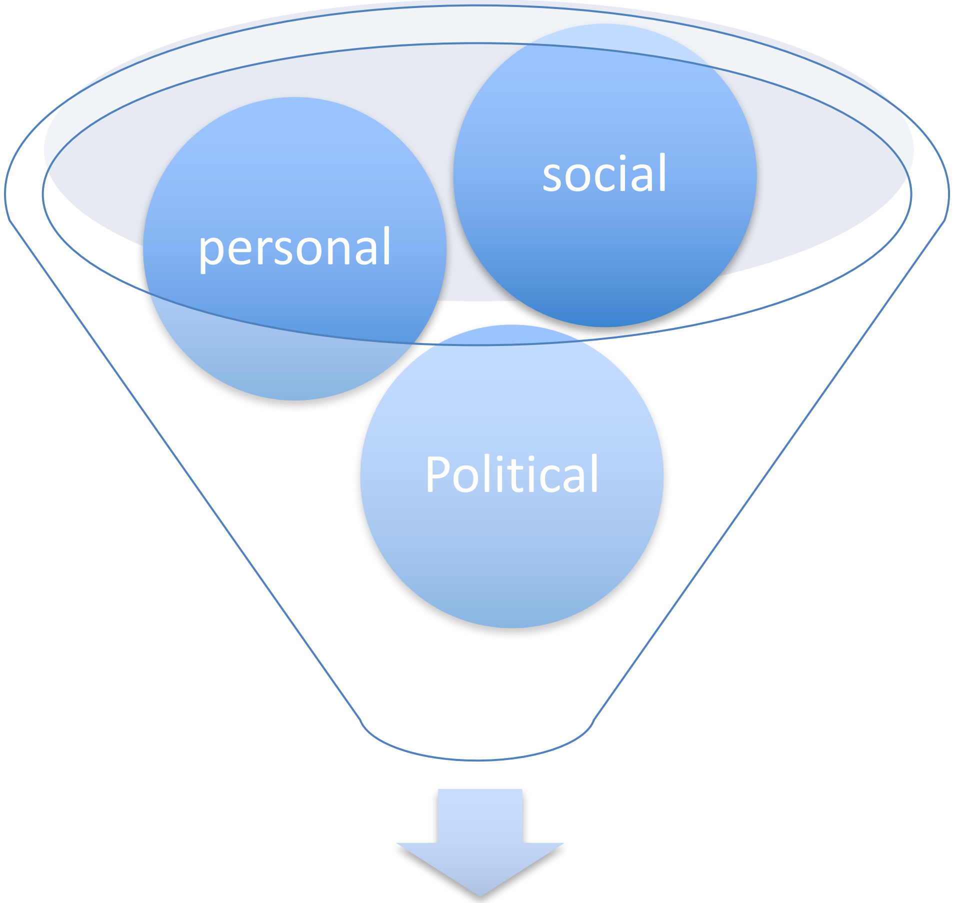
data

information

knowledge

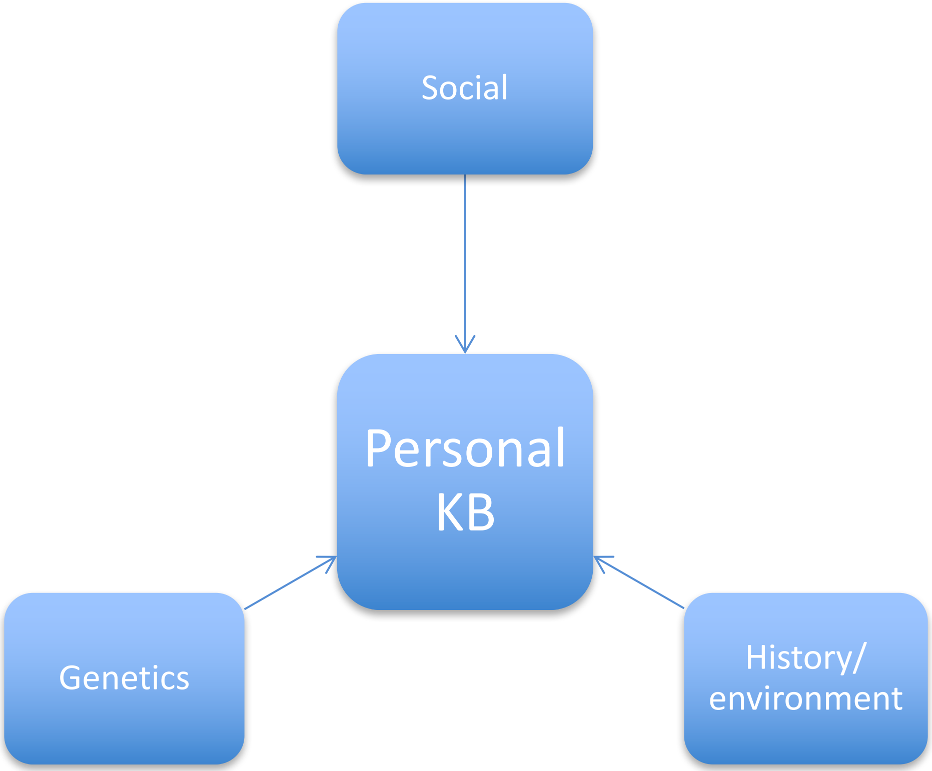
Data

Data Arrives Through An Application Of Filters



Data

The Filters You Apply Originate From Your Personal Knowledge Base (PKB)



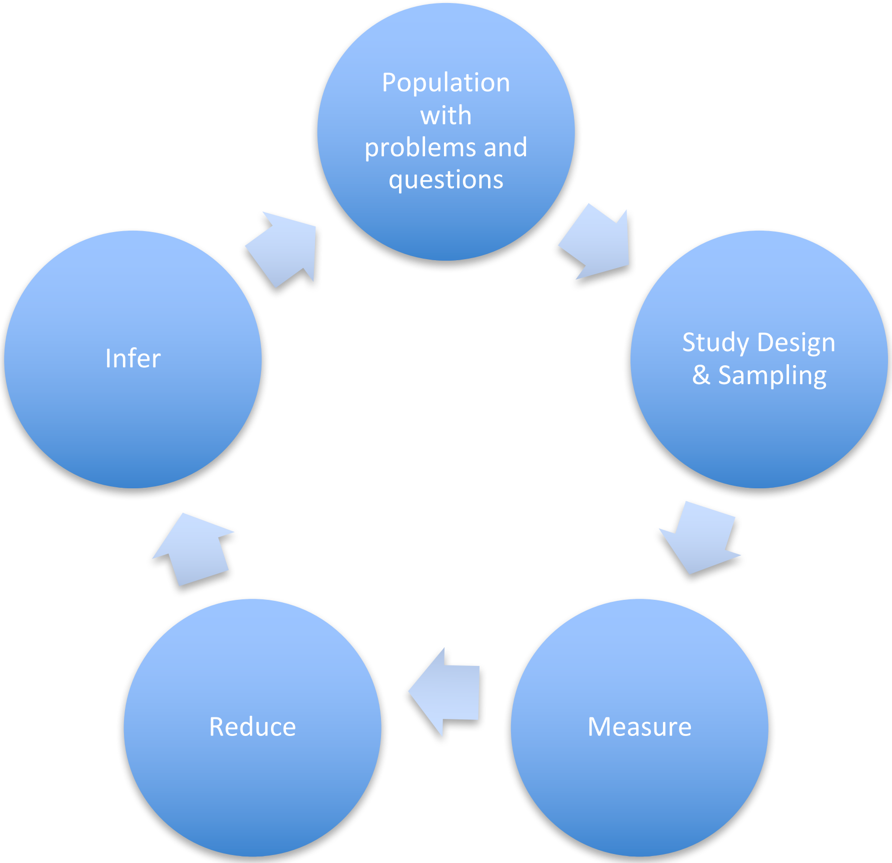
Your PKB

- What is the **health** of your PKB?
  - Old and dated
  - Edgy and raw
  - Static or dynamic
- How was it created?
- How is it built/expanded?
  - How is data turned into knowledge?
- How is it (-is it) tested?
- How is it corrected?

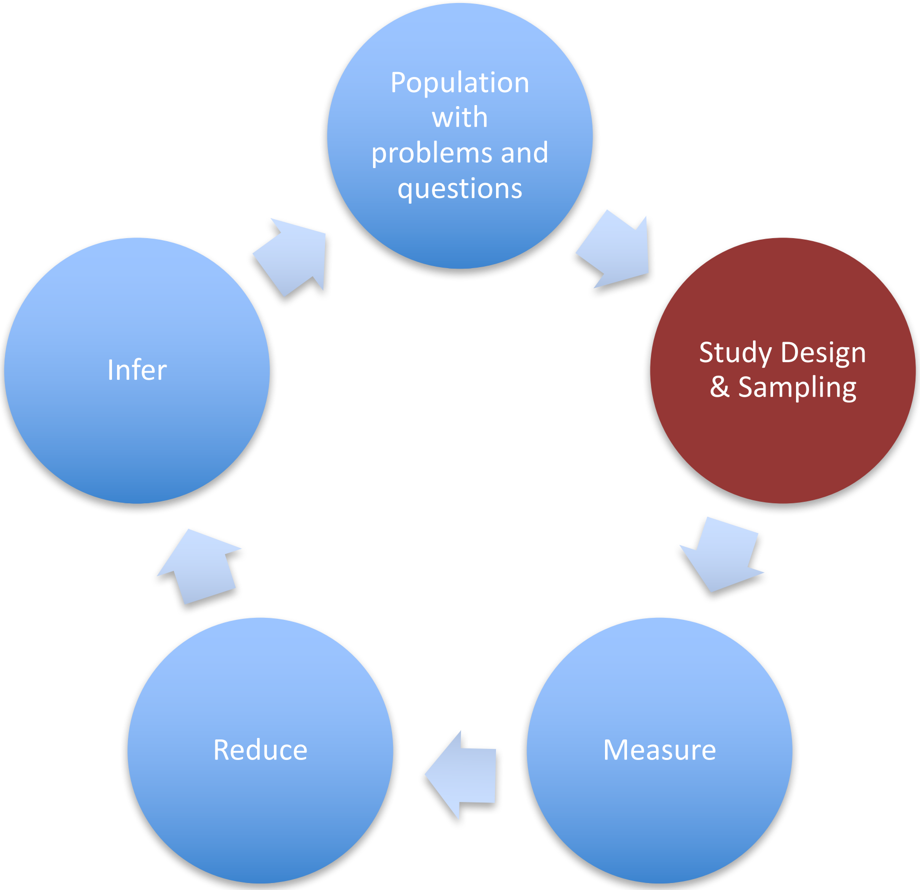
SO LEARN ABOUT THE PHILOSOPHY OF SCIENTIFIC INQUIRY

That is the why of things –
Now, what is the how?

Circle of (Inquiry) Life and Inductive Reasoning



Circle of Inquiry



Study Design (Big & Small)

- Macro design
  - How the parts fit together to provide a solution path towards answering the question/problem/hypothesis
- Micro design
  - Structure of the problem: RQ’s or hypotheses
  - Sampling
  - Use of experimental control: DOE
  - Instrumentation
  - Analytics

Looking at the micro elements

Problem & Purpose

So your students are not completing their homework in a timely manner -----

What is the problem?

Following Problem Formation

- You need to identify an evidence base (population) that contains an expression of the problem (variance) among different sampling units

Sampling

- Theoretical -> Target -> Experimentally Accessible Population
  - All the units to which one whishes to generalize
- Sample Frame
  - List from which a sample is to be drawn in order to represent the population
- Sample
  - All the units of the population that are drawn for inclusion in the study
- Completed Sample
  - All the units that complete the study

Coverage Error In Sampling

- Occurs when not all members of the population have a known, non zero chance of being included in the sample and when those who are excluded are different from those who are included.
  - Example 1: population of interest is single family households and an internet based survey was used
    - Many single family homes may not have access to the internet
  - Example 2: A sampling frame that is out of date

Sampling Error

- Sampling error leads to lack of precision in the estimates because not every person in the population in included in the sample
- Larger samples result in smaller amount of sampling error
- Smaller samples result in larger sampling errors
- The question is HOW MUCH sampling error can be tolerated in answering the RQ?

Estimating Sampling Error

- Probability based sampling
  - Sampling error estimation depends on the sampling design
  - Influences analytical design
  - Inferences to the target population have known precision
- Non-probability based sampling
  - Sampling error estimation ….. Not possible or depends on your assumptions, so ask your favorite sampling statistician!
  - Generally does not affect the analytical design
  - Inferences to the target population are based on logical discourse, assuming a target population is defined

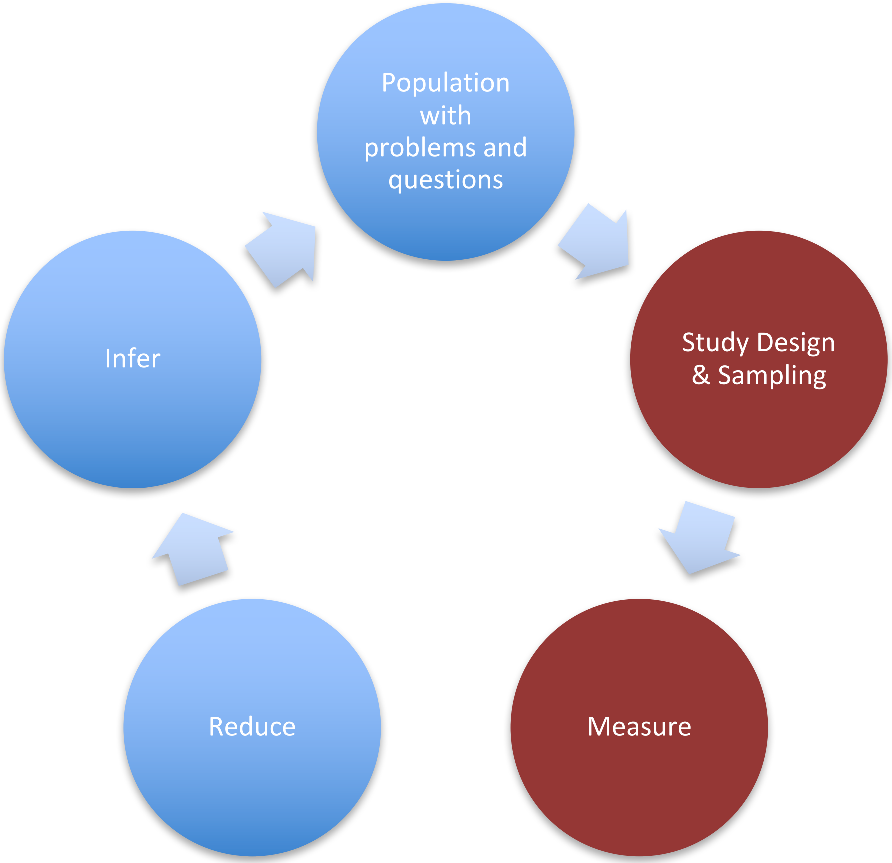
Another Kind of Sampling Error: Non-response Error

- Understanding the nature of missingness
- MCAR, MAR, NMAR
  - Can occur at the level of the assessment
  - Can occur at the level of the item

We Infer from samples: Good samples and poor samples …..

SO IF IT’S NOT TOO LATE, TAKE AS MANY SAMPLING CLASSES AS YOU CAN

Circle of Inquiry



To Create Data We Must Measure - Is there a definition of measurement?

- Stevens (1946) “the assignment of numerals to objects or events according to rules”
- Lord & Novick (1968) & Torgerson (1958) elaborated on this definition by noting that measurement applies to properties of the objects, not the objects themselves
  - If we substitute “symbols” for “numerals” we see that all data representation is a form of measurement:
- A measurement is a **set of rules** for assigning symbols that represent differences (or similarities) of properties of objects, traits, attributes, behaviors, events

Focus On Object Attributes

- Any given object possess multiple attributes which describe and differentiate this object from other objects, both similar and dissimilar
- Many of the “objects” that are of theoretical interest are not directly observable and must be **inferred** from observable properties of the object
  - There is no universal agreement among researchers on a “Gold Standard” property that defines the object

Psychological Constructs

Measurement of a psychological attribute or object (e.g., construct) occurs when a symbolic value is assigned to a collected behavioral sample

- Constructs provide efficient and economical means for studying a number of similar properties
  - Often these “properties” are behaviors that are observed and are inferred to represent a macro characteristic or object that is relevant to study, for example; Self-esteem

Problems Encountered When Measuring Psychological Constructs

- No single approach to measurement of ANY construct is universally accepted (no Gold Standard)
- A “test” represents only a limited number of behavioral samples
- Measurements always contain ERROR
- Psychological constructs derive their meaning and usefulness from two sources
  - Their operational definition
  - Relation to other constructs or observable phenomena

Data Can Come From Many Sources

- Surveys
  - Interviews, questionnaires, assessments
- Observations
- Machines
- Surveillance systems
  - Web-based
- Extent databases
  - Medical records, Student SIS

Measurements Differentiate Samples (of objects, properties of objects, behaviors)

A Measuring system creates VARIANCE

- By selecting properties of a behavioral sample TARGET (like a question stem)
- By characteristics of the RESPONSE scale
- Different response scales = different rules

A Measured Behavioral Sample = Stem + Response Scale

- The question STEM
  - The most important part of the question
  - This is generally where the most explicit and direct information about what is wanted from the respondent is located, e.g., the TARGET
- The response scale
  - Where and how the respondent is allowed to respond

A Stem May Ask

- Physical characteristics
  - Weight, height, BMI, bio sex
- Personal/social characteristics
  - Gender-, ethnic-, occupational-identity
- Attitude
- Opinion
- Perception
- Preference
- Belief
- Behavior or behavioral intent

The Response Scale May Consider

- Time frame
- Undefined structure = constructed response
- Defined structure = discrete
- Defined structure = non-discrete
  - Ordered or continuous
  - Anchors: Unidirectional - bidirectional

A Sampling Unit*Item Interaction

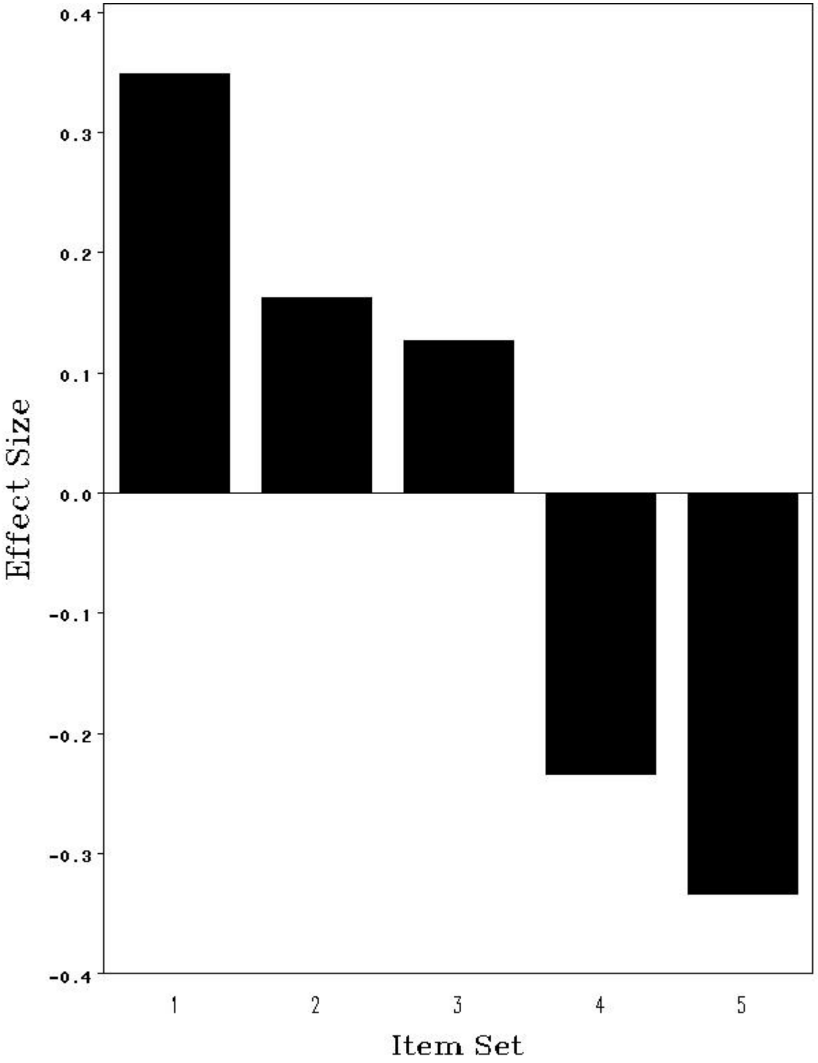
- Item wording, layout, and response structure AFFECTS how respondents respond
- How questions are intrepreted and the available response structure, together with respondent characteristics partially determines the variance observed in the item
- You must consider:
  - What you want to know
  - Who is responding
  - What they read
  - What they interpret
  - What they internalize
  - How they are asked/required to respond

An Example

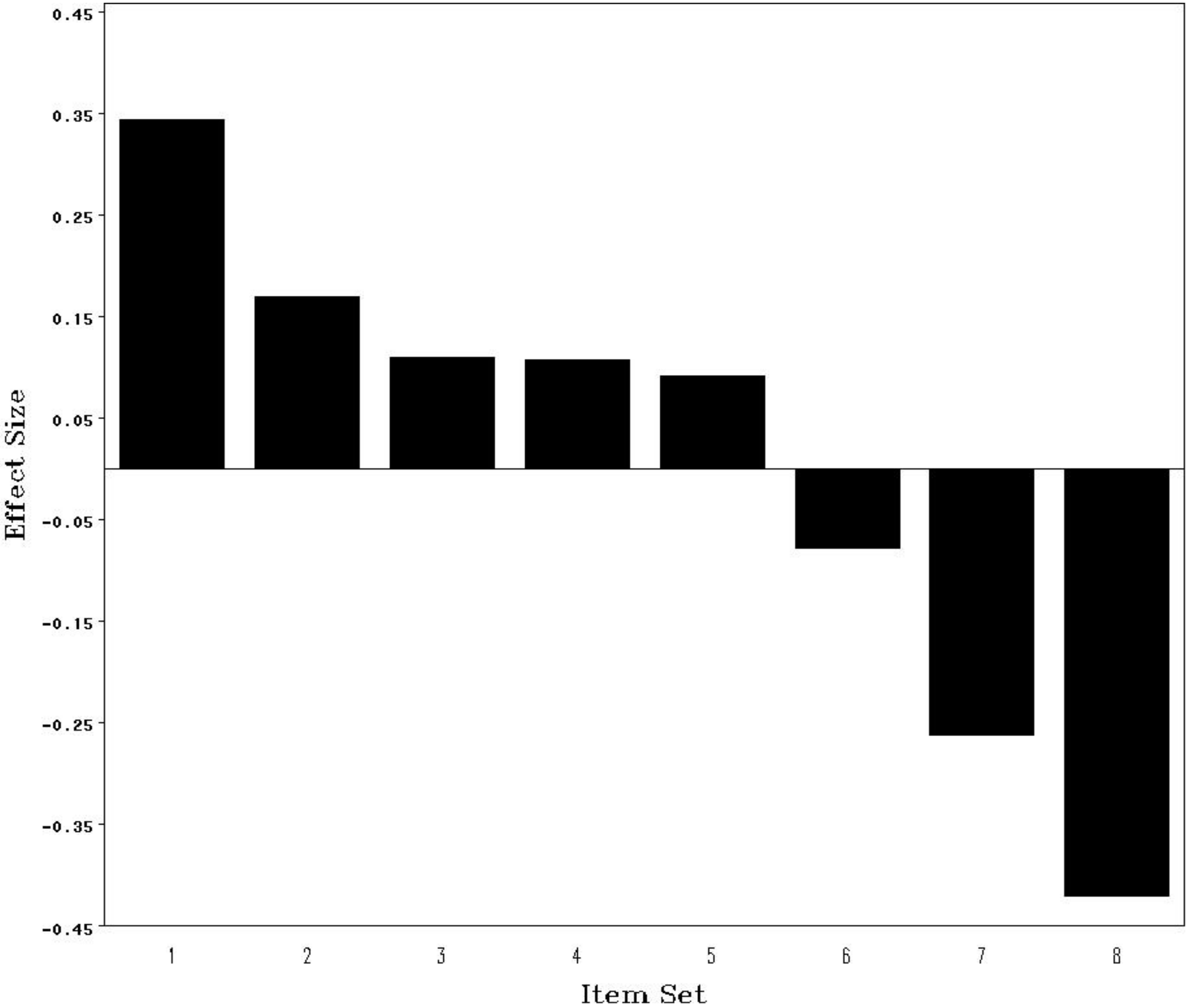
Taken from work focused on the structure of child and adolescent psychopathology as measured via structured diagnostic interview

Looking at item order effects

Mean Effect Size Between Forward and Reverse Presentation Order: Temperment



Mean Effect Size Between Forward and Reverse Presentation Order: Psychopathology Symptoms



For data to be informative (valid) it needs to be reproducible (reliable) and yield variance in a sample

The Need For Test Theory

- Test Theory deals with the measurement, meaning and use ascribed to psychological constructs
- Test Theory describe how inferences from examinees item responses can be made about unobservable char that are measured by the test
- Test Theory, if it is to be useful, must accommodate all forms of measurement
- Test Theory as a discipline deals primarily with
  - The mathematical models linking observable measurements to their inferred meaning
  - Establishing methods for estimating the adequacy of the inferred meaning
  - Establishing methods for estimating things that influence the inferred meanings
  - Provides a general framework for viewing and evaluating the process of test (instrument) development

Classical Test Theory

- X=T+E
  - Formulation of the basic concepts of reliability & validity
- Has limitations

General Theory of Latent Traits

- Assumes that a set of *k* latent traits or abilities underlie examinee performance on a set of test items
- *k* latent traits define *k* dimensional latent space, with each examinee's location in the latent space determined by the examinee's location on each latent trait
- The latent space is complete if all latent traits influencing test performance of a population have been specified

Provides a fusion with VALDITY

Item Response Theory

- An extension of the factor analysis of binary items
  - Begins by fitting a model, estimating item parameters and assessing model fit
  - Second step: estimate the “latent (ability) trait”

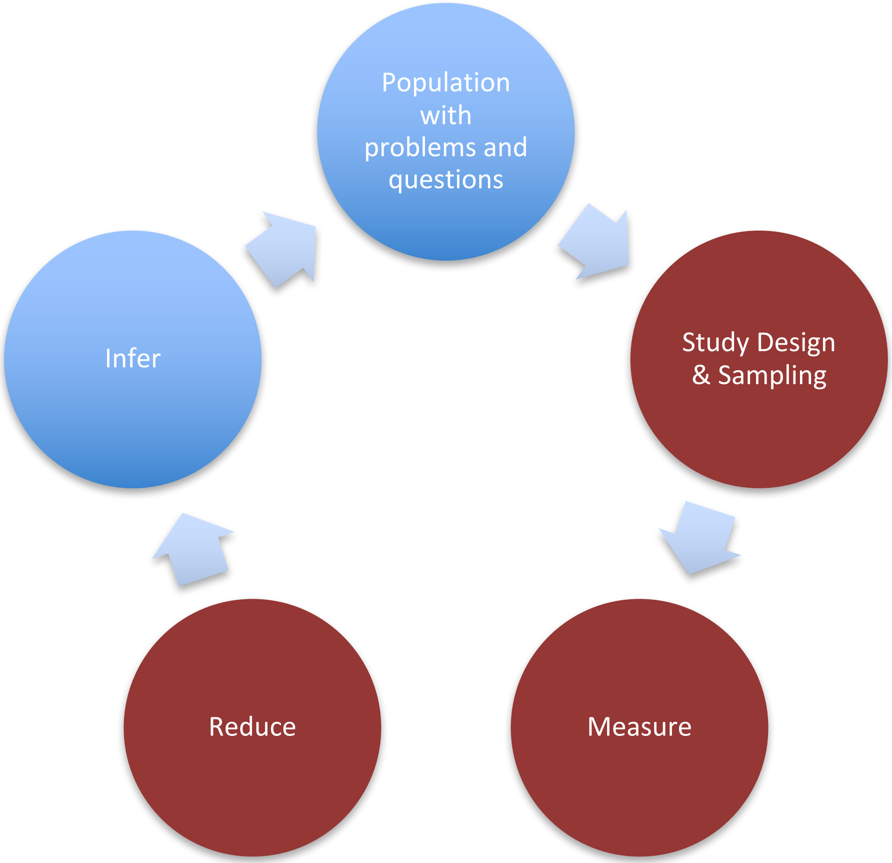
IRT Today

- A ever growing family of models serving a ever growing number of purposes
  - Undimensional and multidimensional
  - Binary, ordered response, nominal response

The scientific knowledge base is grounded to reality through measurement ….

SO IF IT’S NOT TOO LATE, TAKE AS MANY MEASUREMENT CLASSES AS YOU CAN

Circle of Inquiry



Data Reduction & Analytics

- Be sure your analytics ANSWERS the RQ/hypothesis
- Understand the limits of the macro & micro design elements
  - Sampling features
  - Study design and control features
  - Measurement features
  - Model (statistical or logical)
- Evaluate your assumptions

Analytics

- The expansion in analytical methods across the horizon of social science & health science disciplines is striking and increasing at a nonlinear rate
- This is partially fueled by better communications among scientists (e.g., information access on the web) and by an increasing amount of interdisciplinary research

Analytical Tool Expansion

- A professional engaged in the world of inquiry is ethically obligated to bring to their inquiry table an analytical toolbox that has both breadth and depth of skills
- This requires a lifelong commitment to professional development inorder to stay in the forefront of analytics

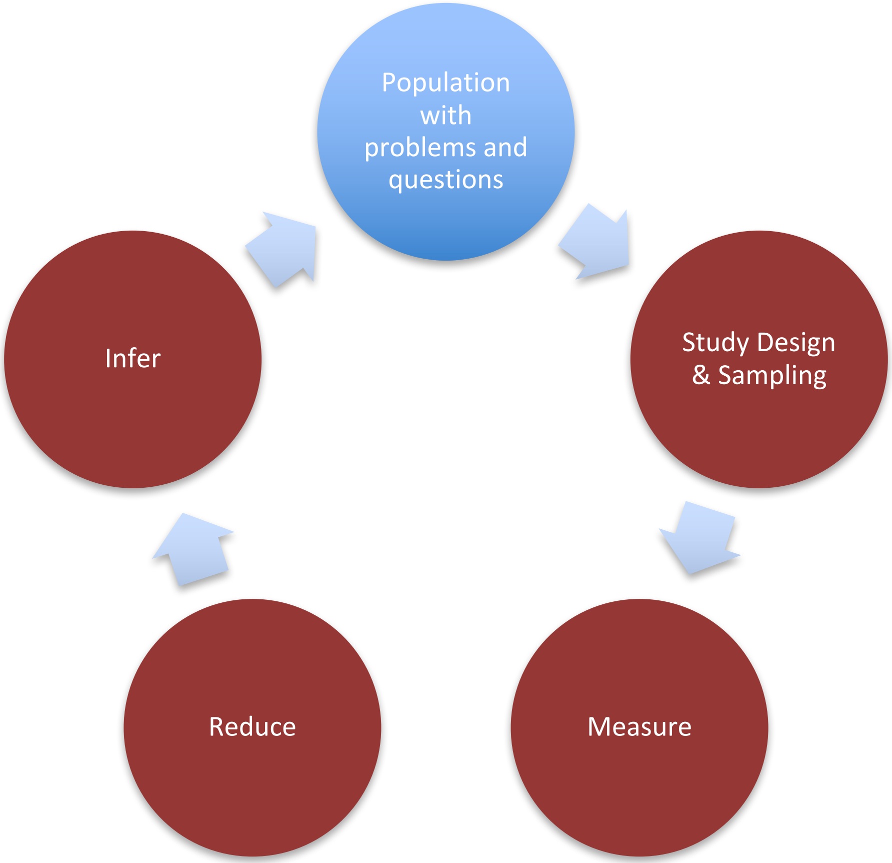
You & Analytics

- There is a strong tendency in many EMR-type graduate programs to narrowly specialize analytical skills
- An outcome of this specialization is the need for large research teams
- This is not so say that developing deep but narrow skills in particular research methods or analytics is not important. Rather, I believe a deep skillset must be developed from a broad-based/general foundation in methods and analytics.

Every study needs data reduction …..

SO IF IT’S NOT TOO LATE, TAKE AS MANY DATA ANALYTIC CLASSES AS YOU CAN

Circle of Inquiry



Population with problems and questions

Study Design & Sampling

Measure

Reduce

Infer

Conclusions and Inference

- Are the original study RQ’s or hypotheses addressed?
- What is the precision in the findings?
- What is the accuracy of the findings?
  - What counterfactuals are (were not) addressed?
- What are the Inferences that can be defended?
- Remember your audience

Needed Skill-Sets

- A actively vetted PKB
- A full methodological tool box
- Listening skills
- Resource allocation skills
- Program management skills
- Macro study design skills (toolbox)
- Micro study design skills (toolbox)
- Communication (oral and written)

Practice, practice, practice ……

SO IF IT’S NOT TOO LATE, GET INVOLVED IN AS MANY STUDIES AS POSSIBLE – YOU NEED VARIANCE!

THANK YOU

Questions