Content Analysis and Archival Research

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  - Study historical documents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

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
    - Sharing of qualitative data more unusual and difficult
  - Inter-university Consortium for Political and Social Research (ICPSR)
    - 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
- 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
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
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.

Content Analysis

Applications

- Describing attributes of messages
- Who says what to whom
  - Text analyzed to make inferences about sender, causes and antecedents
    • Attribute authorship of disputed papers
    • Madison as writer of the Federalist Papers
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 recipient’s behavior
    - message content and message writers’ characteristics

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

Data Coding

- Classifying message units into categories
- Reliability and validity affected by
  - coder qualifications and training
  - the coding process
Creating a Coding Scheme

- 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
  - units of analysis
  - broad versus narrow coding categories

Sources of Coding Categories

- Theory
  - Permits easy comparison of research results with theoretical propositions and previous research
  - Might not fit actual behavior well

Sources of Coding Categories

- 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
- Need an objective basis for coder to determine each unit of analysis
  - However, this is often not possible due to continuous flow of behavior

Recording Units

- Words or terms
- Themes
- Characters
- Paragraphs
- Items = the whole unit the producer of the message employs

Content Analysis

- 5 systems of enumeration
  - Time-space
  - Appearance
  - Frequency
  - Intensity
    - Attitudes and values
  - Direction
    - positive or negative messages, supportive or opposing
Broad versus Narrow Coding Categories

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

<table>
<thead>
<tr>
<th>Broad Codes</th>
<th>Narrow Codes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Expresses anger</td>
<td>Glances at other person</td>
</tr>
<tr>
<td></td>
<td>Glances at location in the room</td>
</tr>
<tr>
<td></td>
<td>Stares over other person</td>
</tr>
<tr>
<td>Expresses sorrow</td>
<td>Wipes tears from eyes</td>
</tr>
<tr>
<td></td>
<td>Sticks back tears</td>
</tr>
<tr>
<td></td>
<td>Rest is tight</td>
</tr>
</tbody>
</table>

Broad Coding Categories

- 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 Coding Categories

- 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

Rules for Developing Coding Schemes

- All terms must be clearly and unambiguously defined
  - Definitions may be quite elaborate
- Each behavior must fit into one and only one category
  - Overlap in categories lowers reliability
Rules for Developing Coding Schemes

- There must be a category for every behavior
  - 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

Coder Qualifications and Training

- 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
  -oders discuss disagreements
  - reach consensus about proper coding of disputed units
  - If the disagreement cannot be resolved, data unit is classified as “uncodable”
  - 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)
    - Lin’s concordance coefficient (for ratio data)

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
- Unit of analysis is objectively defined
- Coding system has small number of categories
- 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)
  - *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


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?


- 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

- **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

- **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


- **Data analysis**
  - Dichotomous – chi-square
    - Three-category nominal-level IV (gender, sexual orientation, gender X sexual orientation)
    - Six-category nominal-level DV (attractiveness, financial security, expressive traits, instrumental traits, sincerity, sexual references)

- **Data Analysis**
  - Continuous – t-test for each category
    - Three-category nominal-level IV (gender, sexual orientation, gender X sexual orientation)
    - Ratio-level DV for each of six categories
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
■ Validity
  - Official stats as proxy for construct
  - Theoretical definitions don’t match
  - Lack of control
■ Reliability
  - Equivalence
  - Representative
■ Missing data

Inferences?

■ Children’s books contain gender stereotypes – what does this mean?
■ Advertisements regarding aging present a stereotype bias. What does this mean?