A Semi-Automated Nonresponse Detection for Surveys (SANDS) model for open-response data

Kristen Cibelli Hibben, PhD; Zachary Smith, PhD; Ben Rogers, MS; Valerie Ryan, PhD; Paul Scanlon, PhD; Travis Hoppe, PhD

Federal Committee on Statistical Methodology Research and Policy Conference
College Park, Maryland
October 24th, 2023

The findings and conclusions in this presentation are those of the authors and do not necessarily represent the official position of the National Center for Health Statistics, Centers for Disease Control and Prevention.
Outline

- Background and context
  - Open-text data: value and challenges
  - Item nonresponse detection: the technology and development of the model

- Evaluating the model
  - Against coded data or human review
  - Comparing performance across key subgroups to detect potential bias
  - Compared with word count and completion time

- How to access and use the model
Background and context
COVID-19 pandemic

- Numerous new COVID-19 related survey items
- Circumstances prevented our usual approach: in-depth cognitive interviewing to inform closed-ended online survey web probes
- Adapted and innovated our methods to include both closed and open-ended probes and experimental designs for post-hoc evaluations
Open-text data: value and challenges

- Range of methodological uses for open-text data (Singer & Couper, 2017)
- Allows for responses without constraint (Schonlau & Couper, 2016) a particular advantage when little is known about a topic (Neuert et al., 2021, Scanlon, 2019; 2020)
- But higher response burden, more prone to item nonresponse, inadequate and irrelevant responses
- Coding and analysis can be labor intensive and time-consuming
- Recent advances in data science offer new efficiencies and opportunities
Item nonresponse detection: prior work

- Traditionally viewed as absence v. presence of data (e.g., Groves et al., 2011)
- More nuanced for open-ends
  - “nonproductive” responses (Behr et al., 2012)
  - Indirect (soft) versus direct (hard) refusals (Meitinger et al., 2021)
  - “useful” versus “not useful” responses (Richards et al., 2022)
  - “problematic” versus “valid” responses (Trejo et al., 2022)
  - “sincere” versus “insincere” responses (Kennedy et al., 2021)
  - “Invalid” (versus valid) responses (Yeung and Fernandes, 2022)
- Ultimately context dependent and subjective (Neuert at al., 2021)
Prior work detecting item nonresponse

- **Rule-based approaches**
  - EvalAnswer* (Kaczmarek et al. (2017); available on GitHub)
    - Complete non-response: blank text box
    - No useful answer: “dfgjh”
    - Don’t knows: “I have no idea”; “DK”; “I can’t make up my mind”
    - Refusals: “no comment”; “see answer above”
    - Other: insufficient to code; “it depends”; “just do”; “just what it is”
    - Single word: “economy”
    - Too fast: < 2 seconds to answer
  - Rapid sensemaking (Etz et al., 2018)

- **Machine learning approaches**
  - Natural language processing (NLP) and bag-of-words to detect “invalid responses” (Yeung and Fernandes, 2022)

* [https://git.gesis.org/surveymethods/evalanswer](https://git.gesis.org/surveymethods/evalanswer)
Limitations of prior work

- EvalAnswer/rule-based approaches
  - Relies on regular expressions (regex)
  - Missed some gibberish and don't know responses: “I dunno”; “no clue”
  - Flagged single word responses that are valid: “quarantine”; “furloughed”; “closings”
  - Flagged valid responses that include one of the rules:
    - “I have not bee unable to travel to see my grandsons who live away from me. I am unsure how this country is going to fare.” [emphasis added]
  - Marked some non-response as valid:
    - “this is not a good question”; “I think my answer is self explanatory”
Limitations of prior work

- NLP/bag-of-words
  - Tends to work best on lengthier and cleaner pieces of text
  - Requires pre-processing and a project-specific training set
Item nonresponse detection: Model development

- Trained NLP model to interpret responses.
  - Fine-tuned a Bidirectional Transformer for Language Understanding (BERT)\(^*\) model using Simple Contrastive Sentence Embedding (SimCSE)\(^**\)
- Refined training via human coding (active learning)
- Semi-automated Nonresponse Detector (SANDS)

Item nonresponse detection: Model development, cont’d

Our working taxonomy:

- **Complete non-response**: Blank text box [Removed in pre-processing]
- **Gibberish** or nonsensical: “dfgjh”
- **Don’t knows**: “I don’t know”; DK; idk
- **Refusals**: “no comment”; “Because”; “none”
- **Other, high-risk**: non-useful response, non-codable
- **Valid**: useful response, codable

The model assigns a score (0-1) for the extent to which a response falls into each of the item non-response categories
Model development: Active learning

- **Round 1**
  - Sample of 3,200 was coded by team of 5 coders. Each researcher coded 1,400 responses: two groups of 600 responses and 200 responses coded by all 5 researchers
  - Good consistency with most categories (gibberish, DKs, refusals)
  - Less consistency between valid versus “other, high risk” item nonresponse
  - Good results for identifying item nonresponse, but flagged many valids as item NR

- **Round 2:**
  - 2 coders reviewed and arbitrated the results to retrain the model
  - Uncertainty retained in the model when warranted
Model evaluation: our approach
Data source

- NCHS’s Research and Development Survey (RANDS) During COVID-19
  https://www.cdc.gov/nchs/rands/index.htm
  - Three-round web/phone survey
  - Focused on health, impacts of pandemic, behaviors

- Conducted using NORC at the University of Chicago’s Amerispeak®, a probability-based panel representative of the US adult, English-speaking, non-institutionalized household population; Rounds 1 and 2 used the non-probability Dynata Panel™ to supplement

<table>
<thead>
<tr>
<th>Round</th>
<th>Complete responses</th>
<th>AmeriSpeak® sample</th>
<th>Dynata™ sample</th>
<th>Fielding dates</th>
<th>Weighted cumulative response rate</th>
<th>Completion rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>13,020</td>
<td>8,663</td>
<td>6,220</td>
<td>6/9/2020 – 7/6/2020</td>
<td>23.0%</td>
<td>78.5%</td>
</tr>
<tr>
<td>2</td>
<td>11,483</td>
<td>8,651</td>
<td>5,502</td>
<td>8/3/2020 – 8/20/2020</td>
<td>20.3%</td>
<td>69.1%</td>
</tr>
<tr>
<td>3</td>
<td>5,458</td>
<td>7,852</td>
<td>0</td>
<td>5/17/2021 – 6/30/2021</td>
<td>11.8%</td>
<td>69.5%</td>
</tr>
</tbody>
</table>
Model development process

Initial Model Coding

Model Training

Arbitrate Uncertainty

Phase 1: Two-probe Analysis

Phase 2: Four-probe Analysis

Phase 3: Model Bias Analysis

Phase 4: Word Count Analysis

Phase 5: Latency Analysis

Model Development

Model Evaluation
Evaluation results
Model evaluation: Phase 2

- Mixed-method evaluation of additional web probe case studies
  - Quarantine
  - Social distancing (new topic)
  - Vaccine hesitancy (new topic)
  - Religion (new topic)
Social distancing probe

- Social distancing survey questions:
  - In the last week, did you socially distance when you were…shopping, eating at a restaurant, etc. (total 7 randomized grid items)
  - [If yes, then] Did you do the following activities inside, outside, or both?
- Social distancing probe: When you were answering about social distancing in the previous questions, what were you thinking about?
### Phase 2 results: Social distancing probe

<table>
<thead>
<tr>
<th></th>
<th>Human-reviewed NR</th>
<th>Human-reviewed Valid</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model NR</td>
<td>450</td>
<td>177</td>
</tr>
<tr>
<td>Model Valid</td>
<td>109</td>
<td>3,876</td>
</tr>
<tr>
<td>Total</td>
<td>559</td>
<td>4,053</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4,612</td>
</tr>
</tbody>
</table>

**False valids (human-coded NR):**
- “Recent activity”
- “EVERYTHING”
- “Being normal”
- “Don’t do it as much”
- “Money”
- “I’m tired and I want to go to bed”

**False NR (human-coded valid):**
- “Safety” (and variations)
- “Save life”
- “lines in the market”
- “It is necessary but a pain.”
- “ Courtesy”
- “IT’S COMMON CERDICY AND GO WITH THE THROW”

**Sensitivity 81% (450/559)**

**Specificity 96% (3,876/4,053)**

**Key take-away:**
- Model did a good job identifying “true” valids; slightly less well identifying “true” item nonresponse.
# Phase 2 results: Additional probes

<table>
<thead>
<tr>
<th>Vaccine Hesitancy</th>
<th>Human-reviewed NR</th>
<th>Human-reviewed Valid</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model NR</td>
<td>151</td>
<td>492</td>
<td>643</td>
</tr>
<tr>
<td>Model Valid</td>
<td>61</td>
<td>4,266</td>
<td>4,327</td>
</tr>
<tr>
<td>Total</td>
<td>212</td>
<td>4,758</td>
<td>4,970</td>
</tr>
</tbody>
</table>

- **Sensitivity:** 71%
- **Specificity:** 90%

<table>
<thead>
<tr>
<th>Religion</th>
<th>Human-reviewed NR</th>
<th>Human-reviewed Valid</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model NR</td>
<td>298</td>
<td>952</td>
<td>1,250</td>
</tr>
<tr>
<td>Model Valid</td>
<td>36</td>
<td>2,314</td>
<td>2,350</td>
</tr>
<tr>
<td>Total</td>
<td>334</td>
<td>3,266</td>
<td>3,600</td>
</tr>
</tbody>
</table>

- **Sensitivity:** 90%
- **Specificity:** 71%

<table>
<thead>
<tr>
<th>Quarantine</th>
<th>Human-coded NR</th>
<th>Human-coded Valid</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model NR</td>
<td>863</td>
<td>239</td>
<td>1,102</td>
</tr>
<tr>
<td>Model Valid</td>
<td>325</td>
<td>4,778</td>
<td>5,103</td>
</tr>
<tr>
<td>Total</td>
<td>1,188</td>
<td>5,017</td>
<td>6,205</td>
</tr>
</tbody>
</table>

- **Sensitivity:** 73%
- **Specificity:** 95%
Proportions of model-coded item nonresponse

- Baseline rates of item nonresponse estimated at 10-20% (Neuert et al., 2021; Lenzer and Neuert, 2017; Meitinger and Behr, 2016)
- Religion: share of responses identified as nonresponse much higher than expected
  - Indicative of potential model difficulties
Distribution by type of item nonresponse

- Model error often concentrated in the High Risk category, as seen for Social Distancing
- More error seen in Refusals for Religion
- More error seen in Unsure for Vaccine Hesitancy
Phase 4: Word count analysis

SOURCE: National Center for Health Statistics Research and Development Survey During COVID-19, Rounds 1 and 3 (n = 34,561)
Phase 5: Latency analysis

SOURCE: National Center for Health Statistics Research and Development Survey During COVID-19, Round 1 (n = 6,377)
Further evaluation results

<table>
<thead>
<tr>
<th>Probe</th>
<th>Sensitivity</th>
<th>Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Over the past three months, what approaches did you use to manage your pain?</td>
<td>97%</td>
<td>89%</td>
</tr>
<tr>
<td>Why {do you/does PERSON} have difficulty doing errands alone?</td>
<td>100%</td>
<td>98%</td>
</tr>
<tr>
<td>When you answered the previous question about difficulty learning how to do things most people {your/their} age can learn, what were you thinking about?</td>
<td>82%</td>
<td>90%</td>
</tr>
<tr>
<td>What do you think the main reason is for these experiences?</td>
<td>88%</td>
<td>81%</td>
</tr>
<tr>
<td>When we asked you how often {you are...}, what were you thinking about?</td>
<td>84%</td>
<td>90%</td>
</tr>
<tr>
<td>What kind of instruction on how to say no to sex were you thinking about in the previous question?</td>
<td>73%</td>
<td>95%</td>
</tr>
<tr>
<td>Please list some things that you associate with being {GENDER}.</td>
<td>71%</td>
<td>90%</td>
</tr>
<tr>
<td>When answering the previous question, what symptoms were you specifically thinking about?</td>
<td>100%</td>
<td>99%</td>
</tr>
</tbody>
</table>

Data from NCHS’s RANDS, rounds 4, 6, and 7, fielded between 2020 and 2022.
Evaluation results summary

- Overall, evaluation results indicate that SANDS performs well in identifying a dataset of likely valid results.

- SANDS also appears to capture item nonresponse and valid responses with substantially more nuance than rule-based approaches (e.g., word/character count or response latency).
Model access and guidance
Model access

- SANDS is currently available for general use on Hugging Face: https://huggingface.co/NCHS/SANDS

- Use via the Hugging Face API or Python with the transformers library

- Model card includes examples, some knowledge of Python is needed

- More information available on NCHS’s site: https://www.cdc.gov/nchs/data-science/SANDS-model-context.htm
Guidance/Best practice tips

- Pre-process hard-coded nonresponse and blank responses
- Evaluate rate of nonresponse detected
- Always review “other, high-risk” responses
- Consider the construct captured by the probe
- Random sample the valid responses
Next steps

- SANDS 2.0: Can we give SANDS information on context and probe type?
- Data quality of open-ended text: is this data useful for question design?
Thank you!!

Questions/comments? Feel free to ask or email kciibelli@cdc.gov

Q-Bank: providing access to survey question evaluation reports, question design and performance https://wwwn.cdc.gov/qbank/

Q-Notes: designed to facilitate the management and analysis of cognitive interviews https://www.cdc.gov/nchs/ccqder/products/qnotes.htm

For more information, contact CDC
1-800-CDC-INFO (232-4636)

The findings and conclusions in this report are those of the authors and do not necessarily represent the official position of the Centers for Disease Control and Prevention.
References


References


The probes for evaluation phases 1 & 2

<table>
<thead>
<tr>
<th>Evaluation phase</th>
<th>Survey question(s)</th>
<th>Open-ended probe questions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phase 1</td>
<td>When do you think the Coronavirus pandemic began? When did the Coronavirus pandemic first affect your daily life?</td>
<td>Why do you say that?</td>
</tr>
<tr>
<td>Phase 1 &amp; 2</td>
<td>Have you isolated or quarantined yourself because of the Coronavirus?</td>
<td>When answering the previous question about isolating or quarantining because of the Coronavirus, what were you thinking about?</td>
</tr>
<tr>
<td>Phase 2</td>
<td>Overall, how hesitant about vaccines in general would you consider yourself to be?</td>
<td>Please list the reasons you say you [are/are not] hesitant about vaccines in general.</td>
</tr>
<tr>
<td></td>
<td>In the last week, did you socially distance when you were...</td>
<td>When you were answering about social distancing in the previous questions, what were you thinking about?</td>
</tr>
<tr>
<td></td>
<td>Currently, how important is religion in your daily life?</td>
<td>Why do you say that?</td>
</tr>
</tbody>
</table>