

# Quantifying data entry as a source of error in survey research

Jonathan Rubright, MS

University of Delaware Education Research and  
Development Center

Presented at the Eastern Evaluation Research Society  
annual meeting, Galloway, NJ, 20 April, 2009

# Despite being in the digital age . . .

- Most of us still enter data from surveys by hand!
- We need to know . . .
  - Data entry error rates
  - Impact on conclusions
  - Ways to mitigate occurrence

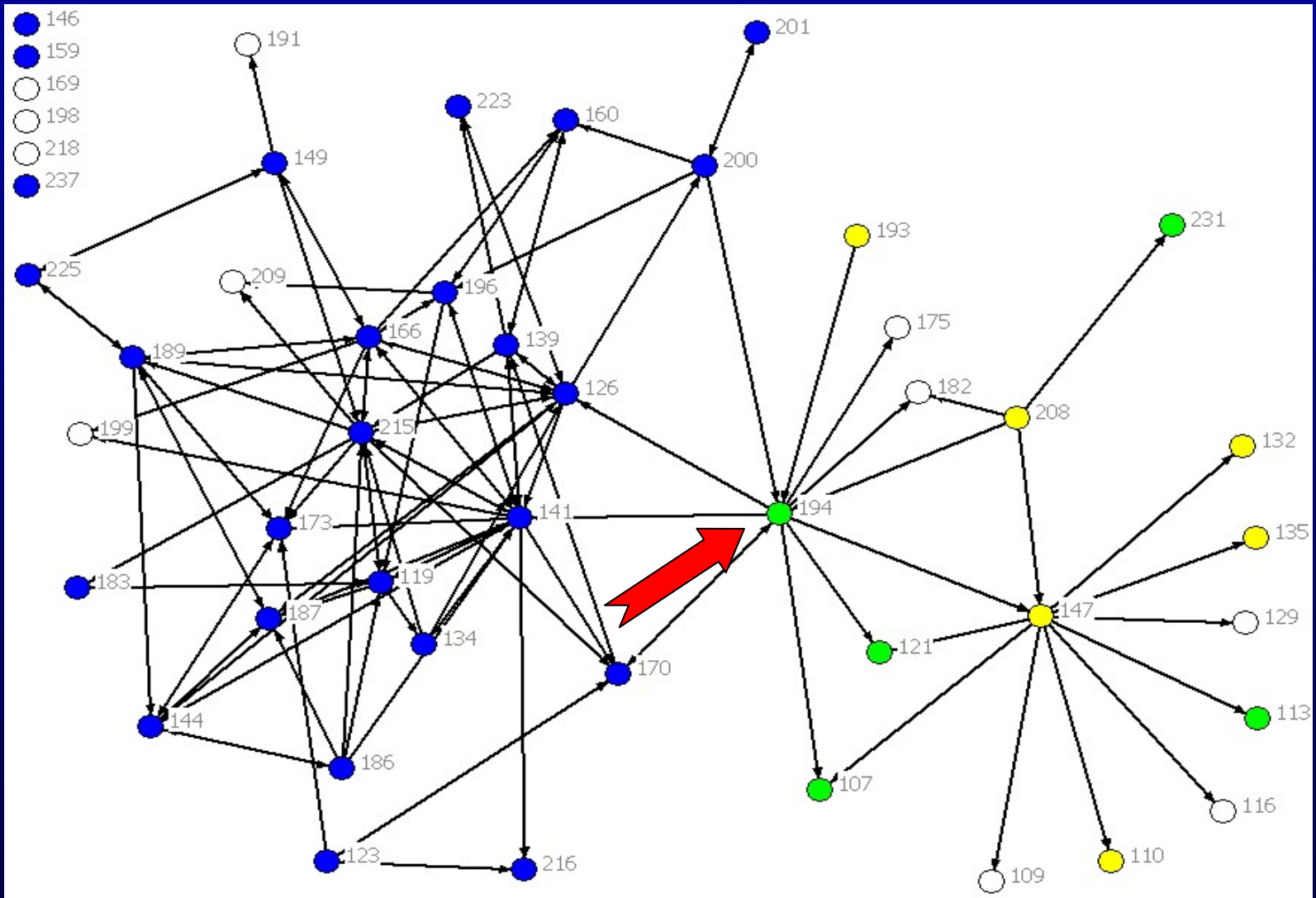
# Attempts to quantify error

- Total survey error
  - Coverage/sampling error
    - May be estimated based on the methodology used
      - Random sampling
      - Convenience sampling
      - Snowball sampling
  - Nonresponse error
    - Recruitment tracking allows some measure
  - Measurement error
    - Can be quantified
  - Editing/processing error?

# Data entry error

- Data entry quality assurance protocols
  - Range checks
    - Including database limitations
  - Stem and leaf plots
  - Read aloud methods
  - Double data entry (DDE)
    - “Independent rekey verification”
- Little data available on manually entered data error (Blumenstein, 1995; Gibson, Harvey, & Parmar, 1996; Gibson, Harvey, & Parmar, 1995; Kleinman, 2001; O’Rourke, 1996)

# Why care about data entry error?



# Performance practice

- Stand practice of 10% rule
- Not known:
  - If errors do exist, will any be found within the 10% sample?
  - How many errors will be accepted?
  - If a certain cut point of errors is reached, what steps will be taken next?

# Method

- NSF funded study of a Translational Research Institute
- 37 surveys entered by 2 research assistants into Excel (DDE)
- Checked for data cells which do not agree
- Count number of entry errors committed on each survey

Survey Number	Errors Student 1	Errors Student 2
1	0	0
2	0	2
3	0	0
4	0	1
5	0	1
6	0	0
7	0	1
8	0	0
9	0	0
10	0	1
11	0	0
12	0	0
13	0	0
14	0	0
15	0	0
16	1	0
17	0	4

# Results

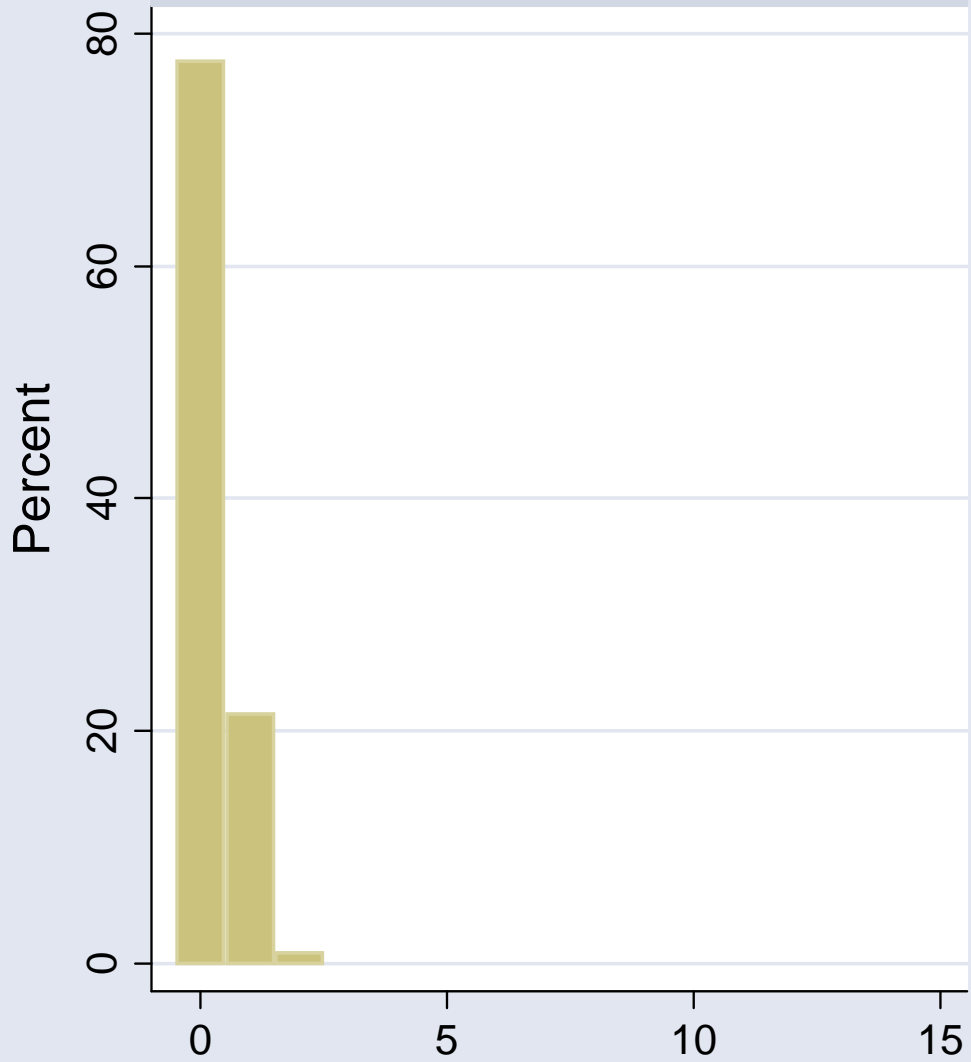
- 37 surveys each containing 168 datapoints
- 6,216 datapoints entered by each research assistant
  - Student 1: committed 2 errors, each on a different survey
    - $0.05 \pm 0.23$  (0-1) errors per survey
  - Student 2 committed 15 errors, with 7 on one survey
    - $0.84 \pm 1.5$  (0-7) errors per survey.

# Bootstrap

- To address whether errors would be located using a 10% rule:
  - Theoretically, there would be 40 chose 4  $[n!/(r!(n-r!))] = 91,390$  possible combinations of unique sample choices
  - Entry error variables sampled 1,000 times by bootstrap method to choose 10% of observations at random
  - Central limit theorem suggests we can approximate the distribution of errors by using a sample of these theoretical possibilities

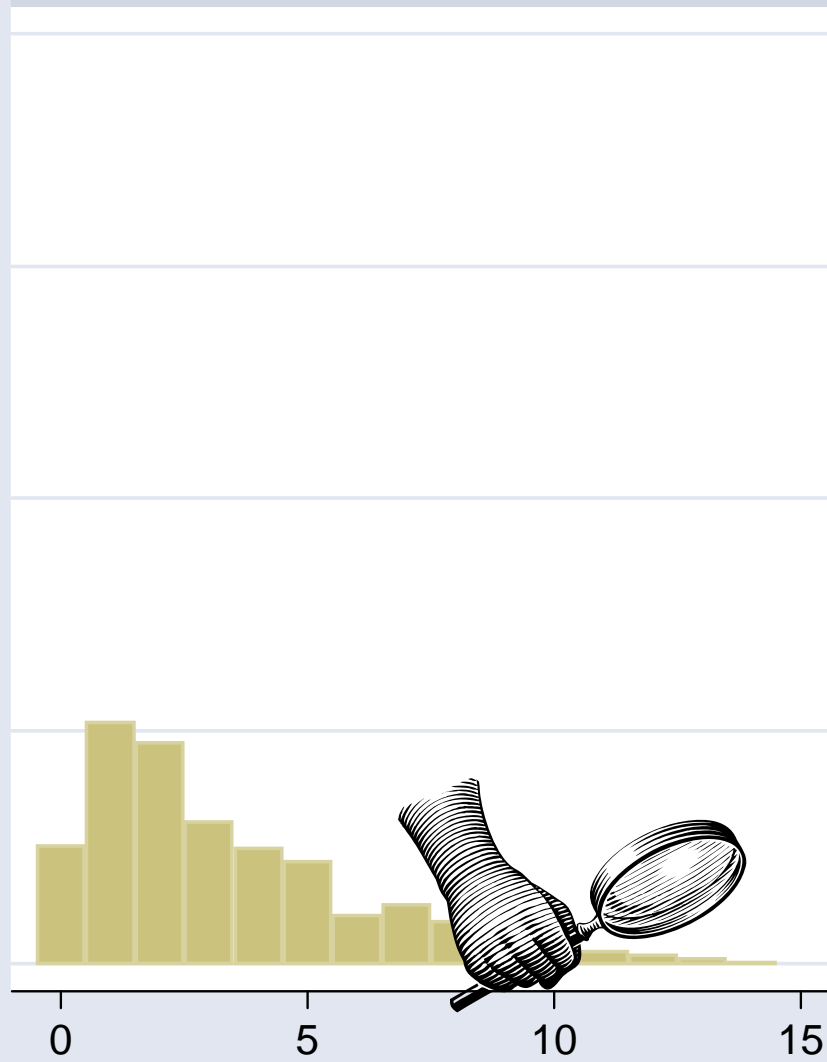
# Bootstrap: Sum of Errors

Subject 1



# Bootstrap: Sum of Errors

Subject 2



# Conclusions

- Small number of errors by student 1, most 10% samples show no errors and miss the errors committed
- Student 2 committed more errors
- We do not know where on the error distribution any one sample will fall
- Any 10% sample taken will likely miss some errors in the remaining data and may mislead evaluators into thinking data is relatively clean

# Conclusions: Part Deux

- Error distribution cannot be known beforehand
- Quality assurance protocols including double data entry should be utilized, especially when data quality requirements are high

# Thank you.

- Questions to [rubright@udel.edu](mailto:rubright@udel.edu).
- Special thanks to Steve Fifield, Katherine Centellas, and Leslie Cooksy for substantial contributions to intellectual content.



This material is based on work supported by the National Science Foundation under grant number SES-0724629 (Steve Fifield, PI).