# Systems and Processes for Assuring Data Quality

M. Rita Thissen

NE AAPOR

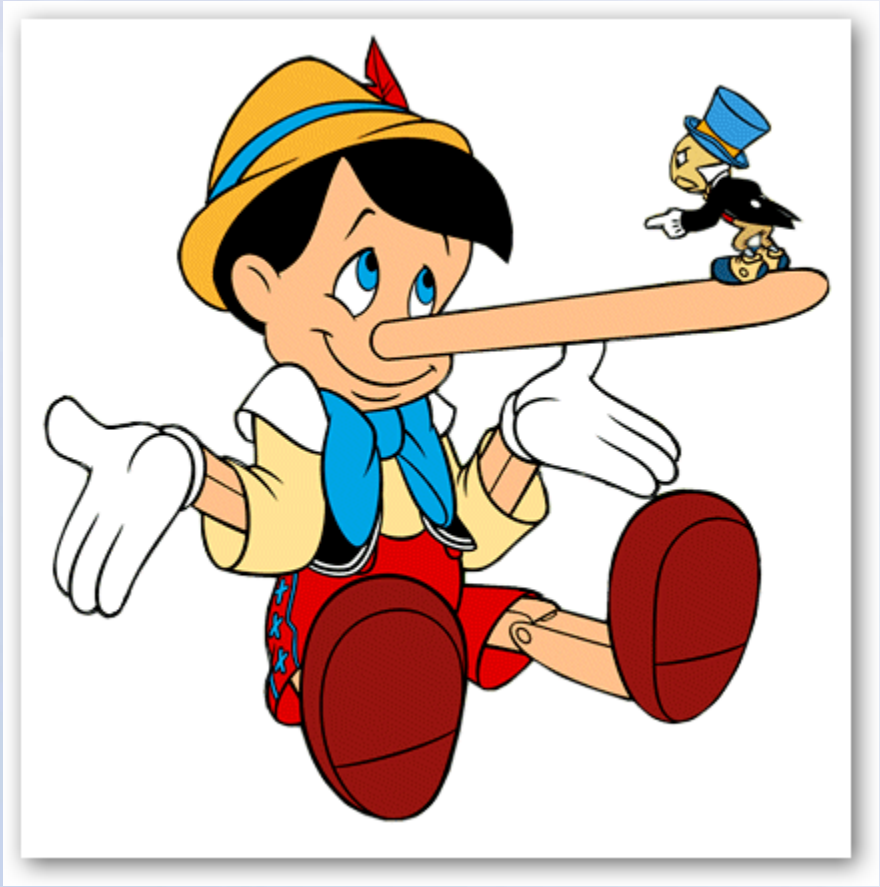
February 13, 2015

Cambridge, MA

# Overview of Topics

- Concepts of deterrence and detection of data falsification
- Survey modes and means of detection and deterrence
- Selecting the tools to use during data collection
  - Audio recording
  - Image capture
  - Geotagging
- Data review processes during and after data receipt
- RTI's practices
  - About the organization
  - Available tools for deterrence and detection of falsifiers
  - Examples from production surveys

# Falsification: Only Cute in Cartoons



http://allainjules.com/2012/05/29/syrie-dans-les-meandres-des-extremes-du-faux-et-de-la-falsification-historique/
http://movies.disney.com/pinocchio

# Deterrence and Detection of Falsification

- Deterrence during data collection
  - Psychological factors: Pride, self-image, reputation, engagement, fear of detection
  - Effort of falsifying vs. collecting valid data
- Detection during and after data collection
  - Observation by authority figures, peers, and respondents
  - Paradata review, including multimedia
  - Response data review, including multimedia
- Support from technology
  - Deterrence: Feedback with positive and negative comments raises awareness. Audio, photos and GPS are harder to fake.
  - Detection: Automated identification and tracking of suspicious situations

# Modes and Means: Methodological View*

- **Paper-and-pencil**: In-person administration by an interviewer or self-administration, followed by data entry
- **CATI** (computer-assisted telephone interviews): Inbound or outbound data collection by phone
- **CAPI** (computer-assisted personal interviews): In-person administration by an interviewer who is equipped with any electronic device
- **Specimen collection**: Physical or sensor data collection by an agent, may allow falsification

**Interviewer-mediated modes, with focus on personal interactions*

# Means and Modes: Systems View*

- **Paper**: Data entry systems, indirect compilation to a centralized database
- **CATI**: Voice-over-IP or analog phones, electronic instruments, desktop computers, computer audio-recorded interviewing (CARI), centralized database
- **CAPI 1**: Laptops with keyboards, electronic instruments, CARI, screenshots, perhaps photos or peripheral devices, direct transmission to a centralized database
- **CAPI 2**: Touchscreen, limited keying, tablets, smartphones, other handheld devices, sensors, direct transmission to a centralized database
- **Specimen collection**: Sensors, scanners, control/tracking systems, direct or indirect compilation of data

** Focus on hardware and software; the "means" determines options for detection*

# The Right Tool for the Job

- **Choice of means (US and abroad) depends in part on…**
  - Availability and consistency of electrical power
  - Access to broadband and/or internet for data transmission
  - Extent of telephone coverage, availability of numbers
  - Budget and logistics
  - Respondent consents and IRB restrictions
- **Choice of detection tools includes…**
  - Real-time checks of paradata and response data in electronic instruments, control/tracking systems and data entry software
  - CARI recording in CATI and all types of CAPI surveys
  - Recording location (geotagging) with most mobile devices
  - Capturing images (screenshots or photos) in most CAPI surveys
  - Collecting and managing verification re-contacting
  - Other sensor data in the future

# Data Capture Tools: Audio Recording

- Background
  - Audio recording has a long history within survey methodology and non-survey interviewing (such as journalism)
  - Digital recording was a “disruptor” technology
  - CARI was introduced by RTI in a production survey in 1999
- Purpose
  - Unarguable evidence of audio characteristics of the interview
  - Spot checks within the interview for authenticity and/or protocol
  - Retention for review, coding, comparison with keyed data, etc.
  - Retention of recordings for retrospective review if needed
- Pervasiveness
  - Used by many US and some non-US survey organizations
  - Standard at RTI

# Data Capture Tools: Images

- Background
  - Screenshots taken during interview
  - Photos taken with built-in cameras in laptops or mobile devices
  - Photos can be of the surroundings, perhaps not the respondent
  - Consent and IRB issues under some circumstances
- Purpose
  - Identification of respondents for follow-up contact (Indonesian study used photos to re-locate nomadic respondents)
  - Confirmation of address or location
- Pervasiveness
  - Not widespread yet
  - Offered on all mobile-device surveys at RTI, not used much yet

# Data Capture Tools: Geotagging

- Background
  - Global Positioning System (GPS) sensors built into many mobile devices, often augmented in smartphones via tower locations
  - Some variation in sensitivity and spurious readings
  - Best outdoors in two dimensions (i.e., not inside tall buildings)
- Purpose
  - Automated matchup and/or distance calculation
  - Geo-fencing
  - Review map for unexpected clustering or lack of clustering
- Pervasiveness
  - Many organizations experimenting
  - More common in rural areas and developing countries that lack western-style addressing
  - Offered on all mobile-device surveys at RTI, used by some

# Automated Data Review Processes

- Review of paradata
  - Potential value of keystroke timing (faster when falsified)
  - Review for internal consistency and completeness, including collected audio or images
  - Comparison with interviewer-population norms
  - Deduplication: ID fields, contact information, key-measure distances
  - Measures of satisficing (higher among falsifiers?)
- Review of response data
  - Non-conforming ID fields
  - Respondent-population outliers
  - Non-response levels
  - Mathematical distance between paired cases
  - Repeated values in open-text fields


*turning knowledge into practice*

# RTI International

***Headquartered in North Carolina, with offices around the world***

RTP, NC

Washington, DC

Rockville, MD

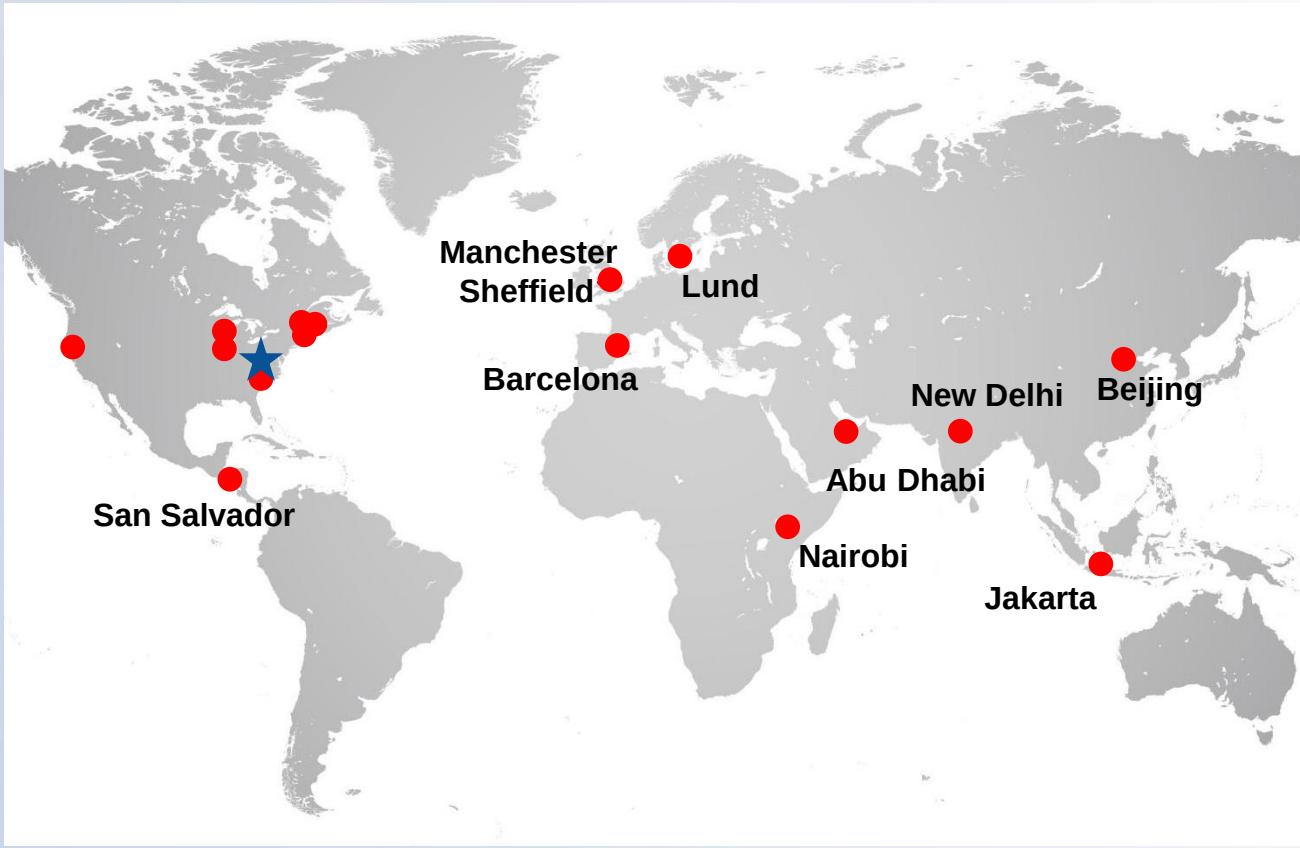
Atlanta, GA

Chicago, IL

Waltham, MA

San Francisco, CA

Ann Arbor, MI



# Survey Research and Services at RTI

- Data collection and management
- Study design
- Sampling
- Quality control
- Weighting and imputation
- Analysis
- 50+ years of experience
- Client-specific approaches as needed

# RTI's Approach: Multi-Sourced, Evidence-Based

- **Paper**: Pre-entry form review, double data entry or scanning with built-in checks and adjudication, post-entry data review
- **CATI**: Within-instrument checks, live monitoring, CARI, paradata review, response data review
- **CAPI 1** (laptop): Within-instrument checks, CARI, paradata review, response data review, verification calls
- **CAPI 2** (mobile): Within-instrument checks, CARI, image capture*, geotagging*, paradata review, response data review
- **Specimens**: Tracking, data review
- **All**: Personal supervision and quality discussions

** Not yet standard but used on some surveys*

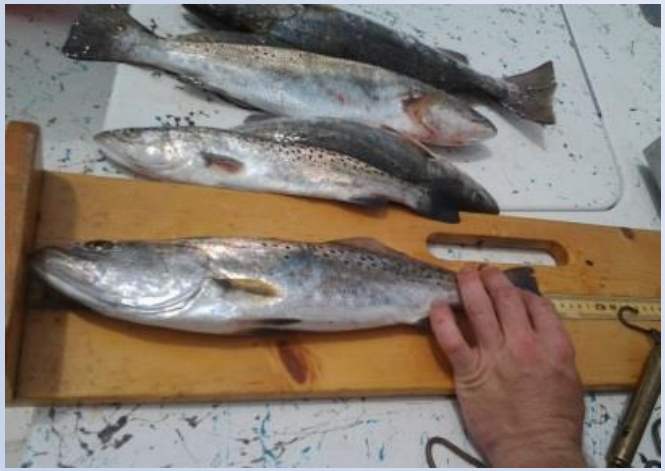
.

# Examples from RTI: CARI Indicators

- Key-clicks or single voices – several falsifiers detected
- Key-clicks prior to response but voices for a later question (shortcutting)
- Same voice for multiple respondents (a friend of the interviewer?)
- Lack of background noise in household interviews
- Household member telling the respondent what to answer
- Protocol violations caught in recording, including one episode of cosmetics sales and another interviewer asking the respondent not to tell about her behavior
- Low percentage of suspicious behavior (few cases per thousand), usually resolved as accidental or excusable situations
- High CARI-refusal rate is often a cause for concern (>15%)

# Examples from RTI: Image Use

- Image capture through built-in cameras
- Requires action by the data collector
- Doorstep address recording (pilot stage); high resolution photo from the street
- Fishing survey required a photo on location

# Examples from RTI: Geotagging

- Usage
  - Collecting GPS coordinates at each address during listing
  - Under investigation for use with doorstep screening
  - Automated capture on several mobile studies in the US and developing areas; capture rate looks adequate but too early for conclusions
- Automated comparisons
  - Collected coordinates can be mapped and displayed
  - Distances can be computed for geo-fencing
  - Coordinates can be matched to expected values (within a tolerance)
  - High missingness (if manual GPS) may be grounds for suspicion
- Visual inspection after mapping (in preliminary use)

# Examples from RTI: Paradata Review

- Outliers suggest problems, trigger investigation
  - Instrument section timers and hours/interview
  - Interview time-of-day
  - High CARI refusal rates
  - Excessive interview completion rates
  - Unusual expense reports (high costs = “needs money”)
- Verification calling (re-contacting a subsample)
  - Each name or phone number for verification contact should be unique
  - Unexpected responses to verification questions or respondent says there was no interview
  - 2011 study of one survey’s verification calls found inexperienced staff falsified more

# Examples from RTI: Response Data Review

- Outliers suggest problems, trigger investigation
  - High frequencies of “don’t know” or refusals (item non-response)
  - Response outliers compared to population values, such as the drug-using senior citizens of a rural state (an accidental error)
  - High rate of “no” on gateway questions (shortcutting)
- Case IDs that occur more than once or do not follow the pattern are rejected automatically by the system
- Checking for duplicates in re-contact information
- High invalid-number frequency in telephone numbers provided for verification calls

# Examples from RTI: Other Sources

- “Informants” among the other field staff and their associates (the jilted boyfriend)
- Paper SAQ forms with similar handwriting (all mailed together from one interviewer)
- In-depth visual inspection of response data: yes-no-yes-no-yes-no or other odd patterns
- Unusually high infection rate in urine samples from the same interviewer

# Thank You for Listening

Questions? Comments?

Contact Information:

Rita Thissen
919-485-7728
rthissen@rti.org

Center for Technology Solutions, Research Computing Division
RTI International
Research Triangle Park, NC 27709
www.rti.org