Content Analysis and Archival Research

Secondary Data Analysis: Content Analysis and Archival Research

- Archival Research
  - *Study historical documents*
- Content Analysis
  - *Measure behaviors in movies or books*
- Same techniques:
  - *Catalog behaviors*
  - *Frequency, duration or interval method*
  - *Inter-rater reliability*
  - *Mean girls*
    - – socially cooperative behaviors, social aggression, consequences

Archival Data

- In archival research, researchers analyze data pulled from existing records, such as census data, court records, personal letters, old newspapers, etc.
  - *Agency records/utilization data*
  - *Existing data*
  - *Actuarial records*
  - *Legislative and Governmental documents*

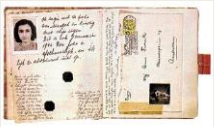
Sources

- Public and Private archives
  - *Murray Research Center*
    - Social science data on human development and social change
    - Sharing of qualitative data more unusual and difficult
  - *Inter-university Consortium for Political and Social Research (ICPSR) at UM*
    - Survey data on all U.S. national elections since 1952

Sources
Private Records

Issues

- Authenticity?
  - *Bogus biography of Howard Hughes (1972)*
  - *Freud*



Types

- Autobiographies
  - *Comprehensive*
  - *Topical*
  - *Edited*

  impersonal
- Diaries
  - *Intimate*
  - *Memoir*
  - *Log*
- Blogs
- Letters

Sources

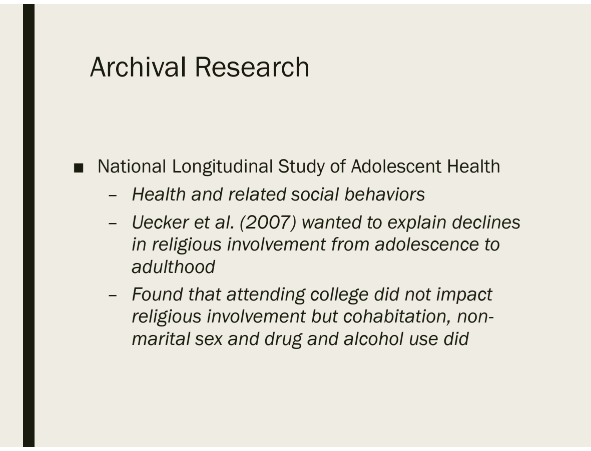
- Published or broadcast media
- U.S. Census etc.
  - *Complete count census – reach every household*
  - *Errors in Coverage*
    - Not covered or covered more than once
  - *Errors in Content*

Archival Data

- Archival data is useful for studying:
  - *Social and psychological phenomena of the past*
  - *Social and behavioral changes over time*
  - *Topics that involve articles, advertisements, or speeches*
  - *Anything that must be studied after it has occurred*
  - *Re-analyze*

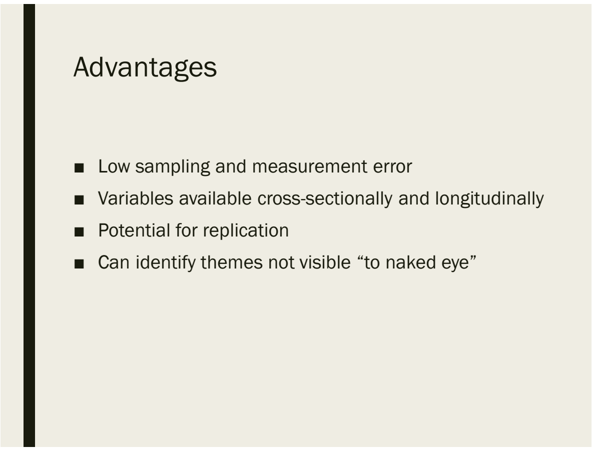
Classic Research

- Emile Durkheim's "Suicide"
  - *Rates higher in Protestant vs. Catholic countries*
  - *Variation between rural and urban societies*
    - Religion, season, marital status, gender
    - Lack of social integration
- Terman's Genetic Studies of Genius
- Gurr's
  - *Civil strife – greater difference between value expectations and value capabilities*



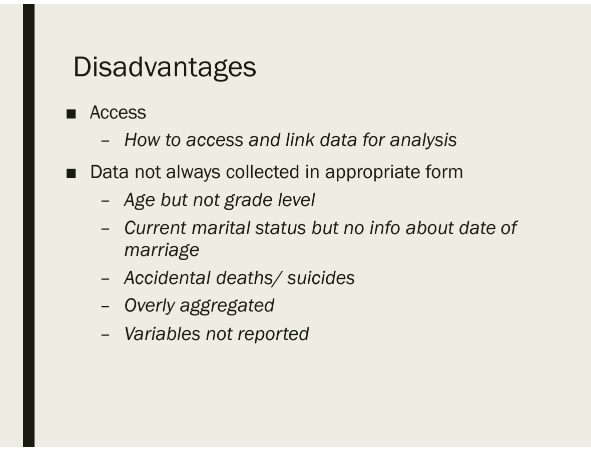
Archival Research

- National Longitudinal Study of Adolescent Health
  - *Health and related social behaviors*
  - *Uecker et al. (2007) wanted to explain declines in religious involvement from adolescence to adulthood*
  - *Found that attending college did not impact religious involvement but cohabitation, non-marital sex and drug and alcohol use did*



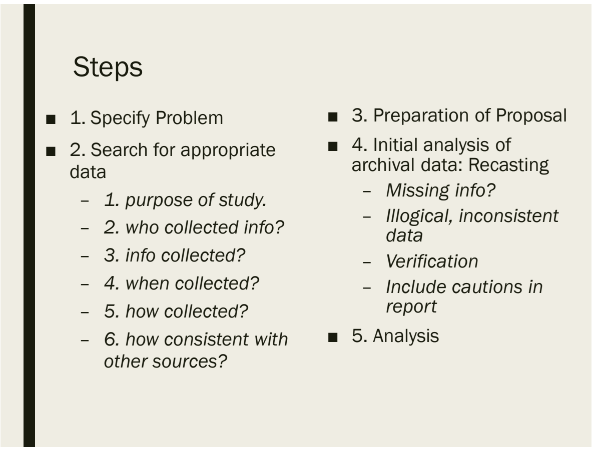
Advantages

- Low sampling and measurement error
- Variables available cross-sectionally and longitudinally
- Potential for replication
- Can identify themes not visible "to naked eye"



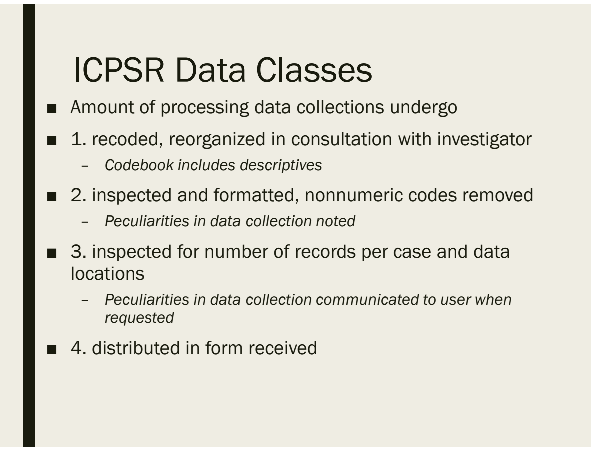
Disadvantages

- Access
  - *How to access and link data for analysis*
- Data not always collected in appropriate form
  - *Age but not grade level*
  - *Current marital status but no info about date of marriage*
  - *Accidental deaths/ suicides*
  - *Overly aggregated*
  - *Variables not reported*



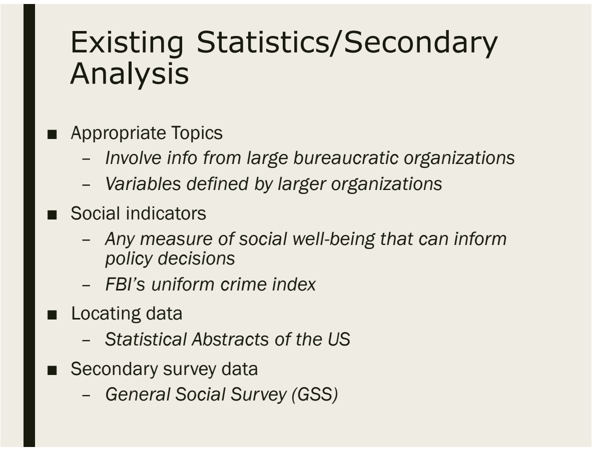
Steps

- 1. Specify Problem
- 2. Search for appropriate data
  - *1. purpose of study.*
  - *2. who collected info?*
  - *3. info collected?*
  - *4. when collected?*
  - *5. how collected?*
  - *6. how consistent with other sources?*
- 3. Preparation of Proposal
- 4. Initial analysis of archival data: Recasting
  - *Missing info?*
  - *Illogical, inconsistent data*
  - *Verification*
  - *Include cautions in report*
- 5. Analysis



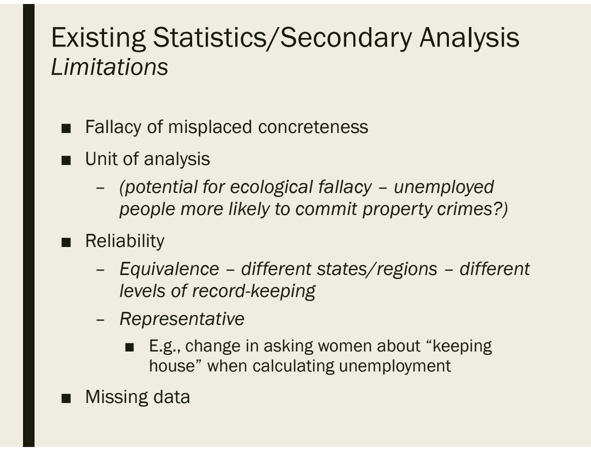
ICPSR Data Classes

- Amount of processing data collections undergo
- 1. recoded, reorganized in consultation with investigator
  - *Codebook includes descriptives*
- 2. inspected and formatted, nonnumeric codes removed
  - *Peculiarities in data collection noted*
- 3. inspected for number of records per case and data locations
  - *Peculiarities in data collection communicated to user when requested*
- 4. distributed in form received



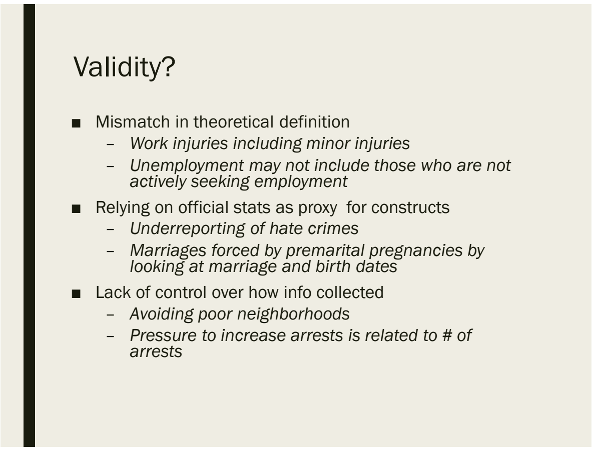
Existing Statistics/Secondary Analysis

- Appropriate Topics
  - *Involve info from large bureaucratic organizations*
  - *Variables defined by larger organizations*
- Social indicators
  - *Any measure of social well-being that can inform policy decisions*
  - *FBI's uniform crime index*
- Locating data
  - *Statistical Abstracts of the US*
- Secondary survey data
  - *General Social Survey (GSS)*



Existing Statistics/Secondary Analysis
Limitations

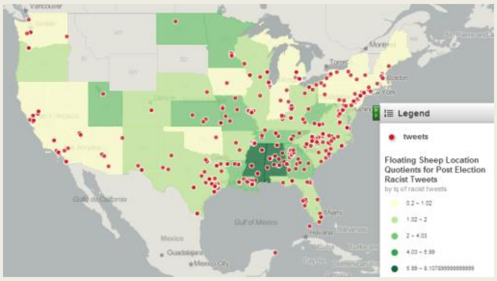
- Fallacy of misplaced concreteness
- Unit of analysis
  - *(potential for ecological fallacy – unemployed people more likely to commit property crimes?)*
- Reliability
  - *Equivalence – different states/regions – different levels of record-keeping*
  - *Representative*
    - E.g., change in asking women about "keeping house" when calculating unemployment
- Missing data



Validity?

- Mismatch in theoretical definition
  - *Work injuries including minor injuries*
  - *Unemployment may not include those who are not actively seeking employment*
- Relying on official stats as proxy for constructs
  - *Underreporting of hate crimes*
  - *Marriages forced by premarital pregnancies by looking at marriage and birth dates*
- Lack of control over how info collected
  - *Avoiding poor neighborhoods*
  - *Pressure to increase arrests is related to # of arrests*

Confidentiality



Content Analysis

- A set of procedures for converting textual information to numerical data
- Make inferences by systematically, objectively identifying specified characteristics of messages
- The goal is to classify words, phrases, or other units of text into a limited number of meaningful categories or to rate those units of text on specified dimensions.
- Non-reactive
- Useful when:
  - *Large amount of text*
  - *Scattered content*
  - *Content difficult to see with casual observation*

Content Analysis

A technique for making inferences by systematically and objectively identifying specific message characteristics

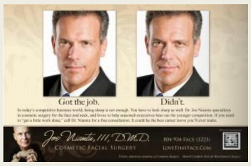
- Messages often referred to as *texts*
  - *Text = anything written, visual or spoken that serves as medium for communication*
- Used to analyze verbal, visual, audio-visual material
- High on naturalism
- Can be conducted with either quantitative or qualitative approaches

Content Analysis
Applications

- Describing attributes of messages
  - *Prevalent themes in Flex (White & Gillett, 1994)*
  - *Position reader as inferior (43%)*
  - *Promises of transformation (64.5%)*
  - *Muscular body as sign of hegemonic masculinity (70.6%)*

  - *Male muscularity served as symbol of male superiority and compensation for diminished privileges in other areas*

Content Analysis
Applications


Got the job.
Didn't.
In today's competitive business world, being sharp is not enough. You have to look sharp as well. Dr. Joe Niamtu specializes in cosmetic surgery for the face and neck, and loves to help seasoned executives beat out the younger competition. If you need to "get a little work done," call Dr. Niamtu for a free consultation. It could be the best career move you'll ever make.
Joe Niamtu, III, D.M.D.
Cosmetic Facial Surgery
804-934-FACE (3223)
LoveThatFace.com

- Calasanti (2007) study of ageism
- Examined websites
- Coded pictures and text into categories:
  - *Problems of aging*
  - *Solutions for problems*
  - *Gendered aspects of old age*
  - *Aspects of aging bodies on the site*
  - *Depictions of class, race and sexual orientation*
  - *Key message: if you can fix your body to forestall aging, you should*
  - *Ideal person was shown as white middle or upper class, heterosexual*
  - *Women shown as sexually alluring*

Content Analysis
Applications

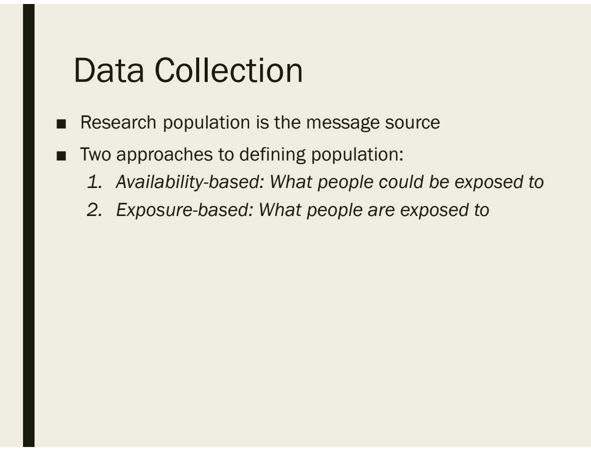
- Making inferences about sender of message, causes and antecedents
  - *Who says what to whom*
    - Attribute authorship of disputed papers
    - Madison as writer of the *Federalist Papers* (not Hamilton) based on word choice

Content Analysis
Applications

- Making inferences about effects of message on recipients
  - *Elements of culture and cultural change*
  - *McClelland – need for achievement*
    - Desire to succeed, non-conforming, enjoys tasks that involve risk
    - Analyzed content of literature in different cultures
    - High proportion of individuals in culture predicts strong entrepreneurial class – society will grow in power and influence

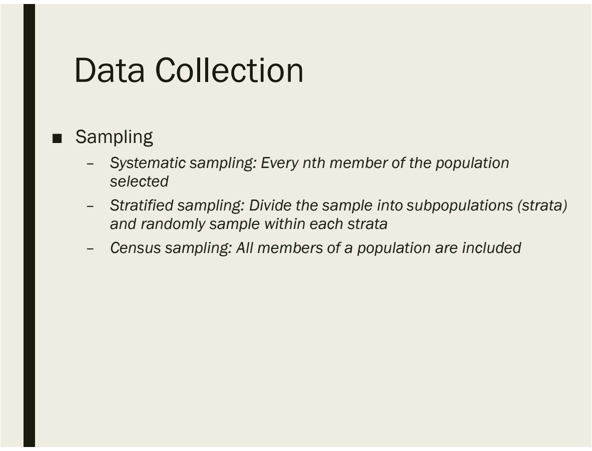
Developing Research Hypothesis for Content Analysis

- Investigate relationships between
  - *variables within each message*
  - *characteristics of the*
    - message source and message content
    - message content and message writers' characteristics
    - message content and message recipient's behavior



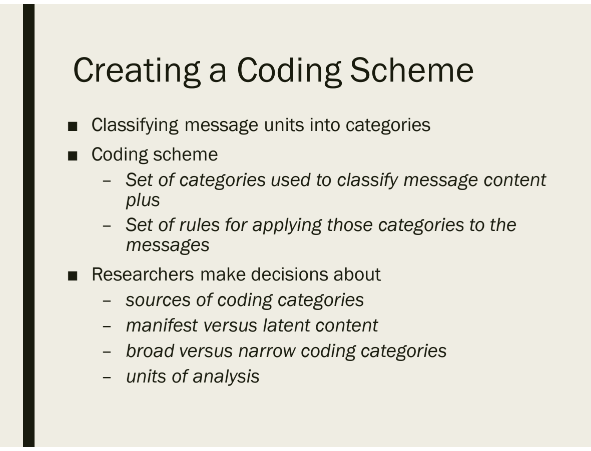
Data Collection

- Research population is the message source
- Two approaches to defining population:
  1. *Availability-based: What people could be exposed to*
  2. *Exposure-based: What people are exposed to*



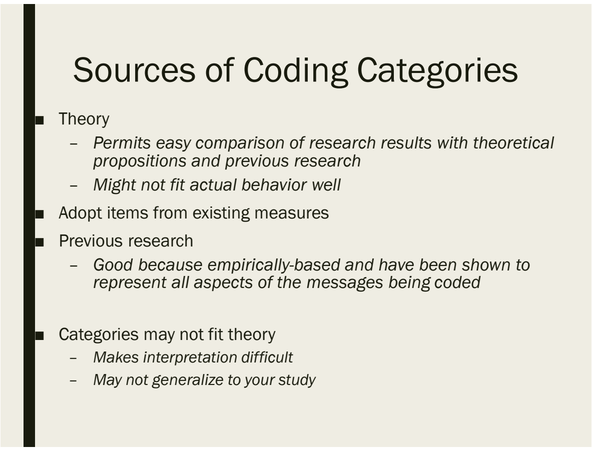
Data Collection

- Sampling
  - *Systematic sampling: Every nth member of the population selected*
  - *Stratified sampling: Divide the sample into subpopulations (strata) and randomly sample within each strata*
  - *Census sampling: All members of a population are included*



Creating a Coding Scheme

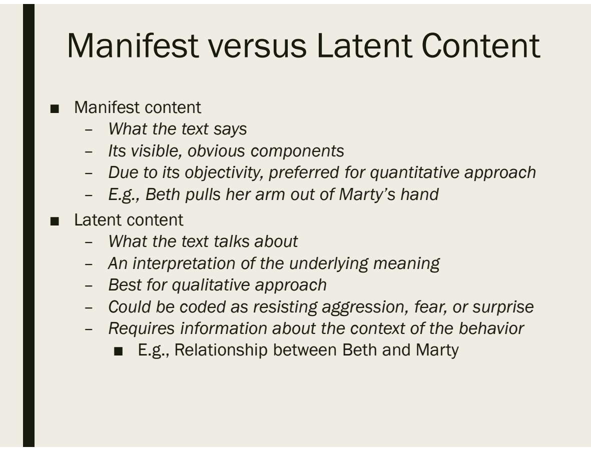
- Classifying message units into categories
- Coding scheme
  - *Set of categories used to classify message content plus*
  - *Set of rules for applying those categories to the messages*
- Researchers make decisions about
  - *sources of coding categories*
  - *manifest versus latent content*
  - *broad versus narrow coding categories*
  - *units of analysis*



Sources of Coding Categories

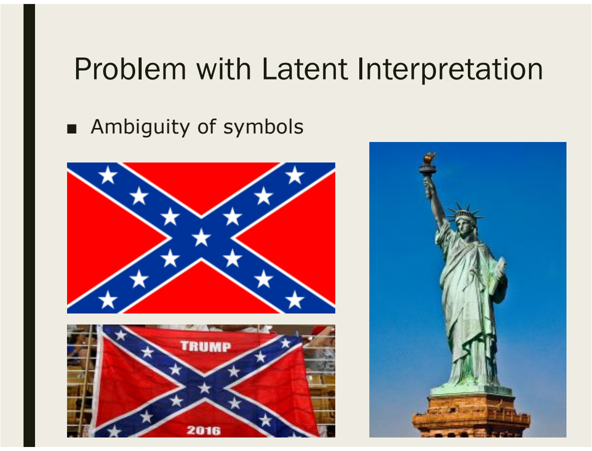
- Theory
  - *Permits easy comparison of research results with theoretical propositions and previous research*
  - *Might not fit actual behavior well*
- Adopt items from existing measures
- Previous research
  - *Good because empirically-based and have been shown to represent all aspects of the messages being coded*

- Categories may not fit theory
  - *Makes interpretation difficult*
  - *May not generalize to your study*



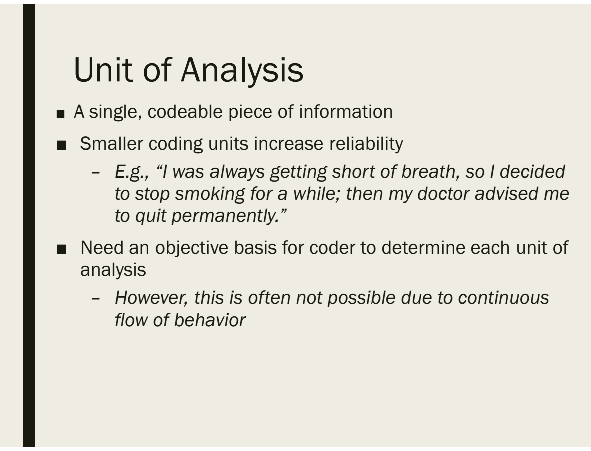
Manifest versus Latent Content

- Manifest content
  - *What the text says*
  - *Its visible, obvious components*
  - *Due to its objectivity, preferred for quantitative approach*
  - *E.g., Beth pulls her arm out of Marty's hand*
- Latent content
  - *What the text talks about*
  - *An interpretation of the underlying meaning*
  - *Best for qualitative approach*
  - *Could be coded as resisting aggression, fear, or surprise*
  - *Requires information about the context of the behavior*
    - E.g., Relationship between Beth and Marty



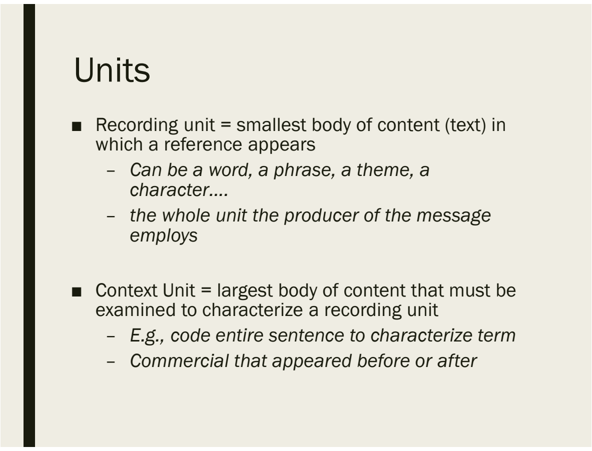
Problem with Latent Interpretation

- Ambiguity of symbols



Unit of Analysis

- A single, codeable piece of information
- Smaller coding units increase reliability
  - *E.g., "I was always getting short of breath, so I decided to stop smoking for a while; then my doctor advised me to quit permanently."*
- Need an objective basis for coder to determine each unit of analysis
  - *However, this is often not possible due to continuous flow of behavior*



Units

- Recording unit = smallest body of content (text) in which a reference appears
  - *Can be a word, a phrase, a theme, a character….*
  - *the whole unit the producer of the message employs*
- Context Unit = largest body of content that must be examined to characterize a recording unit
  - *E.g., code entire sentence to characterize term*
  - *Commercial that appeared before or after*

5 systems of Enumeration to Quantify Content

- Time-space
- Appearance
- Frequency
- Intensity
  - *Attitudes and values*
- Direction
  - *positive or negative messages, supportive or opposing*

Categories

- Recording units combined and coded into categories
  - *What is said*
  - *How it is said*
- All terms must be clearly and unambiguously defined
- Each behavior must fit into one (exhaustive) and only one category (mutually exclusive)
  - *Overlap in categories lowers reliability*
  - *At first stage, too many categories are better than too few*
  - *Rarely used categories can be collapsed together or placed in "Other"*
    - "Other" category should be least frequent

Broad versus Narrow Coding Categories

Broad Codes	Narrow Codes
Expresses anger	Glares at other person Glares at location in the room Looms over other person
Expresses sorrow	Wipes tears from eyes Blinks back tears Face is tight

- Examples of types of coding categories
- Hierarchical coding system
  - *Lower-level coding categories nested within higher level categories*

Broad versus Narrow Coding Categories

- Broad:
  - *Used to categorize:*
    - all behaviors found in a situation
    - all information in a set of messages
  - *Emphasizes comprehensiveness*
    - Need a few high-level categories
    - Aim for high reliability
  - *More useful if little already known about topic*
- Narrow:
  - *Used to categorize a subset of behaviors or messages*
  - *Emphasis is on details*
    - Larger number of categories allow finer distinction between similar behaviors
  - *Coders make more decisions during coding, so reliability may be lower*

Coder Qualifications and Training

- Reliability and validity affected by
  - *coder qualifications and training*
  - *the coding process*
- Qualified coders
  - *can understand the method and coding system*
  - *are conscientious*
  - *maintain consistency*
  - *are familiar with the cultural, social, and intellectual context of the messages*

Two Components of Coder Training

- Explanation stage: Researchers give a comprehensive explanation and discussion of the scheme and how to apply it
- Group practice sessions: Coders code same material
  - *Continue until each coder reaches predetermined level of accuracy*

The Coding Process

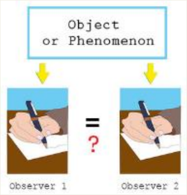
- Coders' decisions must be independent
- After reliability assessment is completed
  - *coders discuss disagreements*
  - *reach consensus about proper coding of disputed units*
  - *if the disagreement cannot be resolved, data unit is classified as "uncodeable"*
  - *a large number of uncodable units indicates poorly designed coding scheme*

Assessing Reliability

- Intercoder (interrater) agreement
  - *Percentage agreement between coders corrected for the probability of chance agreement*
    - Cohen's kappa
    - Krippendorff's alpha (for nominal data)
      - $\alpha = 1\text{-}D_o/D_e$
      - D_o *= observed disagreement,* D_e *= observed disagreement one would expect by chance*
    - Lin's concordance coefficient (for ratio data)
    - Stability – code same content later without reviewing earlier coding

Assessing Reliability

- Assessed twice using independent samples
  - *Pilot testing the coding scheme*
    - Revise coding scheme and continue training until the acceptable level of intercoder agreement is obtained
  - *Report coefficient from final content analysis*



Assessing Validity

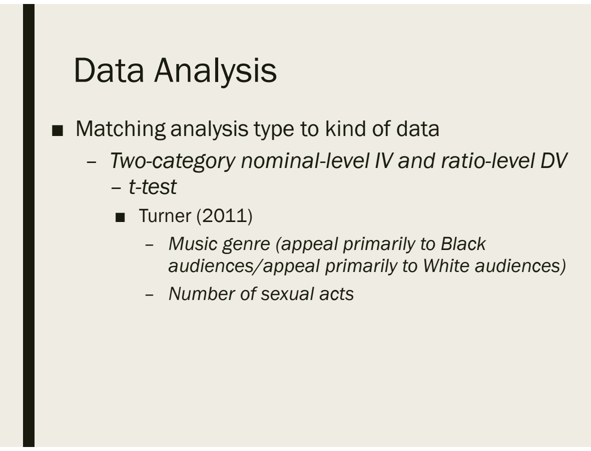
- Content validity most often applied
  - *Expert judgment of relevance and representativeness of coding categories*
    - Does it accurately reflect the concepts?
    - Are all relevant categories included in coding scheme?
  - *Outcome of the coding process*
    - Little overlap between categories (high intercoder agreement)
    - Few uncodeable responses

Reliability and Validity Highest When:

- system has broad rather than narrow focus
- coding system has small number of categories
- unit of analysis is objectively defined
- system focuses on manifest rather than latent content

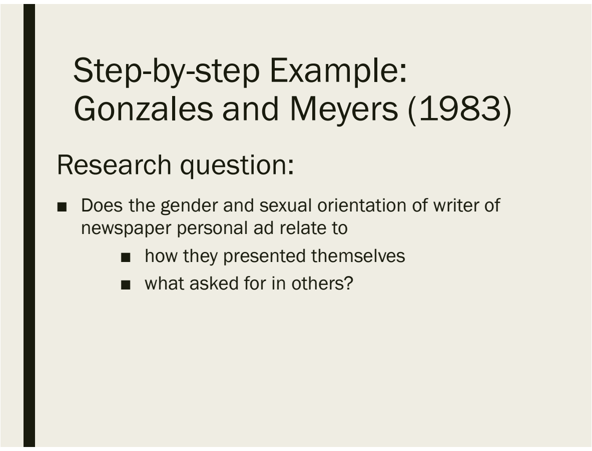
Data Analysis

- Same statistical procedures as other quantitative analyses
- Matching analysis type to kind of data
  - *Two nominal level variables – chi-square*
    - Lan and Russell (1980)
      - *Game outcome (win or lose)*
      - *Type of explanation provided (aspect of self or aspect of situation)*



Data Analysis

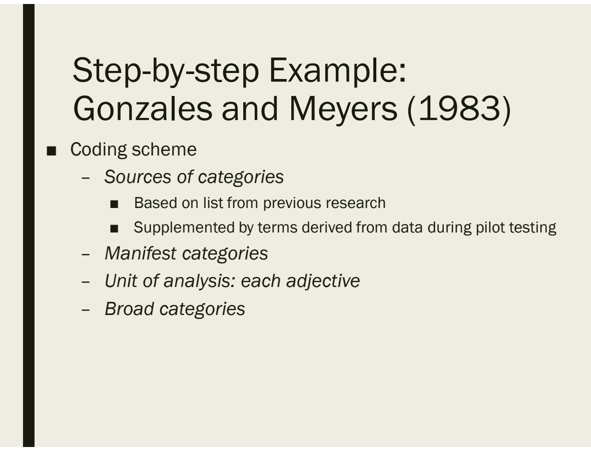
- Matching analysis type to kind of data
  - *Two-category nominal-level IV and ratio-level DV – t-test*
    - Turner (2011)
      - *Music genre (appeal primarily to Black audiences/appeal primarily to White audiences)*
      - *Number of sexual acts*



Step-by-step Example: Gonzales and Meyers (1983)

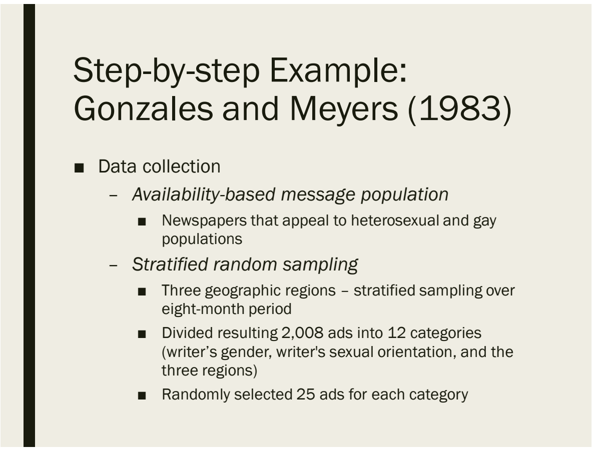
Research question:

- Does the gender and sexual orientation of writer of newspaper personal ad relate to
  - how they presented themselves
  - what asked for in others?



Step-by-step Example: Gonzales and Meyers (1983)

- Coding scheme
  - *Sources of categories*
    - Based on list from previous research
    - Supplemented by terms derived from data during pilot testing
  - *Manifest categories*
  - *Unit of analysis: each adjective*
  - *Broad categories*



Step-by-step Example: Gonzales and Meyers (1983)

- Data collection
  - *Availability-based message population*
    - Newspapers that appeal to heterosexual and gay populations
  - *Stratified random sampling*
    - Three geographic regions – stratified sampling over eight-month period
    - Divided resulting 2,008 ads into 12 categories (writer's gender, writer's sexual orientation, and the three regions)
    - Randomly selected 25 ads for each category

Step-by-step Example: Gonzales and Meyers (1983)

- Data coding
  - *Coder qualifications*
    - Unspecified
  - *Coder training*
    - Unspecified, intercoder reliability scores of 0.84–0.93 for each category indicates was sufficient
  - *Coding process*
    - Dichotomous
      - *Was category mentioned in ad? Yes/No*
    - Continuous
      - *Number of times each category mentioned in each ad*

Step-by-step Example: Gonzales and Meyers (1983)

- Data analysis
  - *Dichotomous – chi-square*
    - three-category nominal-level IV

      (gender, sexual orientation, gender X sexual orientation)

      and
    - six-category nominal-level DV

      (attractiveness, financial security, expressive traits, instrumental traits, sincerity, sexual references)

Step-by-step Example: Gonzales and Meyers (1983)

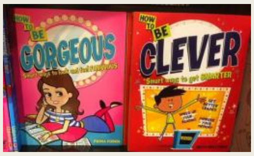
- Data Analysis
  - *Continuous – t-test for each category*
    - three-category nominal-level IV

      (gender, sexual orientation, gender X sexual orientation)

      and

      ratio-level DV for each of six categories

Inferences?



- Cannot determine truth or impact
- Children's books contain gender stereotypes – what does this mean?
- Advertisements regarding aging present a stereotype bias. What does this mean?

IMEDEEN

you can do more

to look as fabulous as you feel