Research on Cognitive Aspects of Classification: Effects on Metadata Practice and Standards

Daniel Gillman
John Bosley
Scott Fricker

8 June 2012

BLS
BUREAU OF LABOR STATISTICS
U.S. DEPARTMENT OF LABOR

Background

- Cognitive Studies in Survey Methodology
  - Question stem
  - Not on response choices
  - Closed-ended questions
    - Classifications
      - Exhaustive
      - Mutually exclusive
    - Question answer choices
    - Communicate
      - Via concepts and categories

Research Question

- What if the concepts are confusing?
- Poor design => poor data
- For closed-ended response option sets,
  - How do cognitive / linguistic relationships
  - Affect respondents interpretation / selection?
- No common interpretive framework =>
  - Confusion =>
  - Poor quality data

Research Question

- Look for test case
  - US Current Population Survey (CPS) Class of Work question
- *"Now I have a few questions about the job at which you worked LAST WEEK.*
- *Were you employed by government, by a private company, a non-profit organization, or were you self employed (or working in the family business)?"*
- Important – Not analyzing CPS

Research Question

- Research on self-employment by
  - BLS
  - Small Business Administration
  - Other researchers
- Show inconsistency
  - SE often equated with IRS
    - Sole proprietorship
    - Schedule C filers
- No definition of self-employment

Study Design

- Hypothesis: SE is confusing
- Response choice variants (options)
  - COW (4 option):
    - Government (G)
    - Private (P)
    - Non-profit (N)
    - Self-employed (SE)
  - 3 option: G, P, N
  - 2 option: SE, not SE

Study Design

- 90 volunteers selected
  - Classify set of 12 vignettes twice
  - Set of follow-up questions
- Developed 20 basic job vignettes
  - Each – a description of a job
    - Primed toward P (10), N (5), or G (5)
    - Ambiguous toward SE (more on this later)

Study Design

- Asked experts to characterize SE
  - Financial control
  - Supervisory independence
    - Altered vignette – total of 100
- Half the volunteers given
  - 3 option (P, N, G) and 4 option (P, N, G, SE)
- Other half given
  - 3 option (P, N, G) and 2 option (SE, not SE)
- Order varied

Results

- All subjects had 3 set option
  - 1st pass results same as 2nd pass
  - 3//4 results same as 3//2
- All 3 option data can be combined
  - Know "truth"

Results

- Compare classification with intention
  - P (87%); G (73%); N (68%)
  - Misclassified
    - G: 22% to P 4% to N
    - N: 20% to P 12% to G
    - P: 6% to G 8% to N
- Bias towards P

Results

- Adding SE to the mix?
  - 4 option 38% to SE
  - 2 option 59% to SE
- By design
  - Basic (4); 2 each variants
  - Basics are ambiguous
  - => 50% should be SE

Results

- Real effects of variants
  - Compare against Basic
  - Use 4 option data
    - Bias toward SE added 3.1%
    - Bias away from SE subtracted 3.1%
- Effects in right direction,
- But unexpectedly small

Results

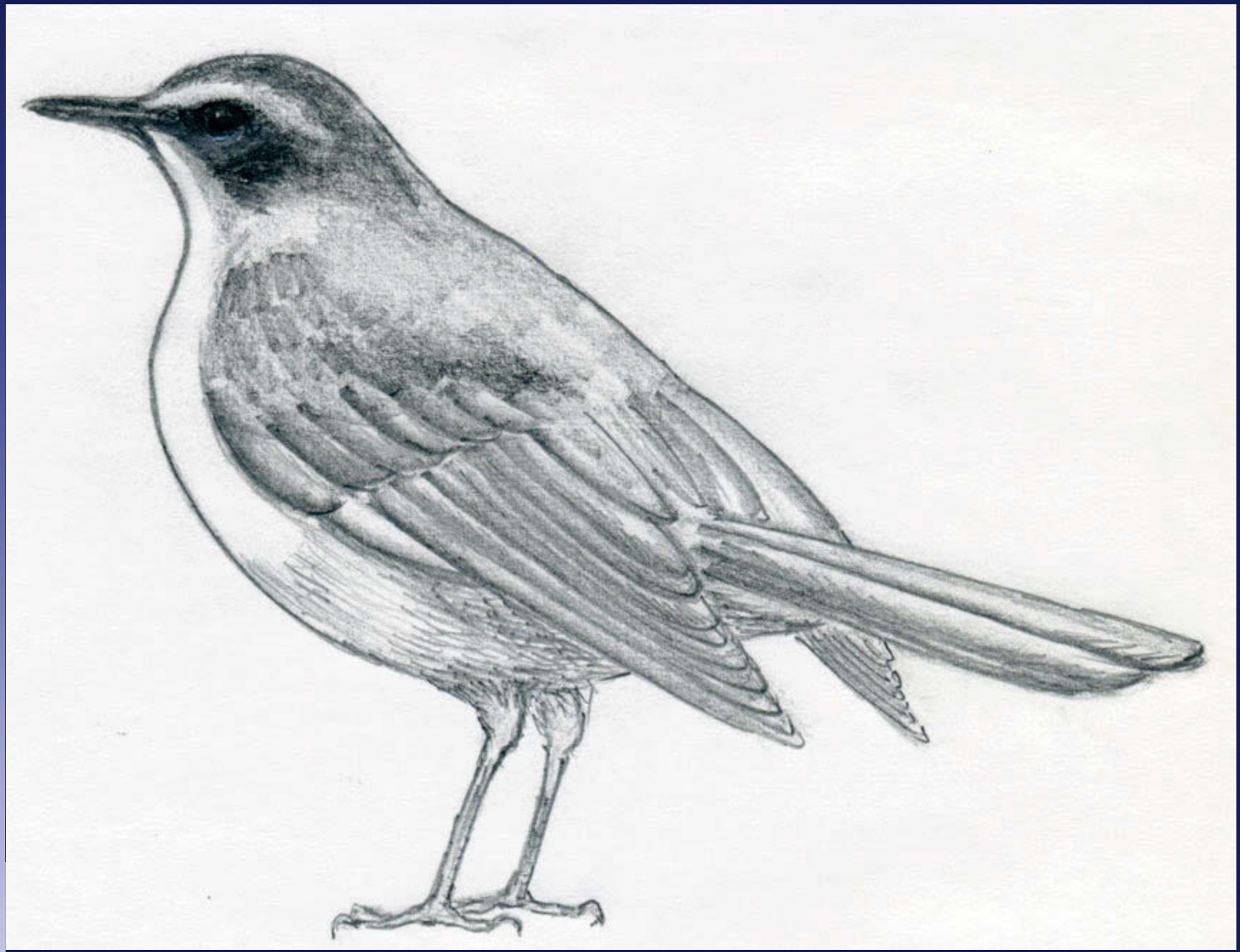
- Follow up questions
  - For each of G, P, N, and SE
    - Provide definition
    - Provide 3 examples
- Far easier for subjects to
  - Provide examples for G, P, N
  - Provide definition for SE
- Far harder for opposites

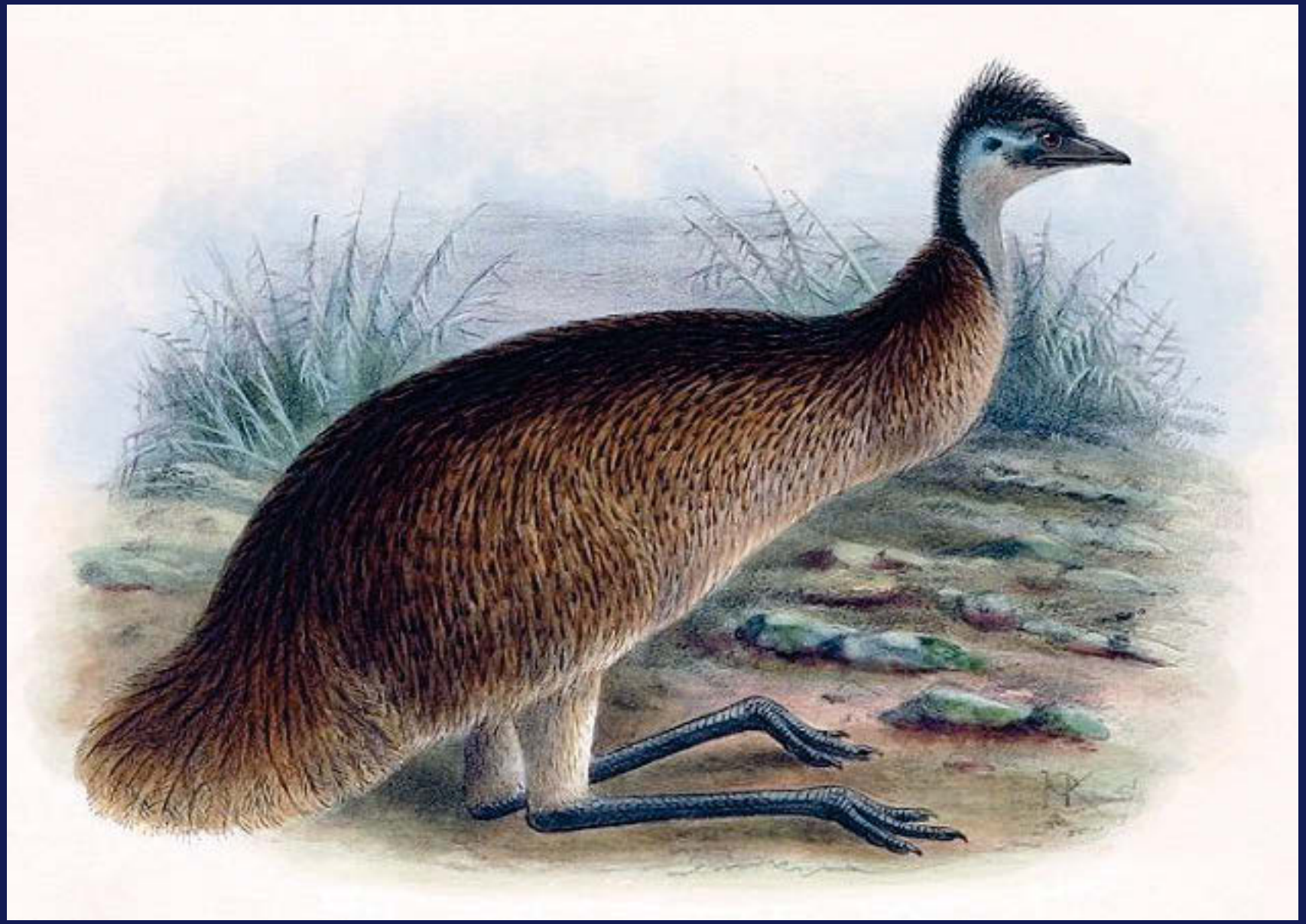
Interpretation

- Prototype theory
  - Rosch (1975) and many others
  - Lakoff (1992 – *Women Fire and Dangerous Things*)
  - Category membership based on
    - Similarity to prototype
    - Graded membership
  - Theoretically best example
- Example - Birds

Interpretation

- Graded membership
  - Typical birds
    - Robin, sparrow, etc
  - Less so
    - Hawk, goose, owl, etc
  - Even less
    - Turkey, chicken, etc
  - Weird
    - Ostrich, emu, penguin, etc





Interpretation

- Prototype theory
- Accounts for
  - 3 option
    - Significant misclassification to P
  - 4 option
    - Lower than expected SE choice
  - 4 option versus 2 option
    - Wide difference in SE selection

Interpretation

- Doesn't account for
  - Low impact of Fin+- and Sup+-
- Relational & Entity Categories
  - Gentner and Kurtz (2005)
- Entity categories
  - Characteristic based
  - Examples
    - Person, Bird, Television, Couch

Interpretation

- Relational categories
  - Role based
  - Not characteristic based
  - Examples
    - Student, Parent, Robbery, Errand
- All concepts have qualities of each

Interpretation

- Results on follow up questions
  - P, N, G behave like entity categories
    - Examples easy
    - Definitions hard
  - SE behaves like relational category
    - Examples hard
    - Definitions easy
- Conclusion
  - SE a much harder concept cognitively

Consequences

- Understanding data concepts
  - Poverty
  - Employment
  - Health well being
  - Education
  - Hard to measure
  - Not based on characteristics
  - Relational

Consequences

- Design
  - Classifications
    - Analyses
  - Response choices
    - Useful answers
- Researcher questions
  - Based on relational concept?
    - Then hard to measure

Consequences

- General
  - Concepts not all based on characteristics
  - Categories not sets (in the math sense)
    - Impact on semantic web?
  - Data have inherent measurement error
- Results here preliminary

Contact Information

Dan Gillman
Information Scientist

202-691-7523
Gillman.Daniel@bls.gov

BLS
BUREAU OF LABOR STATISTICS
U.S. DEPARTMENT OF LABOR