Secondary Data Analysis

Pros and Cons

James B. Schreiber

Introduction

Any data set can be used in secondary data analysis.

- In this FORS Lecture, the positive and problematic aspects of using a secondary data set will be discussed.
- Topics such as public versus secure data sets, IRB issues, aligning constructs with variables in the data set, and weights will be covered.
- Finally, some conversation about different analysis software/systems will be included.

AGENDA

Primary vs. Secondary
Area of interest
Searching for Appropriate Data sets
IRB issues
Public vs Secure data sets
Reading a data book

Aligning Constructs you are interested in with actual variables in the data set
Research Questions based on secondary data set variables and how the data was designed and collected
Missing Data and Weights
Measurement Issues
New Options
Quantitative
Qualitative

Step	Primary research	Secondary data analysis
1.	Formulate research questions and specify tentative hypotheses.	Formulate research questions and specify tentative hypotheses.
2.	Design study. Decide on sample and sample size. Select measures and manipulations.	Search for potential data sets to address research questions. Conduct literature review to avoid duplicating existing work.
3.	Conduct pilot tests. Make design adjustments. Finalize research questions	Obtain data sets and supporting materials. Gain familiarity with codebooks and data structure. Finalize research questions.
4.	Collect data.	Construct and evaluate measures.
5.	Prepare data for analysis.	Create final data set for analyses.
6.	Conduct analyses.	Conduct analyses.
7.	Interpret results	Interpret results
8.	Attend to limitations and unanswered questions	Attend to limitations and unanswered questions
9.	Write report	Write report

Note: Steps modified and expanded from McCall and Appelbaum (1991).

AREA OF INTEREST

- Theory/Conceptual/Framework
- Causality Arguments
- Core Research Questions
- Variables of interest
- These need to be established.
- You will waste time in a sea of data sets if you do not know what you really want.
- This is not the time to think," I will just look and see if I find something great"

Searching

Background and
Location of Data

Data Sets
- Theory Based
- Theoretical/Conceptual Foundation
- Data Collected Based on Foundation

Research Questions

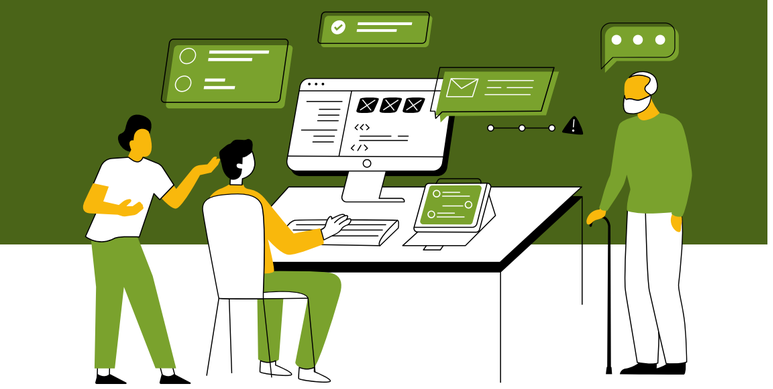
Non-theory based
- Empirical Research Driven
- Conflicting Variables

Governmental/Non-Governmental

Kaggle/Data World/HeathData.gov
- HRSA public health programs

Access
- Public and Private/Restricted Use

Repositories



- NIH: https://dash.nichd.nih.gov/explore/dataset
- NIH RePORT: https://report.nih.gov/databases
- NCIB (Bio_Tech) https://www.ncbi.nlm.nih.gov/datasets/
- ICPSR: https://www.icpsr.umich.edu/
- Need a colleague at a linked institution, they do have some that are free for everyone but the charge is $575
- CDC: https://www.cdc.gov/nchs/data_access/ftp_data.htm
- Long. Studies if Aging, Health Statistics Data, NHANES, NVSS (Vitat Statistics, e.g., Natality

IRB

For our IRB, some data sets have been pre-approved for exempt

Depending on different factors, the public data sets will be exempt/not human subjects needed because the identifying markers have been removed

Some private data sets will have to go through exempt/expedited review

Code Book-Natality Data Set Example

270-271	2	UPREVIS	**Number of Prenatal Visits**	U,R	00-49	Number of prenatal visits
					99	Unknown or not stated
272-273	2	PREVIS_REC	**Number of Prenatal Visits Recode**	U,R	01	No visits
					02	1 to 2 visits
					03	3 to 4 visits
					04	5 to 6 visits
					05	7 to 8 visits
					06	9 to 10 visits
					07	11 to 12 visits
					08	13 to 14 visits
					09	15 to 16 visits
					10	17 to 18 visits
					11	19 or more visits
					12	Unknown or not stated

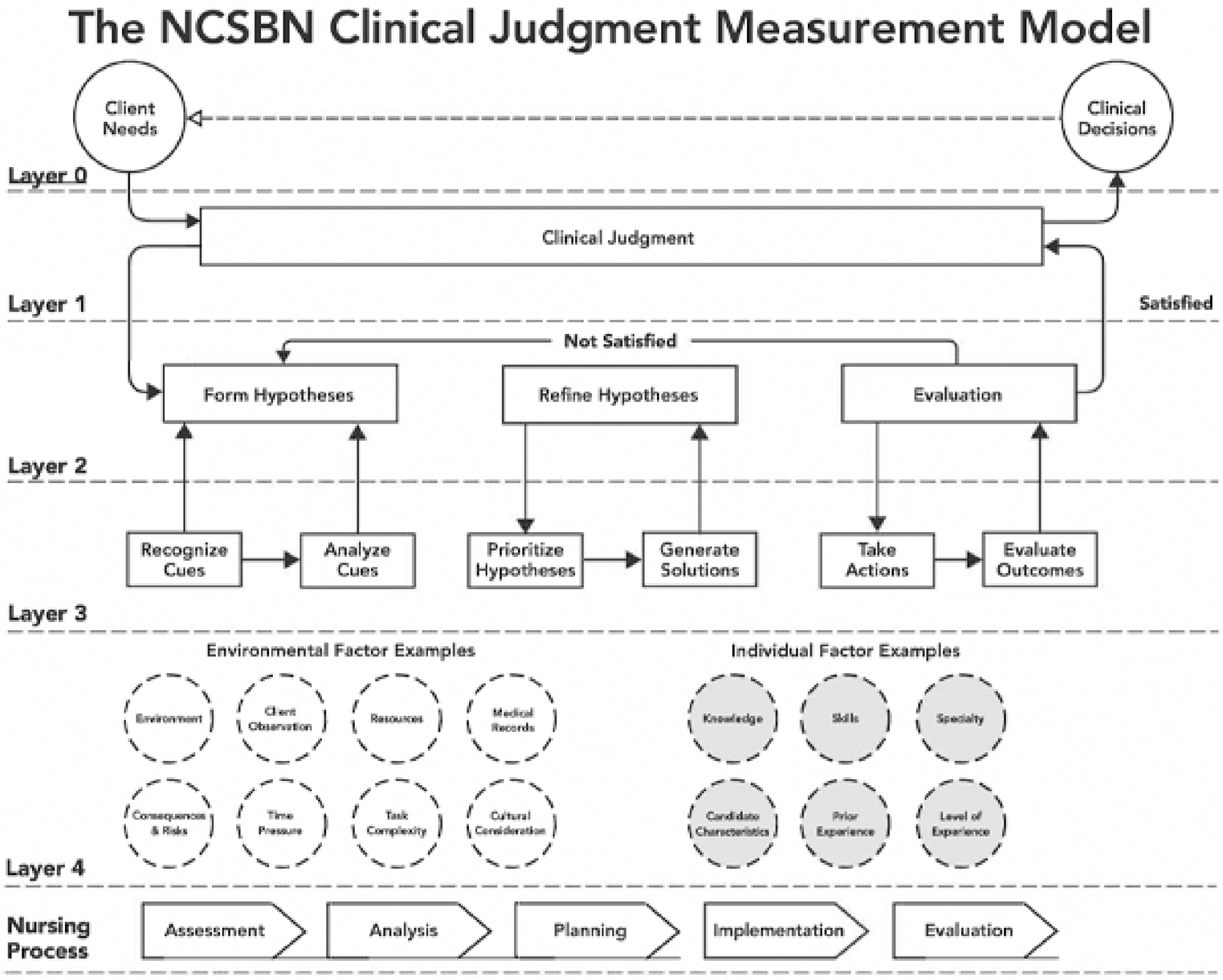
Definition Match

The data

- Alignment with your Operational Definition
  - Threshold, how far is too far for you?
  - You need a deep dive into this area
  - Surveys/Instruments used –
    - Alignment with population
- Believability/Trustworthiness of Data Gathere
  - How the data was gathered
  - Changes to that gathering
  - How the data was input “eh..close enough”
- Traditional Reliability and Validity
  - Is it even discussed in the manual

Theoretical Model/ Statistical Model

The New NCLEX



Statistical Model

Item Response Theory

- Results are not sample dependent-sample independent with linear transformation
- Allows for “Partial Credit” Items/Data
- You do not need to take all of the items

Data Type

Conflict between data collect and what you wanted
Conflict between data collected for a variable and the theoretical argument about that variable
Continuous variable technically/theoretically, gathered in categories for data set you have access to.
This affects the types of analyses you can complete.
Compromise-> Limitations

Data Cleanliness

Missing and Imputation
MICE For Continuous
imputeMCA for categorical
Not all data can be imputed

DataTransformations-
Statistical Assumptions

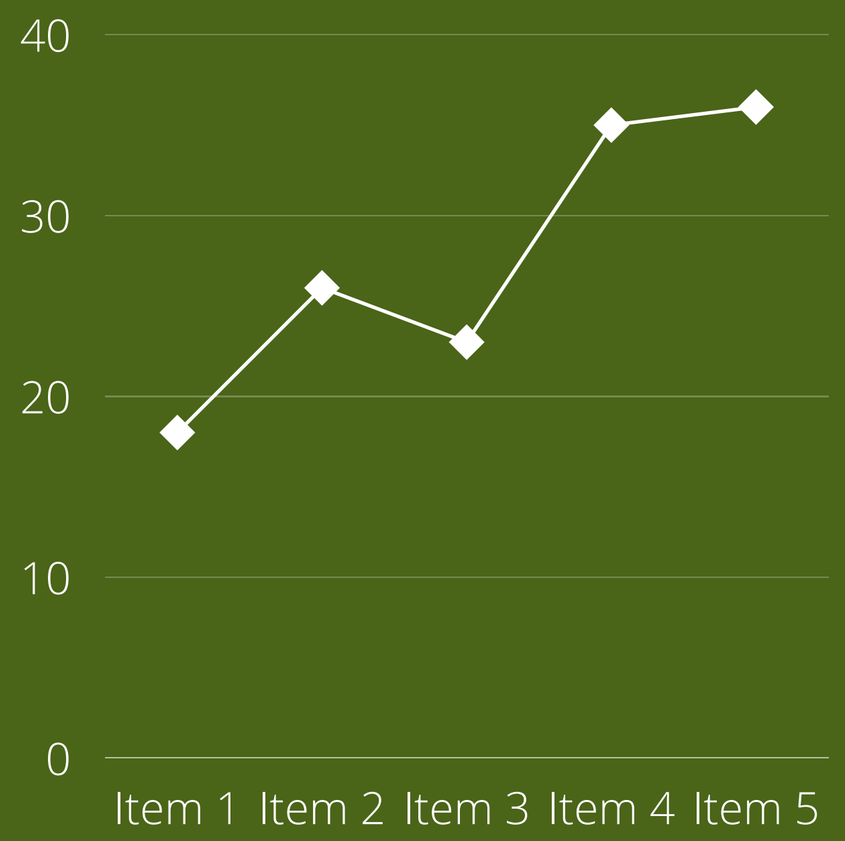
Weights Due to Sampling

Proxy Data

- Variable is closely related to variable you want
- Variable is measuring something besides its title
- High School Employment vs. After School Sports
- ADI-Area Deprivation Index

Statistical Model

You can use any model/technique
You will violate some assumptions
Limits your conclusions



Newer (Kind of)- Easier to Implement Now

Quantitative
Machine Learning at BigML
IRT- JAMOV (Freeware)
BAYESIAN-JASP (Freeware)

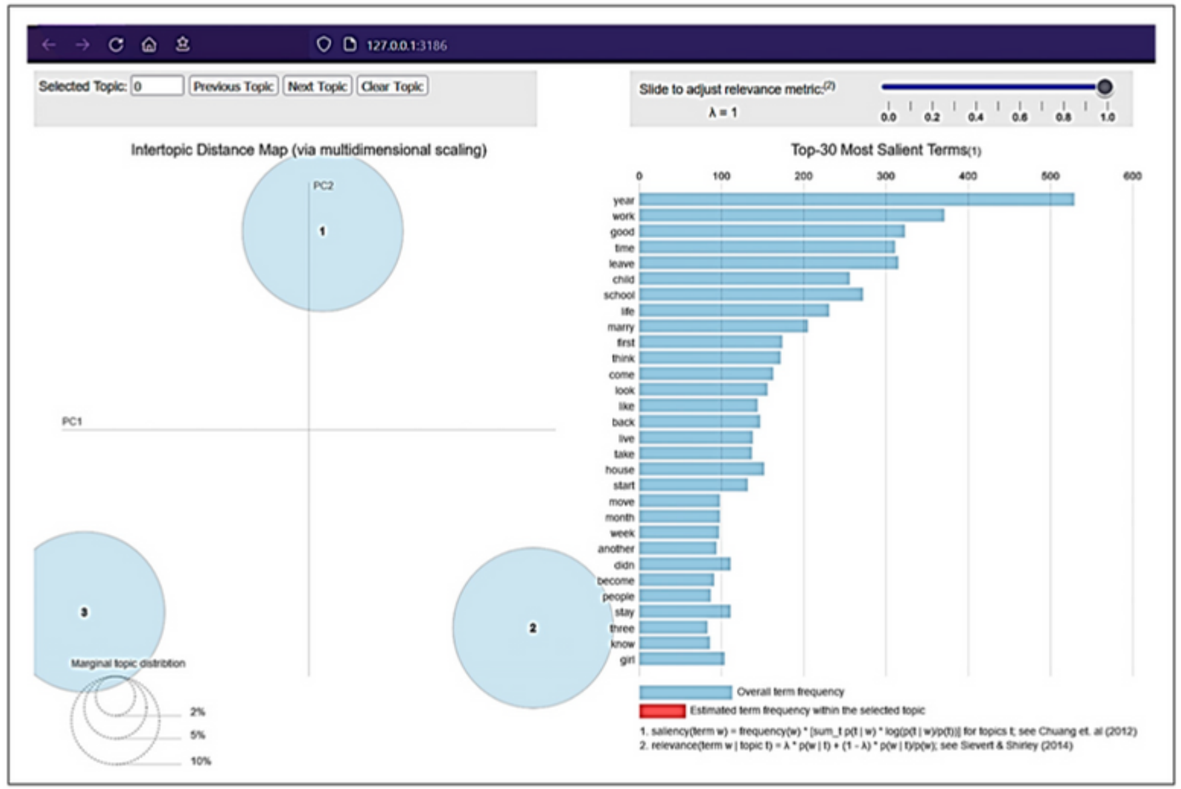
Larger Data sets are in the cloud and accessed from the cloud
Many researchers now use Python or R and manage the data at places like GitHub

Popular Techniques
Linear Regression
Logistic Regression
Structural Equa. Modeling
LCA/LPA (Unsupervised Learning)
Bayes Trees

Newer (Kind of)- Easier to Implement Now

Qualitative

- LACOID for qual, secondary data analysis with qualitative data



References

Goodwin, J. (2012). SAGE secondary data analysis. SAGE secondary data analysis, 1-1408.

Johnston, M. P. (2014). Secondary data analysis: A method of which the time has come. Qualitative and quantitative methods in libraries, 3(3), 619-626.

Vartanian, T. P. (2010). Secondary data analysis. Oxford University Press.

González Canché, M. S. (2023). Latent code identification (LACOID): A machine learning-based integrative framework [and open-source software] to classify big textual data, rebuild contextualized/unaltered meanings, and avoid aggregation bias. International Journal of Qualitative Methods, 22, 16094069221144940.

Ledolter, J., & VanderVelde, L. S. (2021). Analyzing Textual Information: From Words to Meanings Through Numbers (Vol. 188). SAGE Publications.