# Where's the Evidence that Respondents Understand Your Survey Questions?*

Musashi Hinck  Gary King†  Kentaro Nakamura‡
Brandon M. Stewart§

July 2, 2026

**Abstract**

Survey researchers try to align respondent question interpretations with their own by applying theoretical rules, such as writing questions simply, concretely, and without biasing words or double-barreled inquiries, or via post hoc plausibility checks. Few, however, conduct ex ante empirical evaluations of their application of these rules. The widely recognized best empirical approach is "cognitive debriefing" which follows a survey question with a detailed conversation with the respondent. Unfortunately, this time-consuming procedure is rarely used, especially for novel surveys on online platforms. This paradoxically leaves the veracity of the discipline's most-used empirical methodology depending largely on theoretical assumptions. For cognitive debriefing to live up to its intended potential, we offer a unification of the essential interpretation-related statistical assumptions underlying all survey questions. We then automate cognitive debriefing to satisfy these assumptions using a specially-tuned chatbot, apply it to larger numbers than heretofore possible, analyze the collection of transcripts, and summarize the interpretations it reveals. By iteratively rewriting questions and rerunning this procedure, researchers can help ensure respondents understand questions as they do. We also apply our procedure to some of the most common survey questions and demonstrate large disconnects between researcher intentions and respondent understandings. We make available easy-to-use open source software that implements all our suggestions.

*We thank Uma Ilavarasan for her collaboration and research assistance in the early stages of this research, Vineet Bansal for his help developing the software, and Chris Felton for helpful comments. We also gratefully acknowledge funding support from the Princeton Center for Statistics and Machine Learning. The software is available at debriefly.net/ . The current version of this paper and its Supplementary Appendix is at GaryKing.org/debrief.

†Albert J. Weatherhead III University Professor, Institute for Quantitative Social Science, Harvard University; GaryKing.org, King@Harvard.edu.

‡Ph.D. student, John F. Kennedy School of Government, Harvard University. k-nakam.github.io, knakamura@g.harvard.edu.

§Professor of Sociology, Princeton University. brandonstewart.org, bms4@princeton.edu.

# 1 Introduction

If your uncle has chest pain during dinner, why not pick up a steak knife and volunteer to do just a little heart surgery? If you see passengers waiting for a late pilot, why not volunteer to fly the commercial jetliner yourself, just this once? Fortunately, no one takes these absurd steps because these jobs (a) require substantial expertise and (b) obviously require that expertise. Survey question writing fits (a) but not (b): It requires just as much expertise as surgery or piloting, including an understanding of knowledge accumulated in a vast literature over a century, but *seems* so easy that many feel comfortable writing questions and fielding surveys without detailed experience, training, or empirical evidence that any knowledge they may have was applied correctly. After all, healthy people communicate with others regularly, and most of what we all say sounds perfectly clear — to ourselves (with misinterpretations well known to arise from the natural inclination to assume others think about the world as we do; see Robbins and Krueger 2005).

Unfortunately, successful communication is far more difficult than it seems. As William James (1890) wrote long ago,

> The breach from one mind to another is perhaps the greatest breach in nature. The absolute insulation of each mind from all others is a fact of such importance that it deserves to be written in letters of gold. Every mind is a world by itself, and though its neighbors may be similar, they are not the same. Minds do not interpenetrate, nor do they even touch.

Writing good survey questions is much more challenging than ordinary conversation: By following a set of abstract theoretical principles from the literature (e.g., write questions concretely, without biased words, double-barreled inquiries etc.), we compose questions for the general public about our (sometimes obscure) research topics. We then administer them to a few thousand randomly selected strangers hailing from diverse cultures, walks of life, geographic locations, education levels, interests, life situations, and attention levels. And through this, we somehow must ensure that all respondents understand our questions in the same way (1) as each other and (2) as the researcher. (And if this sounds easy, can you really say that you, unlike pretty much everyone else, avoid occasional massive misunderstandings with friends and family you have known for decades?)

The advantages of using standardized questions administered to random selections of individuals are well known, which accounts for why survey research remains the dominant method of learning about the views of large populations across so many areas of academia, business, and government. But we should not think that achieving a common understanding with respondents as easy, as automatic, or as something that can be done well without evidence.

Both the problem and the solution to divergent interpretations among respondents, and between respondents and researchers, seem to have been known throughout the history of the field. Lazarsfeld (1935) identified the problem and suggested that researchers probe what respondents actually mean by their answers and enumerate the multiple ways a given question is interpreted. The foundational study of polling methodology by Cantril (1944) echoed these concerns. Merton (1987) formalized the practice of intensive qualitative interviews as the best way to understand how respondents understand a survey question.

Today, the well-known best practice for gathering empirical evidence during pretesting to discover and correct divergent survey question interpretation is known as *cognitive debriefing* (aka "cognitive interviewing" or "cognitive pretesting"). In this approach, the researcher asks a small sample the proposed survey question, elicits an answer, and then initiates a detailed conversation about how the respondent understood and interpreted the question and what they intended by their chosen answer. In some versions, the respondent is asked to "think aloud" while going over the question a second time. In others, the interview is partially structured, such as by following models that break the survey response into parts (such as comprehending the question, retrieving relevant information from memory, judging information compared to the question, and responding) and then discussing each part (Tourangeau, 1984). See also important contributions by Schwarz (2007), Miller (2014), and Behr et al. (2017).

Unfortunately, as we show in Section 2, cognitive debriefing may be widely recognized as best practice but it is rarely used and even then it is typically applied only to tiny samples of respondents that do not represent the full diversity of types of people to be interviewed. This means that even if it works as designed with small numbers, researchers

may well miss important subsets of the population who interpret questions in divergent ways. In some situations, this misinterpretation-induced bias will reverse key findings of interest.

As a starting point, we offer a unified formalization of what is meant by the proper interpretation or misinterpretation of survey questions, which leads to three assumptions required of all sensible survey questions (See Section 3). We use these assumptions to guide our methods development and applications.

We then enhance this practice by offering what we call "automated cognitive debriefing" (ACD), which uses a specially tuned chatbot to conduct interviews (See Section 4). This enables us to standardize the conversation better than we could with a group of trained human interviewers and to scale it to arbitrarily large numbers of respondents. We then analyze the collection of transcripts from all the cognitive debriefings via automated text analysis procedures to see how different types of people interpret the survey question. By iterating between ACD and rewriting the question, researchers can converge on a single or small number of questions with verified common understandings among respondents and between respondents and the researcher.

Our automation of the venerable cognitive debriefing procedure thus builds on the advantages of the *qualitative* foundations of survey research in ordinary human communication. In addition, since surveys began being used, and cognitive debriefing was known, we have greatly improved *quantitative* formalizations and analyses, such as via statistical inference, measurement, and causality, and so our main goal is to combine these perspectives to get the best of each.

Finally, we apply our ideas, framework, and formalizations to commonly used questions in major social surveys. This provides examples of large disconnects between how we and our research subjects understand our survey questions and examples of how to use these methods to finally enable researchers to bring the powerful ideas behind cognitive debriefing to widespread use (See Section 5).

# 2 Cognitive Debriefing: Best Practice, Rarely Used

Although cognitive debriefing is typically described as best practice for validating or developing survey questions whenever empirical pretesting is discussed (Section 2.1), pretesting of all kinds is skipped entirely in a large fraction of published articles in the social sciences, with proper cognitive debriefing rarely used in practice (Section 2.2).

## 2.1 Best Practice: Use It

Cognitive debriefing is recommended by a broad range of academic, governmental, and professional organizations. GESIS (the Leibniz Institute for Social Sciences, the largest social science research institute in Germany), describes cognitive pretesting in its official survey design guidelines as "generally considered to be an 'essential prerequisite for successful questionnaire development"' (Lenzner, Neuert, and Otto, 2016). The US Office of Management and Budget describes cognitive debriefing as a key technique for "refining collections of information to minimize burden and improve utility" (Office of Information and Regulatory Affairs of Office of Management and Budget, 2016). The US Census Bureau recommends that cognitive debriefing be made a "routine part of questionnaire development and evaluation" (Hess, 1999).

Those who run cross-cultural and multinational surveys are even more outspoken, probably because language and cultural context makes misinterpretation especially likely. The Cross-Cultural Survey Guidelines published by the Institute for Social Research at the University of Michigan describe cognitive interviewing for question pretesting as "important (some would argue essential)" (Hibben and Jong, 2016). The Professional Society for Health Economics and Outcomes Research, writing about survey design in cross-cultural settings, describes cognitive debriefing as an essential step and warns that omitting it risks "missing or inaccurate data resulting from respondents' misunderstanding of items" (Wild et al., 2005).

Major academic surveys that publish pretesting procedures almost always recommend cognitive debriefing. The General Social Survey explicitly describes it as a standard step in the development of new items, to be conducted between initial drafting and a full pilot

(Collaborative for Academic, Social, and Emotional Learning, n.d.). The 2016 American National Election Study methodology report documents qualitative cognitive pretesting for the post-election wave: fourteen one-on-one interviews were conducted by trained survey methodologists, with "findings from the pretest based on respondent reports of issues following scripted probing" (American National Election Studies, 2016). The Pew Research Center states that it "frequently tests survey questions ahead of time through qualitative research methods such as focus groups, cognitive interviews, pretesting (often using an online, nonprobability sample), or a combination of these approaches" (Pew Research Center, 2020). Gallup published a detailed cognitive interviewing protocol (Gallup, 2021). The Latin American Public Opinion Project (LAPOP) uses formal remote cognitive interview protocols, using think-aloud and probing techniques (Latin American Public Opinion Project, 2021). The European Social Survey documents cognitive interviews conducted across multiple countries (European Social Survey, 2023).

## 2.2 Common Practice: Rarely Used

An increasing fraction of surveys used in the academic literature are not designed by major survey organizations and instead are designed and run by individual researchers, often on online platforms (such as Lucid or Mechanical Turk). Thus, to measure common practice, we collected 12,656 articles published across top social science journals and tried to detect any use of empirical pretesting procedures in general or cognitive debriefing in particular.

To do this, we first obtained the full text for ten major social science journals available in JSTOR.[1] To these, we added *Public Opinion Quarterly* (POQ, 2021–25, $n = 287$) via manual download and optical character recognition. We include in our content analysis all articles that mention the term "survey" for the ten JSTOR journals and all articles for POQ, trying to filter out non-article texts based on available metadata. We use a large language model (gpt-5-mini-2025-08-07) to classify each paper on two binary dimensions: (1) whether it mentions any pretest or pilot survey of question wording, and (2) whether it

[1] *American Economic Review* (2000–23, $n = 2632$), *American Journal of Political Science* (1973–2023, 1734), *American Journal of Sociology* (2000-20, 529), *American Political Science Review* (1906–2019, 2718), *British Journal of Political Science* (2000-20, 598), *Comparative Politics* (2000—24, 363), *Journal of Political Economy* (2000—20, 612), *Journal of Politics* ( 1939—2019, 2,035) *Political Analysis* (2000—22, 417), *Political Behavior* (2000—22, 731).

mentions cognitive interviewing techniques. Positive classifications required the model to return a verbatim excerpt from the text, which was checked by us against the source. We manually validated this procedure on a random sample. All cases flagged as mentioning cognitive debriefing were manually reviewed to confirm classification.

In the average journal, only 19.9% of survey related articles report using any type of pretesting procedure, ranging from *Journal of Politics* at 13.4% to the *Journal of Political Economy* with 24.6%. Actual pretesting may be higher than this data on those mentioned, and we suspect that post hoc plausibility checks may be higher still. But a safe conclusion is that empirical validation of the measurement instrument at the question wording stage is mostly undiscussed throughout the social science literature.

To make matters worse, of the papers that mention some form of pretesting, the large majority describe only a pilot or field test — a preliminary survey administered to a small sample for the purpose of checking for operational issues, response rates, or aggregate item performance. These pilots address logistical concerns rather than systematically probing whether respondents interpret questions as intended by the researcher. Versions of cognitive debriefing, or other related techniques designed to ferret out question misinterpretation via qualitative conversation, are reported in *fewer than a dozen articles* across the entire corpus of social science articles.[2]

[2]For example, Munis (2022), "conducted cognitive interviewing to explore the validity of the items.... I recruited a small group of lay individuals living in urban and rural areas in two disparate regions of the United States. Each person read the items both silently and aloud, indicated whether each item made sense to them, and described what they took each item to mean, what considerations came to mind when contemplating each item, how they would answer each item, and why." Hurwitz and Peffley (2007) conducted "in-depth, face-to-face 'cognitive interviews' with a small number of African American respondents and telephone interviews with 25 white and 25 black respondents." In a study of compactness in redistricting, Kaufman, King, and Komisarchik (2021) performed "multiple iterated rounds of pre-testing and cognitive debriefing while adjusting question wording and how the districts appeared." Kaplan, Walker, and Morrison (2025) "used multiple methodologies, including cognitive interviews, card-sort/categorization tasks, and web probing." Pescosolido et al. (2015) "incorporated a cognitive interview lasting two to four hours with native speakers not associated with the mental health sector. This was done explicitly to gauge respondent understanding of the questions, provide a sense of socially sensitive questions and the potential for socially desirable response, and glean a layperson's sense of mental illness and treatment in each society." Hankinson (2018) reports learning from the process: "During cognitive testing of a pilot survey, two miles was a distance that would almost never elicit a NIMBY response." Searing (1978) "tape recorded pre-test sessions with former members of Parliament who were not returned in the 1970 general election and could not therefore appear in the study's sample. They filled in and discussed the value ranking form's penultimate version and suggested how it might best be presented to members of the House." Jerit and Barabas (2023) and Roller and Smith (2025) show how cognitive debriefing can be used during methodological research. Bloom et al. (2019) "adapted [their instrument] to the United States through two years of development

### 2.3 Earlier Automation Attempts

Given the importance of cognitive debriefing, many scholars have worked hard to automate parts of it, unfortunately well before methods of automated text generation were up to the task.[3] An alternative approach has been to automate the interviewer such as via "web probing" (Behr et al., 2017), which uses a single open-ended question, and using an LLM to author a single follow up question designed for the context (Chan et al., 2025). While efforts to use LLMs to automate qualitative interviewers exist (Chopra and Haaland, 2023; Wuttke et al., 2025; Heyde, 2025), to our knowledge no method exists to automate multi-turn cognitive debriefing with human respondents, which is characteristic of traditional cognitive debriefing.

## 3 Mathematical Framework for Survey Interpretation

We now describe a framework for survey question interpretation and misinterpretation intended to clarify essential goals for all sensible survey research. We also specify what it means for cognitive debriefing to help achieve these goals.

### 3.1 Notation

Let $Q$ denote an information set characterizing a survey question, including all information presented to the respondent, such as stem question wording, response options, and other elements of administration such as the survey mode. The survey instrument $Q$ is selected from the set of possible instruments the researcher might field, $\mathcal{Q}$. We also denote $M_i(Q)$ as the substantive *meaning* or interpretation respondent $i$ infers from $Q$, the

and cognitive testing by the Census Bureau." Ameriks et al. (2020) incorporated "exploratory cognitive interviews" with cognitive psychologists, and Tian and Menchik (2008) note that "questions were pretested through a cognitive interview approach."

[3] Graesser et al. (2006) offered an early, rule-based web utility designed to flag comprehension problems for survey takers. Even before, Oksenberg, Cannell, and Kalton (1991) offerd a framework to systematize problems arising in survey pretesting and Bolton (1993) gave a simple dictionary-based coding approach for think-aloud protocols. Others have used large language models to review survey instruments (Olivos and Liu, 2025; Sturgis, Roberts, and Robinson, 2026) and some simulate survey takers for pretesting (Kim et al., 2024; Buskirk, Steiger, and Kennedy, 2025). Some even try to dispense with survey respondents entirely and study how LLM answers differ from humans (Tjuatja et al., 2024; Bisbee et al., 2024; Sturgis, Roberts, and Robinson, 2026).

value of which is within the space of possible meanings $\mathcal{M}$.[4]. The respondent's observed answer to the instrument is then $Y_i(Q)$.

## 3.2 Assumptions

If the researcher prepares and administers a question so that survey effects do not bias responses or otherwise affect the meaning, we write:

**Assumption 1** (No Survey Effects)**.** *For all $Q, Q' \in \mathcal{Q}$ such that $M_i(Q) = M_i(Q')$,*

$$Y_i(Q) = Y_i(Q'),$$

an exclusion restriction implying that the survey response depends only on its meaning and not on other characteristics of the survey instrument. A large literature offers advice to help researchers write questions to satisfy Assumption 1 (or sometimes even to enable the comparison of answers across differently worded questions with the same meaning).

Conditional on Assumption 1, we define the *potential response* $Y_i(m)$ as the answer respondent $i$ would give to any survey instrument with interpretation $m$.[5] The observed response is the potential response which the respondent gives for their actual inferred meaning, $Y_i(M_i(Q))$.

Although a respondent only needs to infer (at most) one meaning to answer to a survey question, researchers almost always allow for a set of meanings, the differences among which are thought not to matter. For example, "Do you have a high school diploma?" has a different meaning if you attended Lynbrook High School rather than Brookline High School, but the two diplomas may be treated identically by the researcher. We formalize this concept by denoting the *set of researcher-intended meanings* as $\mathbf{M} \subset \mathcal{M}$, such that survey responses are the same for all meanings within this set:

**Assumption 2** (Researcher Meaning Coherence)**.** *Given Assumption 1, for all pairs of meanings $m, m' \in \mathbf{M}$,*

$$Y_i(m) = Y_i(m')$$

[4]Elements of $\mathcal{M}$ may be composite, so writing one meaning per respondent is without loss of generality. Respondent $i$ comes from a known population at a specific point in time.

[5]Analogously to the support condition in Fong and Grimmer (2023), $Y_i(m)$ is the common value of $Y_i(Q)$ over $\{Q \in \mathcal{Q} : M_i(Q) = m\}$ and is defined for meanings that some admissible survey instrument can induce in respondent $i$.

*for all $i$.*

Assumption 2 allows the researcher to treat $\mathbf{M}$ as an equivalence class because the specific meaning the respondent chooses does not affect the respondent's reported choice.

We then define researcher-respondent agreement on meaning as membership of the respondent's meaning in the researcher's set:

**Assumption 3** (Researcher-Respondent Alignment)**.** *For all $i$,*

$$M_i(Q) \in \mathbf{M}.$$

Finally, cognitive debriefing is itself a measurement procedure and thus requires an additional assumption to claim it can verify Assumption 3:

**Assumption 4** (Valid Debriefing)**.** *The meaning elicited by debriefing respondent $i$ about survey instrument $Q$ equals $M_i(Q)$, the meaning respondent $i$ used when answering.*

Here we merely clarify the long-held view that cognitive debriefing works. Although we impose this assumption throughout, it can of course be wrong: The interviewer or the analysis of the transcript may be biased, or the respondent may not give sincere answers. Good qualitative interviewing practices help of course, and scholars should continue to study ways of validating this procedure.

## 3.3 Implications

We now show formally how cognitive debriefing can help ensure proper survey analysis. Given our notation, we always know the researcher-intended meaning set $\mathbf{M}$ and, for each respondent, the observed survey outcome $Y_i(Q)$. Under Assumption 4, debriefing additionally reveals each respondent's inferred meaning $M_i(Q)$, so that, given Assumption 1, the observed response is $Y_i(Q) = Y_i(M_i(Q))$. Given these facts, we now give one intuitive but incorrect implication and then a set of suggested best practices.

**Don't Do This** For expository simplicity, suppose each respondent infers one of two possible meanings, $M_i(Q) \in \{a, b\}$ where $a$ is consistent with the researcher's intended meaning, $a \in \mathbf{M}$ and $b$ is not: $b \notin \mathbf{M}$. To take an extreme example for clarity, suppose

we ask the question "Do you approve of the president?" and meaning $a$ is the respondent's approval of the policies and performance of the US president but $b$ is whether the respondent likes President's Choice Peanut Butter (which is popular in some northern US localities). Every person $i$ then has two potential outcomes or answers, $Y_i(a)$ and $Y_i(b)$, only one of which is observed based on whether $i$ infers meaning $a$ or $b$ from the question.

Suppose the quantity of interest is the sample average of responses to the intended question:[6]

$$\theta = \underset{i}{\text{mean}}[Y_i(a)], \tag{1}$$

which, by Assumption 2, takes the same value for all meanings in $\mathbf{M}$.

This point clarifies an intuitive, but incorrect, use of cognitive debriefing. Suppose we use cognitive debriefing to learn the interpretation each respondent assigns to a question and then compute the difference of means across the respondents in groups with interpretations $a$ and $b$:

$$\begin{aligned} d &= \underset{i:M_i(Q)=a}{\text{mean}}[Y_i(M_i(Q))] - \underset{i:M_i(Q)=b}{\text{mean}}[Y_i(M_i(Q))] \\ &= \underset{i:M_i(Q)=a}{\text{mean}}[Y_i] - \underset{i:M_i(Q)=b}{\text{mean}}[Y_i]. \end{aligned} \tag{2}$$

Now, what can we conclude if we observe that mean approving of the "president" is the same in the US president group as in the peanut butter group, so that $d = 0$? Nothing. It may be an interesting coincidence, but concluding that the different interpretations do not matter (which respondents in different groups answer as if they are as unrelated questions) would be wrong. In fact, without researcher-respondent alignment (Assumption 3), bias in an ordinary estimate of $\theta$ can be in any amount or direction. This is the same reason that unobserved common causes confound observational comparisons of responses (Dafoe, Zhang, and Caughey, 2018).

**Do This** The example above emphasizes the essential role of researcher-respondent alignment in all survey research. Without alignment, research conclusions may have little

[6]For set $A$ with cardinality $\#A$, define the mean over $i$ of function $g(i)$ as $\text{mean}_{i\in A}[g(i)] = \frac{1}{\#A}\sum_{i\in A} g(i)$.

to do with the actual opinions of survey respondents. Because authors always carry the burden of proof, no survey analysis can be fully justified without empirical evidence for researcher-respondent alignment. ACD provides a way to do this.

If ACD confirms alignment, then we can of course move forward as usual. If, instead, we detect multiple or divergent interpretations for a question, researchers have at least four options, the first two which require additional rounds of debriefing for validation:

1. Rewrite the original question so all types of respondents have the researcher's intended interpretation;
2. Write separate questions that elicit the different interpretations discovered and use one or more of these in the final survey;
3. Redefine the researcher's intended meaning set to include all interpretations discovered during debriefing, so that Assumption 3 is satisfied by redefinition (which may sometimes work with interpretations less extreme than the President's Choice Peanut Butter example in Section 3.3).
4. Switch from the strict respondent-level invariance in Assumption 3 to less restrictive aggregate conditions, with consequent changes from point estimates to bounds. Appendix A takes this path by deriving new estimators at increasing levels of generality; it also shows formally that ACD can be used solely in the pretest, as intended.

If, during this process, we have a separate question corresponding to each interpretation, we can put the entire set of questions in the same pretest, and effectively observe all the potential outcomes for each respondent, as long as we can satisfy Assumption 1 (thus avoiding carryover effects, which sometimes works by inserting distractors between the questions; see Jenke and King 2026). If we cannot justify putting the questions in the same pretest survey, we can still compare aggregate answers to different questions asked of different randomly assigned groups. Whichever option is chosen, we should be prepared to iterate cognitive debriefing with question rewriting and so practical applications would usually require multiple round to achieve alignment. Either way, this procedure may save some of the rounds of debriefing that would otherwise be necessary. In practice, of course, only one or a subset of respondent interpretations of the original question may be of interest, and so usually fewer questions would be used for the main survey.

# 4 The Need for Automated Cognitive Debriefing

**Without Automation** Cognitive debriefing, in its original purely qualitative format, is simultaneously informative and limiting. For this paper, we performed cognitive debriefing (ourselves) with 50 respondents (using the survey question in Section 5.1, about which more in that section). Our impressions are consistent with the literature: Approaching strangers to ask survey questions is awkward or embarrassing, but more valuable than with family and friends who tend to think more like researchers. Non-native English speakers, even if fluent, and those from different cultures, tend to have divergent interpretations. We found it difficult to avoid steering conversations toward our favorite, expected, or predicted interpretations, or to one of the set of interpretations we had heard in previous interviews. This is especially so when the respondent indicates confusion of any kind and we're naturally drawn to "helping" them understand. Sometimes respondents perceive multiple interpretations and are unsure which should be used to answer the question. The local context of the interview, or the news of the day, seem to prime respondents in choosing interpretations, making standardization across respondents challenging. Finding a diverse array of people, and possibly mapping out the sample space of interpretations, is far easier than anything approaching a representative sample. Trying to scale, as we did by using multiple interviewers, revealed another layer of possible misinterpretation as we worked to explain to each other what we thought we each learned from our own interviews. Like most people who try cognitive debriefing, we were often surprised to learn of interpretations we had not previously considered but, although we took notes after each interview, most of each conversation is lost after the summary is written, and nothing is auditable afterwards.

The infrequent use of empirical pretesting in questionnaire development, and the exceptionally rare use of cognitive debriefing outside of major survey organizations, stands in sharp contrast to the professional consensus about their essential role. This divergence likely has two sources. The first is that cognitive debriefing is logistically challenging and time consuming. The second is the rise of online survey platforms that promise individual researchers the ability to run surveys at a moment's notice, setting up the expectation of

speed and efficiency, but with no guarantee of research-respondent alignment on survey question interpretation. Without the backing and professional practices used by major survey organizations, individual researchers (and most of the commercial world which has moved fast in this area) seem at much greater risk of biased estimates.

**With Automation** Our automation of cognitive debriefing replaces human interviewers with a specially tuned large language model-based chatbot. The base language models we build on enable conversations with the same linguistic fluency as human-to-human conversations and are improving fast. Talking to a bot loses the personal connection, which has obvious disadvantages, but the bot is easier to train, better at keeping the conversation on track, easier to standardize across respondents, makes it easier to record the entire conversation, and scales far more easily. The difficulty of scaling traditional cognitive debriefing probably accounts for why so few applications report using it with more than a few respondents, and the lack of standardization may account for why even those who use it publish so few details about their debriefing protocols, what they learned, what misinterpretations were discovered, what changes were made in response, and whether any additional rounds of debriefing were used to validate the changes. Automated cognitive debriefing (ACD), in contrast, scales to any number of respondents at low cost, makes transparency and replicability far easier, and can greatly reduce the variance across respondents.

To demonstrate ACD, we designed and make available with this paper software, called *Debriefly*, which embeds cognitive debriefing directly into the standard survey flow. We hope other survey providers add ACD facilities to their platforms and thus primarily offer Debriefly as an extensive a proof of concept that may help facilitate adoption of ACD. With Debriefly, researchers can create an entire survey, enabling researchers to choose from a large number of question types (radio buttons, check boxes, slider bars, open ended questions, among others, quite like the commonly used Qualtrics); in addition, they can designate any item to trigger cognitive debriefing. After a respondent answers an item that was flagged for debriefing, the system immediately opens a conversational debriefing module (without the need to click on an external link), so the follow-up occurs in context

while the respondent's interpretation is still fresh. The debriefing module is designed for a multi-turn conversation with the chatbot following up about vague, ambiguous, or incomplete answers. The conversation can continue until the respondent or the model reaches a stopping point. This is closer to cognitive interviewing than a single open-ended survey item because the system can adapt its next question to what the respondent just said.[7]

The respondents' substantive answers are collected in the usual format along with the detailed debriefing transcripts for further analysis. This preserves the logic of conventional cognitive interviewing by eliciting the answer before asking for an explanation of what the question meant and why that answer was chosen. And of course, this strategy removes the need for a team of trained human interviewers, standardizes the qualitative interview in ways that would be impossible with humans, and enables us to scale to arbitrarily large numbers of interviews.

The moment debriefing begins, Debriefly includes in the large language model's context window both the exact wording of the survey item and the respondent's recorded answer. The software renders these elements into a question-specific system prompt and opening assistant message; for categorical questions, it converts stored response codes into human-readable option labels before passing them to the model. The resulting interaction is not merely a generic chat, but a targeted probe tied to a particular item-response pair. In practice, we found that many respondents were so engaged with this process that they continued to chat for as long as our system allowed, and so we now allow researchers to tailor the interaction by setting minimum and maximum debriefing times or, customizing the underlying prompts for a given question. Appendix B offers two examples of the resulting transcripts.

Debriefly allows researchers to inspect responses within the platform, download the full response data, and export debriefing conversations as plain-text transcripts for qualitative coding and downstream analysis. In this sense, this allows ACD to be a full measure-

[7]Debriefly also includes all the "business logic" necessary for running surveys. It allows for multiple researchers, each running multiple surveys, and each with multiple respondents. It includes a preview mode, links to distribute to respondents, ways for researchers to open up previous surveys, the ability to add more question types, among other useful features.

ment workflow, systematically linking substantive answers to respondents' own explanations of meaning, making it feasible to detect recurrent ambiguities, compare interpretations across respondents, and iteratively revise question wording. Debriefly provides the practical workflow for the rewrite-and-verify strategy described in the previous section. Debriefly is easy enough to use that we find iterating with question rewriting simple and fast.

In practice, we use Debriefly to create a hundred or more ACD transcripts, which then become data for further analysis. For example, we often cluster these transcripts into a small, understandable number of interpretations to guide further questionnaire design.

# 5 Applications

We now illustrate our approach to ACD via an analysis of three commonly used survey questions on confidence in the military, democracy, and party identification. In Sections 5.1 and 5.2, we study the first two questions by doing analyses in two stages (and used the CloudResearch Connect survey platform for all three). We adapt our strategy for the venerable party ID question in Section 5.3 because of its unique characteristics.

In the first stage, we field cognitive debriefing interviews targeting 100 respondents for each survey question. These interviews are designed to understand how each respondent interpreted a question. We then treat the set of conversation transcripts as a dataset, which we use to identify clusters of interpretations.[8] We then extract up to five interpretation clusters for each question and use a large language model (gpt-4.1-mini) to classify each conversation into one or more of the discovered categories. We checked random selections of these classifications ourselves by hand to verify classification accuracy and agreement

[8]To do this, we gave the corpus to ChatGPT (browser interface, GPT-5.3) with the prompt: "You are a careful qualitative coder assisting an academic research project. The data consist of complete conversations between a human survey respondent and an AI system. Each conversation reflects how a respondent interprets a widely used survey question. Your task is to identify the most distinctive ways respondents interpret the survey question, based on patterns that recur across conversations. Read all conversations as a set. Treat each conversation as a single unit of analysis. Focus only on how respondents interpret the survey question, as revealed through their answers, questions, clarifications, and comments.

Identify at most five interpretation categories that are conceptually distinct, substantively different in how the question is understood, broad enough to apply across multiple conversations, and defined by how respondents construe the question, not by the answers they provide. If two categories differ only in emphasis or wording, merge them."

with what we intended.[9] See Supplementary Appendix A for the prompts used for ACD and for each classification task.

In the second stage, we use the interpretation clusters identified in the first stage to revise the survey questions. For each original survey question, we generate alternative versions of the question, each carefully aligned with one of the major interpretations. We then field survey experiments (also with a target of 100 respondents per arm) by randomly assigning participants to receive either the original question or one of the reworded versions. (Randomization ensures that selection differences cannot account for differences in results across the questions.) This allows us to examine the causal effect of each revised question on respondents' interpretations and whether the wording changed how they answered it.

Finally, we use the experimental data to evaluate two related questions. First, we evaluate the new questions by examining whether the rewording question shifts respondents toward more consistent understandings. Second, we assess whether such shifts in interpretation translated into differences in observed survey responses. For each original question, we found multiple interpretations and produced multiple verified questions but, in practice, researchers may wish to use only one or a few of the new questions depending on their research goals.

## 5.1 Confidence in Military

We begin by analyzing interpretations of the most commonly used question for measuring confidence in the military: "Please tell me how much confidence you, yourself, have in the military — a great deal, quite a lot, some, very little, or none?" See for example Nichols (2015), Gronke and Feaver (2001), Hill, Wong, and Gerras (2013), Robinson (2019), Kavanagh et al. (2020), and Burbach (2019).

The standard military confidence scale is interpreted in four different ways. For the

[9]To assess classification accuracy, we reviewed the classifications for a randomly selected 20 percent of the conversations in each application and manually verified whether each conversation had been correctly classified into each category. For the psychological identification category in the party identification application, we used a separate prompt specifically designed to identify psychological identification. Because this category is particularly difficult to classify, we then manually reviewed all conversations that the LLM classified as psychological identification. The resulting average classification accuracy was 0.93 for confidence in the military, 0.95 for democracy, and 1.00 for party identification.

original question wording, the most common reading focused on institutional capability and effectiveness (58.2%), but large shares of respondents interpreted the item in terms of personal safety and protection (41.8%), trust in leadership and political direction (40.8%), or moral integrity and ethical conduct (28.6%) (Table 1). Identical response categories on the original question often reflected different latent objects of confidence: some answered the question by evaluating whether the military can accomplish its missions, while others evaluated whether leaders can be trusted or whether the institution acts ethically. See Appendix B for examples of transcripts that reveal two of these interpretations. (We also used traditional in person cognitive debriefing with this question, as described in Section 4; although drawing respondents from the same online population was impossible, the results in Table 1 are fairly close to the overall impressions generated from our qualitative interviews.)

| **Category Name** | **Interpretations** | **%** |
|---|---|---|
| Capability | Whether the military has the strength and ability to carry out its mission | 58.2 |
| Personal Safety | Whether the military makes people feel safe | 41.8 |
| Trust in Leadership | Whether respondents trust military leadership and civilian control, especially the president, top brass, or political appointees | 40.8 |
| Morality | Whether the military does the right thing, makes the right decisions | 28.6 |

Table 1: Distribution of interpretation categories for the military confidence question ($N = 98$). Percentages do not sum to 100 because responses may be assigned to multiple categories.

Based on these results, we created four versions of the question targeted to each of the four interpretations:

- *Capability*: How much confidence do you have in the U.S. military's ability to effectively carry out its missions and defend the country?
- *Personal Safety*: How much confidence do you have that the U.S. military helps keep you and people like you safe?
- *Trust in Leadership*: How much confidence do you have that the U.S. military is led and directed responsibly by its senior military and civilian leaders?

- *Morality*: How much confidence do you have that the U.S. military generally acts ethically and does the right thing, even in difficult or dangerous situations?

Each question comes with responses offered as in the original question: "a great deal, quite a lot, some, very little, or none". We then conduct an experiment by randomly assigning respondents to receive original question wording or one of the four new questions, with a target of about $n = 100$ each. Randomization helps avoid systematic population differences accounting for our results. The original question is a replication of our first cognitive debriefing run, for calibration.

The frequency of the different interpretations appear in Figure 1. The color of each bar is chosen to match the text of the name of one of the four new questions (designed to have a meaning with only that interpretation). We see that the entropy of the interpretation frequencies is high for the original question (i.e., closer to uniform) but extremely low for each of the four newly written questions targeted at a specific interpretation.

The results above suggest that researchers should consider abandoning the original commonly used military confidence question, as it is an unknown function of at least four different meanings. If, for example, respondents are using something close to a weighted average, the survey answers researchers see are an *unknown* function of the weights and their substantive answers. The result is that two respondents can give *divergent* answers to the original military confidence question even though they have given *identical* views on every one of the four component questions. A much safer practice is to try to write survey questions that satisfy Assumption 3 so that respondents are aligned with each other and the researcher on the meaning of the question.

Figure 2 shows that the interpretation differences matter for substantive conclusions. When the questions are rewritten to emphasize capability, reported confidence is high (averaging about 4 on a 1–5 scale), but when rewritten to emphasize civilian control or moral conduct, reported confidence was markedly lower (about 3.25). The interpretation data, reported in Figure 1, tell the same story: each reworded version was far more likely to be understood in the intended way, whereas the original question continued to combine multiple meanings. Unfortunately, researchers who use only the original question cannot know how respondents combine the same judgments or whether they combine them the

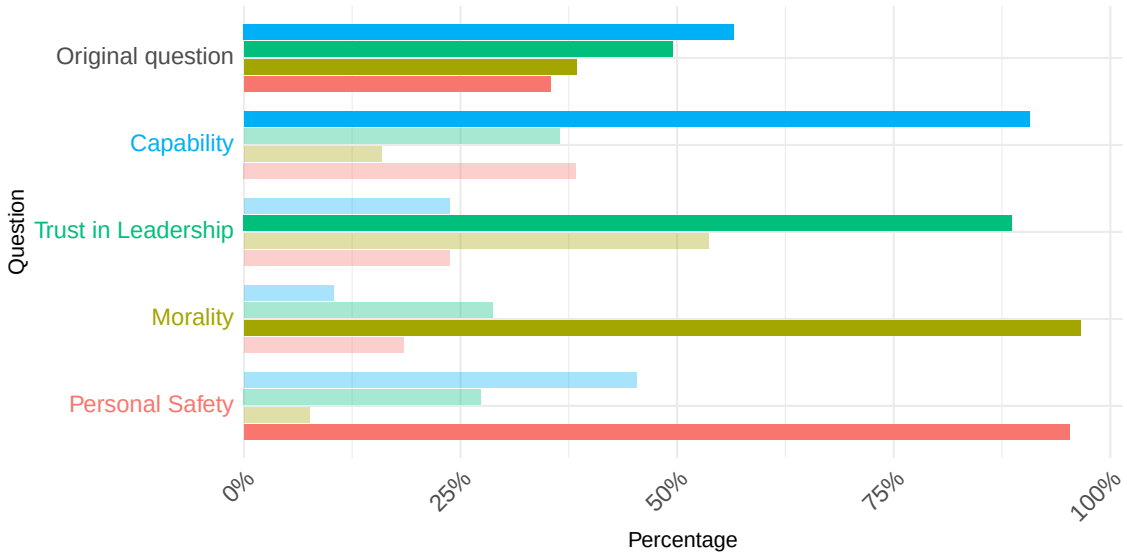


Figure 1: Four Interpretations (colored bars) of the Original and New Questions on the Military

same way and have different judgments of each component; asking one or more of the component questions solves this problem.

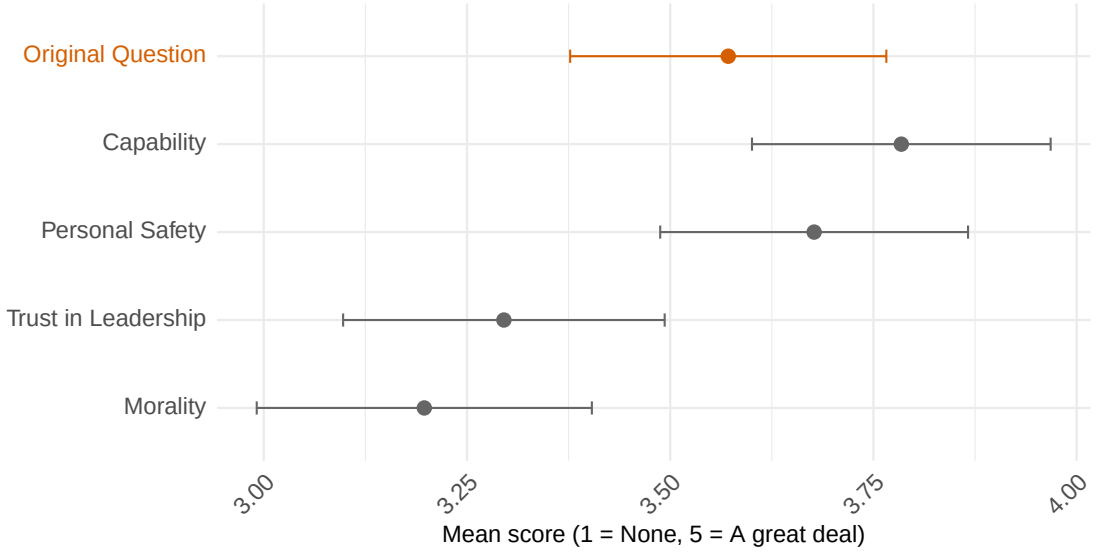


Figure 2: Average 1–5 scores on the military (and 95% confidence intervals) for five randomly assigned question wordings

Finally, because researchers have long known that interpretations can stem from different life experiences (King and Wand, 2007; King, Murray, et al., 2004), we checked whether any of our ACD-classified interpretations was associated with respondent characteristics. In this process, we discovered that 71% of men interpreted the original question

as referring to capability, compared with only 40% of women. To verify that this wasn't merely a multiple comparisons problem, we replicated the analysis of this original question which was also asked in our randomized experiment, and we found close to the same result: 73% of men interpreted the original question as referring to capability, whereas only 50% of women did so. This gender imbalance — 96% for men (53 out of 55) and 90% for women (43 out of 48) — mostly disappears once the question is rewritten and respondents understand the capability interpretation. (All differences in this paragraph are statistically significant except for the last with a p-value of 0.2.)

## 5.2 Democracy

The second commonly asked question we analyze is about beliefs about US democracy: "On a scale from 0 to 100, where 0 is least democratic and 100 is most democratic, how would you rate the political system of the United States as of today?" See, for example, Goldsmith et al. (2025), Spina (2025), and Bergeron-Boutin et al. (2024).

Our initial cognitive debriefing runs reveals clear evidence of divergent and overlapping interpretations of this democracy item. We summarize the main interpretations in Table 2, which reveals that respondents most often understand the question as asking whether government is responsive to what people want (54.5%) or whether the United States has the formal institutions of democracy (49.5%). At the same time, large minorities interpret the question to be about an overall impression (42.4%), the equality of political influence (38.4%), or civil liberties and rule of law (26.3%). (Because conversations could be assigned to multiple categories, these shares do not add to 100%.) The differences among these interpretations are of considerable substantive importance, each of which the subject of important scholarly study. We conclude by noting that, although this measure may have a clear and unitary meaning for researchers, most ordinary citizens share this meaning neither with researchers nor other citizens.

We then wrote a new question for each of the five interpretations in Table 2:

- *Responsiveness*: On a scale from 0 to 100, where 0 means government does not reflect public preferences at all and 100 means government closely reflects what

| Category Name | Interpretations | % |
|---|---|---|
| Responsiveness | Whether government outcomes reflect what most people want. | 54.5 |
| Formal institutions | Whether the United States runs via formal democratic mechanisms. | 49.5 |
| Overall impressions | An overall moral, emotional, or intuitive assessment of the country's political state. | 42.4 |
| Equal influence | Whether political influence on elections and governance is equal across people. | 38.4 |
| Civil Liberties / Rule of Law | Whether civil liberties, due process, and legal constraints on power are upheld. | 26.3 |

Table 2: Distribution of interpretation categories for the democracy question ($N = 99$). Percentages do not sum to 100 because responses may be assigned to multiple categories.

most people want, how democratic is the United States in terms of how responsive government actions are to the will of the people today? Please focus on whether policies and decisions generally follow what a majority of citizens prefer, rather than on formal rules or elections.

- *Formal Institution*: On a scale from 0 to 100, where 0 means not democratic at all and 100 means fully democratic, how democratic is the United States in terms of its formal political institutions and procedures today? Please base your answer on the existence and functioning of elections, representative institutions, courts, and constitutional checks and balances, regardless of whether you approve of current political outcomes.

- *Overall impressions*: Thinking broadly and intuitively, on a scale from 0 to 100, where 0 means not democratic at all and 100 means fully democratic, how democratic does the United States feel to you overall today? There is no right or wrong answer–please give your overall impression, even if it is based on general feelings rather than specific criteria.

- *Equal Influence*: On a scale from 0 to 100, where 0 means no political equality and 100 means full political equality, how democratic is the United States in terms of whether citizens have equal political influence today? Please consider whether each person's vote and political voice carry roughly the same weight, regardless of formal elections or institutions.

- *Civil Liberties / Rule of Law*: On a scale from 0 to 100, where 0 means no meaningful protection of rights and 100 means strong protection of rights and legal limits on power, how democratic is the United States in terms of civil liberties, rule of law, and constraints on government authority today? Please consider whether individual rights are protected and leaders are effectively constrained by law, regardless of elections or public opinion.

We then randomize respondents to one of six questions (the original one plus the five new ones), ask the question, and follow it with cognitive debriefing. Figure 3 presents the results of interpretations (following the same graphical heuristics as for Figure 1).

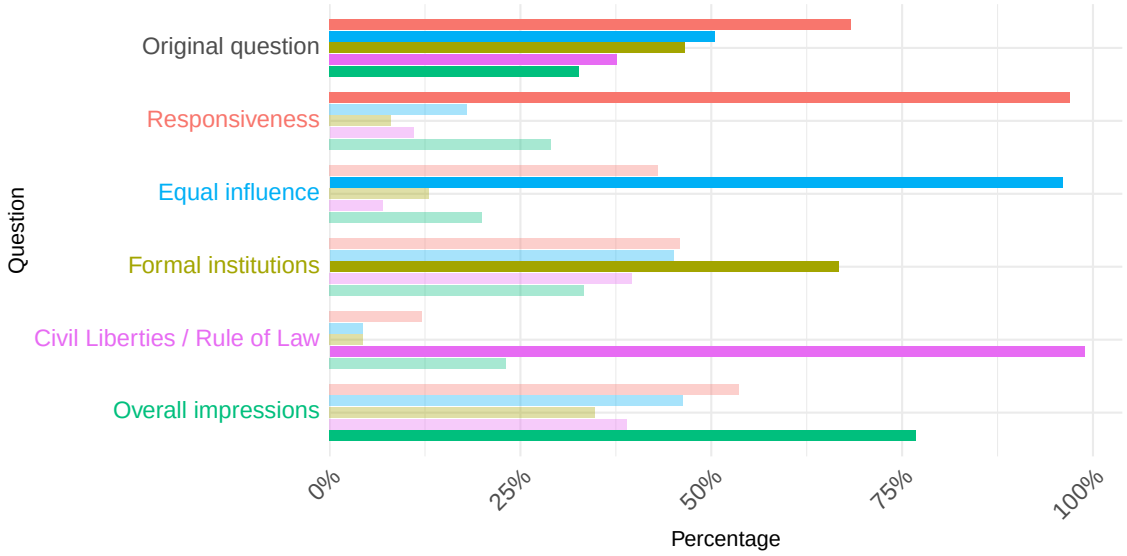


Figure 3: Five Interpretations (colored bars) of the Original and New Questions on Democracy

The key result in Figure 3 is that the much higher entropy (i.e., closer to uniform) for the original question than any of the new ones. The results are particularly dramatic for the final three questions — equal influence, responsiveness, and civil liberties/rule of law. The other two — formal institutions and overall impressions — follow the same pattern but less dramatically, which may suggest the possibility of further improvement via additional iterations if desired.

We then consider the dramatically different substantive conclusions that result from a different choice of survey questions, which we report in Figure 4. This figure gives a point estimate and 95% confidence interval for the average 1–100 score from each of the six questions. A researcher who defines democracy in terms of formal institutions or civil liberties versions produced the highest democracy ratings at about 55, whereas the responsiveness version produced the lowest ratings at only 33. This more than 20 point difference is certainly large enough to account for many descriptive results and causal effects in the literature, and so avoiding the problem altogether and choosing one (or more) of the revised questions, or rewriting and doing further cognitive debriefing to

validate, seems well worth doing.

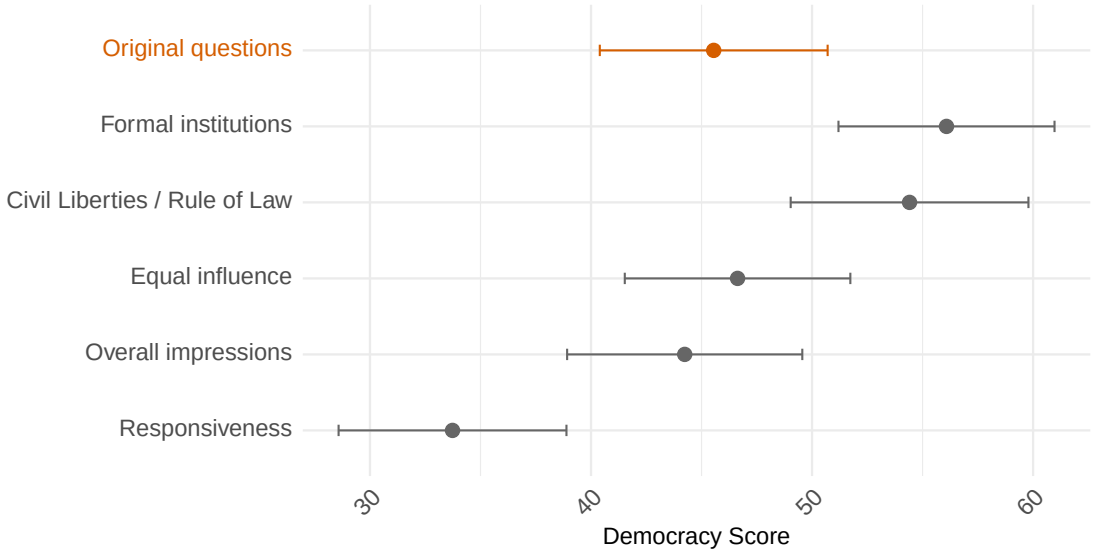


Figure 4: Average democracy scores (and 95% confidence intervals) for six randomly assigned question wordings.

Finally, we offer an example of using ACD to iteratively improve the question wording. We do this with the formal institutions interpretation from Figure 4. Based on the ACD transcripts in our second round, we discovered that some respondents did not think of our new question as referring solely to formal institutions and instead understood it in part as asking about the current administration's performance. To address this issue, we further revised our question as follows: "On a scale from 0 to 100, where 0 means not at all and 100 means completely, to what extent are the United States formal democratic institutions and procedures functioning today? Please consider only elections, Congress, courts, and constitutional checks and balances–not which party is in power, whether you like current policies, or whether leaders follow public opinion." The performance for this improved version appears in Figure 5, along with performance the two previous versions (in separate bars for each of the three survey waves). Although 49.5% of the respondents viewed the question as we intended in Wave 1, 66.7% did with our revised question in Wave 2, and in the final version of the question the percent interpreting the question as we intended was 83.5%.

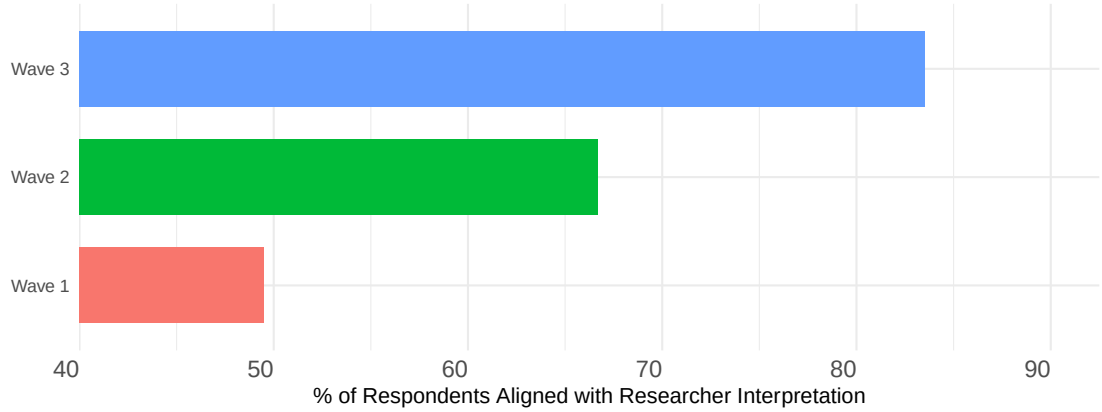


Figure 5: Percent of respondents interpreting the question as intended, about the formal institutions of democracy.

## 5.3 Party Identification

For our final example, we analyze the venerable party identification question: "Generally speaking, do you think of yourself as a Republican, a Democrat, an independent, or something else?", famously introduced in *The American Voter*. Although researcher-respondent alignment was plausibly high (and Assumption 3 likely satisfied) for the 1950s when the idea was conceived, our evidence shows that it is almost nonexistent presently.

In their book, Campbell et al. (1960, p.121) defined the concept as a social group-based psychological identification:

> Only in the exceptional case does the sense of individual attachment to party reflect a formal membership or an active connection with a party apparatus. Nor does it simply denote a voting record, although the influence of party allegiance on electoral behavior is strong. Generally this tie is a psychological identification, which can persist without legal recognition or evidence of formal membership and even without consistent record of party support.

Party ID was thus originally conceptualized as an American-style substitute for European formal party membership. In the 1950s, when the authors were working, the division between American political parties was based primarily on social group membership rather than ideological or policy preferences, and membership in those groups was relatively fixed, often transmitted through political socialization from parent to child and for all practical purposes fixed at birth for life. This interpretation also provided a convenient justification for party ID to be used as a valid *exogenous* causal factor for explaining vot-

ing behavior.

We now use ACD to assess the extent to which survey respondents interpret the party identification question as originally intended. To do so, we first conduct an initial round of cognitive debriefing interviews with 99 respondents. We then classify each conversation following the interpretation topics identified by analyzing the transcripts (following the same procedure as in Sections 5.1 and 5.2). Because psychological identification is a particularly difficult concept to find in our data, we also constructed a separate prompt to try to ferret out interpretations that come reasonably close to the definition in Campbell et al. (1960). We then manually validated or adjusted the classification of each relevant conversation.

Table 3 reports the results of an application of ACD to this question. As the table shows, we were only able to find four of the 99 respondents understanding the meaning of the question as referring to psychological identification. The remaining respondents interpreted the question as either about voting behavior (45.4%) or ideological alignment (90.9%).

| **Category Name** | **Interpretations** | **%** |
|---|---|---|
| Ideological alignment | Which party best matches respondent's core moral, social, or economic values | 90.9 |
| Voting behavior | Which party respondents usually vote for or are registered with | 45.4 |
| Psychological identification (Campbell et al., 1960) | Which party does the respondent psychologically identify with (regardless of formal party membership, registration, or voting record) | 4.0 |

Table 3: Distribution of interpretation categories for the partisanship question ($N = 99$). Percentages do not sum to 100 because responses may be assigned to multiple categories.

We obviously cannot run ACD with a sample of 1950s respondents, but our suspicion is that far more than 4% would have viewed a question about party ID as reflecting a psychological identification. The original interpretation of Party ID fit the American party system of the 1950s well; if you were a union member, you were a Democrat, and predicting other group memberships or voting behavior was comparatively easy. However, that story does not apply to the 1980s, where the parties had little coherence, or the cur-

rent period, where they are sharply divided on partisan and ideological lines but not social groupings.

As researchers, we are drawn to keeping questions constant over surveys so that estimates of change are not biased by question wording adjustments. However, this strategy can fail even if the wording remains the same but the meaning of the question changes over time, as it appeared to do here. This of course reemphasizes a principle of good question writing: willing respondents can answer questions about their world, not necessarily ours.

# 6 Concluding Remarks

Verifying empirically that respondents understand questions in the same way as researchers is essential for all survey research. This obviously applies to new survey questions, but it is also needed for questions that have been used in other populations or times with potentially different or changing norms, linguistic usage, or cultural understandings. Although cognitive debriefing is widely recognized as best practice, it is far too rarely used. Fortunately, the simple automation we suggest here can finally address a problem festering for more than a century. We thus encourage the survey community to use automated cognitive debriefing to help it return to its intended roots and help fulfill the original promise of cognitive debriefing for developing and validating survey questions of all types.

# Appendix A Reducing Restrictive Assumptions

This appendix generalizes our mathematical results in Section 3. In that section, the assumptions require respondent-level alignment and coherence. Since social scientists typically only care about aggregate generalizations, we do not need to get every individual right as long as errors of individuals cancel out or our otherwise accounted for when we make inferences to our population level inferences. We begin with a simple case of the mean, and by reframing misinterpretation as a problem of missing data. We then show how researchers can use ACD only in the pilot survey, and finally we generalize this situation to estimators beyond the mean.

## A.1 Misinterpretation as Missing Data

Misinterpretation can be thought of as a missing data problem for the analyst. By Assumption 2, the potential response under the researcher's intended meaning is well defined:

$$Y_i^* := Y_i(m) \quad \text{for any } m \in \mathbf{M}.$$

Thus, $Y_i^*$ denotes the answer respondent $i$ would give under the researcher's intended interpretation. For respondents with $M_i(Q) \in \mathbf{M}$, Assumptions 1 and 2 imply that the observed answer equals this intended-meaning response,

$$Y_i(M_i(Q)) = Y_i^*.$$

For respondents with $M_i(Q) \notin \mathbf{M}$, however, $Y_i^*$ is unobserved, while the observed answer $Y_i(M_i(Q))$ may correspond to a potentially unrelated interpretation.

For expository simplicity, we focus here on the sample mean and the case where ACD is implemented in the main survey. Appendix A extends these results to more general estimands defined by estimating equations and to the usual setting where ACD is conducted only during a pilot survey. The resulting bias in the mean due to misinterpretation then takes the following form:

**Proposition 1** (Interpretation Bias)**.** *Let $\pi = \frac{1}{n}\#\{i : M_i(Q) \notin \mathbf{M}\}$ denote the misalignment rate, and, among misaligned respondents, let $\bar{y}_{\text{misaligned}} = \text{mean}_{i:M_i(Q)\notin\mathbf{M}}[Y_i]$ denote the average observed answer among misaligned respondents, and $c = \text{mean}_{i:M_i(Q)\notin\mathbf{M}}[Y_i^*]$ the average (unobserved) potential answer under the intended meaning. Under Assumption 1, the observed sample mean $\bar{Y} = \text{mean}_i[Y_i(M_i(Q))]$ satisfies*

$$\bar{Y} - \theta = \pi(\bar{y}_{\text{misaligned}} - c).$$

*Proof.* Let $\bar{y}_{\text{aligned}} = \text{mean}_{i:M_i(Q)\in\mathbf{M}}[Y_i]$. Partitioning the sample by meaning, $\bar{Y} = (1-\pi)\bar{y}_{\text{aligned}} + \pi\bar{y}_{\text{misaligned}}$ and $\theta = (1-\pi)\bar{y}_{\text{aligned}} + \pi c$. Subtracting gives the result. ■

The counterfactual $c$ is unrestricted by the data, so the bias can be of any magnitude and in any direction. When the outcome is bounded, however, debriefing yields partial identification via standard Manski bounds (Manski, 1995): if $Y_i \in [y_L, y_H]$, then $c \in [y_L, y_H]$, and so

$$\theta \in \big[(1-\pi)\bar{y}_{\text{aligned}} + \pi y_L,\ \ (1-\pi)\bar{y}_{\text{aligned}} + \pi y_H\big], \tag{3}$$

an identified interval of width $\pi(y_H - y_L)$, with $|\bar{Y} - \theta| \leq \pi(y_H - y_L)$.[10]

The difference of mean responses across the two self-selected interpretation groups,

$$d = \underset{i:M_i(Q)\in\mathbf{M}}{\text{mean}}[Y_i(M_i(Q))] - \underset{i:M_i(Q)\notin\mathbf{M}}{\text{mean}}[Y_i(M_i(Q))] = \bar{y}_{\text{aligned}} - \bar{y}_{\text{misaligned}}, \tag{4}$$

contains no information about the counterfactual $c$ in Proposition 1, which is why even $d = 0$ is not sufficient to reach the conclusion that the divergent interpretations are innocuous.

## A.2 Limiting ACD to the pilot survey

The bounds derived in the main text assumes that we implement the ACD in the main survey. However, the implementation of ACD in the main survey is often too costly or infeasible and researchers instead want to use ACD in the pilot survey and run the main survey with the validated question without ACD. We derive the bounds in the main text for this setup.

Formally, let $S_i \in \{0, 1\}$ be the sample indicator which takes 1 if respondent $i$ belongs to the main survey and 0 otherwise (i.e., the respondent belongs to the pilot survey). Suppose we implement ACD in the pilot survey, so that we observe $M_i(Q)$ for $S_i = 0$ but not for $S_i = 1$. The key assumption when using ACD as pilot is the transportability. In words, we assume that the sample in the pilot survey is coming from the same population as in the main survey. This assumption is necessary not only for the bounds but also the validity of ACD:

**Assumption 5** (Transportability)**.** *For all $i$, we have*

$$M_i(Q) \perp\!\!\!\perp S_i.$$

Following the previous example, we consider the case in which the target quantity of interest is the sample average of responses to the intended question in the main survey:

$$\theta = \underset{i:S_i=1}{\text{mean}}[Y_i^*]. \tag{5}$$

[10]The probability that at least one of $n$ debriefed respondents reveals an interpretation held by a fraction $p$ of the population is $1 - (1 - p)^n$, so detecting such an interpretation with probability $1 - \alpha$ requires $n \geq \log\alpha / \log(1 - p)$: 59 interviews for an interpretation held by 5% of respondents, and 299 for one held by 1% (at $\alpha = .05$).

The difference from the setting in the main text is that we do not observe $M_i(Q)$ in the main survey, which prevents us from calculating the average among main-survey respondents who interpret the question as intended, $\bar{y}_a = \text{mean}_{i:S_i=1, M_i \in \mathbf{M}}[Y_i^*]$.

To overcome this problem, we combine Manski bounds in Equation (3) with trimming bounds (Horowitz and Manski, 1995; Lee, 2009). Under Assumption 5, the misalignment rate in the main survey is identified from the pilot survey,

$$\pi := \Pr\big(M_i(Q) \notin \mathbf{M} \mid S_i = 1\big) = \Pr\big(M_i(Q) \notin \mathbf{M} \mid S_i = 0\big),$$

where the right side is observed by applying ACD to the pilot sample (Assumption 4). Individual alignment status in the main survey, however, remains unobserved, so the identification problem now has two layers: *which* $(1-\pi)$ share of the observed main-survey answers belongs to the aligned respondents, and *what* the misaligned respondents would have answered under the intended meaning. The first layer is resolved by trimming; the second by the worst-case logic of Equation (3).

Formally, let

$$F(y) = \underset{i:S_i=1}{\text{mean}}\big[\, \mathbb{1}\{Y_i(M_i(Q)) \le y\} \,\big]$$

denote the empirical distribution of *observed* answers in the main survey, and write $q_u(F) = \inf\{y : F(y) \ge u\}$ for the quantile function of a distribution $F$, with $q_u(F) = y_L$ for $u \le 0$ and $q_u(F) = y_H$ for $u > 1$. Under Assumptions 1, 2, 4, and 5, and $\pi < 1$, the sharp bounds on the mean intended-meaning answer among aligned main-survey respondents are

$$\bar{y}_a \in \left[ \frac{1}{1-\pi} \int_0^{1-\pi} q_u(F)\, du, \quad \frac{1}{1-\pi} \int_\pi^1 q_u(F)\, du \right]. \tag{6}$$

The intuition behind Equation (6) is that, by Assumptions 1 and 2, the aligned respondents' observed answers equal their intended-meaning answers, so $\bar{y}_{\text{aligned}}$ is the average of some unknown $(1-\pi)$ share of the answers we observe in the main survey. Without knowing which share, the most pessimistic case is that the aligned respondents gave the *lowest* $1-\pi$ fraction of the observed answers, and the most optimistic case is that they gave the *highest* $1-\pi$ fraction; the two endpoints of Equation (6) are exactly these trimmed means.

Substituting Equation (6) into the decomposition $\theta = (1-\pi)\bar{y}_{\text{aligned}} + \pi c$ from Proposition 1, with the counterfactual mean $c \in [y_L, y_H]$ as in Equation (3), immediately gives the bounds on the quantity of interest:

$$\theta \in \left[ \int_0^{1-\pi} q_u(F)\, du + \pi y_L, \quad \int_\pi^1 q_u(F)\, du + \pi y_H \right]. \tag{7}$$

## A.3 Generalization beyond mean

Proposition 1 and the bounds in Equations (3)–(7) are derived for the sample mean as the quantity of interest. There is nothing special about the sample mean, other than that it is a widely used estimator. Here we briefly demonstrate that the same logic can be used to derive bounds for the wide variety of target quantity defined from the estimating equation. The unifying principle is that the width of the identified set is, to first order, the misalignment rate times the estimator's worst-case sensitivity to a single observation.

To generalize our previous setting, consider the following setup. Suppose that each respondent is drawn from the population of interest, $F$, and as in the previous case, indexed by $i = 1, \cdots, N$. Some of the respondents might be used as a pilot, which we denote $S_i = 0$, and the rest are used for the main survey $S_i = 1$. If the researchers use ACD for the main survey, we have $S_i = 1$ for all $i$. For each respondent, we observe the answer of the survey item of interest $Y_i$ and other variables $\mathbf{D}_i$ that may be used for the downstream analysis. When there are no other variables, $\mathbf{D}_i = \emptyset$.

We assume that $\mathbf{D}_i$ is measured without error, whereas the observed response $Y_i$ might contain some errors due to misinterpretations. As before, by Assumptions 1 and 2, an aligned respondent's observed answer equals the answer under the intended interpretations $Y_i^*$, exactly the value the estimating equation requires, while a misaligned respondent's observed answer is $Y_i(M_i(Q))$ with $M_i(Q) \notin \mathbf{M}$, which may bear no relationship to it.

Our quantity of interest, $\theta$, is defined by the estimating equation

$$\mathbb{E}_F[g(\mathbf{D}_i, Y_i^*; \theta) \mid S_i = 1] = 0. \tag{8}$$

For the sake of explanation, we limit our discussion to the case where $\theta$ is scalar, but this

can be relaxed easily (Beresteanu, Molchanov, and Molinari, 2011). This setup generalizes the previous setup, including many widely used estimators. Indeed, in the case of sample mean, $g(\mathbf{D}_i, Y_i^*; \theta) = Y_i^* - \theta$, $F$ becomes the empirical distribution, and

$$\mathbb{E}_F[g(\mathbf{D}_i, Y_i^*; \theta) \mid S_i = 1] = \frac{1}{\sum_{i=1}^N S_i} \sum_{i=1}^N S_i\big(Y_i^* - \theta\big) = 0.$$

**ACD in the main survey.** When debriefing labels are observed, partitioning the estimating equation by alignment status exactly as in Proposition 1 gives

$$\begin{aligned}&\mathbb{E}_F[g(\mathbf{D}_i, Y_i^*; \theta) \mid S_i = 1] \\ &= (1-\pi)\,\mathbb{E}_F[g(\mathbf{D}_i, Y_i^*; \theta) \mid S_i = 1, M_i(Q) \in \mathbf{M}] + \pi\,\mathbb{E}_F[g(\mathbf{D}_i, Y_i^*; \theta) \mid S_i = 1, M_i(Q) \notin \mathbf{M}].\end{aligned}$$

The first term is computable from the aligned respondents, whose observed answers equal $Y_i(\mathbf{M})$; the second is unrestricted by the data except through the envelopes, exactly as the counterfactual $c$ was in Equation (3). Writing $\bar{g}_{\mathbf{M}}(\theta) = \mathbb{E}_F[g(\mathbf{D}_i, Y_i^*; \theta) \mid S_i = 1,\ M_i(Q) \in \mathbf{M}]$ for the first term, the identified set is therefore

$$\begin{aligned}\Theta_I = \Big\{\theta :\ 0 \in \big[\,&(1-\pi)\,\bar{g}_{\mathbf{M}}(\theta) + \pi\,\mathbb{E}_F[g_i^L(\theta) \mid S_i = 1, M_i(Q) \notin \mathbf{M}], \\ &(1-\pi)\,\bar{g}_{\mathbf{M}}(\theta) + \pi\,\mathbb{E}_F[g_i^H(\theta) \mid S_i = 1, M_i(Q) \notin \mathbf{M}]\big]\Big\},\end{aligned} \tag{9}$$

where $g_i^L(\theta) = \inf_y g(\mathbf{D}_i, y; \theta)$ and $g_i^H(\theta) = \sup_y g(\mathbf{D}_i, y; \theta)$.

**ACD in the pilot survey only.** When alignment status is unobserved in the main survey, $\bar{g}_{\mathbf{M}}(\theta)$ is no longer computable, so we need to use logic similar to that in the previous subsection. The key observation is that, by simple algebra,[11], we can decompose the estimating equation of interest as

$$\mathbb{E}_F[g(\boldsymbol{D}_i, Y_i^*; \theta) \mid S_i = 1]$$

[11] Notice that

$$g(\boldsymbol{D}_i, Y_i^*; \theta) = g(\boldsymbol{D}_i, Y_i; \theta) - \big(g(\boldsymbol{D}_i, Y_i; \theta) - g(\boldsymbol{D}_i, Y_i^*; \theta)\big).$$

Under Assumption 2, the second term is nonzero only for misaligned respondents. Thus, we can write

$$g(\boldsymbol{D}_i, Y_i^*; \theta) = g(\boldsymbol{D}_i, Y_i; \theta) - \mathbb{1}\{M_i(Q) \notin \mathbf{M}\}\big(g(\boldsymbol{D}_i, Y_i; \theta) - g(\boldsymbol{D}_i, Y_i^*; \theta)\big).$$

Taking the conditional expectation given membership in the main sample then yields the desired expression.

$$= \mathbb{E}[g(\boldsymbol{D}_i, Y_i; \theta) \mid S_i = 1] - \mathbb{E}[\mathbb{1}\{M_i(Q) \notin \mathbf{M}\}\big(g(\boldsymbol{D}_i, Y_i; \theta) - g(\boldsymbol{D}_i, Y_i^*; \theta)\big) \mid S_i = 1].$$

In this equation, the left-hand side is the estimating equation of interest. The first term on the right-hand side is observable, whereas the second term is not, because we do not observe which respondents are misaligned. We therefore bound this unobserved term using the same trimming-bounds argument as in the previous subsection.

Formally, define

$$\Delta_i^L(\theta) = g(\mathbf{D}_i, Y_i; \theta) - g_i^L(\theta) \geq 0, \qquad \Delta_i^H(\theta) = g(\mathbf{D}_i, Y_i; \theta) - g_i^H(\theta) \leq 0,$$

and let $F_{\Delta^L(\theta)}$ and $F_{\Delta^H(\theta)}$ denote their distributions among main-survey respondents. From the discussion above, the identified set is therefore

$$\Theta_I = \left\{ \theta : \; 0 \in \left[ \begin{array}{l} \mathbb{E}[g(\boldsymbol{D}_i, Y_i; \theta) \mid S_i = 1] - \displaystyle\int_{1-\pi}^{1} q_u\big(F_{\Delta^L(\theta)}\big)\, du, \\ \mathbb{E}[g(\boldsymbol{D}_i, Y_i; \theta) \mid S_i = 1] - \displaystyle\int_{0}^{\pi} q_u\big(F_{\Delta^H(\theta)}\big)\, du \end{array} \right] \right\}. \tag{10}$$

These bounds use all the information available in the data: the data identify only the misalignment rate $\pi$, leaving both the identity of the misaligned respondents and their intended-meaning contributions unrestricted beyond the envelopes $[g_i^L(\theta), g_i^H(\theta)]$. The lower endpoint is attained by assigning misalignment to the $\pi$ share of respondents with the largest $\Delta_i^L(\theta)$ and setting their contributions to $g_i^L(\theta)$; the upper endpoint follows symmetrically from the smallest $\Delta_i^H(\theta)$. Intermediate values are attainable by intermediate choices, so no narrower interval is valid without further assumptions.

## A.4 Application

We finally return to the randomized experiments for the military and democracy questions, whose results are reported in Figure 2 and Figure 4, respectively. In the main analyses, we used these experiments to examine how alternative question wordings affect responses. One limitation of those analyses, however, is that the revised questions do not achieve perfect respondent–researcher alignment: even after revision, some respondents still interpret a question differently from its intended meaning. Treating the revised-wording estimates

as ordinary means therefore ignores the remaining uncertainty created by misaligned interpretations.

We address this limitation by applying the bounds derived above. For each revised question in the randomized experiment, we treat respondents whose debriefed interpretations do not match the intended meaning of that wording as misaligned. We exclude the original question from this analysis because its intended meaning is ambiguous and therefore cannot serve as a well-defined target for assessing alignment. We then construct uncertainty intervals that incorporate both sampling variability and residual interpretation misalignment by applying the bounds to the endpoints of the conventional 95% confidence intervals. This procedure yields conservative intervals for the average response that would have been obtained if all respondents had interpreted each revised question as intended.

Figure 6 reports the results for the military question, and Figure 7 reports the corresponding results for the democracy question. The bounded estimates show that several interpretation-specific question wordings still produce substantively different responses. However, some differences that appeared meaningful in the main paper's plots are no longer statistically distinguishable once we account for residual interpretation misalignment. This is especially true for question wordings with relatively high misalignment rates, where the bounds become wider and therefore less informative.

These results illustrate both the value and the limitation of the proposed bounding approach. On the one hand, the bounds allow researchers to draw more reliable conclusions by incorporating uncertainty due to respondent misinterpretation, rather than treating revised questions as if they achieved perfect alignment. On the other hand, the informativeness of the bounds depends directly on the quality of the revised questions: when many respondents still interpret a revised question differently from its intended meaning, the resulting bounds may be too wide to support sharp conclusions. The analysis therefore reinforces the importance of iterative question revision with automated cognitive debriefing. Better-aligned question wordings not only improve measurement validity but also yield more informative bounds for downstream substantive analysis.

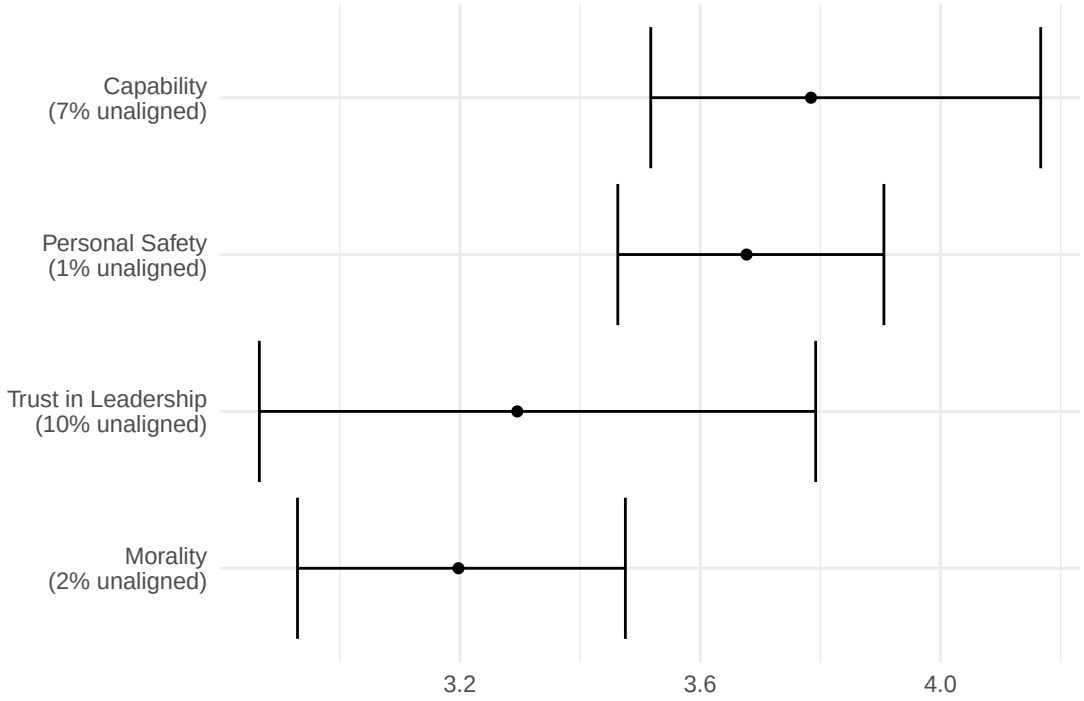


Figure 6: Average 1–5 scores on the military (and 95% confidence intervals with bounds) for four randomly assigned question wordings

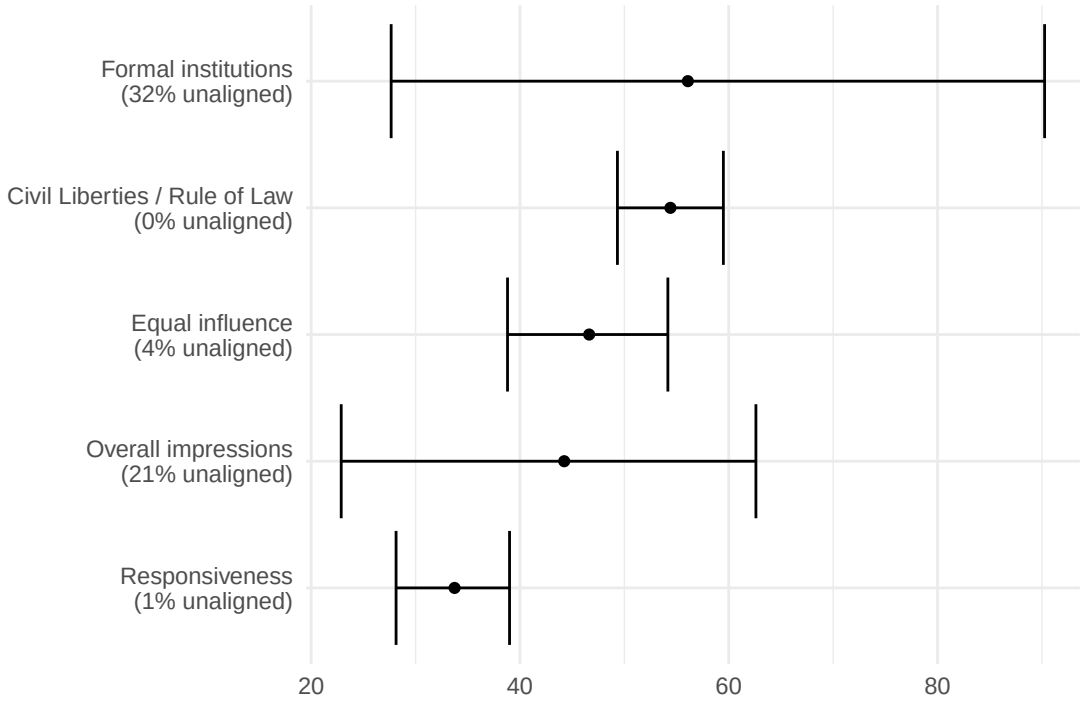


Figure 7: Average democracy scores (and 95% confidence intervals with bounds) for sive randomly assigned question wordings.

# Appendix B Automated Cognitive Interview Examples

This appendix offers two examples of transcripts from our automated cognitive debriefing interviews. They correspond to two of the interpretations from Table 1.

## B.1 Conversation 1: Capability

LLM Hello, I am a conversational AI designed to improve survey research. In the survey you were asked the following question:

> Please tell me how much confidence you, yourself, have in the military – a great deal, quite a lot, some, very little, or none?

You answered *A great deal* to this question. How did you arrive at that answer?

RESPONDENT I have a lot of confidence in the military because they are there to protect the nation and do a good job at it while putting their lives on the line.

LLM Thank you for sharing that. When you read the question, what do you think "confidence in the military" means?

RESPONDENT I think it means how much you trust them to protect the nation.

LLM That's a clear interpretation. When you considered your level of confidence, did you think about any specific events or experiences that influenced your answer?

RESPONDENT I thought about past wars that have been fought and that we have won.

LLM Got it. So, when you think about those past wars, do you also consider the military's role in peacekeeping or humanitarian efforts? How does that impact your confidence?

RESPONDENT I do think of that. I think of all the smaller countries we have lended aid to.

LLM Thank you for that insight. Now, focusing on the answer options, can you tell me what "a great deal" means to you in this context?

RESPONDENT It means doing a lot.

LLM That's helpful to know. How about the other options, like "quite a lot" or "some"? How do you differentiate between them?

RESPONDENT Some is a little bit while quite a lot is a bit more than just a couple.

LLM Thank you for clarifying that. Was it easy or hard for you to choose "a great deal" from the answer options provided?

RESPONDENT It was easy.

LLM That's good to hear. Were there any parts of the question or the answer choices that you found confusing or unclear?

RESPONDENT No.

## B.2 Conversation 2: Morality

LLM Hello, I am a conversational AI designed to improve survey research. In the survey you were asked the following question:

> Please tell me how much confidence you, yourself, have in the military – a great deal, quite a lot, some, very little, or none?

You answered *None* to this question. How did you arrive at that answer?

RESPONDENT I don't really have confidence in the military because I don't trust the people that go into the military. They have a specific personality that makes them attracted to the military and I can't agree with that. They are more prone to violence than other people. Not to mention, the military has not been protecting our constitution with what Trump is doing.

LLM Thank you for sharing your thoughts. When you considered the term "confidence" in the context of the military, what does "confidence" mean to you?

RESPONDENT Confidence in this context means you believe that they are making the right choices and that you aren't afraid of them making any bad decisions.

LLM Got it. So, when you think about "making the right choices," what specific actions or decisions would you expect from the military that would contribute to your confidence?

RESPONDENT When Trump tells the military to attack a country, for example Venezuela, the military would refuse to do so. When Trump is trying to start wars with other countries like Iran, or attempts to go after Canada, the military would refuse to do so. To me, those are making right choices and preserving the respect of this country.

LLM Thank you for clarifying that. You mentioned feeling that the military has not been protecting the constitution. Can you explain what you mean by that and how it relates to your confidence in the military?

RESPONDENT The military is doing exactly what Trump wants. The military should come in and stop ICE from killing it's own citizens. To me, that shows that the military is not protecting the constitution because they aren't doing anything to protect we the people. It makes me feel less confident that the military is protecting us people.

LLM I appreciate your insights. You selected "none" as your answer. Can you tell me if there was anything about the answer options—like "a great deal," "quite a lot," "some," "very little," or "none"—that was confusing or unclear to you?

RESPONDENT No, there was nothing confusing. I really just don't have trust in the

military or it's people who have joined.

LLM Thank you for your openness. Was this question easy or hard for you to answer?

RESPONDENT It was easy.

LLM Great to hear that. Is there anything else you would like to add about your thoughts on the question or your answer?

RESPONDENT No.

# References


American National Election Studies (2016). *ANES Time Series 2016 Methodology Report*. Tech. rep. American National Election Studies. URL: https://bit.ly/4xeFwaq.

Ameriks, John, Joseph Briggs, Andrew Caplin, Matthew D. Shapiro, and Christopher Tonetti (2020). "Long-Term-Care Utility and Late-in-Life Saving". In: *Journal of Political Economy* 128.6, pp. 2375–2451. DOI: 10.1086/706686.

Behr, Dorothée, Katharina Meitinger, Michael Braun, and Lars Kaczmirek (2017). *Web Probing – Implementing Probing Techniques from Cognitive Interviewing in Web Surveys with the Goal to Assess the Validity of Survey Questions*. GESIS Survey Guidelines. Mannheim, Germany: GESIS – Leibniz Institute for the Social Sciences. DOI: 10.15465/gesis-sg_en_023.

Beresteanu, Arie, Ilya Molchanov, and Francesca Molinari (2011). "Sharp identification regions in models with convex moment predictions". In: *Econometrica* 79.6, pp. 1785–1821.

Bergeron-Boutin, Olivier, John M. Carey, Gretchen Helmke, and Eli Rau (2024). "Expert Bias and Democratic Erosion: Assessing Expert Perceptions of Contemporary American Democracy". In: *PS: Political Science & Politics* 57.2, pp. 184–193. DOI: 10.1017/S1049096523000719. URL: https://bit.ly/4a47lIX.

Bisbee, James, Joshua D. Clinton, Cassy Dorff, Brenton Kenkel, and Jennifer M. Larson (2024). "Synthetic Replacements for Human Survey Data? The Perils of Large Language Models". In: *Political Analysis* 32.4, pp. 401–416. DOI: 10.1017/pan.2024.5.

Bloom, Nicholas, Erik Brynjolfsson, Lucia Foster, Ron Jarmin, Megha Patnaik, Itay Saporta-Eksten, and John Van Reenen (2019). "What Drives Differences in Management Practices?" In: *American Economic Review* 109.5, pp. 1648–1683. DOI: 10.1257/aer.20170491.

Bolton, Ruth N. (1993). "Pretesting Questionnaires: Content Analyses of Respondents' Concurrent Verbal Protocols". In: *Marketing Science* 12.3, pp. 280–303. DOI: 10.1287/mksc.12.3.280.

Burbach, David T. (2019). "Partisan Dimensions of Confidence in the U.S. Military, 1973–2016". In: *Armed Forces & Society* 45.2, pp. 211–233. DOI: 10.1177/0095327X17747205. URL: https://bit.ly/4oatACN.

Buskirk, Trent D., Darby Steiger, and Courtney Kennedy (2025). "SAI What?! Ping the Bots Before You Probe the People: Testing Large Language Models for Pre-Cognitive Interviewing in Survey Research". In: *First Workshop on Bridging NLP and Public Opinion Research (NLPOR)*. Montreal, Canada.

Campbell, Angus, Philip E. Converse, Warren E. Miller, and Donald E. Stokes (1960). *The American Voter*. New York: John Wiley & Sons.

Cantril, Hadley (1944). *Gauging Public Opinion*. Research Associates in the Office of Public Opinion Research, Princeton University. Princeton, NJ: Princeton University Press.

Chan, Lee, Virginia Bertelli, Catherine White, Lewis Wooding, Christopher Martin, and Edward Hill (2025). "A Preliminary Investigation of LLM Capability for Cognitive Interviewing". In: *International Journal of Market Research* 67.6, pp. 657–667. DOI: 10.1177/14707853251369238.

Chopra, Felix and Ingar Haaland (2023). *Conducting Qualitative Interviews with AI*. CESifo Working Paper 10666. Munich, Germany: CESifo. URL: https://ssrn.com/abstract=4583756.

Collaborative for Academic, Social, and Emotional Learning (n.d.). *Cognitive Pretesting*. Cross-Cultural Survey Guidelines. URL: https://ccsg.isr.umich.edu/chapters/pretesting/.

Dafoe, Allan, Baobao Zhang, and Devin Caughey (2018). "Information Equivalence in Survey Experiments". In: *Political Analysis* 26.4, pp. 399–416. DOI: 10.1017/pan.2018.9.

European Social Survey (2023). *ESS Round 10 Core Questionnaire Review*. Tech. rep. European Social Survey. URL: https://bit.ly/4x9UoqI.

Fong, Christian and Justin Grimmer (2023). "Causal Inference with Latent Treatments". In: *American Journal of Political Science* 67.2, pp. 374–389. DOI: 10.1111/ajps.12649.

Gallup (2021). *Gallup's Approach to Cognitive Interviews*. URL: https://bit.ly/4g4ICI3.

Goldsmith, Benjamin E., Yusaku Horiuchi, Kelly Matush, and Kathleen E. Powers (2025). "Democratic Backsliding Damages Favorable U.S. Image Among the Global Public". In: *PNAS Nexus* 4.4, pgaf104. DOI: 10.1093/pnasnexus/pgaf104. URL: https://pmc.ncbi.nlm.nih.gov/articles/PMC11983274/.

Graesser, Arthur C., Zhiqiang Cai, Max M. Louwerse, and Frances Daniel (2006). "Question Understanding Aid (QUAID): A Web Facility that Tests Question Comprehensibility". In: *Public Opinion Quarterly* 70.1, pp. 3–22. DOI: 10.1093/poq/nfj012.

Gronke, Paul and Peter D. Feaver (2001). "Uncertain Confidence: Civilian and Military Attitudes about Civil-Military Relations". In: *Soldiers and Civilians: The Civil-Military Gap and American National Security*. Ed. by Peter D. Feaver and Richard H. Kohn. Cambridge, MA: MIT Press, pp. 129–163. URL: https://people.reed.edu/~gronkep/docs/uncertain.pdf.

Hankinson, Michael (2018). "When Do Renters Behave Like Homeowners? High Rent, Price Anxiety, and NIMBYism". In: *American Political Science Review* 112.3, pp. 473–493. DOI: 10.1017/S0003055418000035.

Hess, Jennifer (1999). *The Role of Respondent Debriefing Questions in Questionnaire Development*. Tech. rep. URL: https://bit.ly/4okzYYc.

Heyde, Leah von der (2025). "Who Counts? The Potentials and Pitfalls of Using LLMs in Survey Research". In: *First Workshop on Bridging NLP and Public Opinion Research (NLPOR)*. Montreal, Canada. URL: https://openreview.net/forum?id=ww2KqnPLdK.

Hibben, Kristen Cibelli and Julie de Jong (2016). "Cognitive Interviewing". In: *Guidelines for Best Practice in Cross-Cultural Survey*. URL: https://ccsg.isr.umich.edu/chapters/pretesting/cognitive-interviewing/.

Hill, Andrew A., Leonard Wong, and Stephen J. Gerras (2013). ""Self-Interest Well Understood": The Origins & Lessons of Public Confidence in the Military". In: *Daedalus* 142.2, pp. 49–64. DOI: 10.1162/DAED_a_00203. URL: https://bit.ly/4ohXPI2.

Horowitz, Joel L and Charles F Manski (1995). "Identification and robustness with contaminated and corrupted data". In: *Econometrica: Journal of the Econometric Society*, pp. 281–302.

Hurwitz, Jon and Mark Peffley (2007). "Persuasion and Resistance: Race and the Death Penalty in America". In: *American Journal of Political Science* 51.4, pp. 996–1012. DOI: 10.1111/j.1540-5907.2007.00293.x.

James, William (1890). *The Principles of Psychology*. Vol. 1. New York: Henry Holt and Company. Chap. 9, p. 226.

Jenke, Libby and Gary King (2026). "Who's to Blame for Survey Instability: Respondents with Nonexistent Preferences or Researchers with Flawed Measures?" URL: GaryKing.org/instability.

Jerit, Jennifer and Jason Barabas (2023). "Are Nonprobability Surveys Fit for Purpose?" In: *Public Opinion Quarterly* 87.3, pp. 816–840. DOI: 10.1093/poq/nfad037.

Kaplan, Robin L., Tywanquila Walker, and Rebecca L. Morrison (2025). "Beyond Functional Limitations: A Multimethodological Approach to Questionnaire Pretesting of New Disability Measures". In: *Public Opinion Quarterly* 89.SI. DOI: 10.1093/poq/nfaf029.

Kaufman, Aaron R., Gary King, and Mayya Komisarchik (2021). "How to Measure Legislative District Compactness If You Only Know It When You See It". In: *American Journal of Political Science* 65.3, pp. 533–550. DOI: 10.1111/ajps.12603.

Kavanagh, Jennifer, Katherine Grace Carman, Maria DeYoreo, Nathan Chandler, and Lynn E. Davis (2020). *The Drivers of Institutional Trust and Distrust: Exploring Components of Trustworthiness*. Tech. rep. RR-A112-7. RAND Corporation. DOI: 10.7249/RRA112-7. URL: https://bit.ly/4eqI6ml.

Kim, Sunwoong, Jongho Jeong, Jin Soo Han, and Donghyuk Shin (2024). *LLM-Mirror: A Generated-Persona Approach for Survey Pre-Testing*. arXiv: 2412.03162 [cs.CY]. URL: https://arxiv.org/abs/2412.03162.

King, Gary, Christopher J.L. Murray, Joshua A. Salomon, and Ajay Tandon (Feb. 2004). "Enhancing the Validity and Cross-cultural Comparability of Measurement in Survey

Research". In: *American Political Science Review* 98.1. GaryKing.org/files/abs/vign-abs.shtml, pp. 191–207.

King, Gary and Jonathan Wand (Winter 2007). "Comparing Incomparable Survey Responses: New Tools for Anchoring Vignettes". In: *Political Analysis* 15.1, pp. 46–66. URL: j.mp/CompInC.

Latin American Public Opinion Project (2021). *Remote Pretesting Protocols*. Tech. rep. Vanderbilt University / LAPOP. URL: https://www.vanderbilt.edu/lapop/insights/IMN010en.pdf.

Lazarsfeld, Paul F (1935). "The art of asking WHY in marketing research: three principles underlying the formulation of questionnaires". In: *National marketing review*, pp. 26–38.

Lee, David S (2009). "Training, Wages, and Sample Selection: Estimating Sharp Bounds on Treatment Effects". In: *The Review of Economic Studies* 76.3, pp. 1071–1102.

Lenzner, Timo, Cornelia Neuert, and Wanda Otto (2016). *Gesis Survey Guidelines: Cognitive Pretesting*. Tech. rep. Version 2.0. URL: https://bit.ly/4fAsPRf.

Manski, Charles F. (1995). *Identification Problems in the Social Sciences*. Cambridge, MA: Harvard University Press.

Merton, Robert K (1987). "The focussed interview and focus groups: Continuities and discontinuities". In: *The Public opinion quarterly* 51.4, pp. 550–566.

Miller, Kristen (2014). "Introduction". In: *Cognitive Interviewing Methodology*. John Wiley and Sons, Inc, pp. 19–24.

Munis, B. Kal (2022). "Us Over Here Versus Them Over There...Literally: Measuring Place Resentment in American Politics". In: *Political Behavior* 44.3, pp. 1057–1078. DOI: 10.1007/s11109-020-09641-2.

Nichols, Curt (2015). "Public Opinion and the Military: A Multivariate Exploration of Attitudes in Texas". In: *Political and Military Sociology: An Annual Review* 43.1, pp. 75–105. URL: https://bit.ly/4vdkZlz.

Office of Information and Regulatory Affairs of Office of Management and Budget (2016). *Questions and Answers When Designing Surveys for Information Collections*. Tech. rep. Originally published January 2006. URL: https://bit.ly/4uT5Y7X.

Oksenberg, Lois, Charles Cannell, and Graham Kalton (1991). "New Strategies for Pretesting Survey Questions". In: *Journal of Official Statistics* 7.3, pp. 349–365.

Olivos, Francisco and Minhui Liu (2025). "ChatGPTest: Opportunities and Cautionary Tales of Utilizing AI for Questionnaire Pretesting". In: *Field Methods* 37.4. DOI: 10.1177/1525822X241280574.

Pescosolido, Bernice A., J. Scott Long, Jack K. Martin, Karen Kafadar, Sigrun Olafsdottir, and Tait R. Medina (2015). "The Theory of Industrial Society and Cultural Schemata: Does the "Cultural Myth of Stigma" Underlie the WHO Schizophrenia Paradox?" In: *American Journal of Sociology* 121.3, pp. 783–825. DOI: 10.1086/683225.

Pew Research Center (2020). *Testing Survey Questions Ahead of Time Can Help Sharpen a Poll's Focus*. URL: https://pewrsr.ch/4e2qY6B.

Robbins, Jordan M and Joachim I Krueger (2005). "Social projection to ingroups and outgroups: A review and meta-analysis". In: *Personality and social psychology review* 9.1, pp. 32–47.

Robinson, Michael (2019). *Who Follows the Generals? Polarization in Institutional Confidence in the Military*. APSA Preprints. Version 1 posted 17 September 2019. URL: https://bit.ly/4v4Ok1h.

Roller, Margaret R. and Zachary R. Smith (2025). "Qualitative Research: Advancing the Social and Behavioral Sciences". In: *Public Opinion Quarterly* 89.SI, pp. 477–491. DOI: 10.1093/poq/nfaf035.

Schwarz, Norbert (2007). "Cognitive aspects of survey methodology". In: *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition* 21.2, pp. 277–287.

Searing, Donald D. (1978). "Measuring Politicians' Values: Administration and Assessment of a Ranking Technique in the British House of Commons". In: *American Political Science Review* 72.1, pp. 65–79. DOI: 10.2307/1953599.

Spina, Nicholas (2025). "The Impact of Early COVID-19 Lockdowns on Perceptions of American Democracy". In: *Journal of Elections, Public Opinion and Parties*. Published online 10 October 2025. DOI: 10.1080/17457289.2025.2570379. URL: https://bit.ly/4xvtIRF.

Sturgis, Patrick, Caroline Roberts, and Thomas Robinson (Apr. 2026). *LLMs for Survey Pretesting: How Well Can Large Language Models Identify Flaws in Survey Questions?* Working Paper 15. Survey Futures. URL: https://tinyurl.com/2asgdhmy.

Tian, Xiaoli and Daniel A. Menchik (2008). "Putting Social Context into Text: The Semiotics of E-mail Interaction". In: *American Journal of Sociology* 114.2, pp. 332–370. DOI: 10.1086/590650.

Tjuatja, Lindia, Valerie Chen, Tongshuang Wu, Ameet Talwalkar, and Graham Neubig (2024). "Do LLMs Exhibit Human-like Response Biases? A Case Study in Survey Design". In: *Transactions of the Association for Computational Linguistics* 12, pp. 1011–1026. DOI: 10.1162/tacl_a_00685.

Tourangeau, Roger (1984). "Cognitive Sciences and Survey Methods". In: *Cognitive Aspects of Survey Methodology: Building a Bridge Between Disciplines*. National Academies Press.

Wild, Diane, Alyson Grove, Mona Martin, Sonya Eremenco, Sandra McElroy, Aneesa Verjee-Lorenz, and Pennifer Erikson (2005). "Principles of Good Practice for the Translation and Cultural Adaptation Process for Patient-Reported Outcomes (PRO) Measures: Report of the ISPOR Task Force for Translation and Cultural Adaptation". In: *Value in Health* 8.2. URL: https://bit.ly/4oe6ECC.

Wuttke, Alexander, Matthias Aßenmacher, Christopher Klamm, Max M. Lang, Quirin Würschinger, and Frauke Kreuter (May 2025). "AI Conversational Interviewing: Transforming Surveys with LLMs as Adaptive Interviewers". In: *Proceedings of the 9th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (LaTeCH-CLfL 2025)*. Albuquerque, New Mexico: Association for Computational Linguistics, pp. 179–204. URL: https://aclanthology.org/2025.latechclfl-1.17/.