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* (Kaczmirek 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

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)

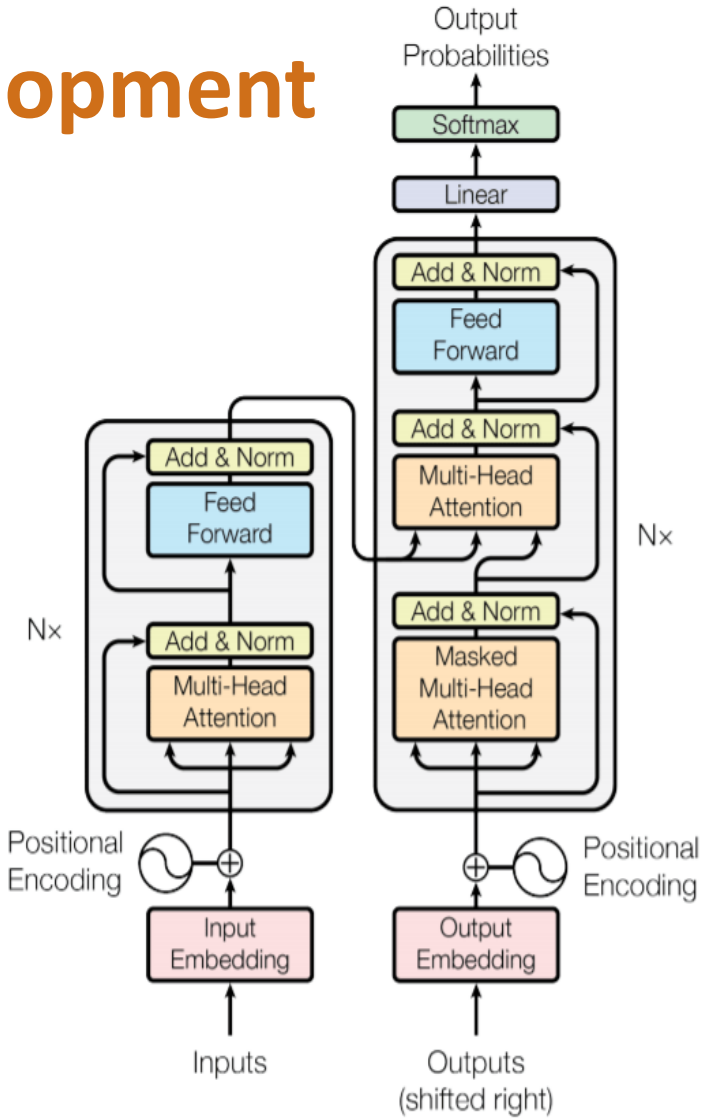


Figure 1: The Transformer - model architecture.

* https://arxiv.org/abs/1810.04805

** https://arxiv.org/abs/2104.08821

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

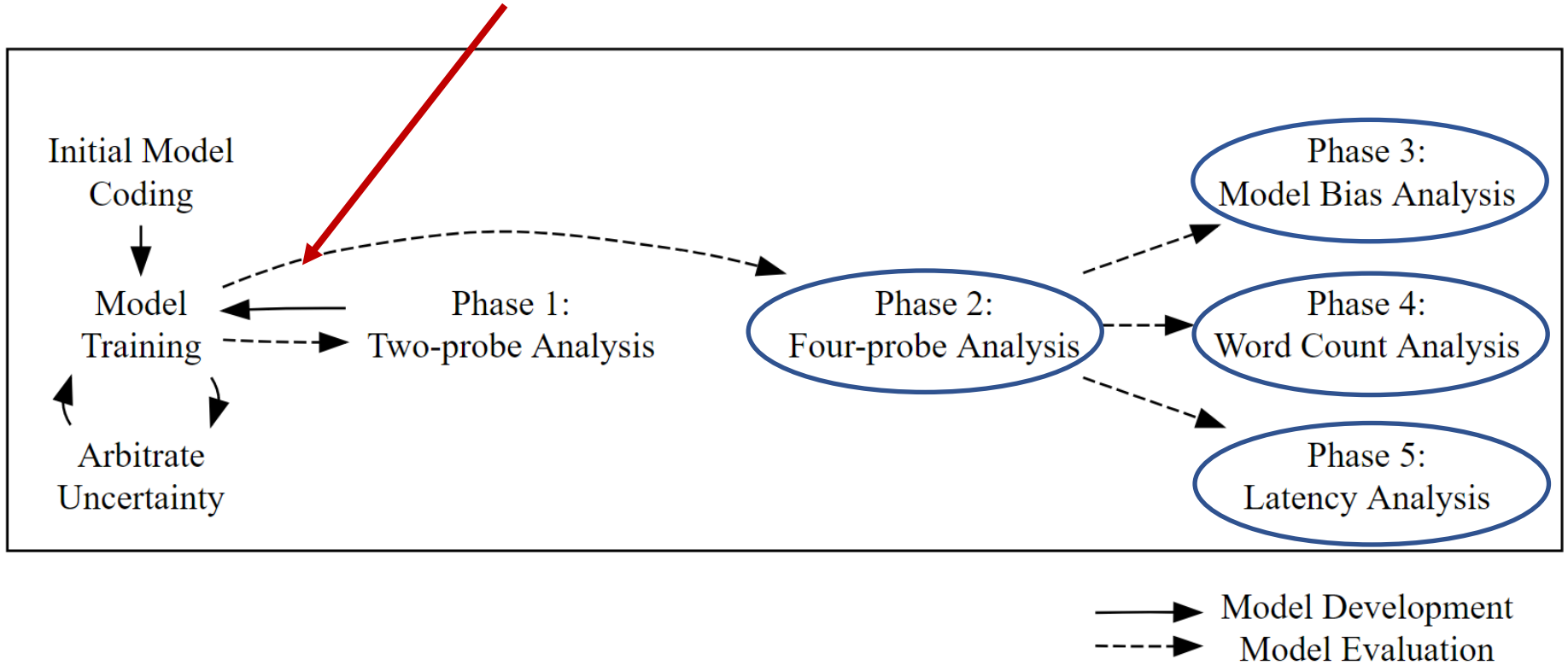
Model evaluation: our approach

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

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

Model development process



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

	Human-reviewed NR	Human-reviewed Valid	
Model NR	450	177	627
Model Valid	109	3,876	3,985
Total	559	4,053	4,612

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

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

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"

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

False NR (human-coded valid):

- "Safty" (and variations)
- "Save life"
- "lines in the market"
- "It is necessary but a pain."
- "Courtesy"
- "ITS COMMON CERDICY AND GO WITH THE THROW"

Phase 2 results: Additional probes

Vaccine Hesitancy	Human-reviewed NR	Human-reviewed Valid	
Model NR	151	492	643
Model Valid	61	4,266	4,327
Total	212	4,758	4,970

- Sensitivity: 71%
- Specificity: 90%

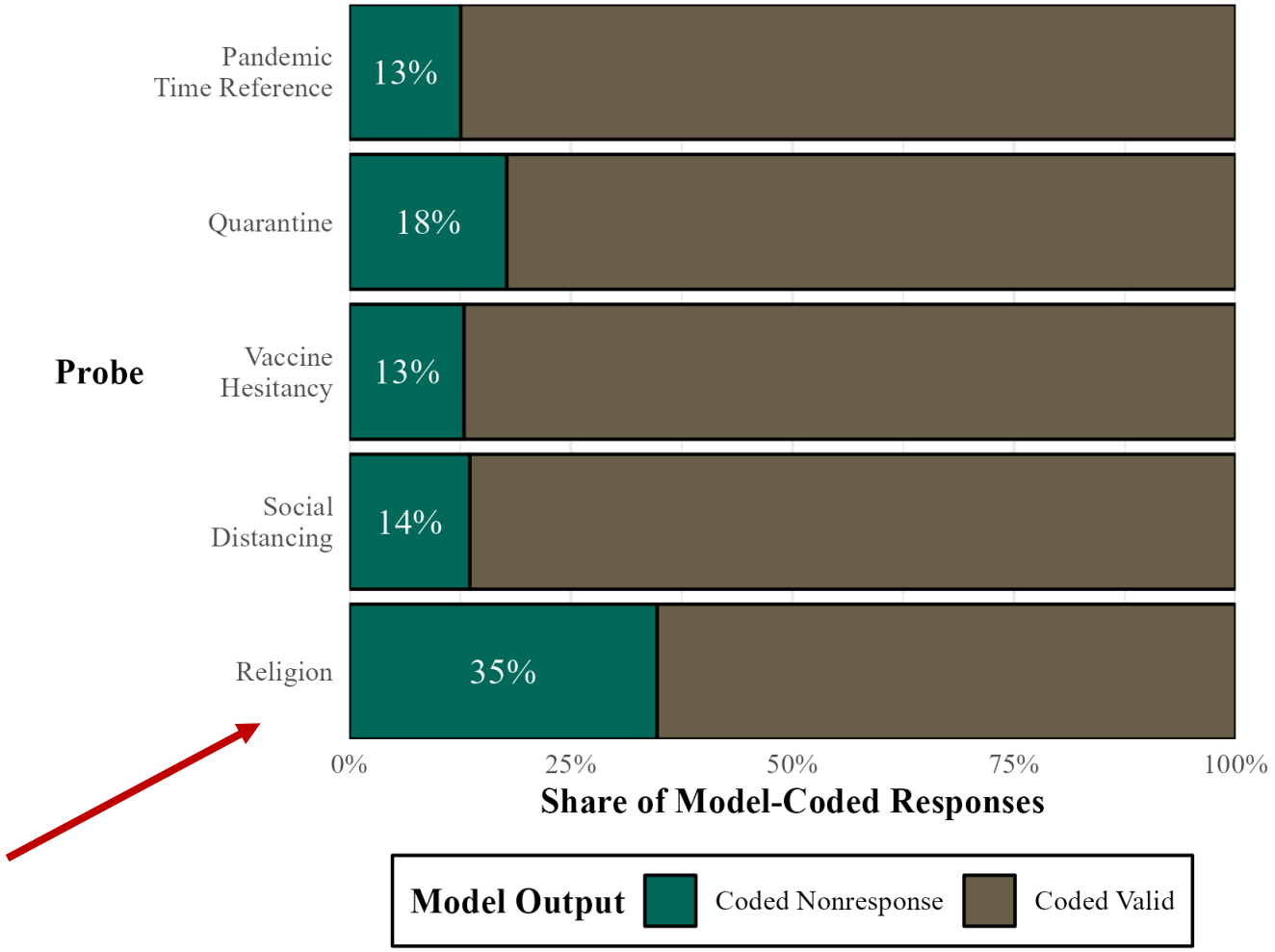
Religion	Human-reviewed NR	Human-reviewed Valid	
Model NR	298	952	1,250
Model Valid	36	2,314	2,350
Total	334	3,266	3,600

- Sensitivity: 90%
- Specificity: 71%

Quarantine	Human-coded NR	Human-coded Valid	
Model NR	863	239	1,102
Model Valid	325	4,778	5,103
Total	1,188	5,017	6,205

- Sensitivity: 73%
- Specificity: 95%

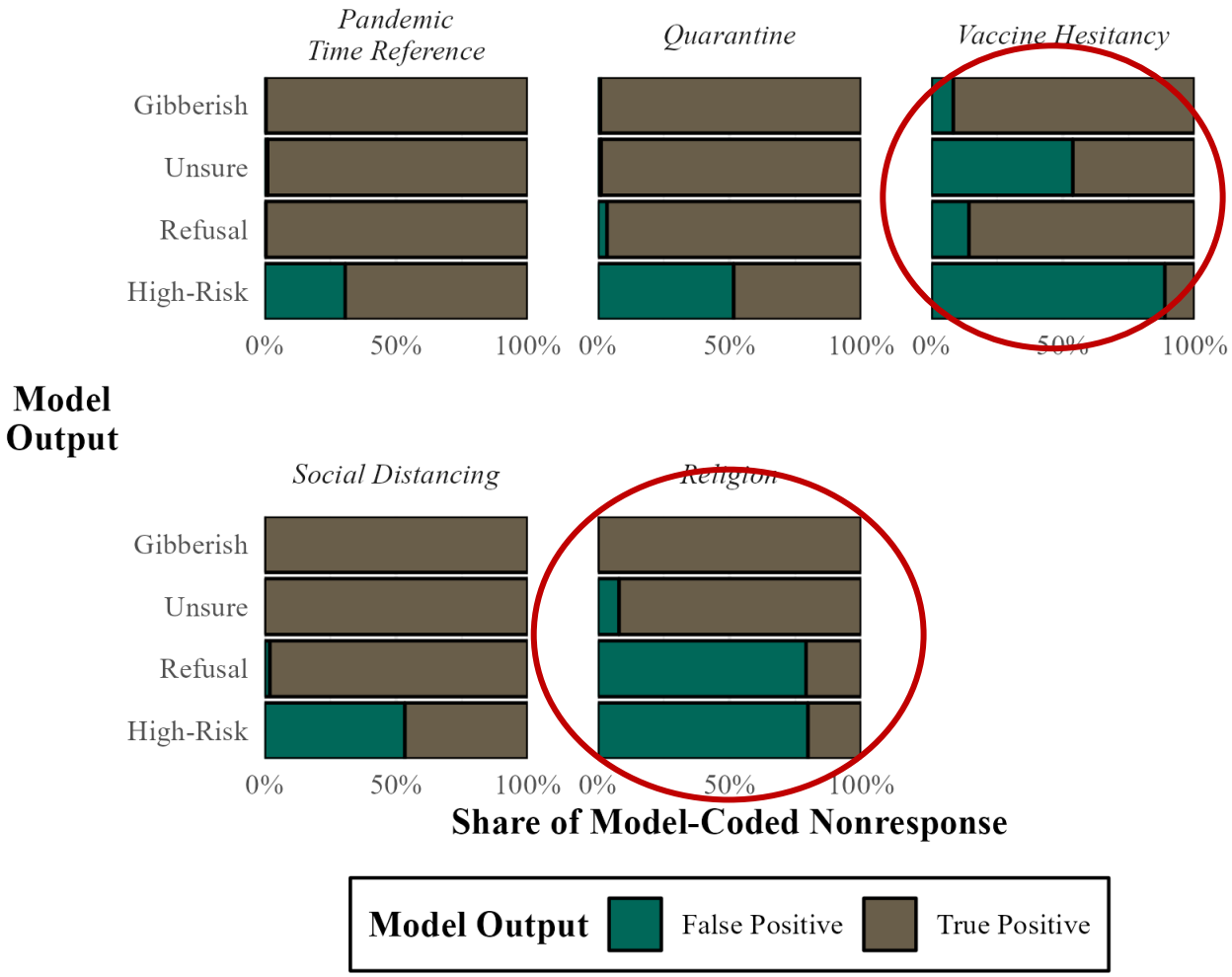
Proportions of model-coded item nonresponse



Output for Pandemic Time Reference is from the first arbitrated model (Phase 1 of Model Evaluation).
Output for all other probes is from the final model version (Phase 2 of Model Evaluation).
Blank responses removed.

- 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

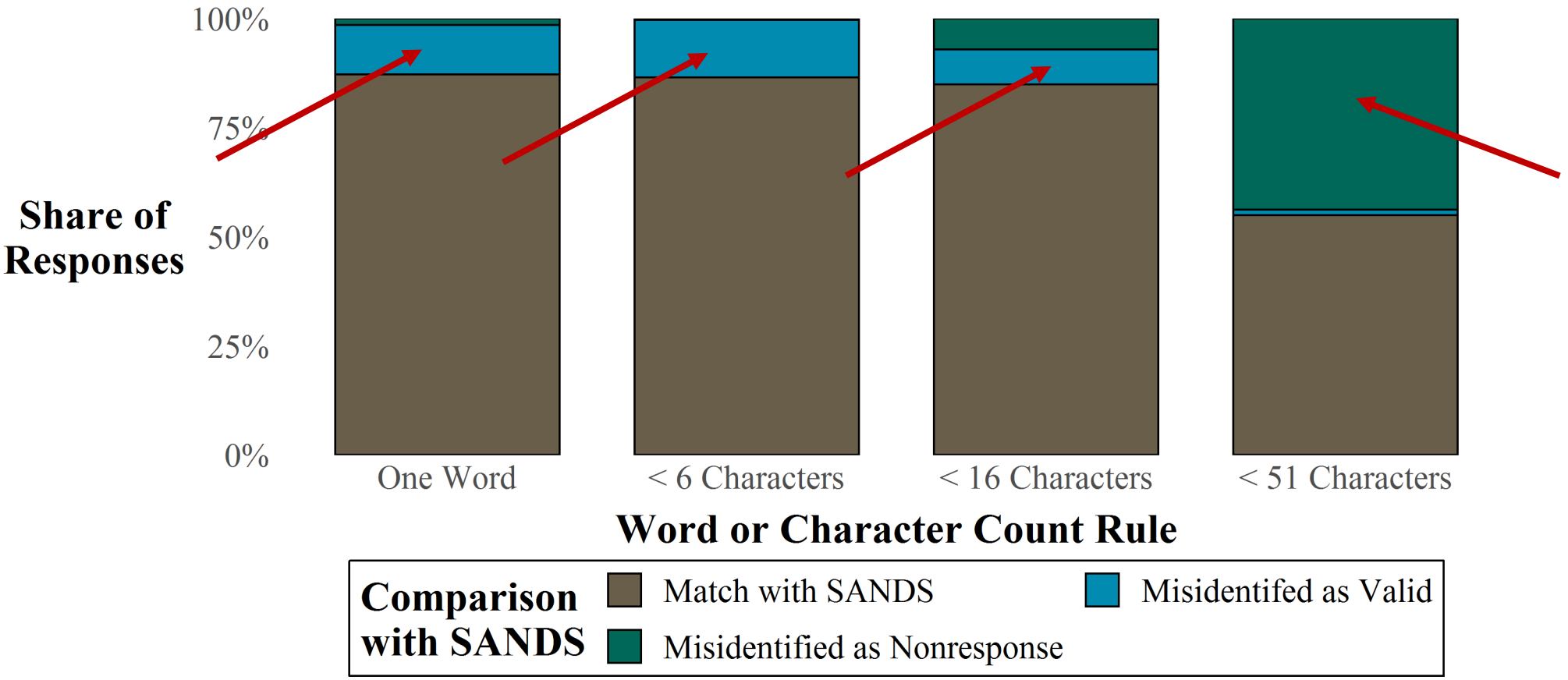
Distribution by type of item nonresponse



Output for Pandemic Time Reference is from the first arbitrated model (Phase 1 of Model Evaluation).
Output for all other probes is from the final model version (Phase 2 of Model Evaluation).

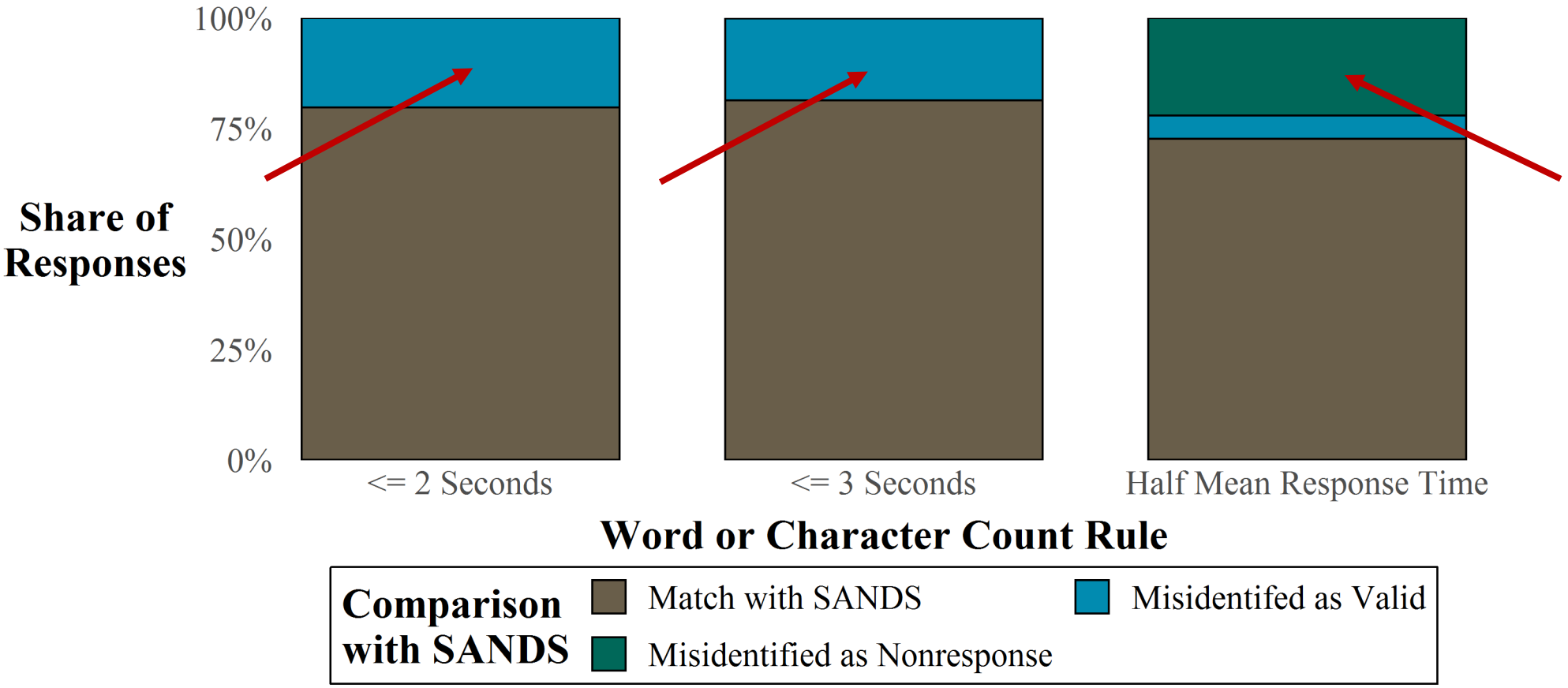
- Model error often concentrated in the High Risk category, as seen for Social Distancing
- More error seen in Refusals for Religion
- More error seen in Unsure for Vaccine Hesitancy

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

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

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
kcibelli@cdc.gov

CCQDER
Collaborating Center for Questionnaire Design and Evaluation Research

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)
TTY: 1-888-232-6348 www.cdc.gov

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

- Behr, D., Kaczmirek, L., Bandilla, W., & Braun, M. (2012). Asking probing questions in web surveys: which factors have an impact on the quality of responses? Social Science Computer Review, 30(4), 487-498.
- Etz, R.S., Gonzalez, M.M., Eden, A.R., & Winship, J. (2018). Rapid sense making: A feasible, efficient approach for analyzing large data sets of open-ended comments. *International Journal of Qualitative Methods*, *17*(1), 1609406918765509.
- Kaczmirek, L., Meitinger, K., Behr., D. (2017). Higher data quality in web probing with EvalAnswer: a tool for identifying and reducing nonresponse in open-ended questions. (GESIS Papers, 2017/01). Köln: GESIS - Leibniz- Institut für Sozialwissenschaften.
- Kennedy, C., Hatley, N., Lau, A., Mercer, A., Keeter, S., Ferno, J., & Asare-Marfo, D. (2021). Strategies for Detecting Insincere Respondents in Online Polling. *Public Opinion Quarterly*, *85*(4), 1050-1075.
- Lenzner, T., & Neuert, C.E. (2017). Pretesting Survey Questions Via Web Probing–Does it Produce Similar Results to Face-to-Face Cognitive Interviewing? *Survey Practice, 10*(4), 2768.
- Meitinger, K., & Behr, D. (2016). Comparing cognitive interviewing and online probing: Do they find similar results?. *Field Methods*, *28*(4), 363-380.
- Meitinger, K., Behr, D., & Braun, M. (2021). Using apples and oranges to judge quality? Selection of appropriate cross-national indicators of response quality in open-ended questions. *Social Science Computer Review*, *39*(3), 434-455.

References

- Neuert, C.E., Meitinger, K., Behr, D., & Schonlau, M. (2021a). The use of open-ended questions in surveys. *Methods, data, analyses: a journal for quantitative methods and survey methodology, 15*(1), 3-6.
- Neuert, C.E., Meitinger, K., & Behr, D. (2021b). Open-ended versus Closed Probes: Assessing Different Formats of Web Probing. *Sociological Methods & Research*, 00491241211031271.
- Schonlau, M. & Couper, M.P. (2016). Semi-automated categorization of open-ended questions. Survey Research Methods 10(2), pp. 143-152
- Singer, E. & Couper, M.P. (2017). Some methodological uses of responses to open questions and other verbatim comments in quantitative surveys. methods, data, analyses 11(2), pp. 115-134.
- Scanlon, P. J. (2019). The effects of embedding closed-ended cognitive probes in a web survey on survey response. Field Methods, 31(4), 328-343.
- Scanlon, P. (2020). Using targeted embedded probes to quantify cognitive interviewing findings. In P. C. Beatty, D. Collins, L. Kaye, J. Padilla, G. B. Willis & A. Wilmot (Eds.), Advances in questionnaire design, development, evaluation and testing, pp. 427–449.
- Trejo, Y.G., Meyers, M., Martinez, M., O'Brien, A., Goerman, P., & Class, B.O. (2022). Identifying Data Quality Challenges in Online Opt-In Panels Using Cognitive Interviews in English and Spanish. *Journal of Official Statistics (JOS), 38*(3).
- Yeung, R.C., & Fernandes, M.A. (2022). Machine learning to detect invalid text responses: Validation and comparison to existing detection methods. *Behavior Research Methods, 54*, 3055-3070.

The probes for evaluation phases 1 & 2

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