Open Responses to Survey Questions: Coding by Humans

Fundamentals of Data Collection II

April 7, 2025

Frederick Conrad

Coding Open Responses

- Most data created by surveys are *closed*
  - categorical response options (e.g., “Strongly Agree,” “Bachelor's degree or higher,” “Never”)
- But not all survey responses are produced in this form
  - “What kind of work do you do, that is, what is your occupation?”
  - “In your own words, what are the most serious problems facing the country?”
- Responses to these questions are *open* and must be transformed, i.e., coded or classified, into closed format so they can be analyzed quantitatively
  - In some cases, they are only used to provide “richness” or quotations for the media as in political polls and, so, are not coded
  - But to be tallied (quantified) they must be coded

Coding and Classification: Outline

1. What are open responses?
2. Error in manually coded open responses
   - Coder idiosyncrasies and “coder effects”
   - Characteristics of the open response affecting coder agreement
3. Hybrid approaches:
   a) Semi-automated classification
   b) Coding during data collection
4. Open responses and survey mode

What are open responses?

- Numerical responses, for which no ranges provided, considered open
  - e.g., frequency of events such as number of doctor visits in last month
  - If responses assigned to a range, e.g., 18-34 years old, no coding needed
- Verbal/linguistic responses (*R*s own words) generally need coding
  - as speech that is coded by *Iwer* in real time, (2) transcribed by *Iwer* for later coding, (3) audio-recorded for later transcription and/or coding
  - as text, i.e., by directly writing on a paper questionnaire or typing into an online questionnaire, both for later coding
- Graphics or photos occasionally collected and require codng

Why Ask Open Questions?

1. Goals of project require more categories with more complex structure than is possible to present as closed options

- While people may not be able to identify their job in a detailed taxonomy, they can describe what they do
- Example: *R*'s occupation
  - e.g., allows tracking which jobs are growing and how much they pay
- Current Population Survey asks two questions about *R*'s job
  1. What kind of work do you do, that is, what is your occupation?
  2. What are your usual activities or duties at this job?
- *R'*s spoken answers to both *Q*s transcribed by *Iwer* and coded by professional coders into one of >500 (Census Bureau) categories

US Census occupational classification system

- ~500 hierarchically structured categories: 11 major groups (e.g.,”Service Occupations”), 23 detailed groups (e.g., “Healthcare Support Occupations”), ~500 specific occupations (e.g., “Medical Assistants”)
- Three fields (“digits”) in numerical code reflect major group, detailed group, specific occupation, e.g., 31-90-94
- Unreasonable for *R*s to find their own occupation in this taxonomy, especially if title slightly different
  - e.g., to find “Medical Technician,” *R* would need to drill down 3 levels: Service Occupations, Healthcare Support Occupations, Medical Assistants
- or if parent groups not clear
  - e.g., “Massage Therapist” is instance in “Healthcare Support Occupations” but plausible in “Personal Care and Service Occupations”)

Why Ask Open Questions (2)

2. Developing closed questions
   - By initially asking *Q* in open form, possible to identify most frequently provided answers and present as options in a closed form of *Q* in future surveys/waves
   - Schuman & Presser (1981) asked about most pressing problem facing country in open or closed form: 22% of open responses concerned “energy shortage” (which had recently become a major issue); closed form did not include “energy shortage”

Why Ask Open Questions (3)

3. To give *R*s an opportunity to explain/think through a closed answer (Singer & Couper, 2017)
   - *R*s presumably more truthful when believe will not be misunderstood
   - Example: Krysan & Couper (2003) observed more negative/racist attitudes from white *R*s when *Q* asked by live vs. video-recorded *Iwer*s; *R*s indicated could explain their thinking to live *Iwer*
   - *R*s may be more positive when consider negative and positive beliefs
   - Example: Couper (2012) observed more positive views (in closed *Q*) toward Dutch immigrants when followed by open *Q;* suggested open *Q* led to deeper thinking about the topic

Coding and Measurement Error

- Open responses created for any of these reasons require coding, i.e., assignment to a category which can be tallied and used in statistical models, etc., just as any variable derived from closed responses might be used
- Although substantial progress continues to be made automating the process, still often relies – at least in part – on manual coding by human judges
  - Human-coded open responses can serve as training examples for ML models
  - Humans outperform automated coding for difficult responses
- But coders’ social and cognitive processes can contribute to a kind of measurement error much as is the case for respondents
  - Correlated error, variance due to coders
  - Characteristics of open responses

Coding: Reliability, Validity and True Value

- May not be possible to establish whether a code is “correct”
  - For a closed question, e.g., *Did you vote in last election? Yes or No*, a video recording or voting records could, in principle, validate a response
  - But there is no analogous process for assigning verbal description to a category
- Makes it hard to discuss validity and true value in coding
- Intercoder reliability is the typical measure of coding quality
  - easy to compute; proportion agreement ($\overline{P}$) vs. kappa (κ); > 2 coders (α)
  - can be high when validity is low, i.e., coders can agree with each other and be “wrong” compared to expert judgment
  - However, if there is disagreement at least one code is incorrect
- “Validity” usually measured as agreement with experts

Correlated Coder Error (2)

- If coders frequently assign different codes to open descriptions than their colleagues, this creates coding error analogous to *Iwer* effects
- Sturgis (2004) measured this, following statistical procedures used to measure interviewer effects, in UK Time Use Study:
  - Coders assign *R*s' descriptions of their activities in 10-minute intervals to activity code
  - ρ (*Rho*) or intraclass correlation is measure of variance due to coders
  - The more variance due to coders, the less precise the estimates, i.e., standard error of estimates is inflated, much like interviewer effects contribute to design effects
  - Coders (n=5) each classified descriptions in same 40 time-diaries
  - Extrapolated from these cases to full sample of ~21,000 diaries
  - Proportion agreement is high (92%) at one digit level but considerable variation in agreement at lower levels for certain times of day
  - Very large workload (~3000 diaries per coder) magnified small idiosyncrasies, leading to large inflation of standard error of estimates for each 10 min window

Impact of coder error on precision of estimates

Activity	Time of Day	Point Estimate	S.E. without coder error*	S.E. with coder error
Personal Care	08:20-08:30	52.4%	0.34	0.42
Employment	14:40-14:50	18.7%	0.27	0.40
Study	11:20–11:30	7.3%	0.18	0.30
Household/family care	16:50-17:00	19.5%	0.27	0.24
Voluntary work	15:30-15:40	2.1%	0.10	0.44
Social Life	22:30-22:40	3.5%	0.13	0.10
Sports/outdoor activities	13:20-13:30	12.0%	0.22	0.75
Hobbies and games	16:20-16:30	5.2%	0.15	0.52
Media Consumption	21:20-21:30	43.4%	0.34	0.61
Travel and Unspecified	16:00-16:10	16.6%	0.26	0.71

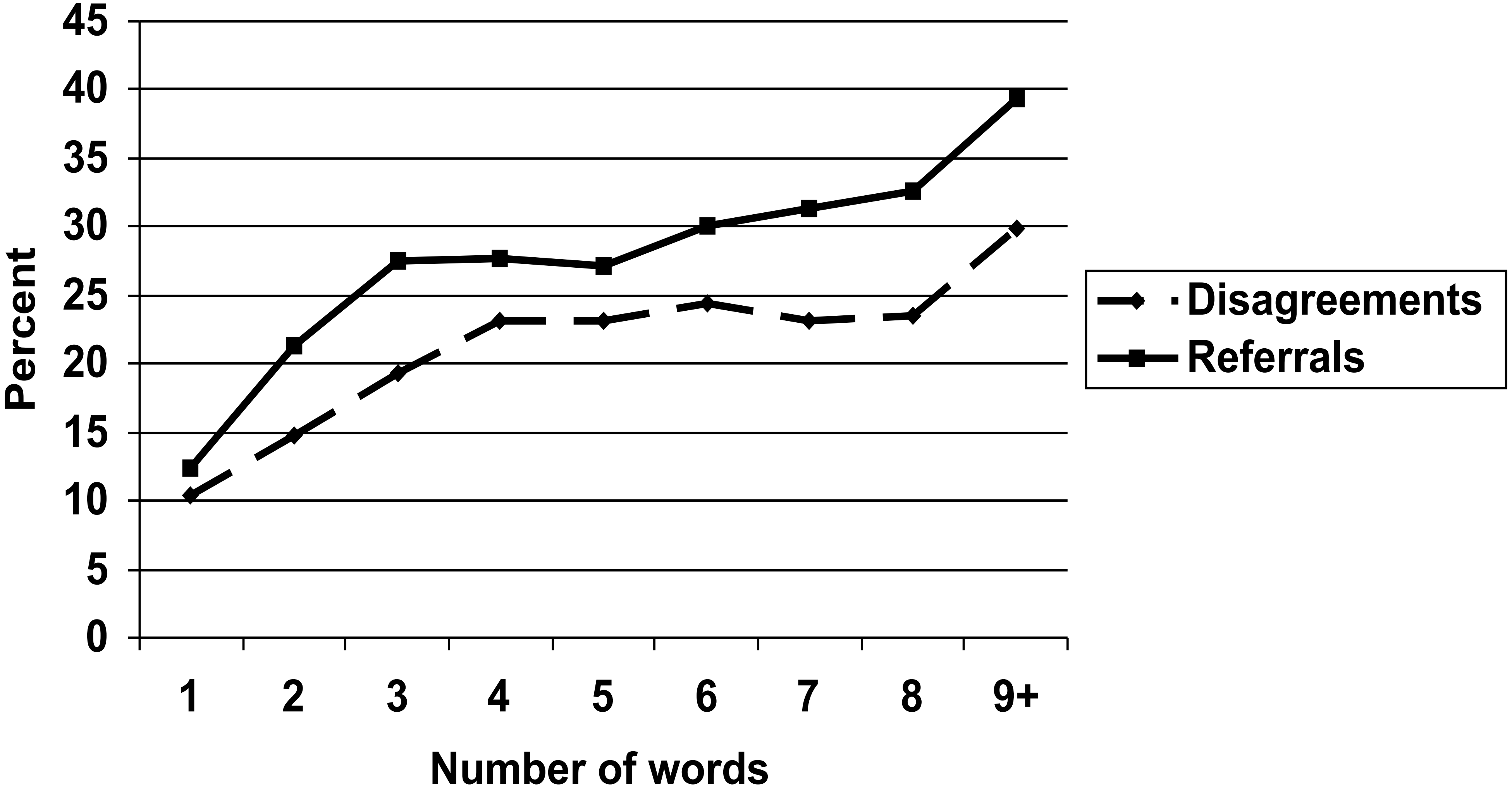
*assumes simple random sample and 0 correlated coder error

- inflated SEs, if ignored, could lead to overconfidence in estimates
- doubling # coders would have reduced variance inflation by 25-35%

Length of description

- Conrad, Couper & Sakshaug (2016) conducted 3-part study of CPS occupation coding
  - (1) What characteristics of occupation descriptions hurt coding reliability?
    - data set of double-coded descriptions (n=32,362) created for Quality Assurance (QA)
  - (2) How do characteristics jointly affect reliability?
    - double-coded experimental descriptions (n=800)
  - (3) What do coders think about while coding?
    - verbal reports while classifying experimental descriptions (n=100)
- Found in (1) that disagreement is higher for longer descriptions but in (2) effect of length depends on *difficulty* of terms in *R*s' descriptions; part (3) indicated that coders develop special purpose (unofficial) decision rules

Length of description and disagreement/referral rates



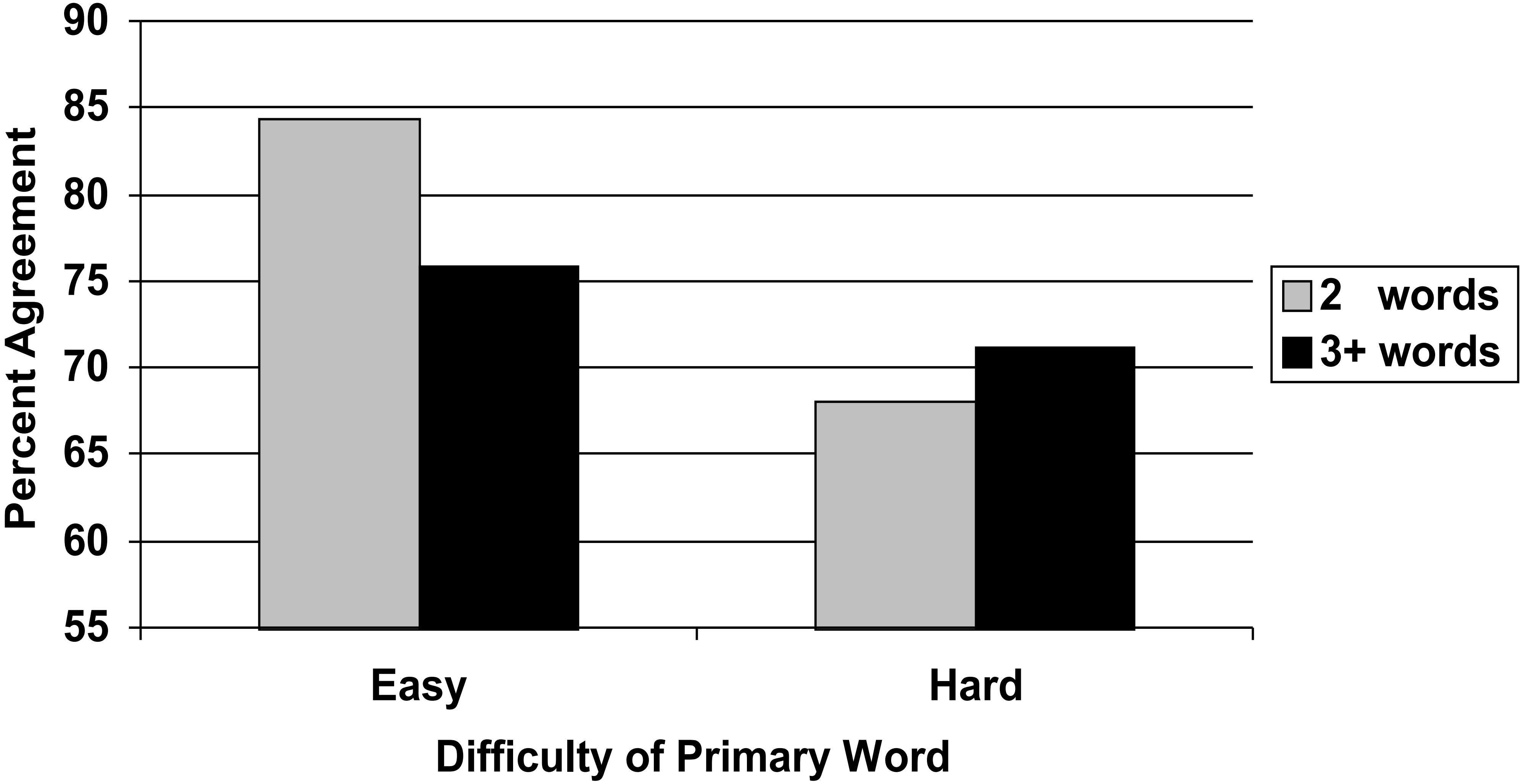
Joint Effects of Difficulty and Length

- Experimentally varied descriptions:
  - length (1, 2 and 3+ words)
  - difficulty of “primary” words (easy, hard)
    - easy=highest ratio agreement to disagreement in QA data:

      secretary, cashier, driver, cook, teacher, nurse, waitress, carpenter
    - hard=highest ratio disagreement to agreement in QA data:

      owner, operator, laborer, director, technician, clerk, supervisor, administrator
- Descriptions selected from QA corpus or crafted to be similar

Increased length lowers agreement when words are *easy*



Special purpose rules

- Coders (n=4) think aloud while coding *R*s' descriptions of their jobs
- All reported using informal, special purpose rules to promote agreement when uncertain what code to assign to a description
    - concerned superficial aspects of descriptions
    - not the definitions of coding categories, i.e., not based in theory
- Could be an acceptable practice when the description is ambiguous if the rules are applied consistently
- But, none of the coders could produce the rules in writing and were developed without supervisors' involvement, so likely used inconsistently

Special purpose rules:

Example More than one occupation

“When two different occupations are described, match to ‘duties’ and go with the first duty listed.”

```
EMP    : NEWBERRYS
IND    : VARIETY STORE
OCC    : CASHIER  STOCKING
DUTIES : STOCKING AND CASHIER
```

```
EMP    : DOMINOS PIZZA
IND    : PIZZA
OCC    : COOK, DRIVER
DUTIES : DELIVERY, COOKING
```

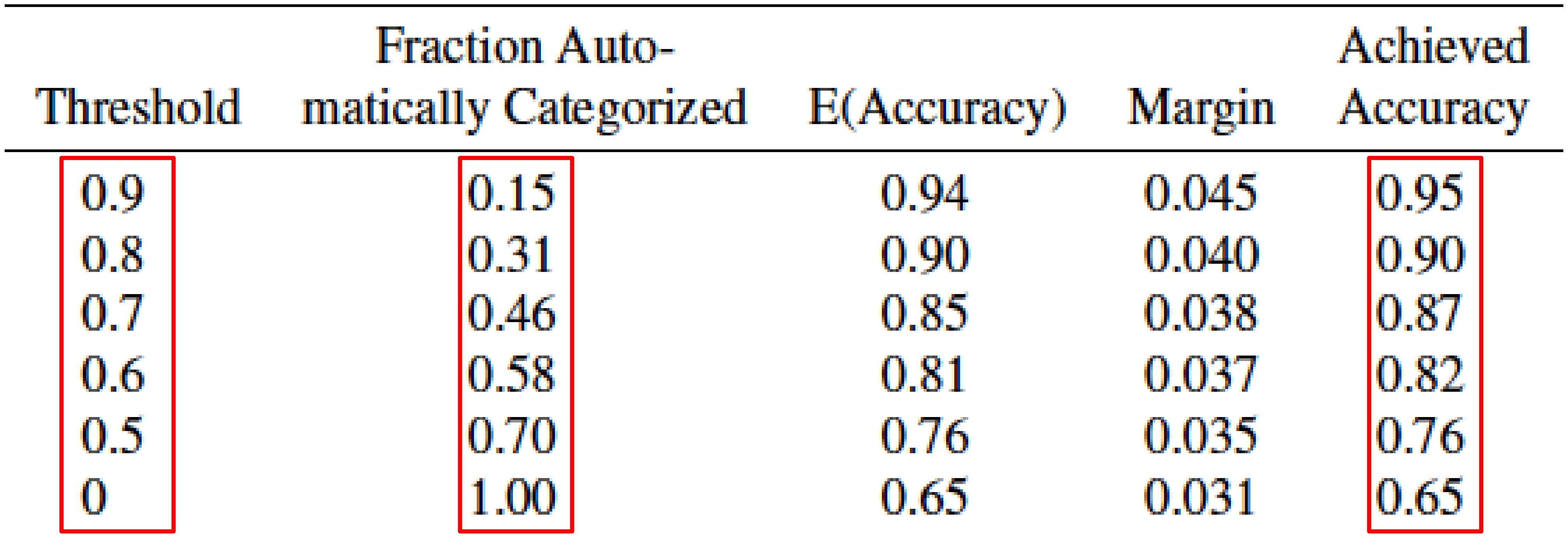
Semi-Automated Coding

- Human coding is resource intensive; automated coding is getting better but, for some open responses, accuracy is lower than needed
- Hybrid approach (Schonlau & Couper, 2016) may reduce required to conserve resources while maintaining acceptable levels of accuracy:
  - "easy-to-categorize" answers are classified automatically; "hard-to-categorize" answers classified by human coders
  - Algorithm trained on 500 randomly selected human-classified answers
  - Algorithm classifies answers in test set and predicts its accuracy for each
  - Algorithm required to classify test answers at level of accuracy determined by a researcher-selected threshold, e.g., 0.8
  - as threshold ↑, fraction answers automatically classified ↓

Semi-Automated Coding (2)

- Example (from Couper et al., 2008): *R*s presented with vignette of research study and asked if would participate; if “yes,” asked “Why would you participate?” and entered textual response
- n=1212; 20 categories of reasons for participation
- Data manually classified by two coders (κ=0.79); disagreements reconciled by an expert
- It was then possible to assess algorithm’s accuracy by computing agreement with manual codes

Semi-Automated Coding (3)



Threshold	Fraction Auto-matically Categorized	E(Accuracy)	Margin	Achieved Accuracy
0.9	0.15	0.94	0.045	0.95
0.8	0.31	0.90	0.040	0.90
0.7	0.46	0.85	0.038	0.87
0.6	0.58	0.81	0.037	0.82
0.5	0.70	0.76	0.035	0.76
0	1.00	0.65	0.031	0.65

- Tradeoff between accuracy and fraction categorized with autocoding
- S&C estimate that for n = 1000 test answers, semi-automated coding will save 14 hours; for 10,000 test answers will save 133 hours

Coding During Survey Data Collection

- Too many codes in a domain like occupation for *R*s to self-classify
- But if a *R'*s open-ended answer can be automatically assigned to the subset of categories to which it is most likely to belong, *R* can choose the category that fits best
  - Transforms original open-ended response task to close-ended task
- Shierholz, et al. (2018) proposed and tested this idea using a supervised machine learning algorithm to generate most likely categories in real time, i.e., during the interview

Coding During Survey Data Collection (2)

- Schierholz, et al. (2018)
  - Trained model on *R*s' open occupation description and their manually assigned codes from previous studies
  - Tested in ~1200 telephone interviews from which about 1064 open occupation descriptions were collected:
    - *Iwer* transcribed and submitted each occupation description to model
    - Model returned up to 5 possible occupation categories from which *R* selected one or "other occupation"
  - 72% of *R*s selected an occupation which largely agreed with human coders' judgments and without increasing interview duration

Open Responses and Survey Mode

- Long assumed that self-administered open questions will produce lower quality responses than when administered by *Iwer*
  - e.g., "Not only do some people find it more difficult to express themselves in writing than orally, but the absence of the interviewer's probes frequently results in answers that cannot be interpreted and sometimes in no answer at all." (Dillman, 1978)
- But until online data collection was common, "self-administered" meant hand-written responses on paper
- Entering text into online forms now ubiquitous; common wisdom about quality of open responses and self-administration should be revisited

Open responses in interviews and online

- Antoun & Presser (2024) compared open responses to three sets of questions in 2016 American National Election Study, administered by *Iwer* in person and self-administered online
  1. Most Important Problems facing US today
  2. Candidate For or Against, i.e., reasons to vote for/against Clinton, Trump
  3. Party likes-dislikes -- Democrats, Republicans
- Main take-away: open responses similar in both modes; historical preference for interviewer-administration may need to be re-evaluated

Open responses in interviews and online (2)

- Antoun & Presser report no differences between modes for
  1. amount of missing data (non-substantive answers)
  2. Substantive responses (codes), e.g., in most important problem Q, answer coded as “Government” given by 19.2% in person and 19.3% online
  3. Incivil responses about candidates; high (27.7%) but same in both modes
- But they report differences between modes for
  1. Number of words per answer: fewer online than in person
  2. Construct validity (prediction of who *R* voted for) based on number of likes and dislikes for the two candidates: predictions better online if favorable to Clinton
     - attribute to more words in person compromising quality of coding

Conclusions

- Open responses provide valuable data
- Automated methods promise to dramatically increase scale, speed and maybe accuracy of coding operations
- Coding and classifying open reports is prone to error whether by human coders or automated system
- Even as coding becomes more automated, human coding will likely be required to:
  - provide training data for statistical learning algorithms
  - code more difficult responses
- Improving the quality of open response data is still an under-explored methodological frontier