ChatGPT (GPT-4o)
ChatGPT (GPT-4o)

Unleashing AI in Survey Research:
Evaluating the Power and Pitfalls of Large Language Models in Evaluating Survey Questions

UWSC
UNIVERSITY of WISCONSIN
SURVEY CENTER

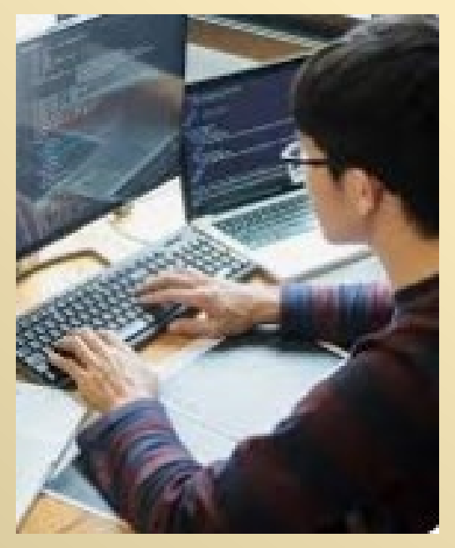
Shengfu Wang[1], Jennifer Dykema[2,3], Nora Cate Schaeffer[2,3], Nadia Assad[2], Dana Garbarski[4], Jiahao Chen[2]

1 National Taiwan University
2 University of Wisconsin Survey Center, University of Wisconsin-Madison
3 Department of Sociology, University of Wisconsin-Madison
4 Department of Sociology, Loyola University Chicago

Background

- **The future? … Large language models (LLMs)**
  - LLMs are types of artificial intelligence programs that use algorithms to perform natural language processing tasks
  - Examples include ChatGPT and Claude
  - LLMs offer potentially unlimited possibilities in survey research, particularly in designing and evaluating survey questions
  - However, critical issues surrounding the implementation and use of LLMs remain unaddressed
  - Foremost among these is
    - How valid and reliable are LLMs as diagnostic tools for reviewing the quality of survey questions?
    - (1) how accurately do LLMs identify problems that would ultimately be associated with data quality
    - (2) how consistent are LLMs across models and with regard to variations in how they are implemented

- **Current practice and the Question Appraisal System (QAS)**
  - Many methods and techniques currently exist for evaluating survey questions (see Maitland and Presser 2016, 2018, 2020)
  - One tool is the Question Appraisal System (QAS; Willis and Lessler 1999)
  - QAS is a framework – a coding manual – specifically designed to identify problems survey questions can pose for respondents (and for a limited number of situations, interviewers)
  - QAS includes check-list of 27 characteristics that may lead to problems in cognitive processing and responding
  - Coding categories include problems with
    - Reading, instructions, clarity, assumptions, knowledge/memory, sensitivity bias, and response categories.
  - To use the QAS, a **human** coder must be trained (by reading a manual) and then apply the coding rules on a question-by-question basis



  - **Coding questions can be very time consuming and unreliable**

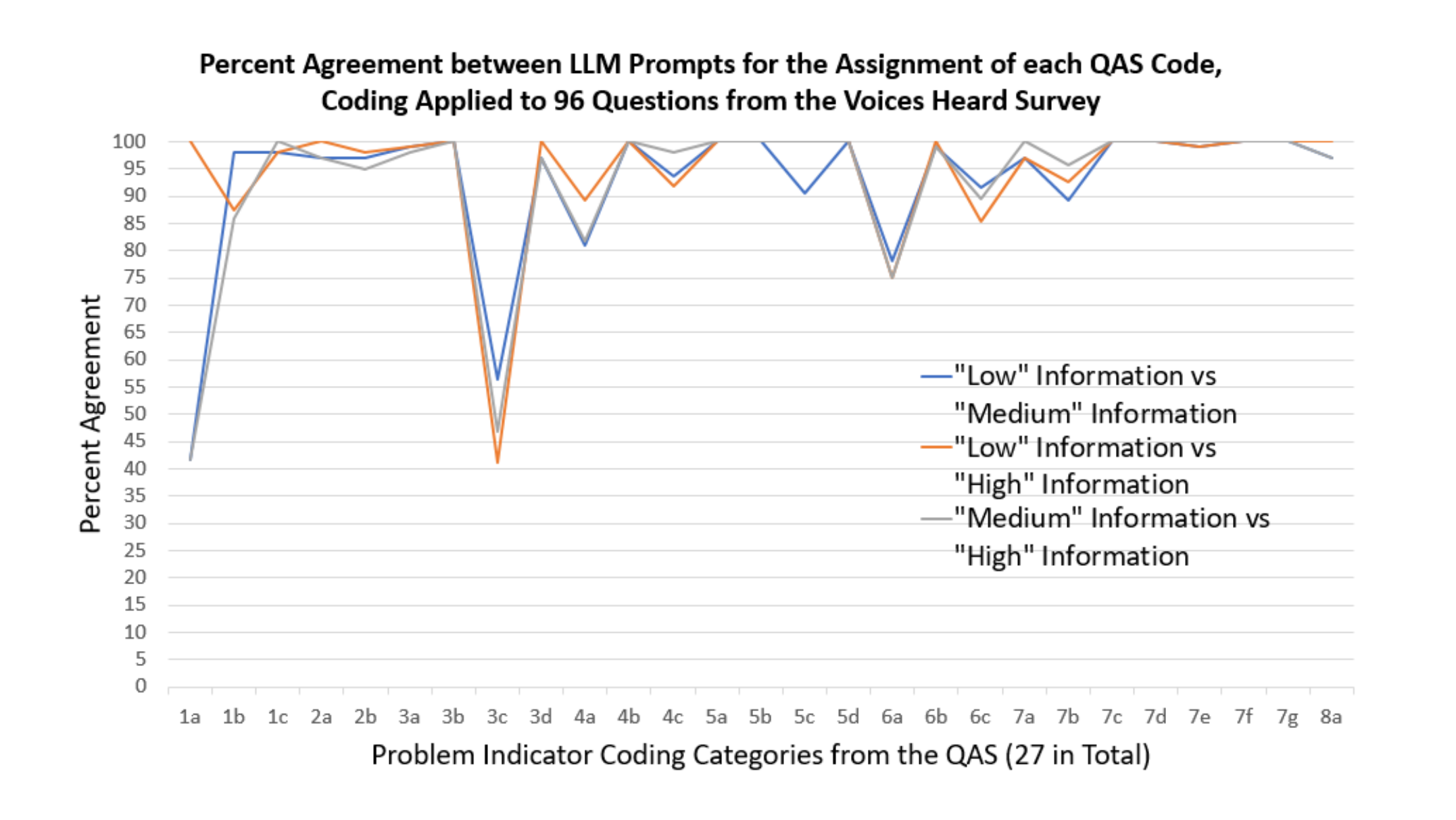
Methods

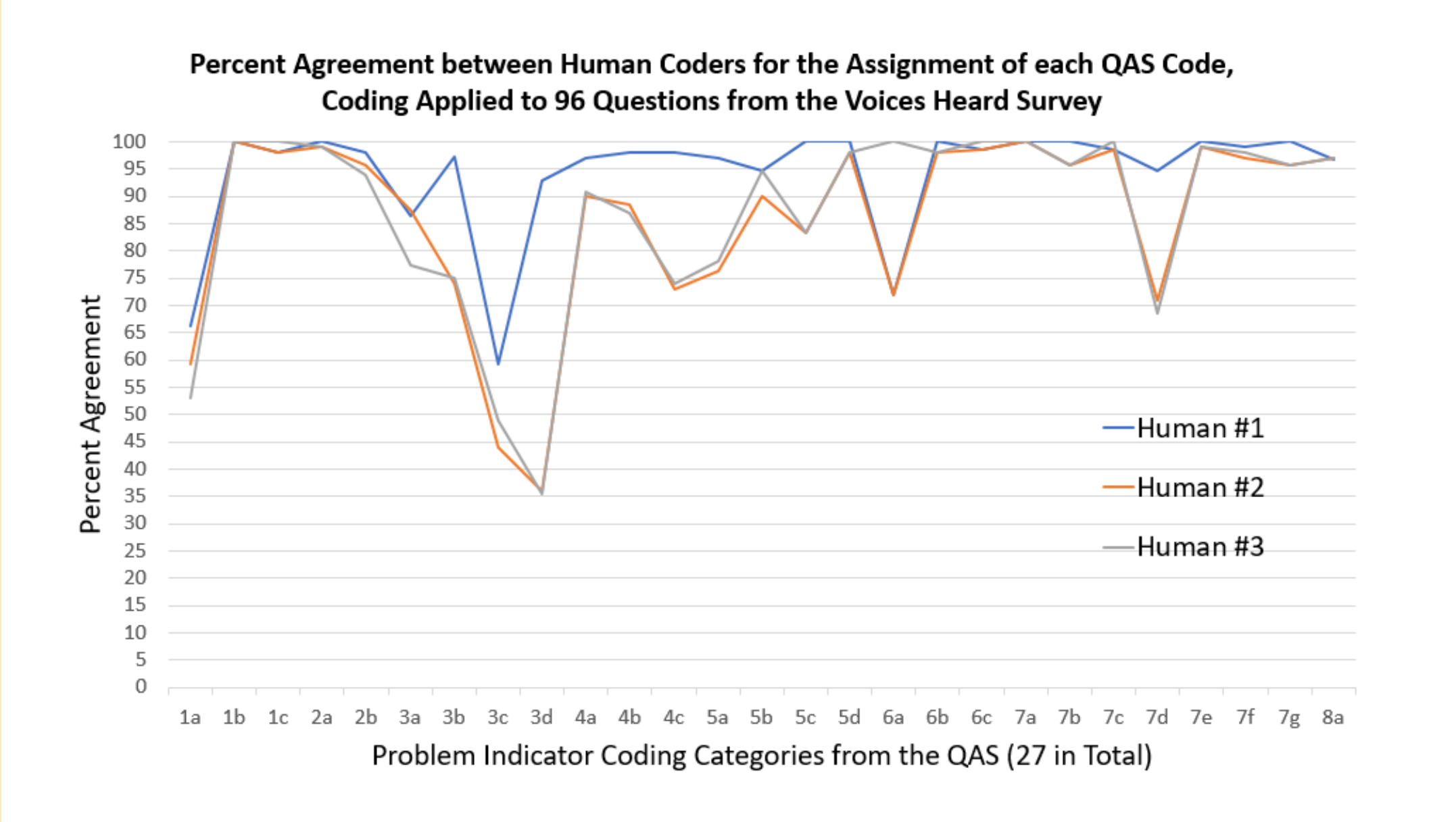
- Questions to be evaluated were drawn from the Voices Heard Survey (N = 96 questions)
  - Questions contained a mix of behavioral, attitudinal, and sociodemographic questions
- We examined how consistently the LLM coded questions using the QAS for 3 prompts that varied in the amount of information they provided (used ChatGPT, GPT-4o)

LLM INFORMATION PROMPTS		
"Low" Information Prompt	Medium Information Prompt	High Information Prompt
Included only brief descriptions of the QAS coding categories	Included detailed information about the QAS coding categories and examples	Used a PDF of the QAS manual to code the questions

- LLM coding was completely automated so that all 96 questions were coded in a batch file and output from the LLM was written to an Excel file for analysis
- We compare results from the LLM coding to 3 human coders
  - Human coders received minimal training; they are probably similar to the Medium LLM information prompt
  - Time for human coders to code was substantially longer

Coding Consistency across LLM Prompts and Humans





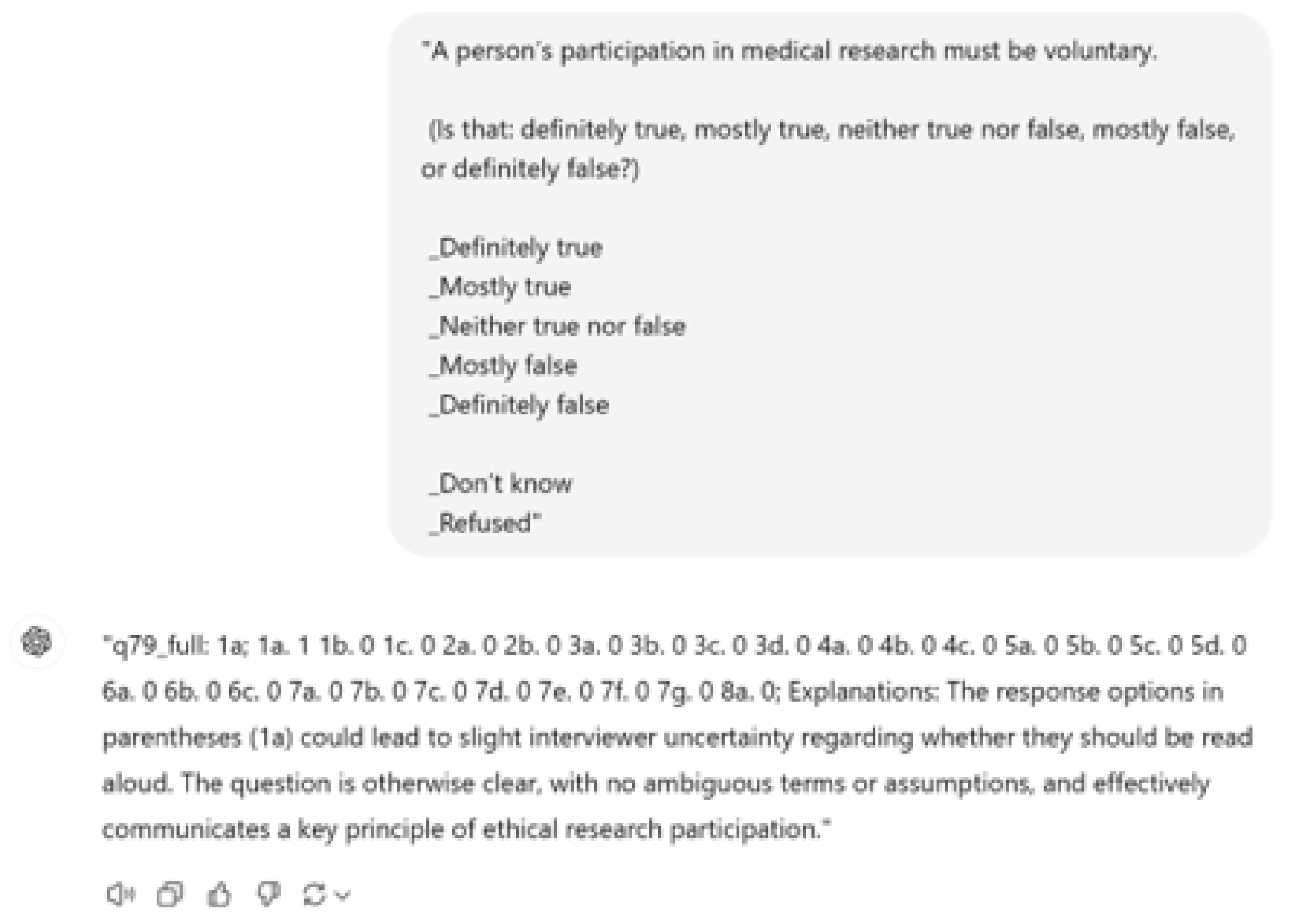
- Extensive consistency in coding across LLM prompts and across human coders
- Average rates of consistency are high for majority of QAS codes
- Much of the consistency is for agreement that a problem did not exist
- Does not address the question of whether LLMs or Humans are more accurate coders
  - Need to create a criterion – values from a SuperHuman Coder -- for such an assessment

QAS Codes Associated with Lower Agreement (Reliability) Across Coders

		Lower Levels of Agreement	
Code	Description	Among LLM Prompts	Among Human Coders
1a: WHAT TO READ	Interviewer may have difficulty determining what parts of the question should be read	X	X
3c: VAGUE	There are multiple ways to interpret the question or to decide what is to be included or excluded	X	X
3d: REFERENCE PERIODS	Are missing, not well specified, or in conflict		X
4a: INAPPROPRIATE ASSUMPTIONS	Are made about the respondent or about their living situation	X	
4c: DOUBLE-BARRELED	Contains more than one implicit question		X
6a: SENSITIVE CONTENT	The question asks about a topic that is embarrassing, very private, or that involves illegal behavior	X	
7d: VAGUE	Response categories are subject to multiple interpretations		X

Next Steps

- The current reliability analysis is part of a larger project to determine how well LLMs predict actual survey data quality
- Have three studies with very good data quality indicators, including
  - General Social Survey
    - Reliability scores (three-wave indicators) for the core questions
  - Wisconsin Longitudinal Study and the Voices Heard Study
    - Indicators of respondent comprehension problems
    - Indicators of interviewer errors (e.g., question-reading and follow-up errors)
  - Includes over 300 questions that vary substantially based on their characteristics and the likely problems they would pose for respondents and interviewers
- LLM coding using the QAS will be compared to coding from a SuperHuman Coder
- SuperHuman coding values are created by having three experts in questionnaire design code the questions and resolve differences collaboratively
  - Intent is that these values are accurate indicators of problems identified by the QAS
- SuperHuman coding scores will be compared to LLMs
  - Comparison will include ChatGPT and Claude
  - Assess accuracy along with reliability
- LLMs and SuperHuman coding scores will be analyzed to see which is a better predictor of data quality for the three studies



"A person's participation in medical research must be voluntary.

(Is that: definitely true, mostly true, neither true nor false, mostly false, or definitely false?)

_Definitely true
_Mostly true
_Neither true nor false
_Mostly false
_Definitely false

_Don't know
_Refused"

"q79_full: 1a; 1a. 1 1b. 0 1c. 0 2a. 0 2b. 0 3a. 0 3b. 0 3c. 0 3d. 0 4a. 0 4b. 0 4c. 0 5a. 0 5b. 0 5c. 0 5d. 0 6a. 0 6b. 0 6c. 0 7a. 0 7b. 0 7c. 0 7d. 0 7e. 0 7f. 0 7g. 0 8a. 0; Explanations: The response options in parentheses (1a) could lead to slight interviewer uncertainty regarding whether they should be read aloud. The question is otherwise clear, with no ambiguous terms or assumptions, and effectively communicates a key principle of ethical research participation."