Toward a semi-automated item nonresponse detector model for open-response data

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

AAPOR Annual Conference
Chicago, IL
May 11th, 2022
Outline

- Background and context
  - COVID-19 pandemic
  - Open-text data: value and challenges
  - Item nonresponse detection: the technology and development of the model
- Evaluating the model: our approach
- Evaluation results
- Discussion/Next steps
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

- Categorizing item non-response
  - “nonproductive” responses (Behr et al., 2012)
  - Indirect (soft) versus direct (hard) refusals (Meitinger et al., 2021)
Detecting item non-response

- **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

* [https://git.gesis.org/surveymethods/evalanswer](https://git.gesis.org/surveymethods/evalanswer)
Item nonresponse detection: prior work, cont’d

- Limitations of EvalAnswer
  - 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"
Item nonresponse detection: Model development

- Trained a natural language processing (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)

Item nonresponse detection: Model development, cont’d

- Our working taxonomy:
  - **Complete non-response**: Blank text box
  - **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**
  - 5 coders hand-coded 1,400 each, 200 overlapping with one other coder; full overlap for 500
  - 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 more valids than we wanted

- **Round 2:**
  - 2 coders reviewed and arbitrated the results to retrain the model
  - Uncertainty retained in the model when warranted
The data

- NCHS’s Research and Development Survey (RANDS)
  https://www.cdc.gov/nchs/rands/index.htm
- RANDS During COVID-19 – Multi-round web/phone survey
- Topics: health, impacts of pandemic on health care access, COVID-19 related health care and behaviors
- Round 1 fielded June-July 2020: 13,020 Completes
  - 6,800 NORC’s AmeriSpeak probability-based sample = 23.0% weighted cumulative response rate/78.5% completion rate
  - 6,220 Dynata opt-in panel
Model evaluation: our approach
Model evaluation: our approach

- Mixed-method evaluation of two web probe case studies
  - Quarantine probe
  - Pandemic time reference probe
Evaluation results
Round 1: model was initially very uncertain

Model certainty is calculated by subtracting the highest nonresponse prediction score from the valid score. Negative scores indicate model-predicted nonresponse. Positive scores indicate model-predicted valid response.
Round 2: model is now much more uncertain

Model certainty is calculated by subtracting the highest nonresponse prediction score from the valid score. Negative scores indicate model-predicted nonresponse. Positive scores indicate model-predicted valid response.
Quarantine probe

- Quarantine survey question: Have you isolated or quarantined yourself because of the Coronavirus? Yes/No

- Quarantine probe: When answering the previous question about isolating or quarantining because of the Coronavirus, what were you thinking about? (half the sample received (n=6,308), other half received a closed-ended version)

- Comparison with “source of truth”: human coding (July-September 2020)
  - Sensitivity and specificity calculations
Quarantine probe: evaluation results

<table>
<thead>
<tr>
<th></th>
<th>Coded NR</th>
<th>Coded Valid</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model NR</td>
<td>848</td>
<td>288</td>
</tr>
<tr>
<td>Model Valid</td>
<td>392</td>
<td>4768</td>
</tr>
<tr>
<td>Total</td>
<td>1240</td>
<td>5056</td>
</tr>
</tbody>
</table>

Key take-away:
Model did a good job identifying “true” valids; less well identifying “true” item nonresponse

Sensitivity **68%** (848/1240)
False valids (human-coded NR):
- “None” (61)
- “Quarantine” (10)

Specificity **94%** (4768/5056)
False NR (human-coded valid):
- “If I had symptoms”
- “Did I need to quarantine because of a possibility of Coronavirus”
- “If I was knowingly exposed to the virus”
- Almost all “other, high risk”
Pandemic time reference probe

The probes:

1. When do you think that the Coronavirus pandemic began? Your best guess is fine.
2. When did the Coronavirus pandemic first affect your daily life? Your best guess is fine.
3. Why do you say that? (n=12,662)

Different “source of truth”; hand-review but not full coding

Full review of model-identified nonresponse (n=1,619); random sample (n=1,000) of valids

“Implied” sensitivity and specificity calculations
## Pandemic time reference probe: evaluation results

<table>
<thead>
<tr>
<th></th>
<th>Coded NR</th>
<th>Coded Valid</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model NR</td>
<td>1372</td>
<td>247</td>
<td>1619</td>
</tr>
<tr>
<td>Model Valid</td>
<td>199</td>
<td>10844</td>
<td>11043</td>
</tr>
<tr>
<td></td>
<td>(= (18/1000)*11043)</td>
<td>(= (982/1000)*11043)</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>1571</td>
<td>11091</td>
<td>12662</td>
</tr>
</tbody>
</table>

**Key take-away:**
Model did a good job identifying “true” valids; slightly less well identifying “true” item nonresponse

- **Sensitivity 87%** \((1372/1571)\), 95% CI [83%, 93%]
- **Specificity 98%** \((10844/11091)\), 95% CI [98%, 98%]

**False valids (human-coded NR):**
- “None”
- “Because it just doesn’t”
- “I’m fine”
- “Best guess”
- “You asked”

**False NR (human-coded valid):**
- “because i sdyat yhome”
- Almost all “other, high risk”
Discussion/next steps
Discussion/next steps

- Evaluation results show promise for our semi-automated item nonresponse detection model

Next steps:

- Further evaluation on additional open-text data on wider range of topics
- Analysis to better understand the types and patterns of item nonresponse and possible subgroup differences
- Work toward release of a generalized model (possibly web-based) to share with others
Thank you!!

- Please contact us with any questions
  - Kristen Cibelli Hibben - kcibelli@cdc.gov
  - Zachary Smith – zsmith@cdc.gov
  - Travis Hoppe – thoppe@cdc.gov

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.
For more information contact: Amanda Wilmot awlimot@cdc.gov

**Q-Bank:** providing access to survey question evaluation reports, question design and performance [https://wwwn.cdc.gov/qbank/](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](https://www.cdc.gov/nchs/ccqder/products/qnotes.htm)

Centers for Disease Control and Prevention
1600 Clifton Road NE, Atlanta, GA 30333
Telephone: 1-800-CDC-INFO (232-4636)/TTY: 1-888-232-6348
Visit: www.cdc.gov | Contact CDC at: 1-800-CDC-INFO or www.cdc.gov/info

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


