Correlates of Item Nonresponse to Open-Ended Web Probes

Zachary Smith[1,2,3], Kristen Cibelli Hibben[1,2,3], Benjamin Rogers[2,3], Valerie Ryan[1,2,3], Paul Scanlon[1,2,3], and Travis Hoppe[3]

[1]Collaborating Center for Questionnaire Design and Evaluation Research

[2]Division of Research and Methodology

[3]National Center for Health Statistics

Federal Committee on Statistical Methodology Research and Policy Conference

Washington, DC

October 25-27, 2022

Outline

- Item nonresponse in open-ends: the problem
- A semiautomated model for detecting item nonresponse
- Web probes evaluated and model validity
- Subgroup variation in model-coded nonresponse
- Implications for question evaluation

Item nonresponse in open-ends: the problem

Value and challenges of working with open-text data

- Wide range of methodological uses (Singer & Couper, 2017)
- Because responses are unconstrained, particularly useful when little is known about a topic (Neuert et al., 2021, Scanlon, 2019; 2020)
- However:
  - More burdensome to respondents
  - Prone to nonresponse or inadequate, non-codable responses
  - Coding and analysis is time- and labor-intensive for researchers

A semiautomated model for detecting item nonresponse

Building a model to detect nonresponse

- Prior work:
  - Categorizing item nonresponse (Behr et al., 2012; Meitinger et al., 2021)
  - Detecting item nonresponse via rule-based approach (Kaczmirek et al. (2017); available on GitHub)
- Leveraging advances in data science to build a more accurate detector
- Trained a natural language processing (NLP) model
- Refined with human coding (active learning)

- EvalAnswer: https://git.gesis.org/surveymethods/evalanswer
- Bidirectional Transformer for Language Understanding (BERT): https://arxiv.org/abs/1810.04805
- Simple Contrastive Sentence Embedding (SimCSE): https://arxiv.org/abs/2104.08821

Taxonomy of responses

- The model assigns a score (0-1) for the extent to which a response falls into each of the item non-response categories
  - **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
- Several rounds of arbitration and refined training produced latest model version

Web probes evaluated and model validity

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
- Round 3 fielded May-June 2021: 5,458 Completes
  - 7,852 NORC's AmeriSpeak probability-based sample = 11.8% weighted cumulative response rate / 69.5% completion rate

Web probes evaluated

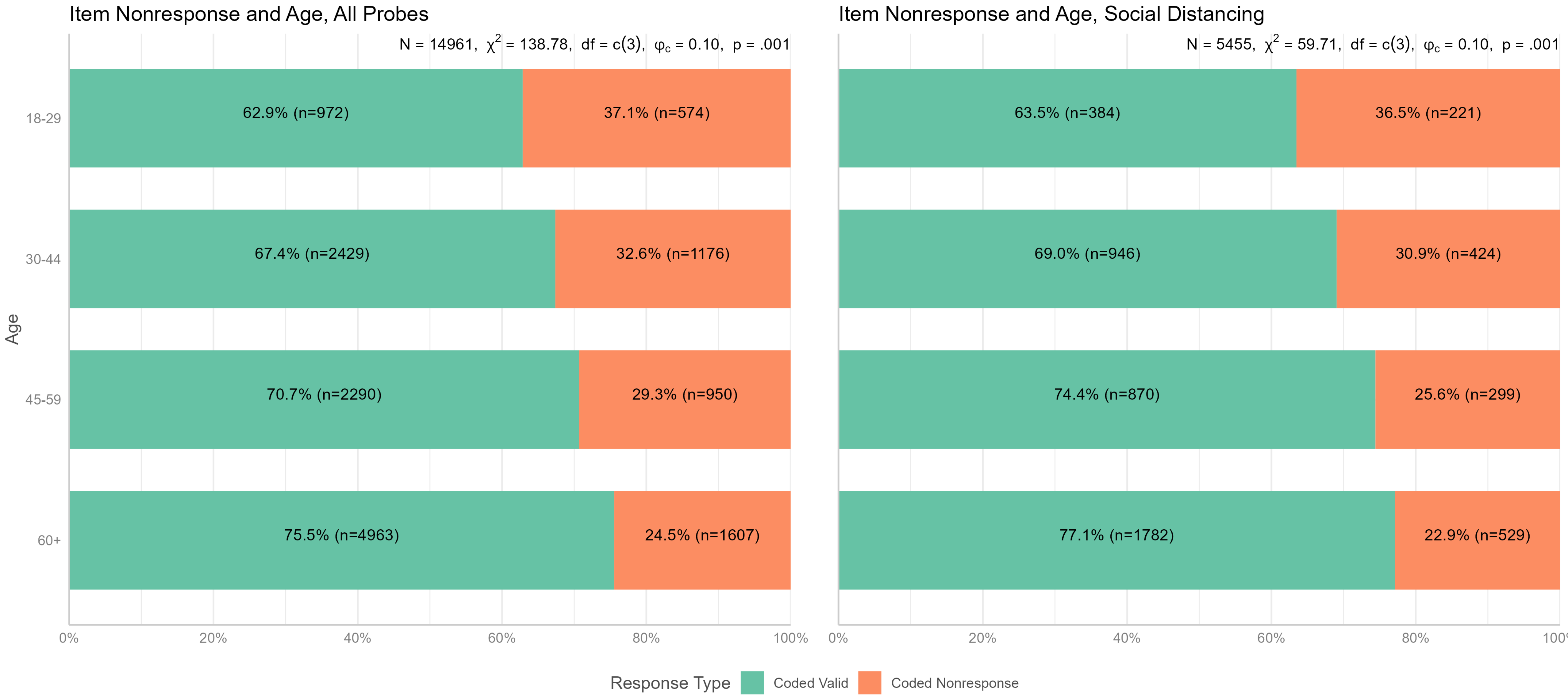
- Vaccine hesitancy:
  - Please list the reasons you say you [are / are not] hesitant about vaccines in general.
- Social distancing:
  - When you were answering about social distancing in the previous questions, what were you thinking about?
- Religion:
  - Currently, how important is religion in your daily life? (Very, somewhat, not important)
  - Why do you say that?

Model validity

- Manually evaluated all coded soft nonresponses and 1,000 randomly sampled coded valids to determine model sensitivity and specificity
- Overall:
  - Sensitivity: 83.6%
  - Specificity: 86.5%
- Hesitancy:
  - Sensitivity: 77.7%
  - Specificity: 89.7%
- Distancing:
  - Sensitivity: 81.9%
  - Specificity: 95.6%
- Religion:
  - Sensitivity: 90.1%
  - Specificity: 70.9%

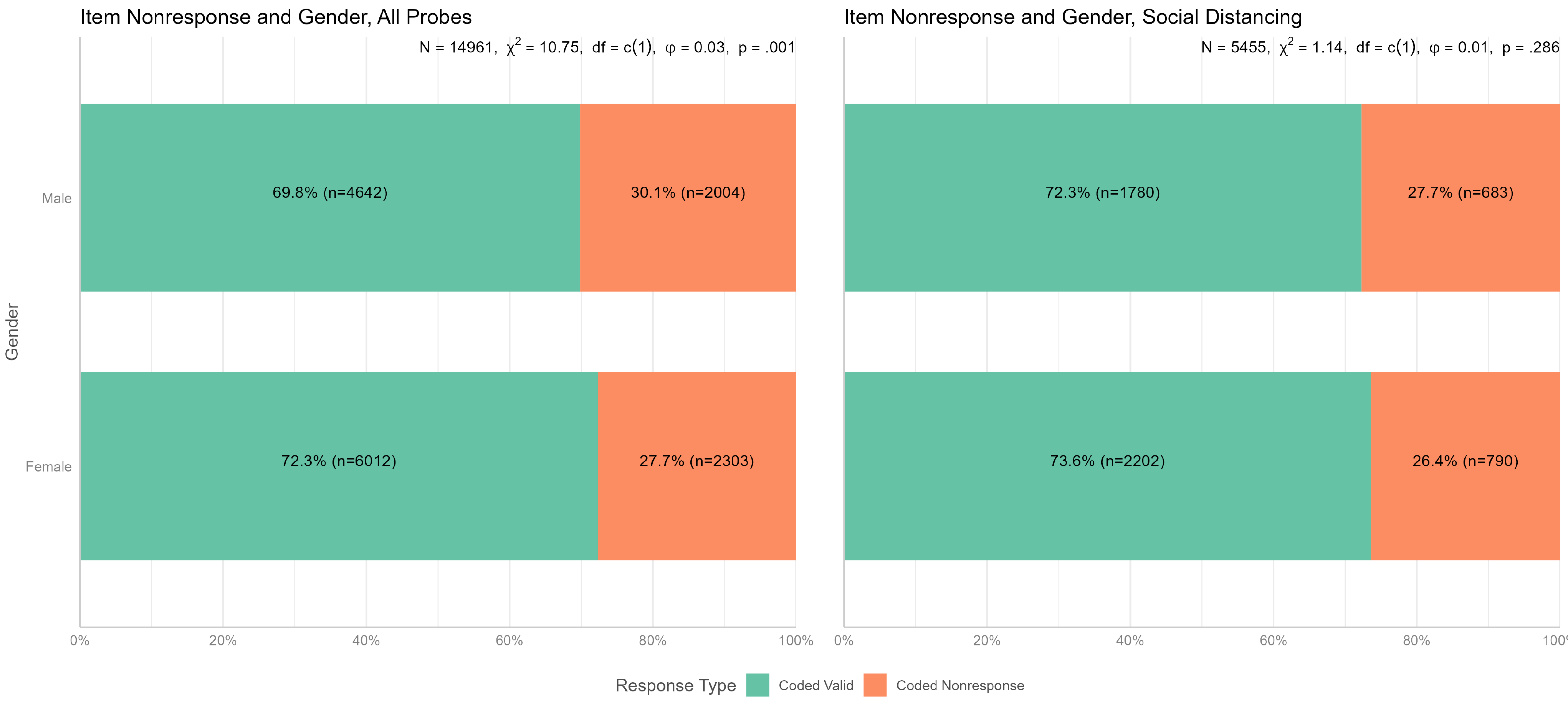
Subgroup variation in model-coded nonresponse

Subgroup variation: age



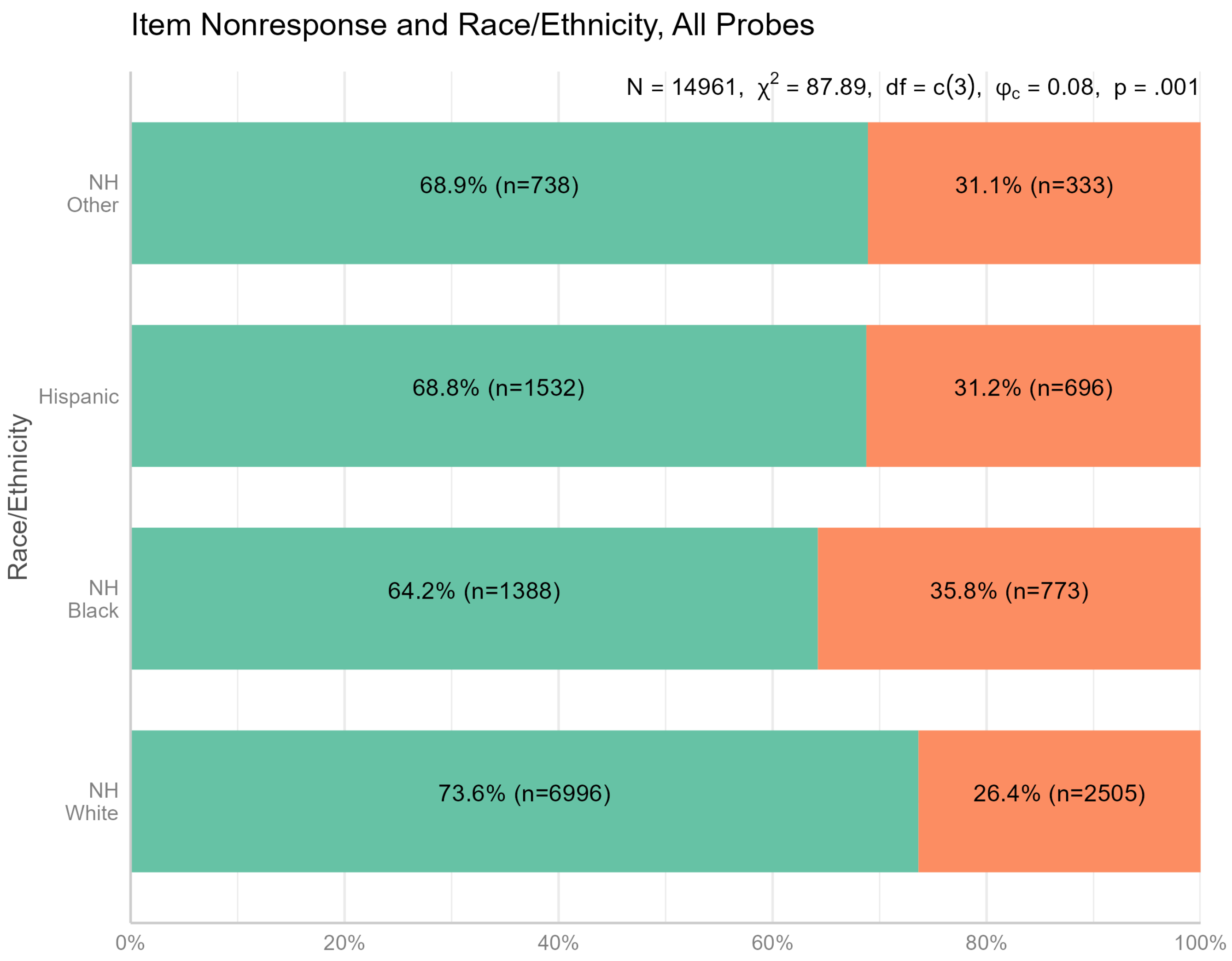
SOURCE: National Center for Health Statistics Research and Development Survey During COVID-19, Round 3

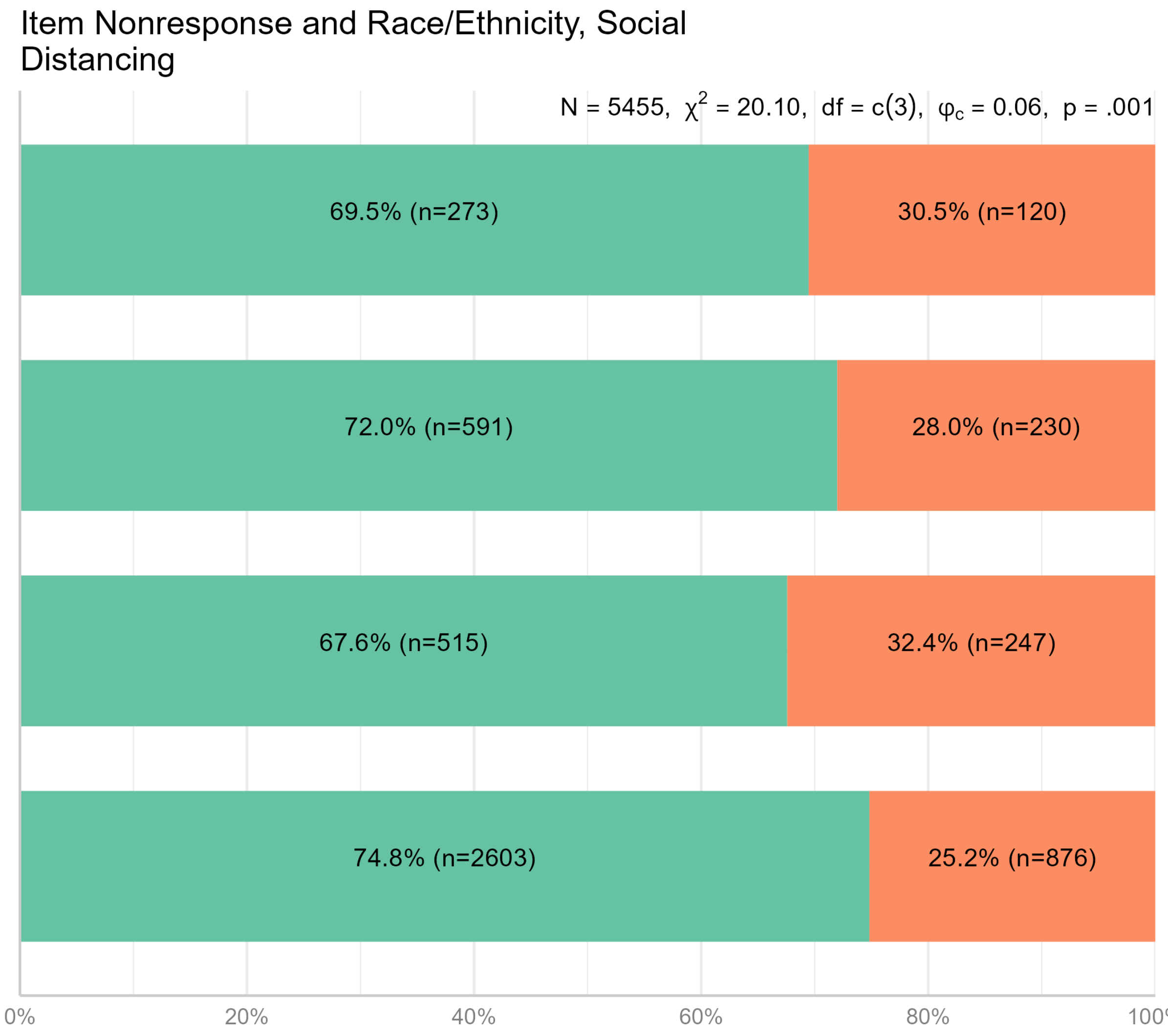
Subgroup variation: gender



SOURCE: National Center for Health Statistics Research and Development Survey During COVID-19, Round 3

Subgroup variation: race/ethnicity

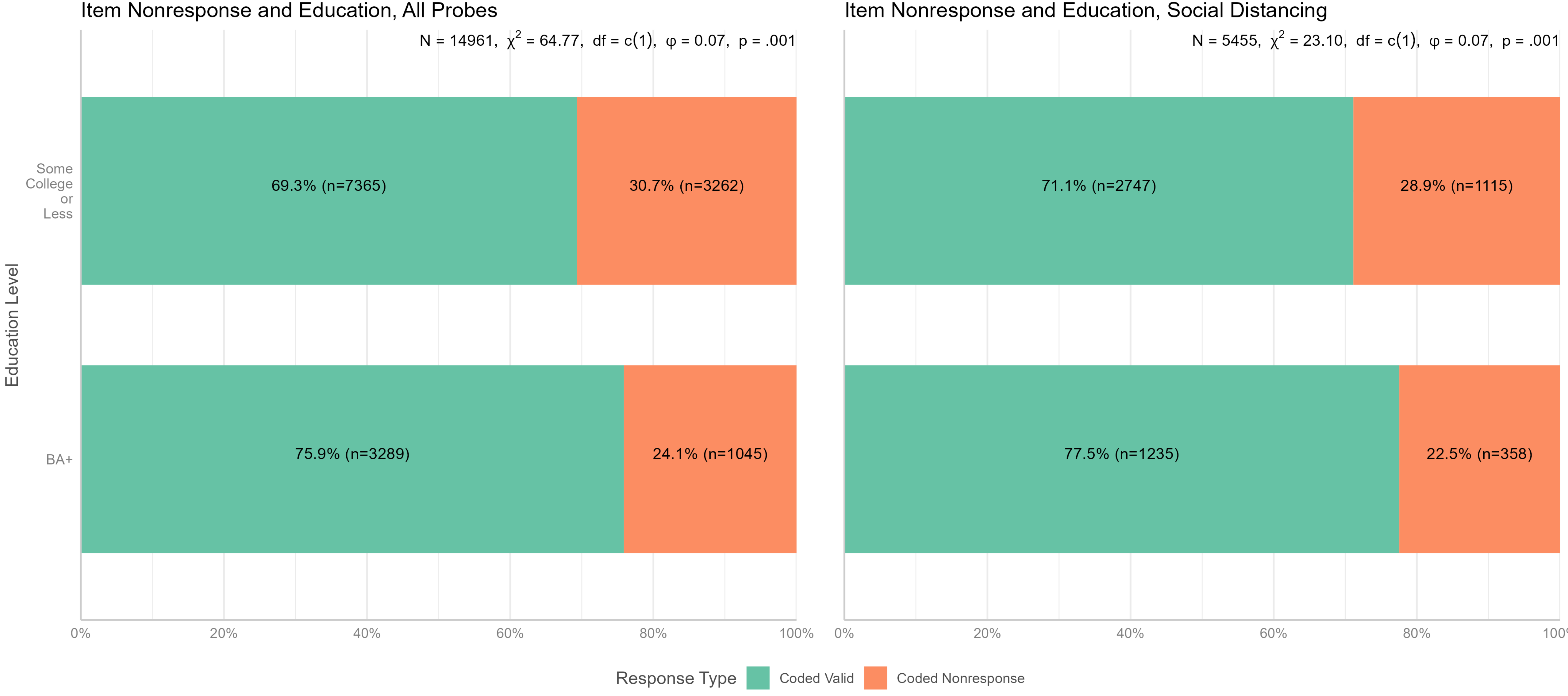




Response Type Coded Valid Coded Nonresponse

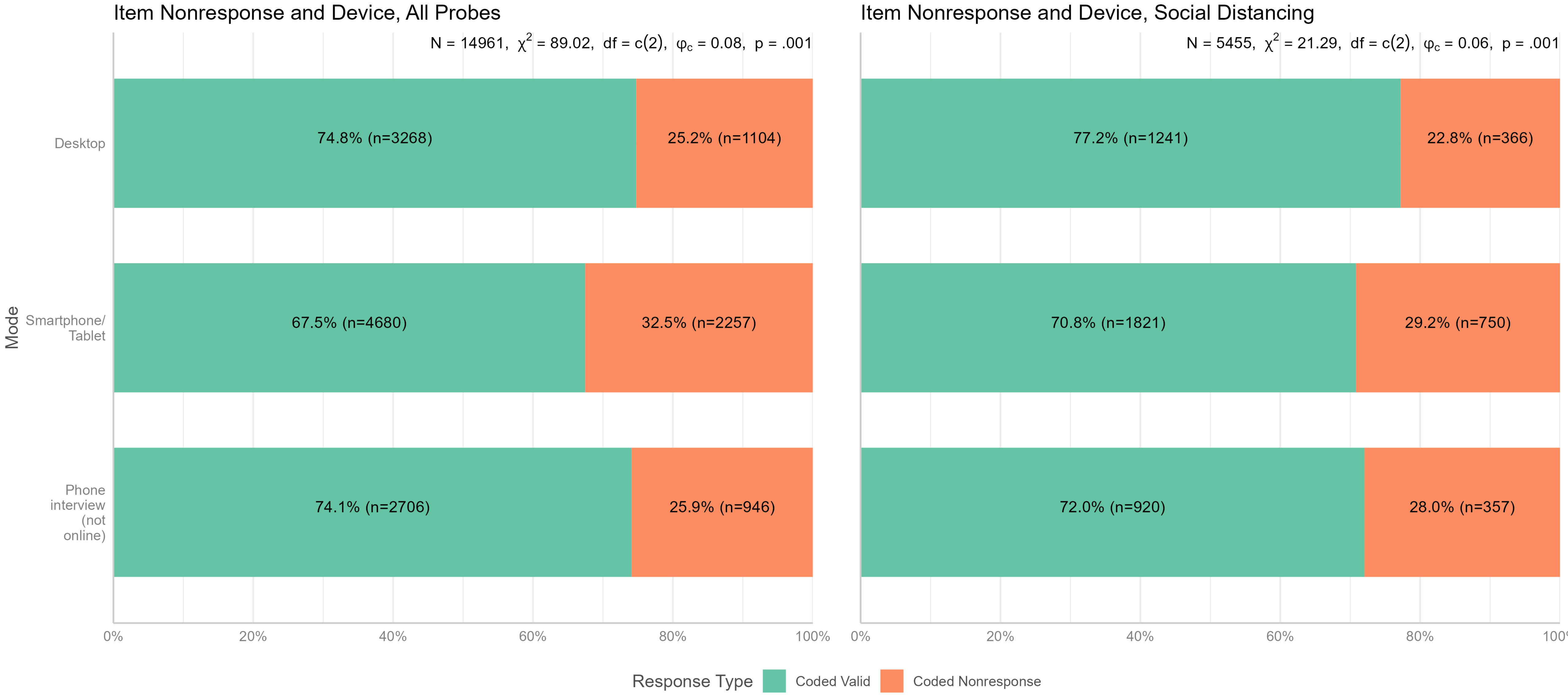
SOURCE: National Center for Health Statistics Research and Development Survey During COVID-19, Round 3

Subgroup variation: education



SOURCE: National Center for Health Statistics Research and Development Survey During COVID-19, Round 3

Subgroup variation: device

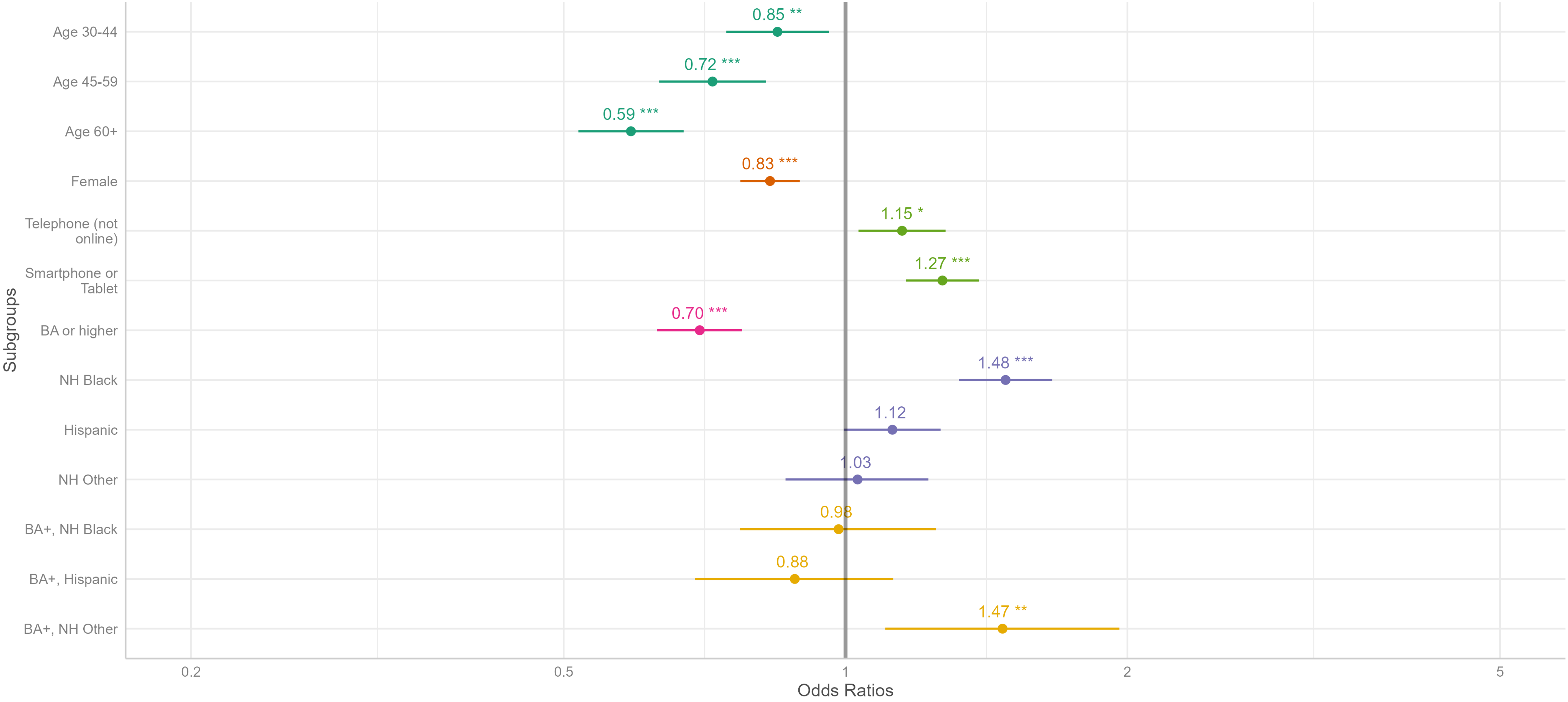


SOURCE: National Center for Health Statistics Research and Development Survey During COVID-19, Round 3

Pulling all the data together

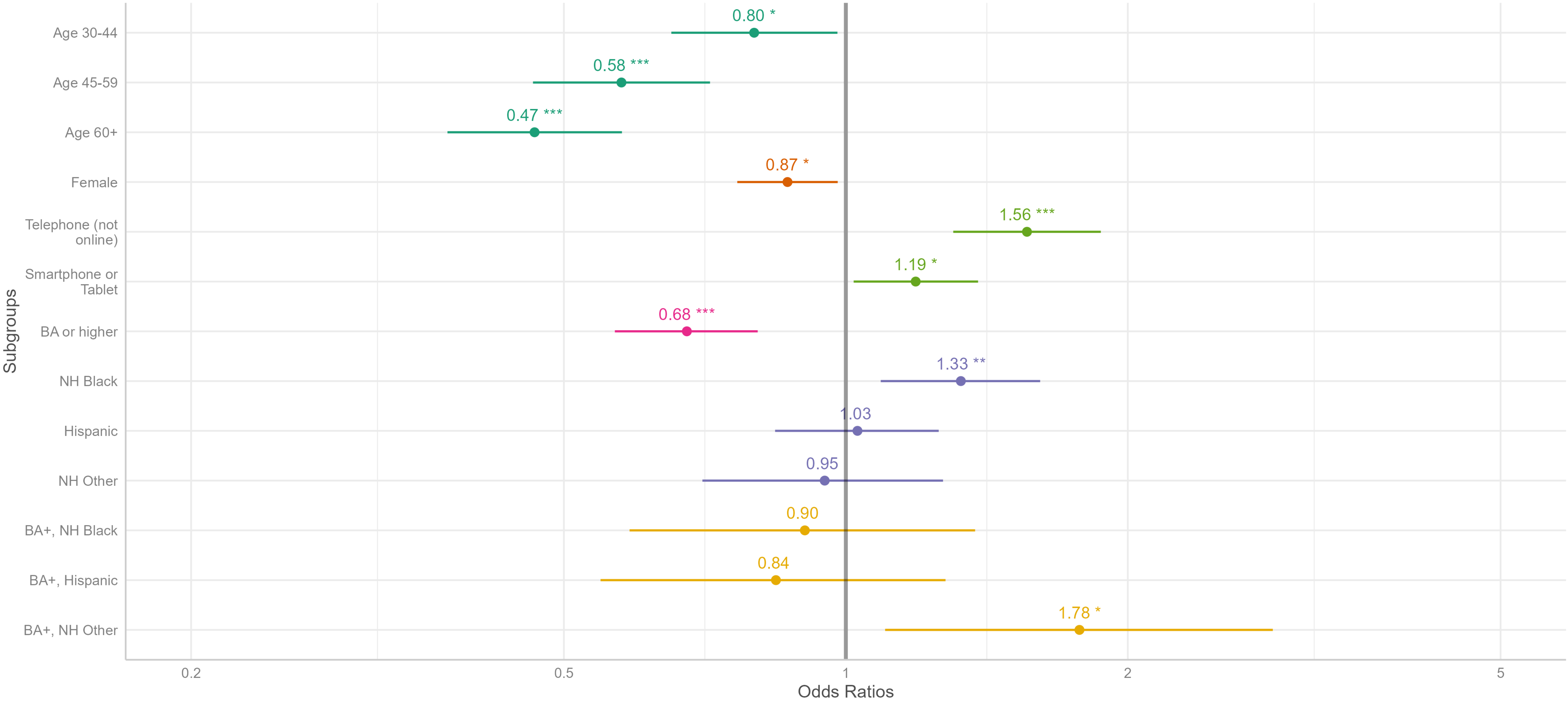
- Logistic regression to estimate odds of nonresponse by subgroup for a) all probes and b) the social distancing probe
- Reference categories:
  - Age: 18-29
  - Gender: Male
  - Race/ethnicity: Non-Hispanic White
  - Education: Some college or less
  - Device: Desktop computer
  - Interaction of race/ethnicity and education: NH White, Some college or less
- 95% confidence intervals shown
- Analysis run in R 4.2.0 and Rstudio 2022.02.03 using *tidyverse* and *sjPlot* packages

Odds of Nonresponse by Subgroup, All Probes



* p < 0.05, ** p < 0.01, *** p < 0.001; SOURCE: National Center for Health Statistics Research and Development Survey During COVID-19, Round 3, N = 14,961 responses

Odds of Nonresponse by Subgroup, Social Distancing



* p < 0.05, ** p < 0.01, *** p < 0.001; SOURCE: National Center for Health Statistics Research and Development Survey During COVID-19, Round 3, N = 5,455 responses

Implications for question evaluation

How can this model assist in question evaluation?

- Speedily categorizes open-text data with reasonable sensitivity and specificity
- Clear understanding of demographics of non-responders – potential for insight into patterns of nonresponse that can improve question design
- But, some dangers: reliance on the coded valid dataset excludes:
  - Potentially valid responses missed by the model (false negatives)
  - The voices of groups systematically more likely to be categorized as nonresponse

Thank you!

For more information contact: Zachary Smith, zsmith@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