AI-Assisted Conversational Interviewing: Effects on Data Quality and Respondent Experience

Online Appendix

A1. Additional Experiment Details

Figure A1. Probing Mechanism in Treatment Conditions

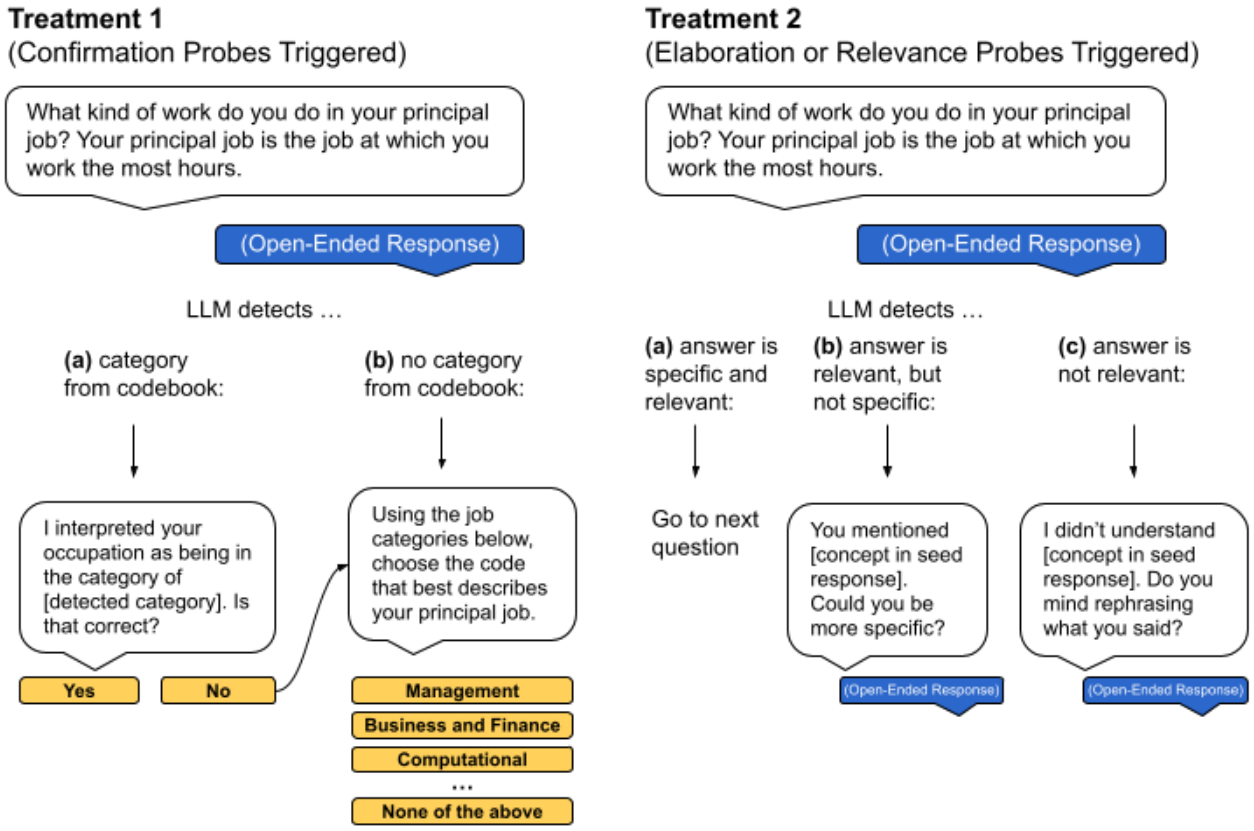


Table A1. Full Questionnaire (continued on following pages)

Q0. Consent Screener		
Control (No Probes)	**Treatment 1** (Confirmation Probes)	**Treatment 2** (Elaboration or Relevance Probes)
In this study, researchers at a nonprofit research institution are interested in understanding the American public's opinions on issues of public importance. In addition to asking respondents about their attitudes on these issues, this survey will ask about demographic and employment characteristics in order to accurately create a representative profile of these attitudes. Respondents may be randomly assigned to received follow-up questions generated by an AI chatbot. By selecting "I agree", you - as a participant in this study - agree to your responses being collected and analyzed as part of this study. You may refuse to participate by selecting "I do not agree." 1 I agree 2 I do not agree		

<table>
<tr><th colspan="3">Q1. Most Important Issue</th></tr>
<tr><th>Control
(No Probes)</th><th>Treatment 1
(Confirmation Probes)</th><th>Treatment 2
(Elaboration or Relevance Probes)</th></tr>
<tr><td colspan="3">[Seed Question] What do you think is the most important problem facing this country today? _____ (open-ended)</td></tr>
<tr><td>Go to next question</td><td>If LLM detects close-ended category in seed response:

➔ [Binary Confirmation Probe] I interpreted your answer as generally being about [sampled detected category]. Is that correct?
1 Yes
2 No

If response to binary confirmation probe is ‘No’ or LLM does not detect category in seed response:

➔ [Categorical Confirmation Probe] Which of the following topics did you mention in your answer to the previous question?
1 Economy
2 Cost of Living
3 Federal Budget
4 Jobs
5 Wages
6 Taxes
7 Economic Inequality
8 Corporate Power
9 International Trade
10 Immigration
11 Poverty
12 Elections and Democracy
13 Crime
14 Foreign Policy
15 Abortion
16 Race Relations and Racism
17 Climate Change
18 Education
19 Guns/Gun control
20 LGBTQ Issues
21 None of the above</td><td>If LLM detects that response is relevant, but not specific:

➔ [Elaboration Probe] (example) You mentioned [vague concept detected by LLM]. Could you tell me more?

Else if LLM detects that response is not relevant:

➔ [Relevance Probe] (example) I’m not sure what you mean by [irrelevant concept detected by LLM]. Could you rephrase what you said?

Else:

(Go to next question)</td></tr>
</table>

<table>
<tr><th colspan="3">Q2. Economic Conditions (Sentiment)</th></tr>
<tr><th>Control
(No Probes)</th><th>Treatment 1
(Confirmation Probes)</th><th>Treatment 2
(Elaboration or Relevance Probes)</th></tr>
<tr><td>[Seed Question] How would you rate economic conditions in this country today?
1 Very Good
2 Good
3 Neither Good Nor Bad
4 Bad
5 Very Bad</td><td colspan="2">[Seed Question] How would you rate the economic conditions in this country today?
_____ (open-ended)</td></tr>
<tr><td>Go to next question</td><td>If negative sentiment detected in seed response & positive sentiment not detected:

➔ [Binary Confirmation Probe] Just to confirm, are you saying that economic conditions are more negative than positive?
1 Yes
2 No

If positive sentiment detected in seed response & negative sentiment not detected:

➔ [Binary Confirmation Probe] Just to confirm, are you saying that economic conditions are more positive than negative?
1 Yes
2 No</td><td>If LLM detects that response is relevant, but not specific:

➔ [Elaboration Probe] (example) You mentioned [vague concept detected by LLM]. Could you tell me more?

Else if LLM detects that response is not relevant:

➔ [Relevance Probe] (example) I’m not sure what you mean by [irrelevant concept detected by LLM]. Could you rephrase what you said?

Else:

Go to next question</td></tr>
</table>

<table>
<tr><th colspan="3">Q3. Economic Conditions (Reason)</th></tr>
<tr><th>Control
(No Probes)</th><th>Treatment 1
(Confirmation Probes)</th><th>Treatment 2
(Elaboration or Relevance Probes)</th></tr>
<tr><td colspan="2">[Seed Question] What are the main reasons for your rating of economic conditions in this country?
_____ (open-ended)</td><td rowspan="2">Triggered during Q2</td></tr>
<tr><td>Go to next question</td><td>If LLM detects close-ended category in seed response:

➔ [Binary Confirmation Probe] I interpreted your answer as generally being about [sampled detected category]. Is that correct?
1 Yes
2 No

If response to binary confirmation probe is ‘No’ or LLM does not detect category in seed response:

➔ [Categorical Confirmation Probe] Which of the following reasons matches your answer to the previous question?
1 Employment/Jobs
2 Layoffs
3 Inflation
4 Wages
(cont’d on next page)
5 Stock Market
6 Economic Growth
7 Government Spending
8 Gas Prices
9 Democratic / Biden Administration Policies
10 Interest Rates
11 Illegal immigration
12 Wealth inequality
13 Corporations or Corporate greed
14 Poverty or homelessness
15 Taxes
16 Politicians</td></tr>
</table>

<table>
<tr><th colspan="3">Q4. Preferred News Source</th></tr>
<tr><th>Control
(No Probes)</th><th>Treatment 1
(Confirmation Probes)</th><th>Treatment 2
(Elaboration or Relevance Probes)</th></tr>
<tr><td>[Seed Question] What is your main source of news?
1 Fox News or FoxNews.com
2 Local TV
3 CNN or CNN.com
4 Facebook
5 NBC
6 ABC or ABCNews.com
7 NPR
8 The New York Times
9 Local radio
10 CBS or CBSNews.com
11 MSNBC
12 The Washington Post
13 Another newspaper
14 Newsmax, OANN, Daily Wire or Daily Caller
15 Another TV network
16 Other</td><td colspan="2">[Seed Question] What is your main source of news?
_____ (open-ended)</td></tr>
<tr><td>Go to next question</td><td>If LLM detects close-ended category in seed response:

➔ [Binary Confirmation Probe] I interpreted your answer as generally being about [sampled detected category]. Is that correct?
1 Yes
2 No

If response to binary confirmation probe is ‘No’ or LLM does not detect category in seed response:

➔ [Categorical Confirmation Probe] Do any of the following news sources match your previous answer?
1 Fox News or FoxNews.com
…
16 None of the above match my answer</td><td>If LLM detects that response is relevant, but not specific:

➔ [Elaboration Probe] (example) You mentioned [vague concept detected by LLM]. Could you tell me more?

Else if LLM detects that response is not relevant:

➔ [Relevance Probe] (example) I’m not sure what you mean by [irrelevant concept detected by LLM]. Could</td></tr>
</table>

		you rephrase what you said? *Else:* *Go to next question*

Demographics Module			
Q#	**Control** (No Probes)	**Treatment 1** (Confirmation Probes)	**Treatment 2** (Elaboration or Relevance Probes)
Q5	**[Age]** Thank you for answering those questions. Now I have a few questions about you… What is your age? ___ (numeric input)		
Q6	**[Gender]** How would you describe your gender? 1 Male 2 Female 3 Other		
Q7	**[Educ]** What is the highest level of education you have completed? 1 Less than high school 2 High school graduate or equivalent 3 Some college/associate's degree 4 Bachelor's degree 5 Postgraduate study/professional degree		
Q8	**[Employment]** Are you currently working? 1 Yes 2 No		

<table>
<tr><th colspan="3">If response to Q8 is Yes:
Q9. Main Occupation</th></tr>
<tr><th>Control
(No Probes)</th><th>Treatment 1
(Confirmation Probes)</th><th>Treatment 2
(Elaboration or Relevance Probes)</th></tr>
<tr><td>[Seed Question] What kind of work do you do in your principal job? Your principal job is the job at which you work the most hours. Select from the following list of job codes from the Bureau of Labor Statistics that best describes your principal job.
1 Management Occupations
2 Business and Financial Operations Occupations
3 Computer and Mathematical Occupations
4 Architecture and Engineering Occupations
5 Life, Physical, and Social Science Occupations
6 Community and Social Service Occupations
7 Legal Occupations
8 Educational Instruction and Library Occupations
9 Arts, Design, Entertainment, Sports, and Media Occupations
10 Healthcare Practitioners and Technical Occupations
11 Healthcare Support Occupations
12 Protective Service Occupations
13 Food Preparation and Serving Related Occupations
14 Building and Grounds Cleaning and Maintenance Occupations
15 Personal Care and Service Occupations
16 Sales and Related Occupations
17 Office and Administrative Support Occupations
18 Farming, Fishing, and Forestry Occupations
19 Construction and Extraction Occupations
20 Installation, Maintenance, and Repair Occupations
21 Production Occupations
22 Transportation and Material Moving Occupations
23 Military Specific Occupations</td><td colspan="2">[Seed Question] What kind of work do you do in your principal job? Your principal job is the job at which you work the most hours.
_____ (open-ended)</td></tr>
<tr><td>Go to next question</td><td colspan="2">If LLM detects close-ended category in seed response:

➔ [Binary Confirmation Probe] I interpreted your answer as generally being about [sampled detected category]. Is that correct?
1 Yes
2 No

If response to binary confirmation probe is ‘No’ or LLM does not detect category in seed response:</td></tr>
</table>

	➔ **Categorical Confirmation Probe:** Which of the following reasons matches your answer to the previous question? 1 Management Occupations ... 24 None of the above

Respondent Experience Module			
Q#	**Control** (No Probes)	**Treatment 1** (Confirmation Probes)	**Treatment 2** (Elaboration or Relevance Probes)
Q10	Finally, please answer the following questions about your survey experience. **[Tone]** Overall, how formal was the tone of questions in this survey? 1 Very formal 2 Somewhat formal 3 Not very formal 4 Not formal at all		
Q11	**[Quality]** Overall, how would you rate the quality of your responses to questions in this survey? 1 Very high quality 2 Somewhat high quality 3 Somewhat low quality 4 Very low quality		
Q12	**[Ease]** How easy was it to complete this survey? 1 Very easy 2 Somewhat easy 3 Neither easy nor difficult 4 Somewhat difficult 5 Very difficult		
Q13	**[Frustration]** How frustrating was this survey experience? 1 Very frustrating 2 Somewhat frustrating 3 Not very frustrating 4 Not frustrating at all		

Our survey experiment was fielded on the conversational AI platform Inca, with an interface structurally identical to that depicted in Figure 1 in the manuscript. The large language model (LLM) used to administer probes and conduct classification in our experiment is SmartProbe (Seltzer et al., 2023), a model fine-tuned from the InstructGPT family (Ouyang et al., 2022) which are a collection of models themselves fine-tuned from the GPT-3 foundational model (Brown et al., 2020).

Structure. To generate probes in a real-time survey, the model is given an instructional prompt consisting of (1) the seed question, (2) the respondent's response, along with (3) structural information about the dialogue context—such as whether the prior response was vague, off-topic, or already specific—provided by a lightweight semantic classifier trained to detect relevance and elaboration needs.[7] This classifier helps the system determine whether a follow-up probe should seek clarification or elaboration, enabling the prompt to convey both the content and conversational intent of the prior exchange. Multiple candidate probes are generated, screened for clarity, safety, and contextual relevance, and then ranked. If none pass a quality threshold, a rule-based backup probe was used.

While the exact prompt format used by the proprietary SmartProbe system is not publicly available, the following simplified structure reflects the key components used to generate probing questions:

```
Seed Question: "What is your current occupation?"
Respondent Answer: "I work in a lab."

[Dialogue Context]:
- Answer Type: Short
- Relevance: Relevant
- Specificity: Low
- Requires Elaboration: Yes

[Optional Metadata]:
- Target Codebook Category: Medical or Scientific Occupations

Instruction: Generate a follow-up question that asks the respondent to
elaborate on their job role using clear and concise language.
```

Fine Tuning and Evaluation. InstructGPT models, upon which our LLM was based, were trained using a multi-step fine-tuning process designed to align language model behavior with human preferences (Ouyang et al., 2022). This process began by collecting human-written examples of desired outputs in response to user prompts. Human annotators were then asked to compare different model outputs and indicate which they prefer. These preferences are used to train the model to better follow instructions and produce high-quality responses. The resulting models are evaluated both by human raters and on standardized benchmarks. For instance, on the `TruthfulQA` benchmark – a set of questions designed to test whether models give factually accurate answers – InstructGPT produced truthful responses about twice as often as GPT-3. On the `RealToxicityPrompts` dataset – a collection of prompts used to assess how often

[7] As SmartProbe is a proprietary system developed by the vendor, the exact instructional prompts used to generate probes are not publicly available.

models generate harmful or offensive content – InstructGPT reduced toxic completions by approximately 25% when prompted to be respectful. Notably, human raters preferred outputs from the smaller InstructGPT model (1.3 billion parameters) over the much larger GPT-3 model (175 billion parameters) in 70–85% of comparisons, demonstrating the effectiveness of fine-tuning.

SmartProbe built on this foundation through additional fine-tuning on domain-specific examples from the vendor's proprietary Market Research Knowledge Base. This corpus includes question–answer pairs written by professional researchers across topics such as advertising testing, brand perception, and customer experience. In prior vendor evaluations, 69% of SmartProbe-generated probe questions – elicited via the augmented seed question/response/dialogic context prompts described above – were rated 4 or 5 (on a 5-point scale) by expert annotators, with fewer than 1% rated as poor quality. In a parallel field test with over 900 North American respondents, 76% of responses elicited by SmartProbe were rated high quality, compared to only 25% for responses to standard, generic probes.

Deployment. In our study, SmartProbe was deployed with a temperature setting of 0 to ensure reproducibility and a relatively high presence penalty to discourage verbatim copying of training examples.

Table A2. Completions by Treatment Condition

Condition	Completes	Dropouts
Control	601	43
Treatment 1 (Conf. Probing)	601	64
Treatment 2 (Elab/Rel. Probing)	601	88

Table A3. Summary Statistics of Interview Duration by Treatment Condition

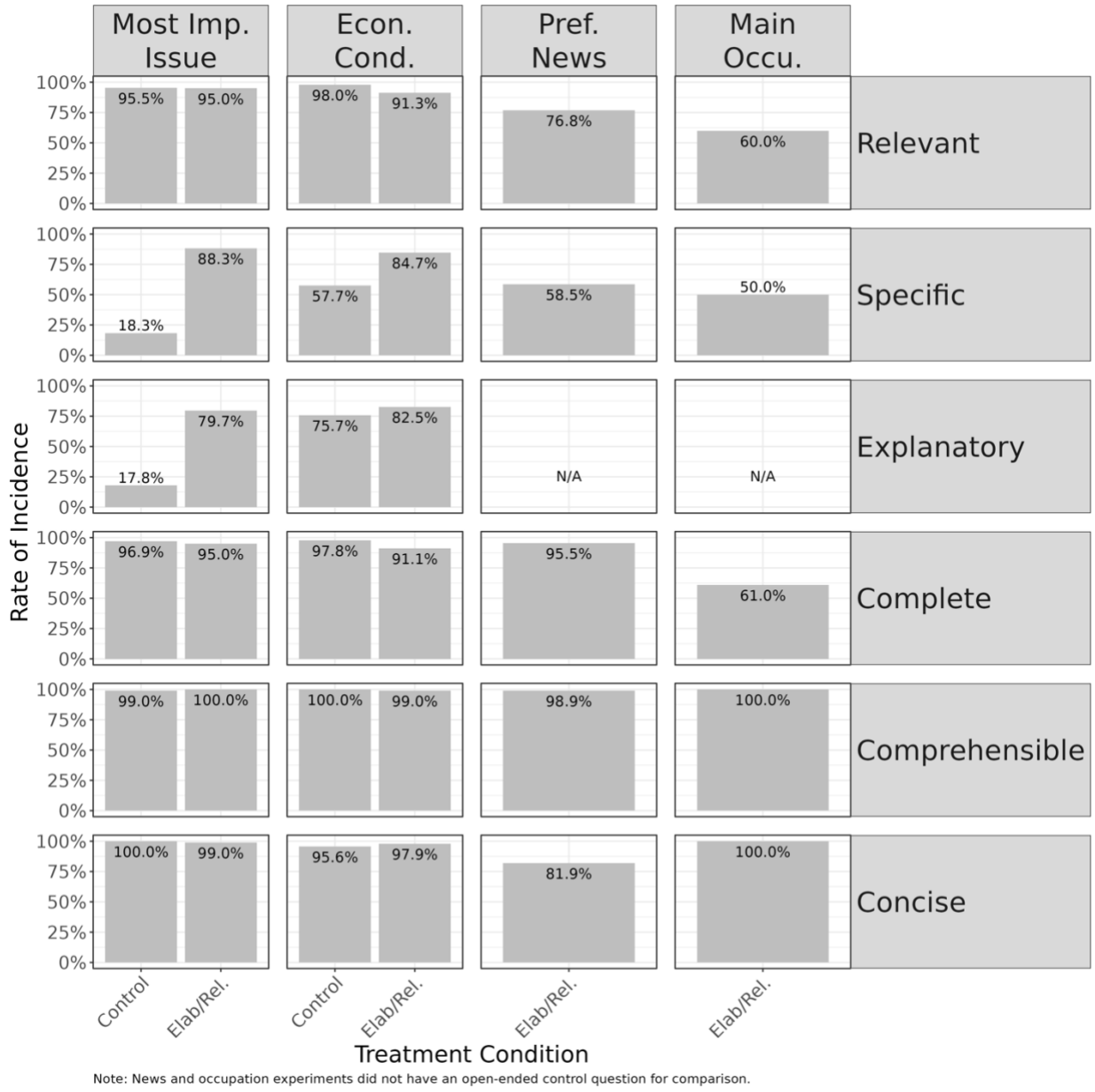
Condition	Duration of Interview (Minutes)					
	1%	25%	Median	Mean	75%	99%
Control	0m	1.9m	2.4m	3.1m	3.2m	14.2m
Treatment 1 (Conf. Probing)	0m	2.5m	3.3m	4.1m	4.7m	15.8m
Treatment 2 (Elab/Rel. Probing)	0m	3.1m	4.5m	5.9m	6.8m	29.3m

Table A4. Summary Statistics of Completion Time by Treatment Condition

Condition	Duration of Interview Among Non-Dropouts (Minutes)					
	1%	25%	Median	Mean	75%	99%
Control	1.3m	1.9m	2.5m	3.3m	3.3m	16.1m
Treatment 1 (Conf. Probing)	1.8m	2.7m	3.5m	4.4m	4.8m	16.2m
Treatment 2 (Elab/Rel. Probing)	1.9m	3.4m	4.9m	6.4m	7.1m	29.3m

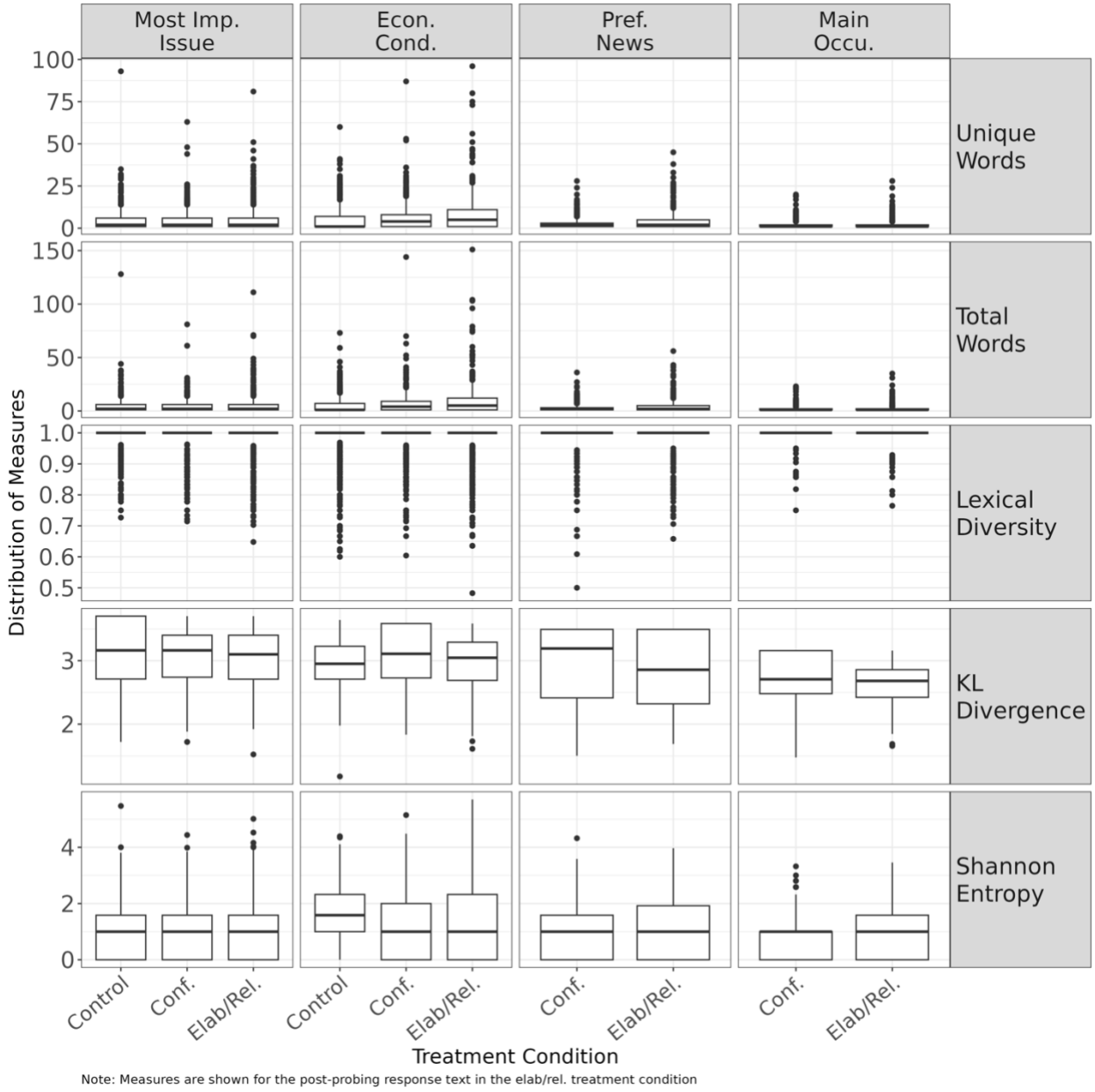
Figure A1. Sample Composition

Figure A2. Summary of Human-Coded Quality Criteria



Note: News and occupation experiments did not have an open-ended control question for comparison.

Figure A3. Summary of NLP-Based Informational Measures



Note: Measures are shown for the post-probing response text in the elab/rel. treatment condition

Figure A4. Summary of Respondent Attrition

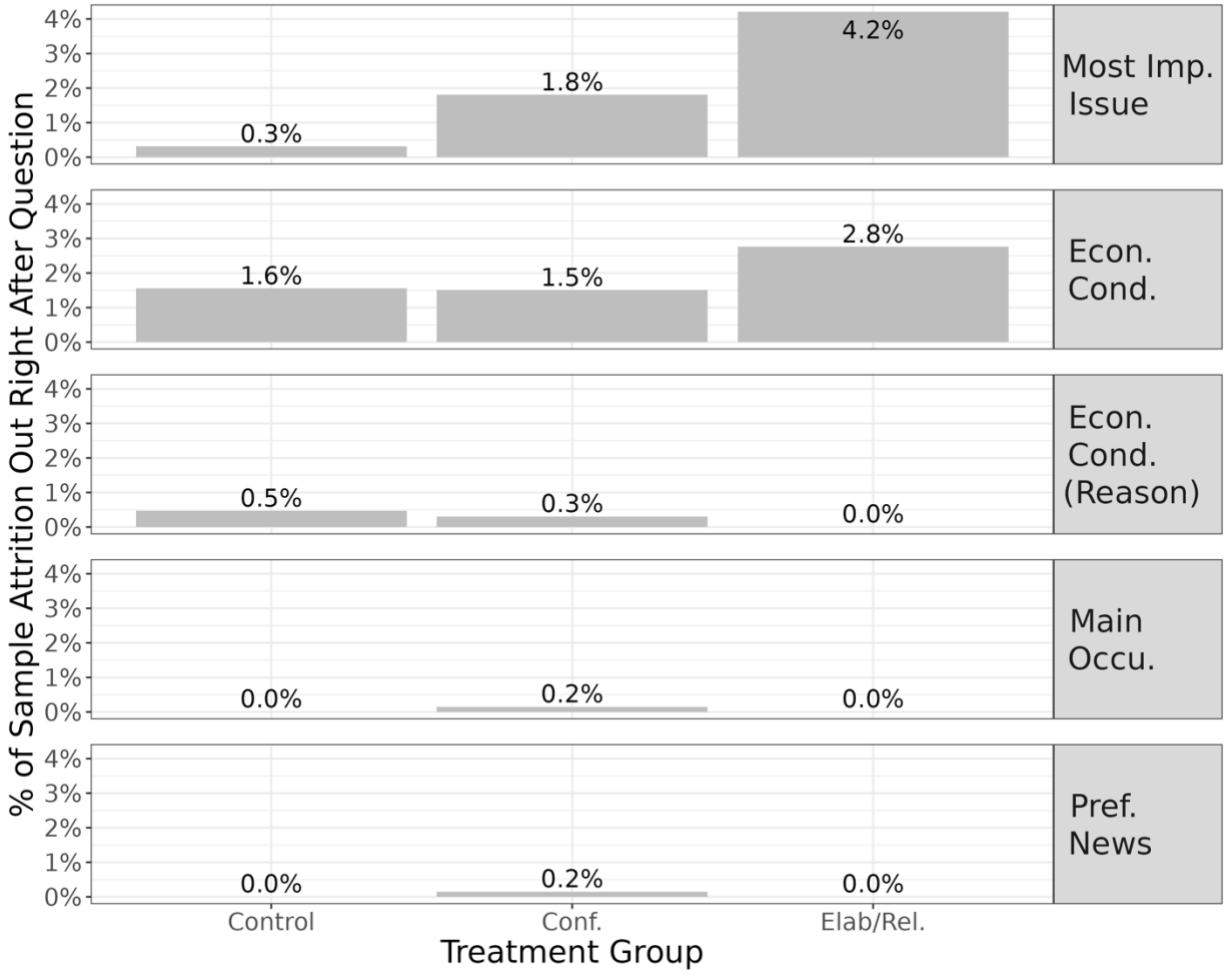
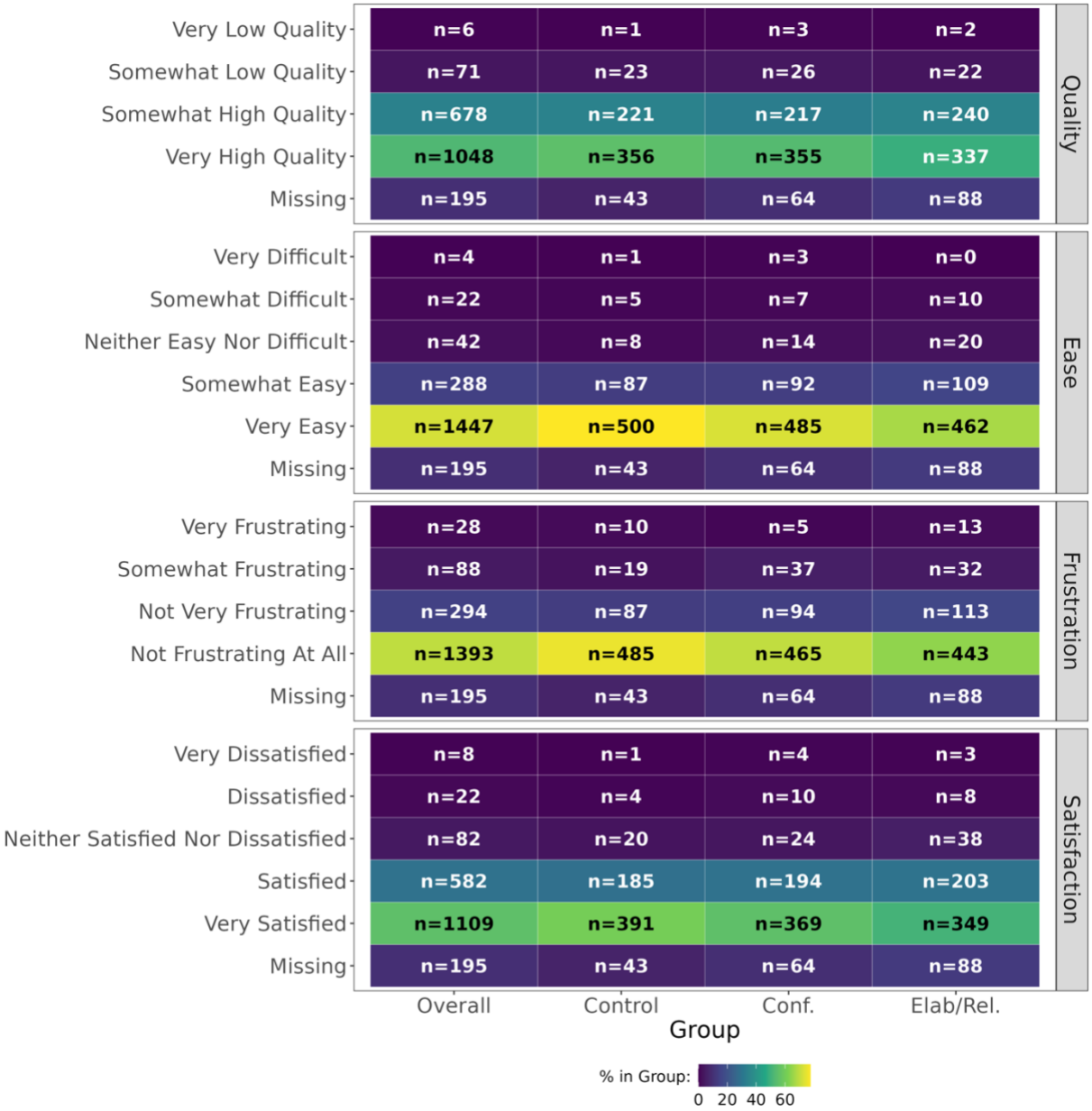


Figure A5. Summary of Self-Reported Respondent Experience



A2. Additional Live Coding Results

Figure A6. Evidence of Acquiescence Bias in Confirmation Probing: Comparison of Response Categories (Coded by Coders vs. Confirmed by Respondent)

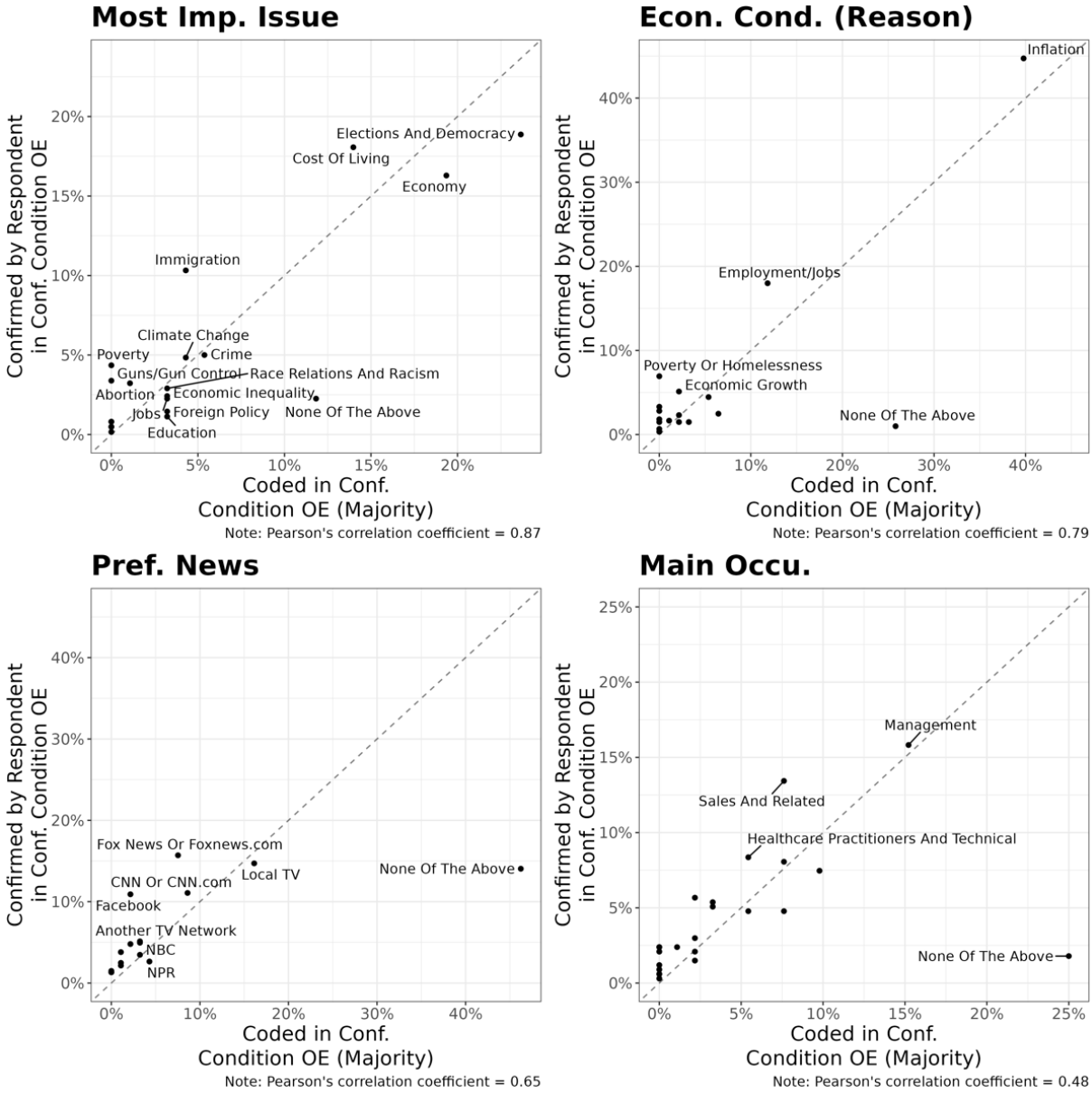
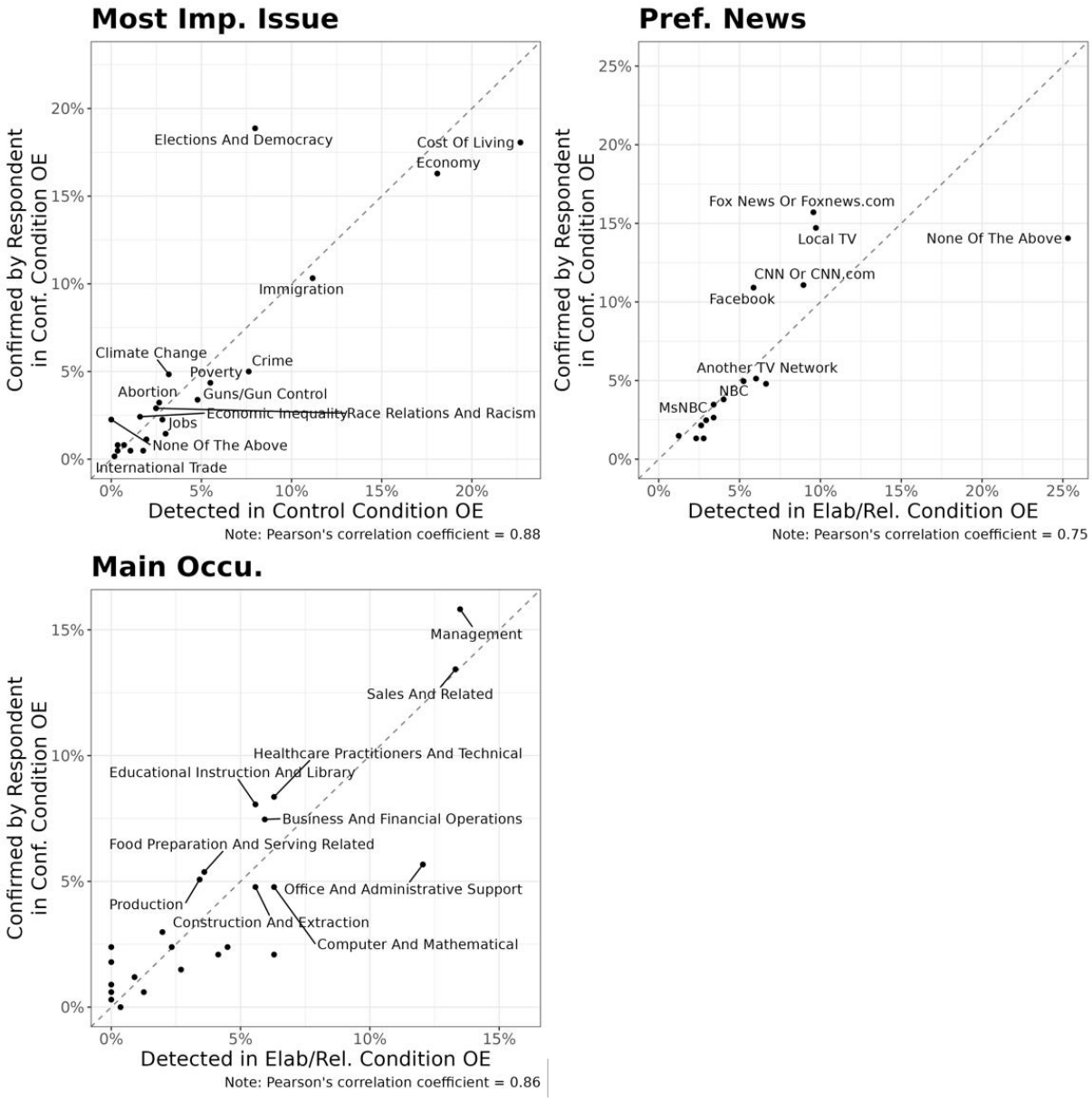
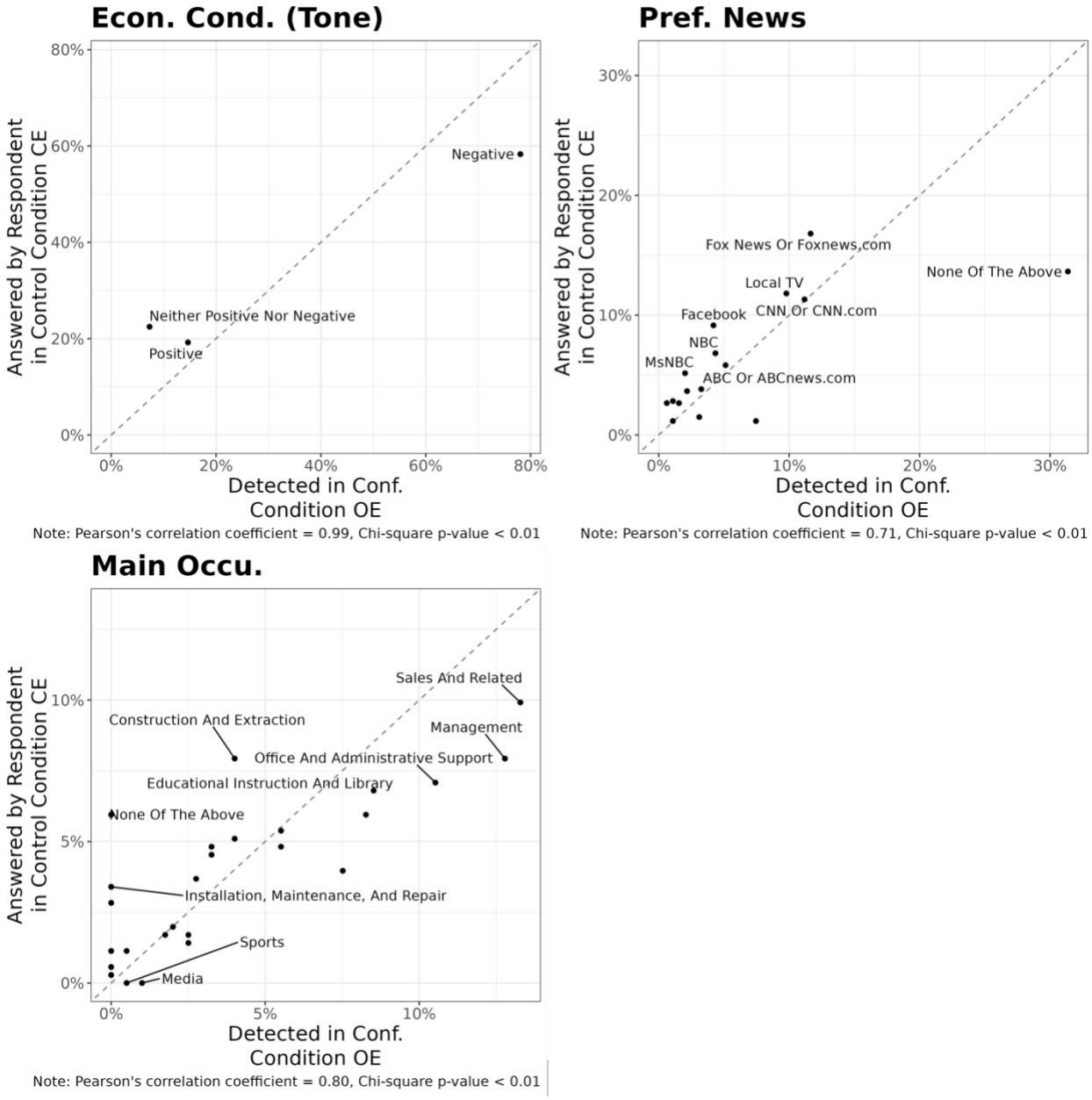


Figure A7. Comparison of Response Categories (Confirmed vs. Coded)[8]



[8] Due to the way the economic conditions (reason) question was asked in the control condition (close-ended) versus the elaboration/relevance conditions (as a probe, rather than a standalone static question), coding of categories was not possible and is therefore omitted from Figure 4.

Figure A8. Comparison of Response Categories (Live Coded by Chatbot vs. Answered in Control Close-End)[9]



[9] "Tone" here is used interchangeably with "sentiment".

Figure A9. Comparison of Response Categories (Live Coded by Chatbot vs. Coded in Other Open-Ended Conditions)

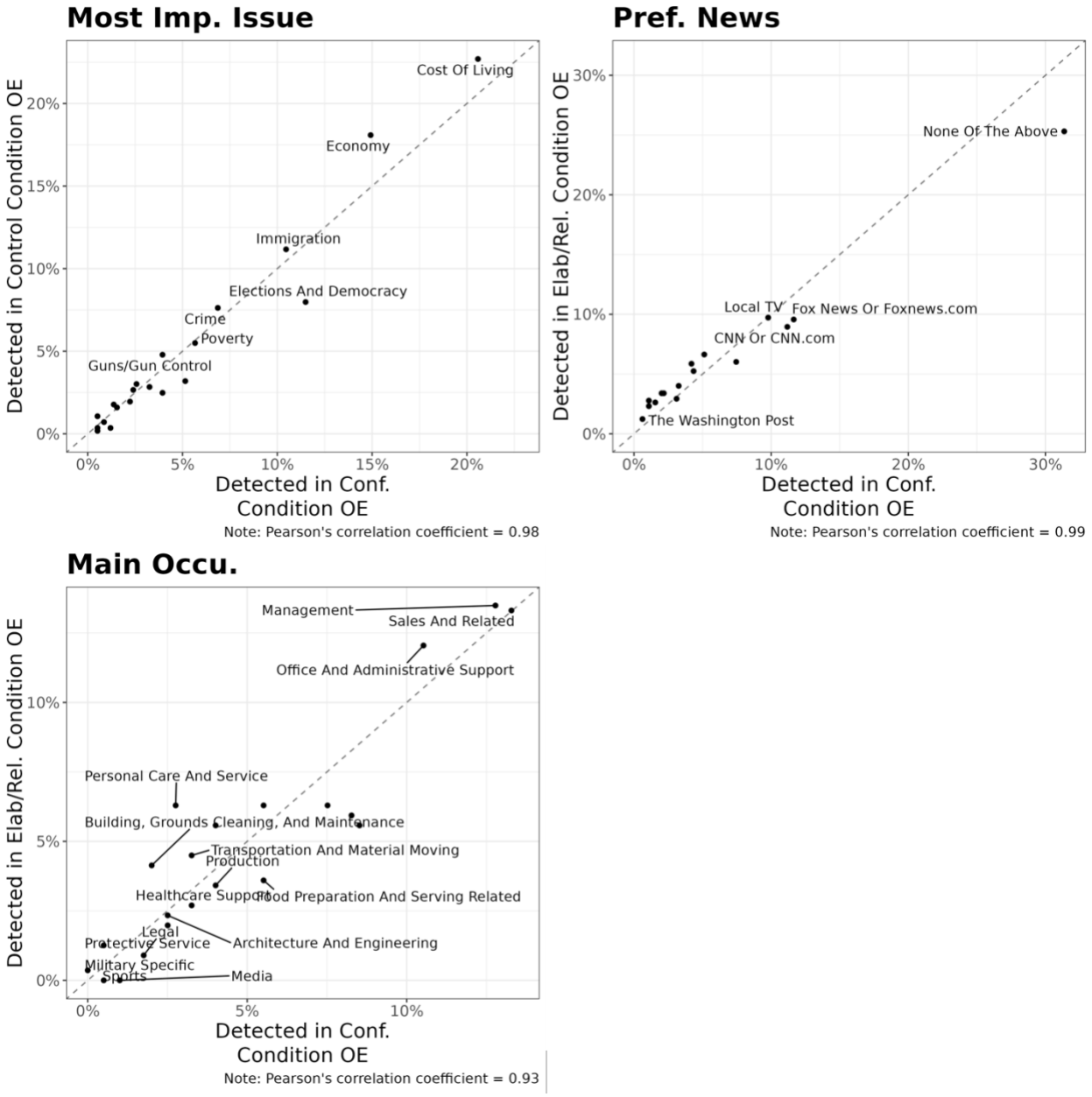
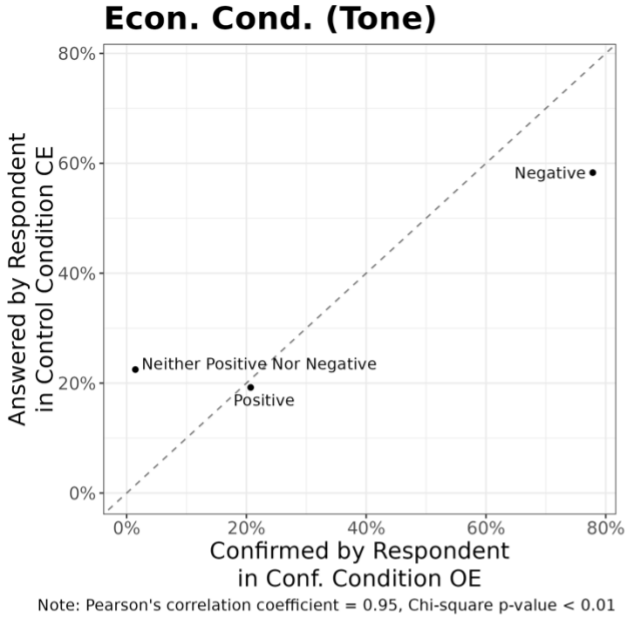
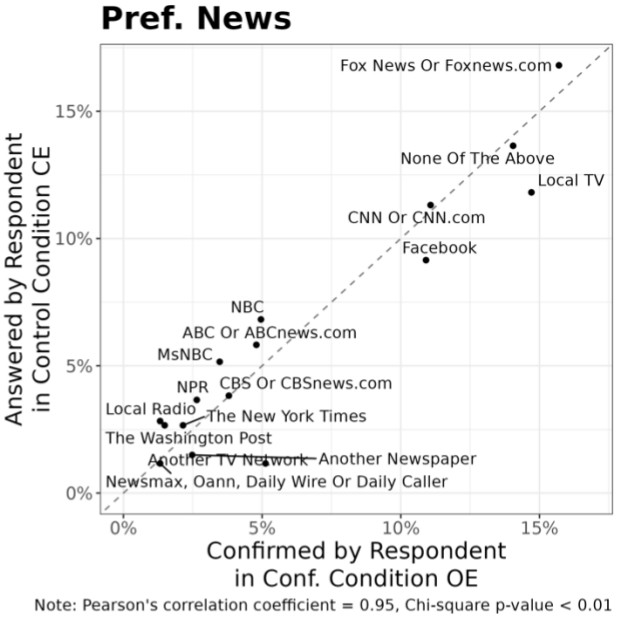


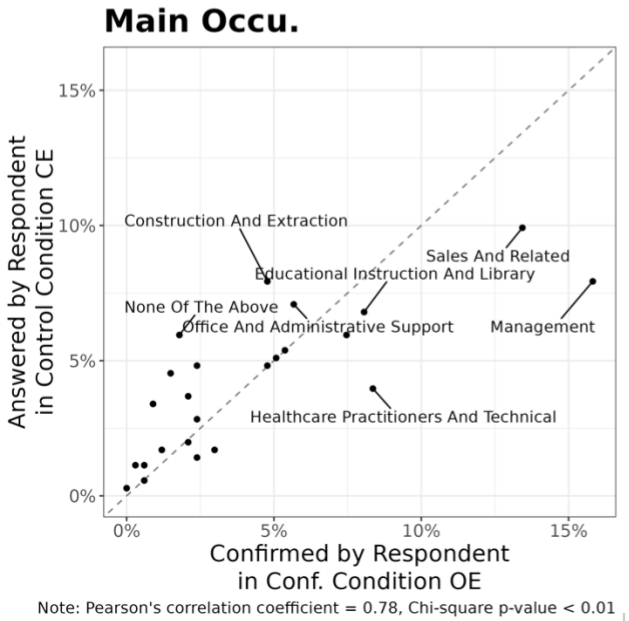
Figure A10. Comparison of Response Categories (Confirmed vs. Answered in Close-End)



Note: Pearson's correlation coefficient = 0.95, Chi-square p-value < 0.01



Note: Pearson's correlation coefficient = 0.95, Chi-square p-value < 0.01



Note: Pearson's correlation coefficient = 0.78, Chi-square p-value < 0.01

Figure A11. Comparison of Response Categories (Confirmed vs. Live Coded by Chatbot)

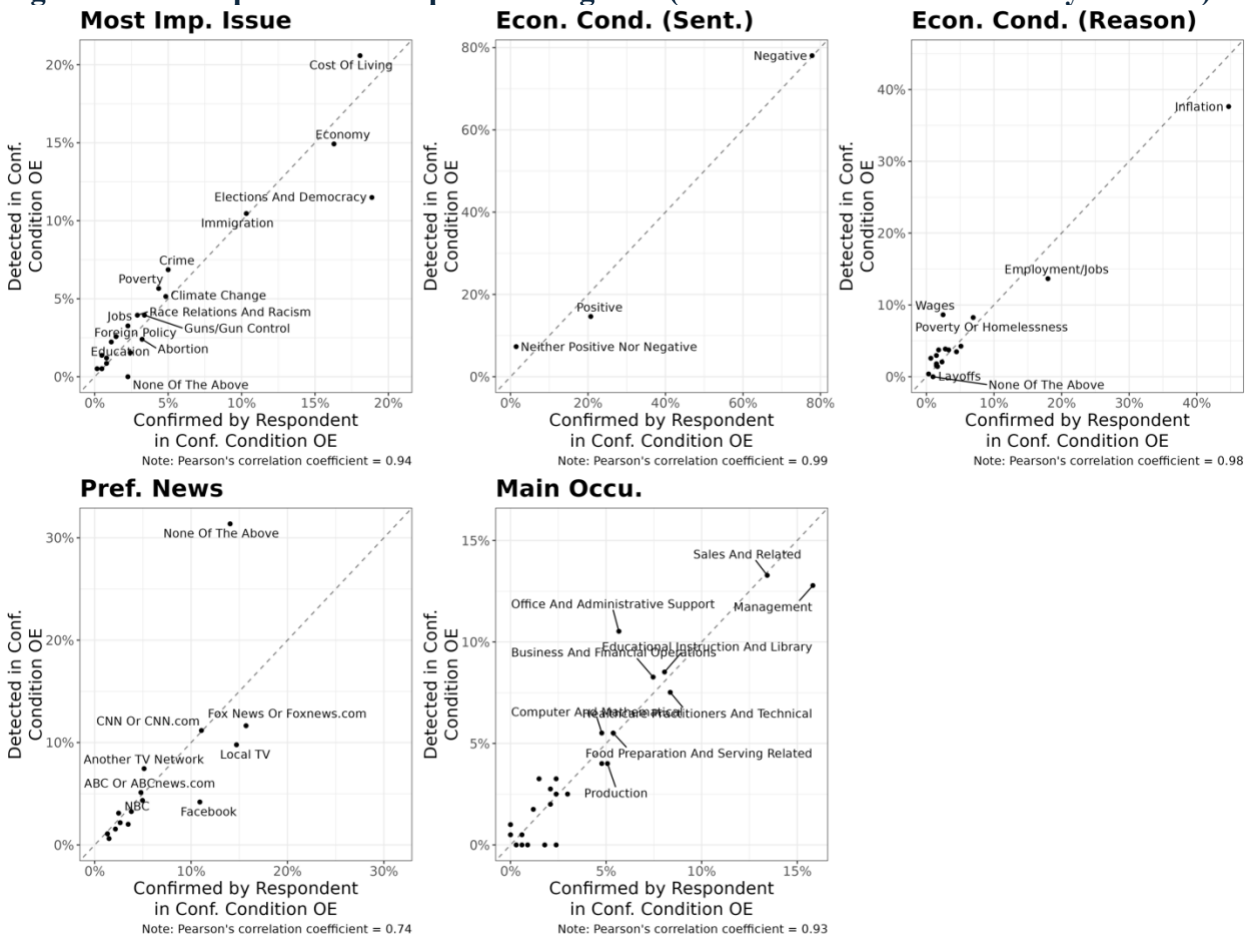
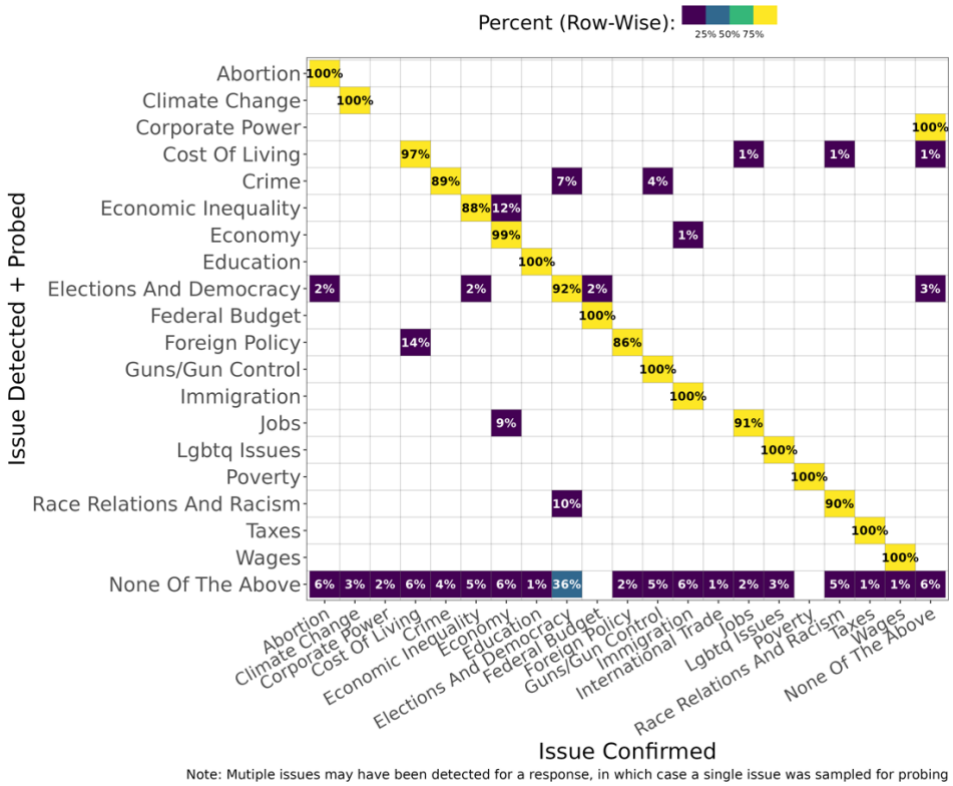


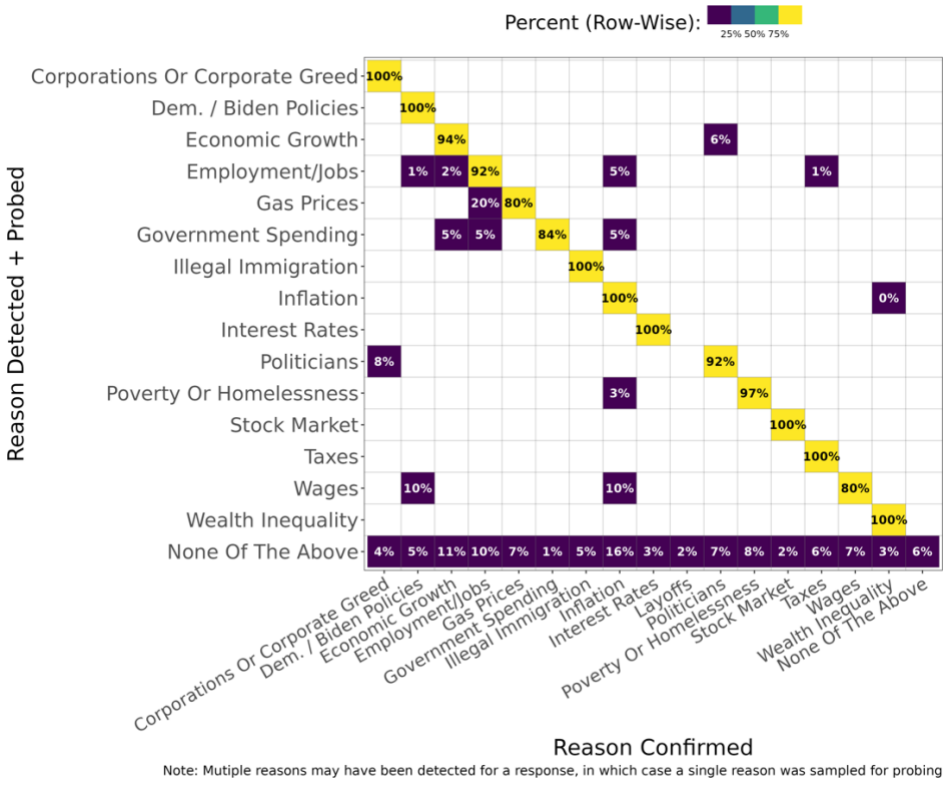
Table A5. Respondent Confirmation (Treatment 1) by Device Type

Device	Most Imp. Issue	Econ. Cond. (Sentiment)	Econ. Cond. (Reason)	Pref. News	Main Occu.
Smartphone (n=445)	74.0%	95.9%	77.1%	62.8%	87.7%
Desktop (n=194)	73.3%	97.2%	88.8%	74.0%	76.8%
Tablet (n=24)	69.6%	95.0%	91.3%	65.2%	90.9%

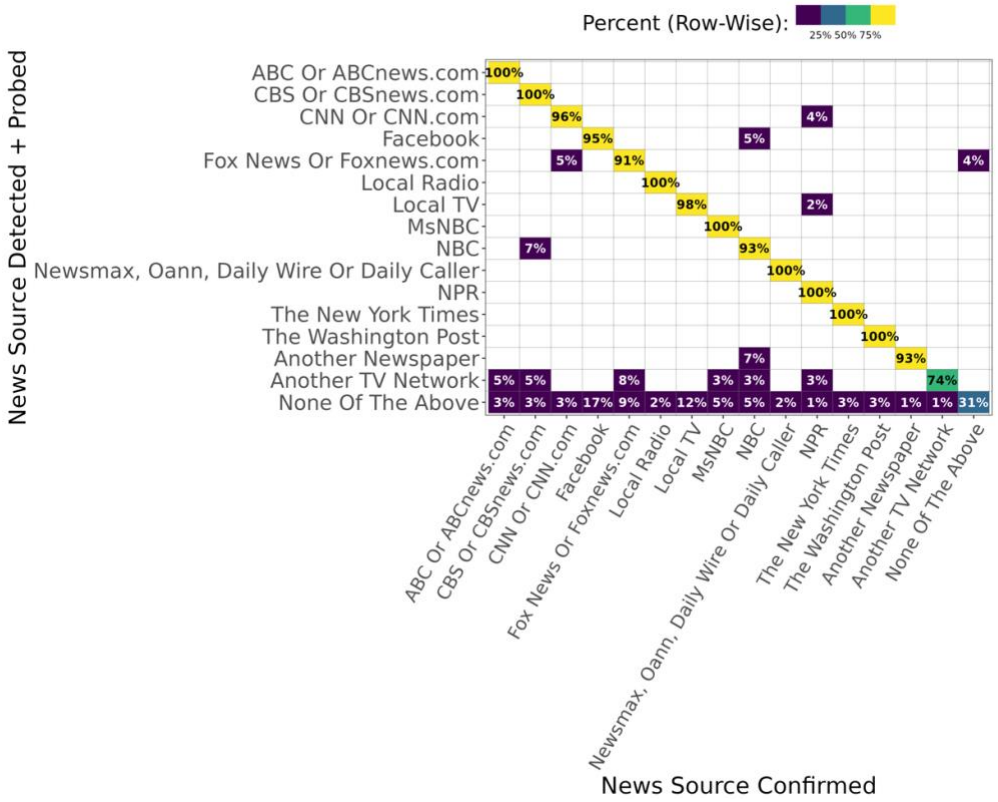
Figure A12. Live Coding (Treatment 1) Confusion Matrices



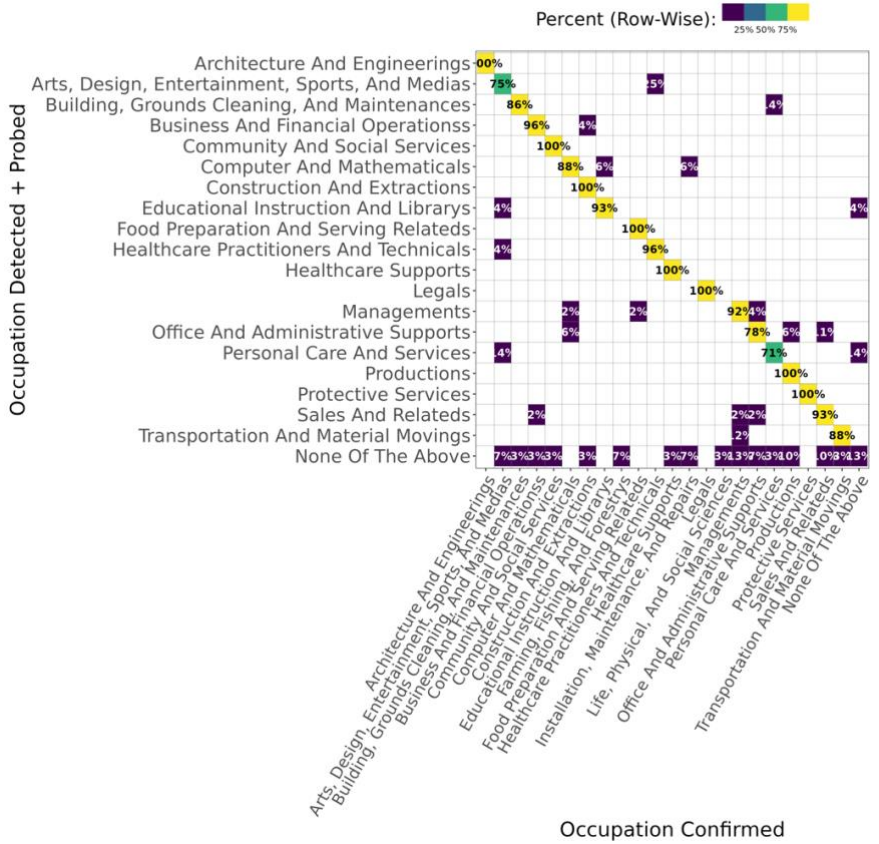
(a) Most Important Issue



(b) Economic Conditions (Reason)



(c) Preferred News Source



(d) Occupation

A3. Additional Probing Results

Regression results in this section follow the same corrections and specifications as in the main text.

Table A6. Examples of Elaboration/Relevance Probes

Probe Triggered	Example Question Sequence			
	Most Imp. Issue	**Econ. Cond.**	**Pref. News**	**Main Occu.**
Elaboration Probe	**[Seed Response]** Economy and inflation **[Chatbot Probe]** Thank you for sharing your thoughts. Could you please elaborate on why you think the economy and inflation are the most important problems facing the country today?	**[Seed Response]** Our food bank needs more donations since more people are in need **[Chatbot Probe]** It sounds like you're saying that there are more people in need of assistance. Can you tell me more about how the economy is affecting people's ability to provide for themselves?	**[Seed Response]** New York Times and the Washington Post. **[Chatbot Probe]** That's interesting. Can you tell me more about why you prefer these sources over others?	**[Seed Response]** Contractor **[Chatbot Probe]** That's interesting. Could you tell me more about the type of contracting work you do?
Relevance Probe	**[Seed Response]** Democracy not what religion **[Chatbot Probe]** I'm sorry, I'm not sure I understand your response. Could you please elaborate on what you mean by "democracy not what religion"?	**[Seed Response]** good **[Chatbot Probe]** Hmm I think I know what you mean, but could you explain a bit more?	**[Seed Response]** Internet and my husband **[Chatbot Probe]** Thank you for sharing. Could you please clarify which specific websites or online platforms you use to consume news? And when you say your husband, do you mean he shares news articles with you or do you discuss news topics together?	**[Seed Response]** per week 35 hours **[Chatbot Probe]** I'm sorry, but I think there might have been a misunderstanding. I was asking about the type of work you do in your principal job. Could you please provide more information about the kind of work you do?

Figure A13. Comparison of Seed Response Quality (Treatment 2 Probe Triggered vs. No Probe Triggered)

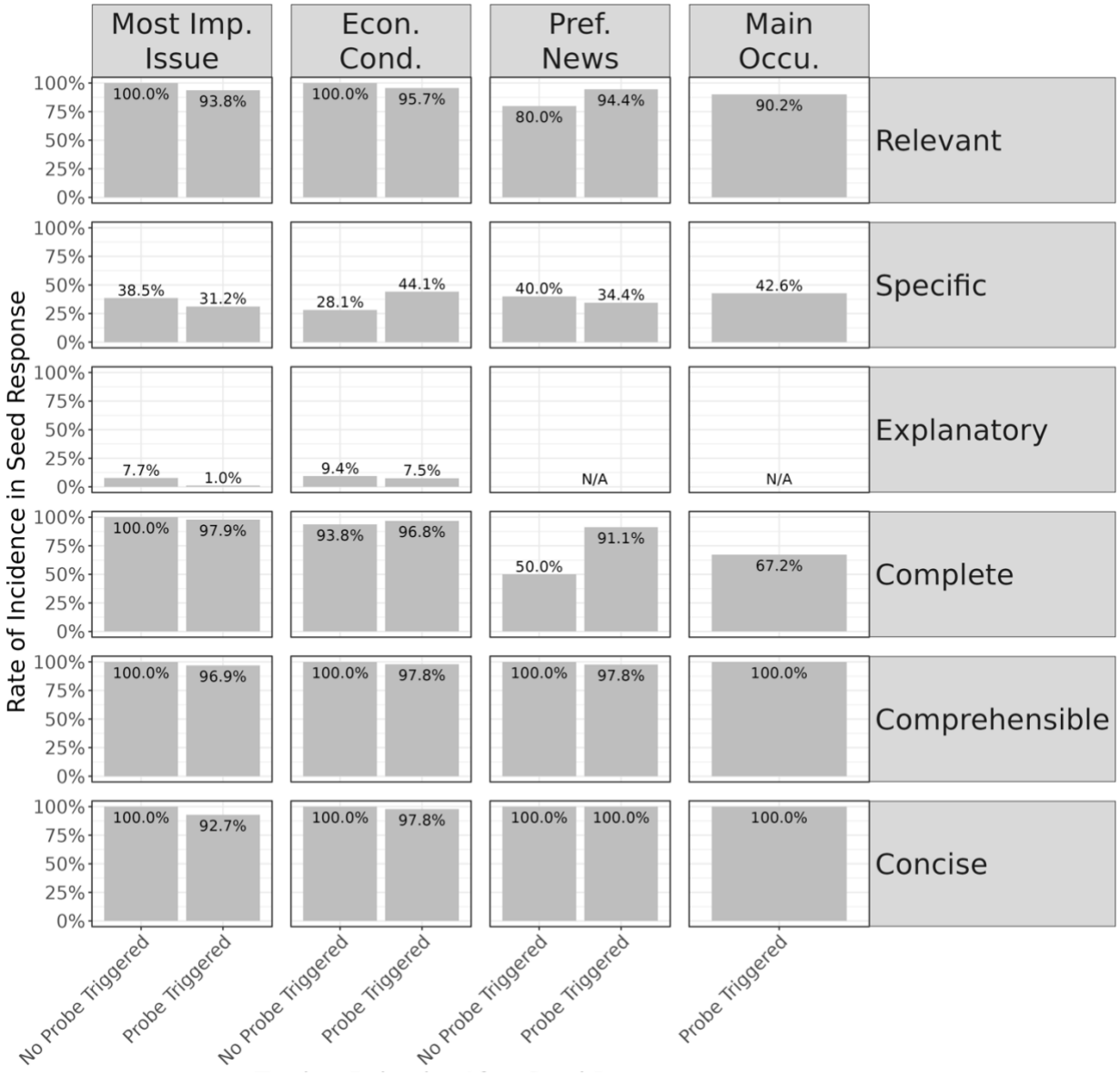
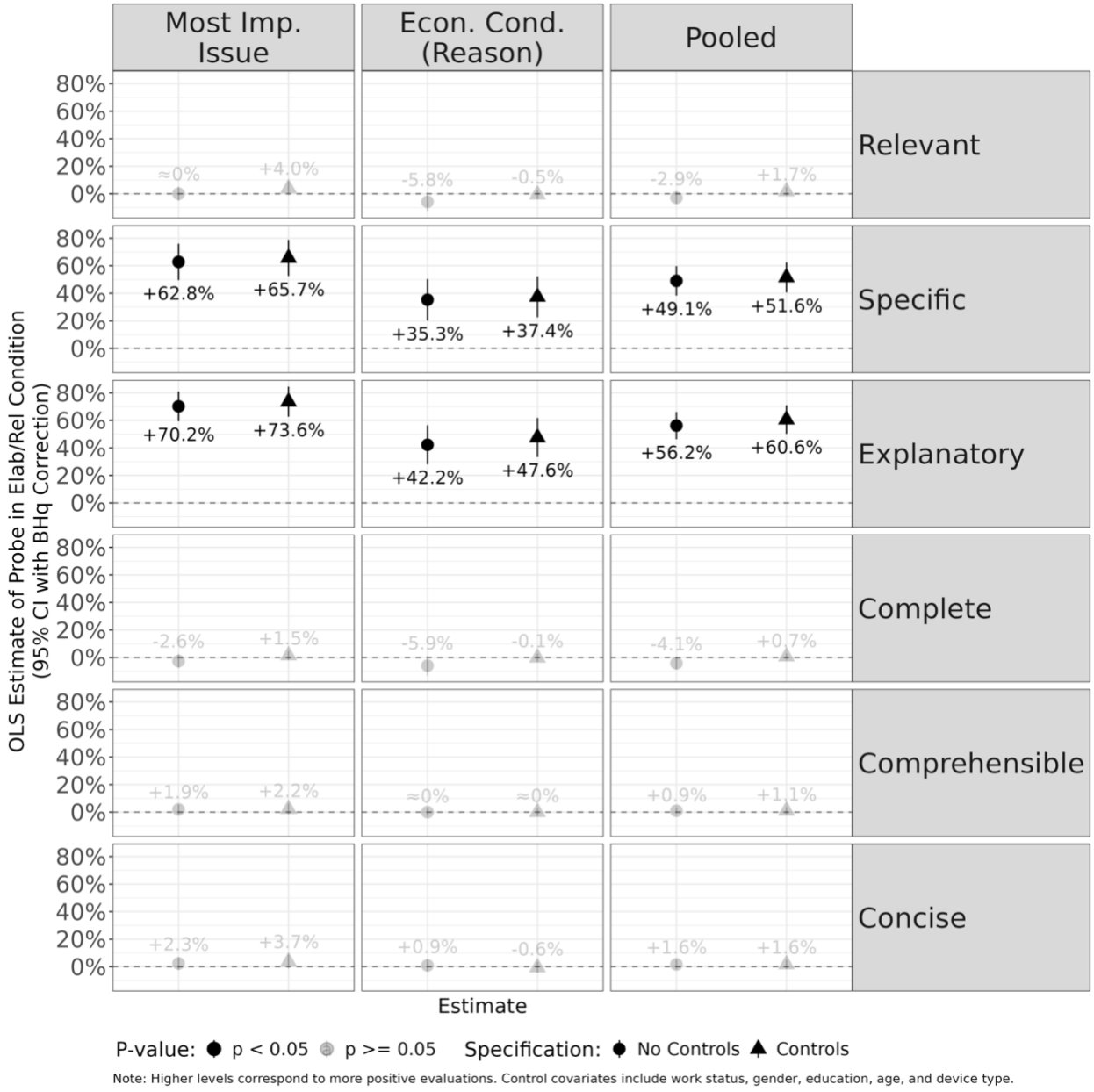


Figure A14. Effects of Elaboration/Relevance Probing (Treatment 2 Post-Probe vs. Pre-Probe) on Human-Coded Quality Criteria

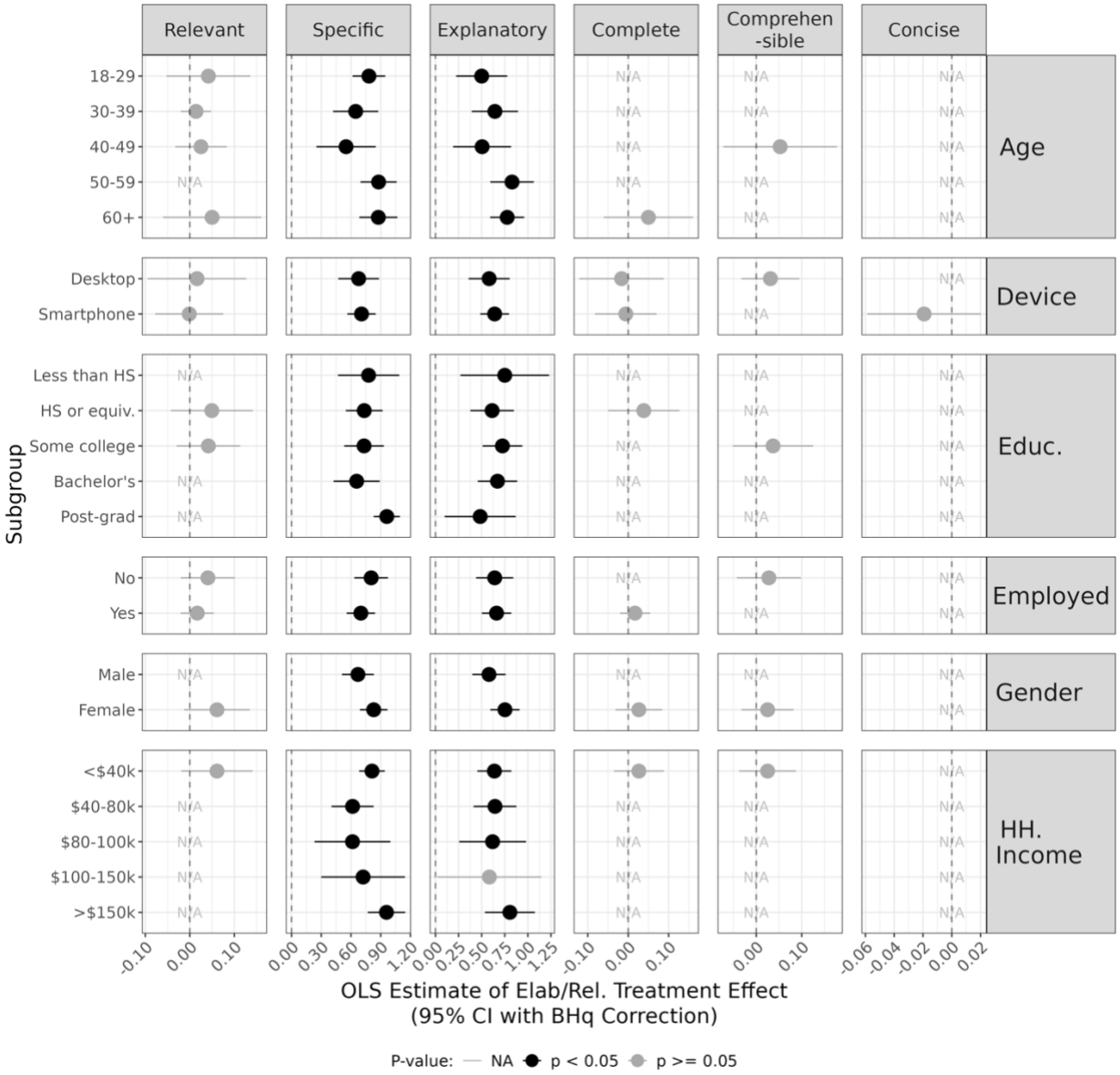


A4. Heterogeneous Treatment Effects of Probing

To capture heterogeneity in response quality effects, we collected demographic information and asked about the survey experience through a series of Likert-scale questions at the end of the survey. For a consistent respondent experience and for validation purposes, we included demographic questions in the survey, even though similar variables were available from the panel provider. Analyses requiring demographic variables (e.g., heterogeneity analyses and multivariate regression models of treatment effects) prioritized the use of panel-provided data when available. In instances where panel data were unavailable, survey-collected variables were used. Our comparison of these sources revealed only minor discrepancies, with a few exceptions. Age was exclusively captured through the survey, introducing the possibility of post-treatment bias if used as a control variable in analyses of treatment effects (Montgomery et al., 2018). We, however, find little to no differences in the regression estimates of treatment effects when including or excluding age as a covariate.

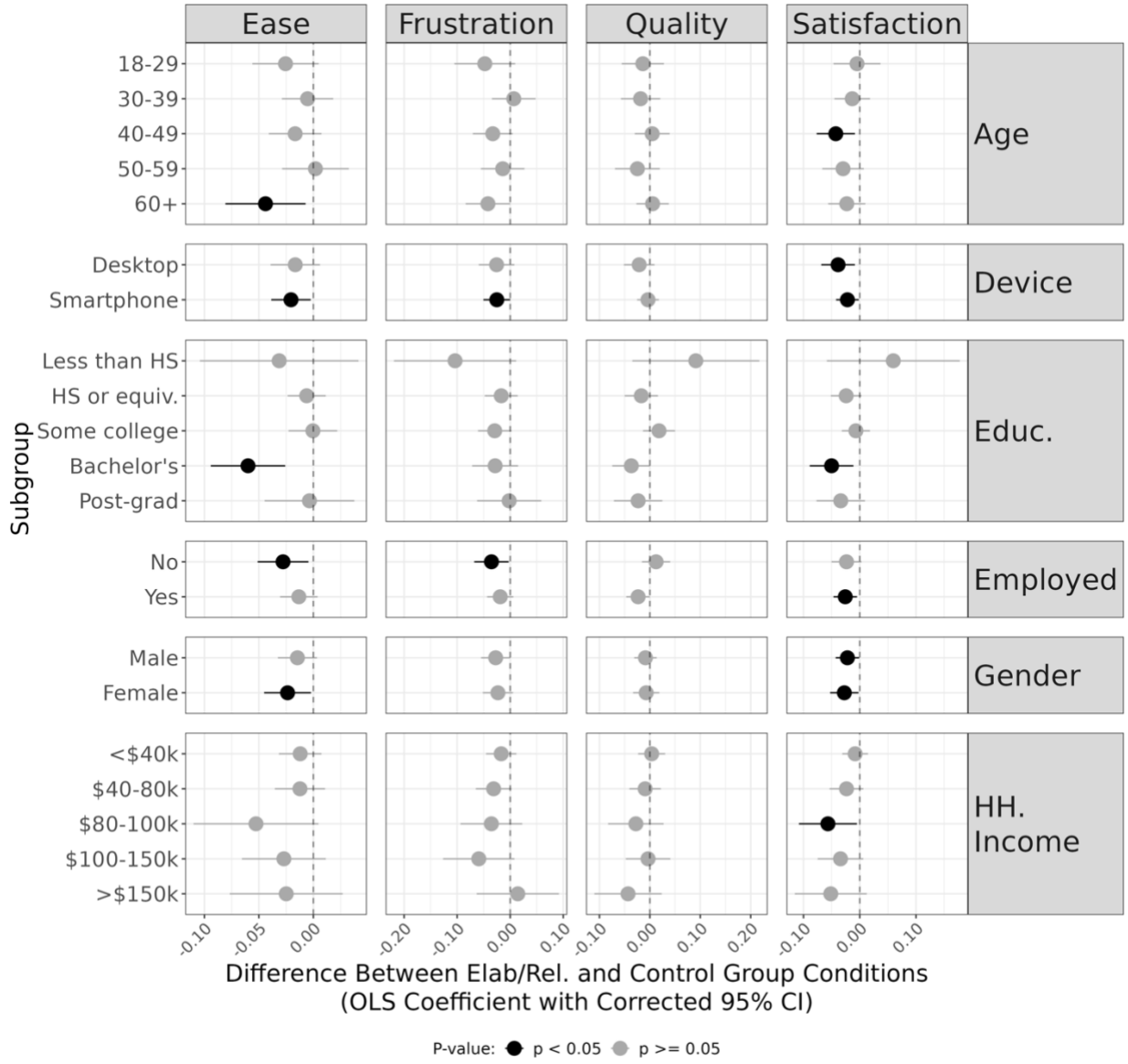
Regression results here follow the same corrections and specifications as in the main text.

Figure A15. Heterogeneous Effects of Elaboration/Relevance Probing (Treatment 2 vs. Control) on Human-Coded Quality Criteria



Note: Higher levels correspond to more positive evaluations. Control covariates include work status, gender, education, age, and device type. N/A indicates not a large enough subgroup or enough outcome variation to estimate heterogeneity.

Figure A16. Heterogeneous Effects of Elaboration/Relevance Probing (Treatment 2) on Self-Reported Respondent Experience



Note: Higher levels correspond to more positive evaluations. Control covariates include work status, gender, education, age, and device type. Outcomes are all normalized to the [0-1] range. Excluded subgroups in 'Other' categories and who answered by tablet due to small sample sizes.