Journal of Economic Literature 2025, 63(4), 1244–1280
https://doi.org/10.1257/jel.20251780

Understanding Economic Behavior Using Open-Ended Survey Data†

Ingar Haaland, Christopher Roth, Stefanie Stantcheva, and Johannes Wohlfart*

We survey the recent literature in economics using open-ended survey data to uncover mechanisms behind economic beliefs and behaviors. We first provide an overview of different applications, including the measurement of motives, mental models, narratives, attention, information transmission, and recall. We next describe different ways of eliciting open-ended responses, including single-item open-ended questions, speech recordings, and artificial intelligence–powered qualitative interviews. Subsequently, we discuss methods to annotate and analyze such data with a focus on recent advances in large language models. Our review concludes with a discussion of promising avenues for future research. (*JEL* C83, C90, D83, D91)

1. *Introduction*

Consider an individual who does not invest in the stock market or who subscribes to a dominated health insurance plan. Or think of a firm manager who refrains from cutting employee wages even as the economy enters a recession. At first glance, these behaviors may seem puzzling—yet there could be plausible reasons behind them. How should one measure these reasons? One potential solution to this problem involves simply asking respondents open-ended questions about why they engage in the behavior of interest. Unlike structured survey questions offering respondents a set of predetermined response options, open-ended

* Haaland: NHH Norwegian School of Economics, FAIR, and CEPR. Roth: University of Cologne and ECONtribute, Max Planck Institute for Research on Collective Goods, CEPR, and NHH Norwegian School of Economics. Stantcheva: Harvard University, NBER, and CEPR. Wohlfart: University of Cologne and ECONtribute, Max Planck Institute for Research on Collective Goods, CEBI, and CESifo. We thank David Romer and four anonymous referees for extremely constructive and helpful feedback. Maximilian Fell, Paul Grass, Hrishikesh Iyengar, Sarthak Joshi, Apoorv Kanoongo, Julian König, Nicolas Röver, and Gregory Tham provided excellent research assistance. An earlier version of this review was circulated under the title "Measuring What Is Top of Mind" (Haaland et al. 2024). Roth and Wohlfart: funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy—EXC 2126/1-390838866. Financial support from CEBI and the Danish National Research Foundation, grant DNRF134, is gratefully acknowledged. This work was partially supported by the RCN through its CoE Scheme, FAIR project No 262675.
† Go to https://doi.org/10.1257/jel.20251780 to visit the article page and view author disclosure statement(s).

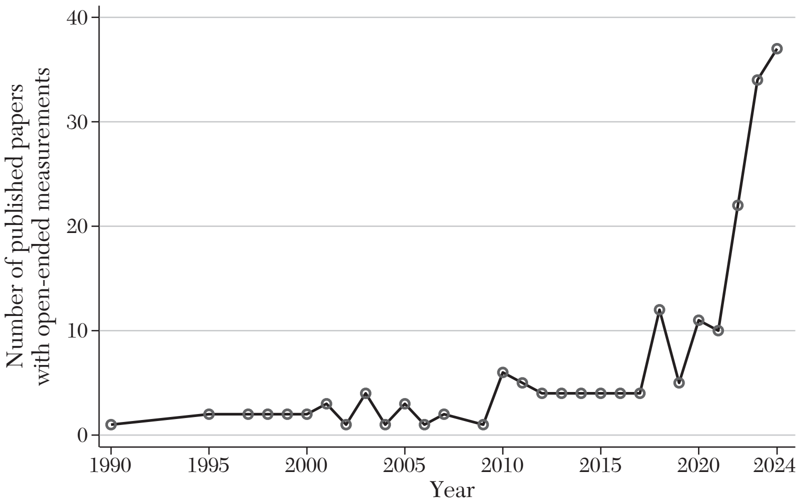

Figure 1. Number of Studies with Open-Ended Measurements (Including Qualitative Interviews) Published in Leading Journals and Working Paper Series between 1990 and 2024

Notes: This figure shows the number of studies with open-ended and qualitative text measurements published in leading journals since 1990. For 2024, publications and forthcoming papers as of mid-November are included. The figure is based on publications in leading journals (the *American Economic Review*, *American Economic Journal: Applied Economics*, *American Economic Journal: Economic Policy*, *American Economic Journal: Macroeconomics*, *Econometrica*, the *Economic Journal*, *Journal of Development Economics*, *Journal of Political Economy*, *Journal of Public Economics*, *Journal of the European Economic Association*, *Review of Economics and Statistics*, and the *Review of Economic Studies*), the *AEA Papers and Proceedings*, and working paper series (CEPR, CESifo, and NBER). To identify articles, we used Google Scholar to search for all articles published in these journals since 1990 containing the words "open-ended," "open," "qualitative interview," "qualitative survey" and then verified which of the search results featured an actual open-ended measurement. We also supplemented the figure with papers covered in our review that were not captured using this search algorithm.

questions do not prime individuals on any particular potential answer. These questions also do not require researchers to have prior knowledge about all relevant options. The qualitative text data resulting from open-ended response formats therefore provide a detailed lens into respondents' self-reported considerations.

In this paper, we review an emerging literature in economics that uses open-ended questions to better understand the mechanisms behind economic behaviors and expectations. As shown in Figure 1, open-ended survey data have become increasingly common in economics. Our review focuses on open-ended questions included in large-scale surveys, where participants are typically asked to write down their considerations in the context of a particular issue, decision, or prediction problem. Open-ended questions are applied to study topics such as attention allocation, reasoning, mental models, or verbal communication, and thus help to gain a deeper understanding of the mechanisms underlying economic choices and expectations.

Open-ended questions allow researchers to test predictions of influential theories of human behavior, such as theories of associative memory (Bordalo et al. 2024) or of the role of salience in attention allocation (Bordalo et al. 2025). They also let us paint a more realistic picture of the variables that individuals consider relevant when forming beliefs or making choices (Chinco, Hartzmark, and Sussman 2022). In addition, open-ended survey questions enable researchers to measure the motivations behind particular decisions (Braghieri, Schwardmann and Tripodi 2024) or the perceived motives driving others' behaviors (Bursztyn et al. 2023). Open-ended questions are also crucial to understanding mental models and economic narratives (Andre et al. 2024; Andre et al. 2022; Stantcheva 2024; Colarieti, Mei, and Stantcheva 2024).

The techniques presented in this review can also be used to address policy questions, such as understanding which concerns loom the largest in voters' minds (Ferrario and Stantcheva 2022), capturing the salient issues in public opinion (Geer 1991), and characterizing people's first-order considerations when thinking about redistribution (Cappelen, Falch, and Tungodden 2025; Andre 2025).

In this review, we proceed in three steps regarding open-ended survey data: its applications, how to collect it, and how to analyze it. Regarding the different applications, we cover ways of measuring motives and the reasoning behind decisions, mental models, narratives, and attention allocation. We also outline the usefulness of open-ended data to quantify the first stage in priming interventions and to measure information transmission, recall, and beliefs about experimenter expectations. Throughout our discussion, we provide examples from existing studies, explain the core advantages of open-ended questions for particular applications, and highlight specific points regarding the design and analysis in the different applications.

We next discuss the data collection process and important design considerations as well as the advantages and disadvantages of different ways of asking open-ended questions. We cover both survey questions to which participants respond in written text as well as new methods enabled by recent technological advances. These methods aim to collect rich data on people's considerations *at scale*. In particular, we discuss how speech recordings (Graeber, Roth, and Schesch 2024; Graeber, Noy, and Roth 2024) and qualitative interviews powered by artificial intelligence (AI) (Chopra and Haaland 2023; Geiecke and Jaravel 2024) can be used to measure considerations. Compared to written text responses, speech recordings provide richer data, as they also contain nonverbal cues, such as emotions, while AI-powered qualitative interviews allow for clarification and follow-up questions, leading to more depth in responses.

Subsequently, we discuss different approaches to analyzing open-ended data collected through surveys or interviews. We start by discussing human coding, including the design of coding schemes, coding with the help of research assistants, and calculation of inter-coder reliability. We then proceed to large language models (LLMs), which provide new opportunities for characterizing unstructured data in a nuanced and cost-effective way. Among other topics, we discuss practical issues related to annotating unstructured data using LLMs, including the use of application programming interfaces (APIs). Lastly, we briefly review fully automated text analysis methods.

We conclude by laying out a series of open questions and avenues for future research. Specifically, we discuss the role of incentives in increasing effort, voice-based AI interviews, as well as the possibility of combining methods from neuroeconomics with open-ended questions to better understand conscious thought processes and the attentional foundations of decision-making.[1] We will provide regular updates of relevant methodological developments in the supplemental materials accompanying the paper.

[1] Our review also relates to overview articles on attention in economics (Enke 2024; Loewenstein and Wojtowicz 2025; Bordalo, Gennaioli, and Shleifer 2022; Gabaix 2019).

Before proceeding, we address an important question: To what extent can mechanisms uncovered in open-ended responses, such as mental models, be considered genuine explanations of behavior? A long-standing critique in the social sciences and psychology argues that individuals' explanations of their decisions often represent ex post rationalizations rather than authentic accounts of their decision-making processes (Machlup 1946; Nisbett and Wilson 1977; Berger et al. 2016). For example, Jerolmack and Khan (2014) highlight that survey-based explanations frequently reflect what respondents perceive as reasonable rather than the true drivers of their behavior. This perspective cautions against treating verbal responses as direct windows into decision-making mechanisms, as doing so risks confusing narratives constructed after the fact with the actual underlying factors. We thus caution readers to remain mindful of this caveat, recognizing that individuals may sometimes not be able to fully explain the reasons behind their behavior.[2]

Nevertheless, a growing literature in psychology and behavioral economics suggests that individuals are often able to accurately articulate essential elements of their decision-making processes (Ericsson and Simon 1980, 1993; Sloman 2009; Morris et al. 2025). For example, Hanna, Mullainathan, and Schwartzstein (2014) illustrate that people's mental models significantly shape their choices. Handel and Schwartzstein (2018) emphasize that deeper insights into people's reasoning can inform the design of more effective economic policies and interventions. For instance, Handel and Schwartzstein (2018) highlight the importance of accounting for miscalibrated mental models when designing policies to help consumers avoid costly mistakes—such as selecting dominated health insurance plans or high-fee mutual funds. Even when the reasoning expressed in open-ended responses primarily reflects rationalizations, it still offers valuable insight into how individuals interpret their own behaviors and beliefs (Bruner 1991).

This review builds on seminal work in economics that pioneered the use of qualitative interviews and open-ended questions in the context of wage dynamics (Bewley 1995, 1999), price setting (Blinder et al. 1998), inflation (Shiller 1997), financial budgeting decisions (Morduch and Schneider 2017) and home price expectations (Case and Shiller 2003; Case, Shiller, and Thompson 2012).[3] Our paper also relates to research in survey methodology and on public opinions, where open-ended questions have been discussed as an alternative to closed-ended ones (Geer 1988, 1991; Krosnick 1999; Lazarsfeld 1944).

Our review also builds on important work in other disciplines that have used qualitative data for a long time. For example, research in anthropology and sociology uses open-ended data obtained from qualitative interviews (Knott et al. 2022; Patton 2002; Denzin and Lincoln 2017; Emerson, Fretz, and Shaw 1995; Kvale 1996). Sociologists and anthropologists have explored individuals' "mental models" through ethnographic methods, aiming to understand how cultural norms and social interactions shape cognition (Becker 1998; Geertz 1973; Holland and Quinn 1987). We do not go in-depth into this type of research. The distinguishing feature of the work we cover in this review is the large-scale aspect of the data.

[2] Another concern with surveys is that they may not be predictive of real-world behaviors. As Colarieti, Mei, and Stantcheva (2024) show, individuals are quite good at predicting their behaviors in familiar, everyday scenarios, and provide rationales for them, making their stated choices reliable indicators of actual actions.

[3] For excellent reviews on qualitative interviews in economics, see Piore (2006) and Starr (2014). For reviews on the design of surveys and information provision experiments in economics, see Haaland, Roth, and Wohlfart (2023); Stantcheva (2023); and Fuster and Zafar (2023).

2. *Major Applications*

In this section, we discuss different types of applications of open-ended survey data. Table 1 presents an overview of various applications, including example questions. Supplemental Appendix Tables A1–A7 provide an overview of papers in economics that use open-ended questions, organized by field of study.

2.1 *Reasoning and (Perceived) Motives behind Decisions*

Open-ended questions are particularly useful for exploring the reasoning and motives behind specific decisions. The open-ended nature of the questions allows respondents to express their motivations and lines of reasoning in a natural and unconstrained manner.

First, open-ended questions can be used to understand the considerations behind *decisions within experiments*. Such decisions include redistribution decisions (Andre 2025) or the willingness to pay for interacting with others (Braghieri, Schwardmann, and Tripodi 2024). Typically, researchers first ask respondents to make the decision or the prediction in question. Subsequently, respondents are asked to report the main considerations or reasons underlying their decision in an open-text box.

Second, open-ended questions can be used to characterize motives underlying *real-world decisions*. Applications include protesting property taxes (Nathan, Perez-Truglia, and Zentner 2025), stock market nonparticipation (Chopra and Haaland 2023), the consumption of goods with externalities (Kaufmann, Andre, and Kőszegi 2024), spending and saving decisions (Colarieti, Mei, and Stantcheva 2024), or gun ownership (Alsan, Schwartzstein, and Stantcheva 2025). Researchers usually first elicit the behavior of interest using a structured survey question and subsequently pose an open-ended question asking participants why they behave in a specific way.

Third, open-ended responses can be leveraged for the measurement of inferences about others' motives. In particular, respondents can be asked to explain why another respondent decided in a particular way. For example, Bursztyn et al. (2023) ask respondents why they think another respondent made a public posting on social media.

2.2 *Narratives and Mental Models*

Another key application concerns the narratives and mental models individuals invoke in economic contexts. According to a common definition, narratives are causal accounts for why a specific event occurs (Shiller 2017). Mental models are beliefs about the co-movement between different variables and the underlying mechanisms driving this co-movement (Andre, Schirmer, and Wohlfart 2024).

Open-ended measurements can be a powerful tool for understanding narratives and mental models. The most common applications are the following: first, asking respondents to explain the causes of a given phenomenon, e.g., asking respondents "which factors caused the increase in inflation" (Andre et al. 2024; Binetti, Nuzzi, and Stantcheva 2024); second, asking respondents about their perceived mechanisms underlying the relationship between different given variables, such as the effects of interest rate hikes on inflation (Andre et al. 2022) or the effects of "old news" on expected stock returns (Andre, Schirmer, and Wohlfart 2024); and third, measuring considerations about the broader consequences of a given change in a variable

TABLE 1
Applications of Open-Ended Survey Questions

Application	Type of measurement	Example questions
Reasoning and motives	Respondents are asked about reasons behind their decisions.	Why did you consume this news? Why did you engage in a political campaign?
Narratives and mental models	Respondents explain the relationship between two given variables, the evolution of a given variable, or the consequences of movements in a given variable.	Why do you think inflation increased to 8 percent? How do you think interest rate hikes affect subsequent inflation?
Attention allocation	Respondents list concerns, considerations, or issues that come to mind when thinking about a topic.	What comes to mind when you think about government policies? What economic issues concern you most?
Measuring priming interventions' effects	Respondents describe which considerations are top of their mind after having been primed on a particular issue.	What considerations are on your mind right now? What is the first thing that comes to mind about immigration policy?
Priming through open-ended questions	Respondents are asked to think about an issue from a specific (priming) angle.	When you think about immigration in the US, what makes you angry? What is the first thing that comes to mind about immigration policy?
Recall	Respondents are asked to recall past real-world experiences or information seeded in a baseline experiment.	What past stock market episode first comes to your mind? Please describe what you remember about this scenario.
Information transmission	Respondents communicate information (e.g., an explanation) to another respondent.	Please explain the reasoning behind your choice to another respondent.
Experimenter demand	Respondents guess the purpose or hypothesis of the study.	What do you think is the hypothesis that is being tested in this study?
Testing respondents' knowledge	Respondents are asked to define or explain something.	How would you define [concept] in your own words?

without specifying any particular outcome variable, such as the effect of inflation on the economy at large and on people's lives. For example, Stantcheva (2024) asks respondents, "What were the most important impacts of inflation on your life?"

Researchers measuring subjective models and narratives are often interested in comparing respondents' reasoning with textbook models. The analysis of the data then requires an especially careful design of coding schemes that are able to capture subtle differences in reasoning (see Section 4).

2.3 *Attention Allocation*

Open-ended measurement approaches can be used to measure people's attention allocation—that is, their allocation of cognitive resources across different topics or issues. For instance, such measurement can be applied to better understand which characteristics of an asset investors attend to when making investment decisions (Chinco, Hartzmark, and Sussman 2022; Wekhof 2024), people's attention to different aspects of a statistical problem (Bordalo et al. 2025), or households' and firm managers' attention allocation across different economic variables (Link et al. 2024). Open-ended questions are attractive for measuring attention allocation, as respondents' attention is not mechanically drawn to topics that appear in response options.

While measuring motives, mental models, or narratives typically requires measuring respondents' full *explanations* for a certain decision, belief, or event, eliciting attention allocation often merely requires capturing which *topics or issues* are top of respondents' minds in a given context (Ferrario and Stantcheva 2022). This can be done by confronting survey respondents with a prompt on the context of interest. For instance, Link et al. (2024) measure households' and firm managers' attention to different economic topics by asking them, "What topics come to mind when you think about the economic situation of your company/household?" On policy issues, Ferrario and Stantcheva (2022) ask, "What are your main considerations about the federal income tax?" The participants then write their responses into an open-text box. The topics raised in the open-text responses provide insights into respondents' attention allocation to different issues. Given that the interest lies in the topics or issues that are mentioned rather than in more concrete arguments and explanations, open-ended data on attention allocation often requires less nuanced coding, favoring automated methods (see Section 4).

2.4 *Priming Interventions*

Priming interventions are widely used in economic research (Cohn and Maréchal 2016). Such interventions are typically used to exogenously draw respondents' attention to a particular issue or aspect of a decision problem. This allows researchers to study the causal effect of attention to a particular issue on beliefs, decisions, or behaviors elicited later in the survey. The mechanisms underlying the effects of priming interventions, however, have been widely criticized for being a black box (Cohn and Maréchal 2016). Open-ended questions open up the possibility of measuring how priming interventions affect attention allocation (Henkel, Oslislo, and Schwerter 2024).

A common approach to priming is to order survey questions differently so as to generate variation in the contextual cues treated and control respondents have been exposed to when making the decision or prediction of interest. For example, Alesina, Miano, and Stantcheva (2023) use such an approach to study how attention to different issues affects attitudes toward immigration. An open-ended question can then be used to measure which issues are top of respondents' minds, for instance, when taking a specific decision within the experiment. The resulting text data then allows the researcher to estimate the "first-stage" effect of the priming intervention on the respondents' attention allocation. Structured questions are less suited for this purpose, as the included response options might themselves change respondents' focus, potentially interfering with the treatment variation created by the contextual cues. Structured questions may also be more likely to induce experimenter demand effects.

Open-ended questions cannot only be used to measure the impact of priming interventions—they can be the priming interventions. For instance, Stantcheva (2022) asks randomly selected subsamples of respondents a series of open-ended questions to consider the impacts of trade on their consumption bundles or on their jobs. The goal of these questions is to prime respondents to focus on either the consumption or the job impacts from trade. Algan et al. (2025) induce emotions such as anger or fear in respondents by asking open-ended questions about what makes them angry or happy regarding specific policy issues such as immigration, trade, or tax policy.

2.5 *Recall*

Open-ended questions can also be used to study memory and recall (Bordalo et al. 2024). Such applications proceed in similar ways as applications to attention allocation (discussed in Section 2): The respondents are exposed to a prompt or cue and are asked to write down their thoughts; however, the prompt concerns things that happened in the past. Open-ended measures appear particularly attractive for studying recall: closed-ended questions with response options indicating particular past events remind the respondents of these events by construction, making them inherently unsuited to understanding respondents' natural recall processes (Connor Desai and Reimers 2019). Moreover, open-ended data reveal additional nuance about what is being recalled and provide the opportunity to detect memory distortions and confusion.

On the one hand, open-ended questions can be used to measure recall of *real-world experiences* people made in the past. For example, Jiang et al. (2025) ask retail investors to write down a past stock market episode that first comes to mind. They show that the stock market performance on the survey day shapes investors' recall, which in turn influences their beliefs about future returns.

On the other hand, open-ended questions can be used to study the recall of *information seeded in a baseline experiment*. For example, Graeber, Roth, and Zimmermann (2024) use the following open-ended question: "Please tell us anything you remember about this product scenario. Include as much detail as you can. Most importantly, please describe things in the order they come to mind, that is, the first thought first, then the next one etc." This enables the authors to study selective recall of stories versus statistical information. Reassuringly, their open-ended data yields similar conclusions to a structured incentivized task, suggesting that unstructured open-ended elicitations are a reliable measure of recall even in the absence of incentives for accuracy.

2.6 *Information Transmission*

Given that most communication relies on natural language, open-ended questions also lend themselves to studying information transmission. For example, questions asking subjects to record their response in a voice message have been used to study the causal impact of verbal explanations on social learning (Graeber, Roth, and Schesch 2024) and information transmission (Graeber, Noy, and Roth 2024). Graeber, Roth, and Schesch (2024) tell respondents to "record an explanation that helps the other participant select the correct answer." Similarly, it is possible to study communication through writing in an open-text box (Grunewald et al. 2024).

Open-ended measurement mimics key features of communication in the real world and allows individuals to express their considerations, feelings, and experiences in their own words, without being constrained by predefined options. One special feature of open-ended questions in the context of communication is that one can provide respondents with incentives in a straightforward manner. For example, Graeber, Roth, and Schesch (2024) incentivize people's recorded explanations by telling them that their payoff will depend on the accuracy of the choices made by the other respondent who will receive their recorded explanation before making their choice. As such, respondents have aligned incentives to communicate the most informative explanation to the other respondent.

2.7 *Experimenter Demand Effects*

Experimenter demand effects are an important concern in survey-based research (de Quidt, Haushofer, and Roth 2018). Open-ended questions are increasingly used to mitigate concerns about experimenter demand effects. Specifically, respondents can be asked to guess the hypothesis that the researchers are testing in an open-ended question included at the end of the experiment. For example, participants are asked: What do you think is the hypothesis that the researchers aim to test? or What do you think is the purpose of this study? The open-ended nature of such questions ensures that respondents do not simply tick response options that are socially desirable, potentially giving a false impression of the prevalence of demand effects.

2.8 *Testing for Knowledge and Understanding*

Open-ended survey questions can also be used to measure respondents' knowledge. Closed-ended questions with specific answer options often provide information that might inadvertently influence responses, making them less effective at gauging respondents' underlying knowledge (Brosius, Hameleers, and van der Meer 2022). For example, Stantcheva (2024) asks respondents to define what inflation is. When analyzing the data in such applications, it is important to specify clear criteria for what counts as a correct or false response.

3. *Collecting Open-Ended Survey Data*

In this section, we discuss design considerations for open-ended questions in the context of surveys and qualitative interviews. We start with a discussion of single-item open-text boxes—the most common way of collecting open-ended survey data and the main focus of this review—covering their advantages and disadvantages compared to more traditional closed-ended survey questions. A complementary review by Stantcheva (2023) covers guidance on the complete survey process. We then highlight the recent methodological advance of using speech recordings to measure considerations, which allows researchers to capture rich contextual data such as nonverbal cues. Subsequently, we provide a brief overview of human-led qualitative interviews, before providing a more in-depth discussion of AI-powered qualitative interviews. Lastly, we briefly discuss approaches used in other social sciences.

3.1 *Measuring Written Considerations in an Open-Text Box*

The most common approach to eliciting open-ended considerations in surveys is to invite participants to describe, in an open-text box, the key factors they consider when reflecting on a particular issue, forming a specific belief, or making a specific decision.

3.1.1 *Design Considerations*

Cognitive Costs.—One key issue to consider when designing open-ended questions is that they demand more cognitive effort from respondents than structured formats. This can lead to higher nonresponse rates and lower response quality if not managed carefully (Dillman 2007; Millar and Dillman 2012). To preserve high data quality, it is important to reduce the mental load on respondents as much as possible.

One potential strategy is to position open-ended questions early in the survey. Early placement may help engage respondents when they are likely to be more attentive and reduce fatigue, which can compromise data quality if open-ended questions are asked later. Indeed, research has shown that placing open-ended questions at the start of a survey leads to longer responses (Galesic and Bosnjak 2009; Miller and Lambert 2014). Additionally, limiting the number of open-ended questions in a single survey prevents overburdening respondents, which should reduce dropouts and help preserve response accuracy.

Mode Effects.—It is also important to consider the device respondents may use to complete the survey. Open-ended questions can be particularly challenging for respondents on mobile devices due to limited screen space and typing difficulties, which can lead to shorter, less thoughtful responses (Mavletova 2013). If researchers are worried about these effects, they can encourage respondents to complete the survey with a computer.[4] Nevertheless, a major advantage of online surveys is the ability to reach broad and diverse samples, which is largely due to the use of mobile technologies. Thus, such restrictions need to be carefully weighed against the benefits of allowing surveys on mobile devices.

Gauging Willingness to Write.—Many studies using open-ended responses begin their surveys with a question to gauge respondents' general willingness to write, often using this question as a screening tool (e.g., Graeber, Roth, and Zimmermann 2024). This approach serves to filter out respondents who are disinclined to invest the necessary effort in open-ended responses or who are inattentive. While closed-ended questions can also be used as attention checks, they are more susceptible to participants guessing the right answers and allow little variation in the measured degrees of engagement and attention (Ziegler 2022). Attention checks based on open-ended questions thus impose stricter screening criteria, enhancing overall data quality. The use of such screeners has to be weighed against the additional selection they produce.

Increasing Respondent Effort.—Smyth et al. (2009) provide evidence that simple motivational prompts, such as "This question is very important to our research. Please take your time answering it," increase response length and lead to a higher fraction of respondents giving more elaborate answers. Subjects that receive the motivational prompt also take more time when responding. Additionally, providing clear guidance on the expected length and format of responses should help reduce heterogeneity in response behavior (Züll 2016). When participants know approximately how much text is expected (e.g., "Please respond in full sentences" or "Please spend 1–2 minutes on this question"), they should be more likely to respond thoughtfully and consistently.

[4] Note that some survey platforms allow restricting the survey to desktop participants, e.g., Prolific.

Lastly, tailoring the visual layout of response boxes to the type of input needed enhances clarity and ease of response. For single-answer questions, a single entry box is usually sufficient. For multiple answers, offering separate fields for each response creates a structured layout, reducing the cognitive load associated with parsing multiple ideas in a single text box. When longer written responses are desired, using a larger text box can make the task feel more manageable and signal to respondents that a more detailed answer is appropriate. Indeed, Smyth et al. (2009) and Israel (2010) document that larger text boxes somewhat increase the response length for some groups of respondents.

Ex Post Rationalization.—Another important consideration is minimizing the potential for ex post rationalization, which can introduce biases (Nisbett and Wilson 1977). In particular, participants might retrospectively make up reasons to justify their earlier responses rather than reporting genuine reasons. This concern about rationalization is particularly relevant when image concerns make specific reasons more desirable than others. To reduce this risk, one approach is to ask open-ended questions directly on the decision screen, potentially even before participants make their choice. This setup prompts participants to articulate their reasoning in real time, rather than rationalizing retrospectively. However, as noted by Imas, Kuhn, and Mironova (2022), such a prompt may inadvertently influence behavior by increasing deliberation time, which could impact the naturalness of the decision-making process.

LLM-Generated Responses.—One concern with open-text data collection is that participants could use LLMs to generate their responses, potentially compromising data integrity. Since these responses may not genuinely reflect participants' own thoughts or experiences, they could skew research results. While there exist detection tools to flag AI-generated text, these detection tools often have low accuracy (Akram 2023) and might perform even worse as LLMs become more advanced. One attempt to limit LLM-generated responses is to prevent copy-pasting into response boxes, as implemented in Chopra and Haaland (2023). However, in some cases, respondents might legitimately use LLMs to edit or refine their open-ended text responses, making it potentially counterproductive to try to limit their use. A bigger concern in this regard is that responses from AI-based bots may become indistinguishable from human responses. As shown in Höhne et al. (2024), AI-based bots are able to pass common attention checks and populate open-ended questions in a meaningful way. They are also able to pass common bot protection measures, such as CAPTCHAs and hidden "honey pot" questions. These developments make it very important to recruit survey respondents from platforms with effective measures to deal with bots.

3.1.2 *Advantages and Disadvantages of Open-Ended Questions*

We next outline the main advantages and disadvantages of questions with an open-text response box compared to more traditional questions with structured response options. Table 2 provides an overview of the trade-offs involved when choosing between open-ended and closed-ended response formats.

Advantages of Open-Ended Questions.—Compared to structured approaches, open-ended measurement of considerations offers several advantages.

TABLE 2
Comparison of Open-Ended and Closed-Ended Questions

Aspect	Open-ended questions	Closed-ended questions
Depth of response	Allows detailed, nuanced responses, capturing subtle insights (Hansen and Świderska 2023)	Constrains depth, produces standardized responses
Ease of analysis	Complex to analyze, often requiring coding (Geer 1991; Singer and Couper 2017)	Easier to analyze quantitatively
Respondent fatigue	Can increase fatigue due to longer, more involved answers (Dillman, Smyth, and Christian 2014)	Can be completed more quickly, reducing fatigue risk
Flexibility	Adaptable to various contexts, uncovering unexpected themes	Limited flexibility; predetermined answers may miss key insights (Krosnick 2018)
Scalability	Scalability more challenging due to analysis complexity (Dutta and O'Rourke 2020).	Highly scalable for large samples and repeated measurements
Bias–variance trade-off	Lower bias but higher variance due to more noise in the open-ended data (Sinkowitz-Cochran 2013)	Possibly biased by the options provided but lower variance (Reja et al. 2003); responses are standardized and consistent
Effort variability	High; participants may invest varying levels of effort, leading to heterogeneous response quality (Miller and Lambert 2014)	Low; structured format limits variability in participant effort, ensuring more consistent quality
Nonresponse bias	High; participants who choose not to answer open-ended questions may differ systematically from those who do (Reja et al. 2003)	Low; participants are more likely to respond due to the simplified format

First, open-ended questions allow respondents to freely express their considerations, not restricting them to a predefined set of structured response options (RePass 1971; Geer 1988; Kelley 1983). This is especially important in settings where the researcher wants to discover novel factors and in settings where it is difficult for the researcher to predict respondents' spontaneous considerations ex ante (Krosnick 2018). Open-ended responses may also reveal misunderstandings or confusion on the part of participants and allow for qualitative insights that cannot be achieved with structured measures. Compared to closed-ended questions, which constrain the depth of responses but simplify and standardize them, open-ended questions allow for detailed, nuanced responses, often uncovering unexpected themes and providing a richer understanding of the respondents' perspectives (Hansen and Świderska 2023; Krosnick 2018).

Second, open-ended questions do not change people's considerations by informing them about potential lines of reasoning or drawing their attention to particular issues through the displayed response options.[5] This feature should alleviate concerns about potential confounds, such as social desirability bias or ex post rationalization (Singer and Couper 2017). For instance, when eliciting memories, asking respondents to write down events they remember allows them to provide genuine recollections, whereas a structured list of response options may confound whether participants truly recall the event or are simply reminded of it (Connor Desai and Reimers 2019). In the case of questions related to knowledge, open-ended measures do not prime respondents about magnitudes or signs and can thus better capture the participants' actual knowledge (Brosius, Hameleers, and van der Meer 2022). This flexibility contrasts with closed-ended questions, where predetermined answers might miss key insights (Krosnick 2018).

Third, open-ended questions can be asked directly on the screen eliciting the prediction or decision of interest, which allows researchers to document the respondents' considerations immediately after they have made their prediction. This potentially allows for a more precise measurement and might further mitigate ex post rationalization. Structured questions are unsuited for being asked on the decision screen, as the content of the response options might change respondents' decision. As mentioned above, one caveat is that even the mere presence of the open-ended question might change the decision-making process by inducing more deliberation (Imas, Kuhn, and Mironova 2022).

Disadvantages of Open-Ended Questions.—Open-ended measurement techniques also have a series of disadvantages. First, as a result of their unstructured nature, there is likely more scope for classical or nonclassical measurement error, as some respondents may be unwilling to exert effort when describing their considerations. The willingness to exert effort may vary systematically across different groups (Miller and Lambert 2014). Even when respondents exert substantial effort, some responses may still be ambiguous and hard to interpret.[6] Open-ended questions are also more time intensive, potentially increasing respondent fatigue, whereas closed-ended questions reduce fatigue risks by being quicker to complete (Dillman, Smyth, and Christian 2014).[7]

Another source of measurement error arises from the potentially large variation in the way individuals understand and respond to open-ended questions. This variation may affect the content of the answer and its length (Dunn and Gómez 2023). For instance, consider the setting in Andre et al. (2022), where respondents describe the considerations underlying their predicted effects of macroeconomic shocks on inflation and unemployment. In this setting, a respondent may write that they used their knowledge of economics without indicating which specific economic mechanism they had in mind. Another respondent may describe the full propagation channels of the shocks. This variability is reflected in a higher variance of responses compared to closed-ended questions, which offer standardized and consistent

[5] Of course, the question itself—even when presented merely with a text box—could prime subjects on its topic. However, this issue is common across any type of survey question and seems hard to avoid. Structured questions *additionally* prime subjects on potential responses to the question.

[6] As we review in detail in Section 3.3.2, AI interviews offer a promising avenue for partly dealing with this source of measurement error.

[7] These issues potentially introduce a bias–variance trade-off: Compared to structured formats, open-ended formats may be less subject to biases, e.g., due to priming, but feature a higher variance due to increased noise (Sinkowitz-Cochran 2013).

answers but may introduce bias due to constrained options (Sinkowitz-Cochran 2013; Reja et al. 2003).

Next to measurement error, selective nonresponse bias is a prominent challenge for open-ended questions, as participants who choose not to answer open-ended questions may differ systematically from those who do, potentially skewing results. For instance, Reja et al. (2003) show that open-ended questions produce more missing data and inadequate answers than closed-ended ones. By contrast, Geer (1988) shows that almost all subjects in their setting respond to open-ended questions and that nonresponse is driven by disinterest in the specific question posed, rather than an inability to answer such questions in general. Miller and Lambert (2014) provide a systematic analysis of nonresponse, which reveals that factors such as age, employment status, and race are related to the likelihood of responding to open-ended items. For instance, older, unemployed or retired respondents are more likely to provide answers. One way to mitigate nonresponse bias could be higher participation incentives (Dutz et al. 2022). Additional systematic evidence on how nonresponse bias to open-ended questions varies across different economic decision contexts would be helpful.

Finally, open-ended responses involve challenges in the analysis stage. While text analysis methods are straightforward to implement, it is often necessary to develop a coding scheme to exploit the full richness of the data (Saldaña 2021). Developing a coding scheme is a costly process and requires the researcher to make subjective choices that might not be fully replicable and could also be prone to potential researcher biases (Geer 1991; Singer and Couper 2017). There are also subjective judgments to be made when coding the responses according to the scheme, potentially introducing additional noise and measurement error (O'Connor and Joffe 2020; Saldaña 2021). Although LLMs can reduce the costs of annotating open-ended text data (Törnberg 2024; Gilardi, Alizadeh, and Kubli 2023), they could also introduce biases. Therefore, comparing the LLM coding with human coders remains important, particularly for responses that require nuanced judgments beyond current LLM capabilities. This complexity contrasts with the ease of analyzing closed-ended questions.

When to Use Open-Ended versus Closed-Ended Questions.—Understanding the distinct advantages of each question type can guide researchers in selecting the most appropriate tool for their study objectives. Open-ended questions are particularly effective in exploratory phases, such as when generating hypotheses or gathering initial insights into a new topic (Krosnick 2018). They allow respondents to provide insights that the researcher might not anticipate, revealing context-specific information or unique perspectives. However, they have traditionally been more resource-intensive, given the high dimensionality of such data (Geer 1991; Singer and Couper 2017). While this has traditionally limited scalability, recent technological advances in online surveys and LLMs greatly reduce the cost of collecting open-ended survey data at scale.

In contrast, closed-ended questions offer streamlined, quantifiable data that can be readily analyzed and that is easier to compare across samples (Reja et al. 2003). This format is advantageous when the research objective is to test specific hypotheses or examine patterns across large groups. Open-ended questions are most appropriate when the diversity of potential answers cannot be captured through predefined options, when priming through provided response options is a concern, when responses require narrative detail that resists reduction to brief categories, and when gauging knowledge, as open-ended formats minimize the influence of random correct answers inherent in true/false setups (see the discussion in Fowler 1995).

3.2 *Measuring Considerations with Speech Recordings*

A recent innovation is to ask participants to record their considerations instead of writing them down (Graeber, Roth, and Schesch 2024; Graeber, Noy, and Roth 2024). Participants are prompted to verbalize their considerations in response to an open-ended question. Speech recordings thus capture real-time thought processes and articulation in a dynamic manner.

Design Considerations.—Collecting speech recordings as part of a data collection process requires careful consideration of several critical factors. Providing participants with an initial practice opportunity or incorporating a speech recording into an attention check can help familiarize them with the recording process and ensure their equipment is functioning properly. For example, Graeber, Roth, and Schesch (2024) ask respondents to record a voice message with a duration of at least 15 seconds as an initial attention check. This step allows participants to address any technical issues early, fostering confidence and reducing the likelihood of errors during data collection. Moreover, clear guidance should be provided on how to set up an optimal recording environment.

Another key consideration is addressing participants' privacy concerns and error recovery needs. Sharing voice data can feel more intrusive than submitting written responses, making it vital to communicate clearly about how recordings will be securely stored and used. Explicit consent must be obtained, with an emphasis on anonymizing data where possible to protect participant identity. Moreover, one potentially useful design option could offer participants the opportunity to correct errors in their recordings through a simple and accessible "rerecord" feature. These steps can make the speech recording process more user-friendly and trustworthy.

Advantages of Speech Recordings.—Speech recordings have several advantages relative to written text responses. Speech recordings capture the spontaneity and natural flow of considerations, which are often lost in written communication. Speech may be particularly effective at capturing what initially comes to mind. Beyond text alone, speech recordings capture more features than just text, including information about emotions, tone, emphasis, and natural disfluencies.[8] For instance, when eliciting narratives—the stories people tell to explain a specific event (e.g., the rise in inflation or past financial crises)—documenting the broader thought process and the emotional tone people use to discuss different relevant factors might give nuanced insights into their thinking. Galasso, Nannicini, and Nozza (2024) study differences in responses to open-ended questions when they are either given through text or audio. Respondents who provide audio answers give longer, though lexically simpler, responses compared to those who type. Galasso, Nannicini, and Nozza (2024) also document that oral responses offer more information and contain more personal experiences than written responses.

Disadvantages of Speech Recordings.—However, there are also potential disadvantages to this method. One potential concern is participant self-consciousness: awareness of being recorded might influence how participants express themselves, possibly leading to altered or restrained responses, though data from Graeber, Roth, and Schesch (2024) suggest that participants feel

[8] For an excellent review on prosody (rhythm, stress, and intonation patterns of speech), see Wagner and Watson (2010).

comfortable recording themselves. Additionally, analyzing speech data can be more complex and time consuming than written responses due to the need to interpret nonverbal cues, such as hesitation markers or disfluencies. Finally, technical limitations, such as poor audio quality, or speech impediments can pose challenges in ensuring clarity and usability of the recordings, although this rarely matters in practice (Graeber, Roth, and Schesch 2024; Graeber, Noy, and Roth 2024).

3.3 *Qualitative Interviews*

3.3.1 *Qualitative Interviews in the Social Sciences*

Qualitative interviews allow individuals to articulate, in their own words, their perceptions and interpretations of the world around them (Knott et al. 2022). Designed to be "flexible, iterative, and continuous," these interviews evolve naturally rather than adhering to a rigid, preplanned structure (Rubin and Rubin 2011). Their adaptive nature and ability to explore emerging themes in depth make them particularly effective for hypothesis generation and understanding personal experiences, cognitive processes, and mental models.

In-depth interviews involve extended discussions with research subjects and can be structured, semi-structured, or unstructured, depending on the level of adherence to predetermined questions. Larger projects often use structured or semi-structured formats for comparability, while smaller projects may impose less structure to capture respondents' views more naturally. Ideally, interviews are recorded and transcribed for comprehensive analysis, though note taking can be used if recording is not possible. Consistent, detailed records are recommended for systematic analysis. Interviews can be conducted in person or through video, voice, or text. Namey et al. (2020) show that there are no differences in the quality and quantity of information communicated through face-to-face compared to text-based interviews. At the same time, a text-based approach might have several advantages, such as a greater sense of privacy and control of the interview (Gibson 2022).

Advantages and Disadvantages Compared to Open-Ended Questions.—Qualitative interviews offer many advantages over predefined open-ended survey responses, particularly in their ability to capture depth and context (Patton 2002; Knott et al. 2022). Unlike online survey questionnaires, interviews enable iterative questioning, allowing researchers to probe and clarify answers in real time. This process often uncovers hidden nuances, complex mechanisms, and contextual factors that simpler methods cannot address. Additionally, interviews provide a holistic view of participants' perspectives, situating responses within broader life circumstances, attitudes, or cultural contexts, making them especially valuable for understanding subjective experiences or intricate decision-making processes (Denzin and Lincoln 2017). While qualitative interviews offer more richness than predefined open-ended questions, they also have some downsides. First, they are expensive and time consuming to conduct, making them very difficult to scale. Second, they are prone to potential interviewer biases (Himelein 2016). For instance, Stefkovics and Sik (2022) show that, when using qualitative interviews to understand the overall happiness of individuals, the interviewer's own happiness strongly correlates with the respondents' measured levels of happiness.

Prominent Applications in Economics.—Qualitative interviews have been used to study questions in labor economics and macroeconomics for a long time. In a classic study, Blinder et al. (1998) conduct interviews with business leaders to differentiate between different theories of price stickiness. Relatedly, Bewley (1999) conducts interviews with a large sample of executives, labor leaders, and other professionals to understand why businesses are not willing to cut wages during recessions when labor demand is low. While there are more recent prominent examples of qualitative interviews in economics (e.g., Bergman et al. 2024; Bustos et al. 2022; Duraj et al. 2024), they are still not commonly used among economists, who typically collect open-ended responses using survey questions without any adaptive probing.

Compared to other social scientists, economists typically would like to conduct interviews with larger and more representative samples, making the collections much more expensive. For instance, in an early discussion of interviews in economics, Bewley (2002) writes the following: “It is important that the sample be large, both to be confident of conclusions and because of the need for variety and key informants.”

3.3.2 *AI-Conducted Qualitative Interviews*

With recent advances in generative AI, it is now possible to conduct high-quality qualitative interviews with AI, making them low-cost and compatible with large-scale surveys. Chopra and Haaland (2023) introduce a framework for conducting qualitative interviews using an AI interviewer, leveraging the advanced capabilities of transformer-based LLMs. Relying on API integration with OpenAI’s GPT-4, they conduct text-based interviews using a chat interface that mirrors text messaging applications. Their application can easily be integrated into standard survey software, such as Qualtrics, and allows researchers to conduct unlimited interviews in parallel at a marginal cost of less than ten cents per interview for the API calls.

The AI interviewer is programmed to adhere to the methodological best practices inherent in qualitative research, such as using open-ended, non-leading questions. The key advantage of using AI-conducted interviews compared to a series of predefined open-ended questions is the capability for adaptive probing. Probing questions have two main purposes. First, they can resolve ambiguities when respondents provide answers that are vague or difficult to interpret. Second, they can be used to achieve breadth and depth of the conversation.

Quality of AI-Conducted Interviews.—A key question is whether AI-conducted interviews can be of high quality. Chopra and Haaland (2023) evaluate the quality of their AI-conducted interviews using several complementary strategies, including respondent engagement, human evaluation of the interview transcripts, the potential for making novel discoveries, and the predictive power for economic decisions. They find high respondent engagement throughout the interviews. Participants write 29 words per minute, almost 50 percent more than typical benchmarks from chat-based interviews with human interviewers (Namey et al. 2020).

Furthermore, in contrast to surveys with a series of open-ended questions, effort does not decline over the course of the interview. Respondents also rate the interview experience highly, with 96 percent indicating a preference to participate in another interview. Additionally, 53 percent of the respondents report preferring an AI interviewer over a human one, while 21 percent express a preference for a human-led interview. A separate correspondence study that randomizes whether a study invitation is to a “40-minute survey,” a “40-minute survey with a text-based interview with a human,” or a “40-minute survey with a text-based interview with an AI” shows that the high satisfaction is not driven by selection effects: While there is no differential selection into an AI-conducted text-based interview compared to a regular survey, respondents are significantly less likely to sign up for a human-conducted text-based interview (Chopra and Haaland 2023).

Another important quality metric is to what extent the AI interviewer aligns with its instructions. A team of trained human evaluators systematically hand-coded 12,000 interview questions and responses, showing a high alignment with the instructions: 95 percent of questions are open-ended, 94 percent are non-leading and neutral, and 84 percent of questions are considered high quality according to textbook standards. The human evaluation also shows that hallucination of the AI is a close to nonexistent problem, happening in only 0.01 percent of cases.

The hallmark of qualitative research is its ability to discover novel findings. In their main application, where they collect 385 AI-conducted interviews on stock market nonparticipation, Chopra and Haaland (2023) demonstrate that AI-led interviews can be used to generate novel hypotheses. They also compare AI-led interviews to a series of predefined open-ended questions, showing that AI-conducted interviews lead to richer insights that are qualitatively different from those observed from predefined open-ended questions. For instance, narratives, mental models, and subjective experiences are frequently discovered during interviews but are almost absent in responses to predefined open-ended questions. Furthermore, they run a follow-up study that shows that factors mentioned by respondents in the interviews predict economic behavior eight months later, suggesting that common concerns about uninformative "cheap talk" dominating the discourse in qualitative interviews are unwarranted.

Robustness across Different Interview Settings.—Geiecke and Jaravel (2024) further demonstrate the robustness of the method through a series of AI-conducted interviews on measuring meaning in life, people's political preferences, and decision-making in the context of educational and occupational choices. Respondents consistently rate the interview experience favorably, underlining how AI interviews can be flexibly adopted in different settings with high interviewee satisfaction. To assess the quality of their AI-conducted interviews, Geiecke and Jaravel (2024) work with trained sociologists who rate the quality of the AI-conducted interviews relative to a hypothetical human expert. The sociologists perceive the quality of the AI-conducted interviews as similar to what a hypothetical human expert could have achieved under similar circumstances, again demonstrating the robustness of the method.

Flexible Implementation.—The implementation of AI interviews is flexible and the number of questions included depends on the goal of the interview. For instance, when surveying less literate populations or those with lower educational backgrounds, who may find it more difficult to articulate their considerations, it can be important to clarify ambiguous responses. Allowing an AI interviewer to ask a follow-up question to resolve ambiguities in the initial top-of-mind response might significantly increase the quality of the qualitative data at a relatively low cost (Chopra and Haaland 2023). In other settings, a full interview with several follow-up questions to achieve additional breadth and depth might be desirable, but this depends on the setting, time budget, and other factors.

It is worth emphasizing that the AI-conducted interviews by Chopra and Haaland (2023) and Geiecke and Jaravel (2024) relied on carefully tested prompts that reflect best practices. Interview quality might deteriorate if interviews are conducted using prompts that have not been developed and validated according to best practice methods. However, both Chopra and

Haaland (2023) and Geiecke and Jaravel (2024) use prompts that require only minimal adjustments to adapt them to different settings, making it possible for other researchers to conduct AI interviews without developing their own prompts from scratch. Open-source platforms for AI-conducted qualitative interviews are provided by both Geiecke and Jaravel (2024) and Chopra and Haaland (2023).[9]

Advantages and Disadvantages Compared to Human-Led Interviews.—AI-conducted interviews inherit many of the same challenges as human-led qualitative interviews, such as a lack of comparability between respondents, a factor that is magnified compared to single open-ended questions. In addition to these, AI-conducted interviews face unique challenges, such as potential algorithmic biases (Rozado 2024) and potential concerns about data privacy (Dell 2025). Finally, an important concern is that selection into participating in an AI interview depends on trust in AI. This might induce significant selection biases by making less tech-savvy and more conservative respondents less keen to participate in AI interviews. Yet, Chopra and Haaland (2023) show that this does not seem to be a major concern in practice, as respondents in a correspondence study are equally likely to sign up for a text-based AI interview as for a regular survey. Moreover, standard demographic characteristics do not predict differential selection into surveys compared to AI interviews. Taken together, these data counteract concerns about severe selection effects specific to AI interviews.

While AI-conducted interviews offer high-quality data at a low cost, there are still some cases where a human-led interview may be a better choice. For example, in sensitive topics where privacy concerns are heightened and interviewees may need emotional support, a human-led interview may be necessary for ethical and practical reasons. Human-led interviews are also preferable in situations requiring responsiveness to emotional language and facial expressions.[10] However, as video-based AI-conducted interviews with human avatars become feasible and low-cost, the AI-based interview experience might further improve. Video-based interviews might also give rise to new dimensions of analysis, for instance, facial expressions and emotionality of language. Finally, while an experienced human interviewer may ask higher-quality questions and have more flexibility to develop and pursue hypotheses during interviews, these potential advantages must be weighed against the higher costs of human-led data collections. Furthermore, AI interviews might be a better choice when high consistency between interviews is considered important.

3.4 *Similarities and Differences to Other Social Sciences*

In what follows, we briefly discuss how other social sciences approach the collection of open-ended data compared to economics. Anthropologists focus on cultural context, using open-ended interviews as part of immersive fieldwork to understand beliefs and practices within specific cultural settings (Bernard 2018; Spradley 1979).[11] Anthropological interviews tend to

[9]The platforms are available on the following links: https://github.com/fchop/interviews and https://github.com/friedrichgeiecke/interviews.

[10]While Namey et al. (2020) do not find quality differences between face-to-face and text-based interviews, this may depend on the interview's context.

[11]Anthropological approaches have also been applied to economic questions. For example, Ho (2009) studies the culture of Wall Street investment banks, Venkatesh (2008) analyzes career ladders in urban gangs, and Levitt and Venkatesh (2000) examine the financial activities of a drug-selling street gang.

be unstructured and exploratory, allowing researchers to adapt questions based on participant responses and emerging insights. In contrast, economists typically use interviews in a more structured manner, often designed to align with specific hypotheses or to complement quantitative data collections.

Sociologists emphasize social structures and patterns, often employing semi-structured interviews to link individual experiences to broader societal forces and to enable comparative analysis (Bourdieu 1990; Weiss 1994). These interviews often explore themes such as inequality, social mobility, and institutional barriers, providing rich qualitative data on lived experiences. Economists, while also interested in these topics, often approach interviews with a focus on understanding specific mechanisms or obtaining actionable policy insights.

Psychologists prioritize individual cognition and emotions, using structured interviews to explore psychological processes, often integrating them with experimental or clinical approaches (Kvale and Brinkmann 2007; Kazdin 2022). Behavioral and experimental economists have increasingly drawn on similar methods. For instance, interviews can be used to probe how individuals perceive probabilities, understand incentives, or frame decisions, complementing experimental approaches. However, economists often prioritize comparability across participants and replicability of findings, which can limit the depth of open-ended responses compared to the approaches used in psychology.

Despite these disciplinary differences, all four fields—anthropology, sociology, psychology, and economics—use open-ended data as a tool to gather rich qualitative insights. Moreover, all four fields rely on these methods for hypothesis generation, offering researchers a means to identify emergent themes and develop theories grounded in empirical observation.

4. *Analyzing Open-Ended Survey Data*

In this section, we review different methods for analyzing open-ended data. We begin with the most comprehensive approach—human coding based on a qualitative codebook—and discuss how the previously time- and resource-intensive process of coding open-ended data can be implemented at significantly lower costs by leveraging the capabilities of LLMs. We then discuss more traditional text analysis methods, such as keyness procedures that compare word frequencies across groups. The choice of method often depends on how important it is to preserve richness and capture subtleties in the data compared to the time and resource costs associated with managing a detailed coding procedure. We also highlight how to choose the best method for a given research application based on these considerations. Thereafter, we cover data analysis approaches from other social sciences. Finally, we discuss best practices for reproducibility when working with open-ended survey data.

4.1 *Human Coding*

The most comprehensive way to analyze open-ended survey data involves human coding of scripts. This requires developing a coding scheme that can be applied to the data.

4.1.1 *Creating a Coding Scheme*

There are two main approaches for creating a coding scheme from qualitative data. The first approach—inductive coding—starts with the data and creates the codes based on insights that emerge directly from the open-ended responses. The second approach—deductive coding—uses existing knowledge and theory to create a coding scheme. Whether to employ an inductive or deductive coding scheme depends on the goal of the study. The inductive approach, in which codes emerge from the data, is particularly useful for discovery and hypothesis generation. The deductive approach, in which codes might correspond to predictions from different economic theories, is better suited for hypothesis testing. It is also possible to have a coding scheme that includes both theoretically relevant codes as well as additional codes emerging from the data.

To create an inductive coding scheme, it is necessary to read through the open-ended responses to find common themes. During this process it is common practice to create codes that stay close to how respondents talk about the concepts themselves (Corbin and Strauss 2015). For instance, if respondents commonly talk about how they fear making large losses in the stock market, it is natural to include a code for "fear of making large losses" rather than frame it as "a high degree of risk aversion" (Chopra and Haaland 2023). Or if respondents express concerns that "high taxes hurt the economy," that is a more intuitive code than "efficiency costs" (Stantcheva 2021).

Depending on time, resources, and the purpose of the analysis, it might be advantageous to have two researchers independently read through responses to identify good codes and then work together on creating a set of final codes. During this process, it is common practice to combine closely related codes. The granularity of the final coding scheme depends on the research question and on how important it is to distinguish between subtle concepts in the data. After converging on a set of codes, the next step is to create a qualitative codebook that includes the final set of codes that have the desired depth and sufficient support in the data. The codebook should include example responses that illustrate how the codes should be applied. A comprehensive qualitative codebook—with both positive and negative examples of how to apply the code as well as a clear definition of the code—helps reduce disagreement between different coders and is especially important for applications with subtle distinctions. It is also key for LLM applications, as discussed in the next subsection.

4.1.2 *Manual Coding of the Data*

When the codebook is created, the next step is to hand code the data according to the codebook. This process is often done by a team of research assistants. To ensure a high quality of the manual coding of the open-ended data, we recommend implementing the following steps. First, all coders should carefully go through the codebook and be encouraged to ask clarifying questions if there are any ambiguities. Ideally, all involved coders participate in a joint training session and subsequently hand code the same data independently to ensure that everyone has a similar understanding of the coding scheme. During this process, it is common to make refinements to the coding scheme based on feedback from the coders, for example, by collapsing similar codes into a broader category or refining the definition of some codes to reduce ambiguity about how to apply the codes. Second, ensuring that coders are unaware of the research hypothesis can reduce the potential for biases in coding. In the case of hand coding open-ended data collected following an intervention, it is important that the human coders are blind to the treatment assignment. Lastly, double-coding responses and resolving discrepancies through a third coder can reduce measurement error and mitigate the effect of biases of individual coders. Alternatively, conflicts in the initial codes can also be resolved by discussions among the initial set of coders, though frequent conflicts may indicate that the codebook might need refinement or that the coders need better guidance and training.

As we discuss in Section 4.2, it is becoming increasingly common to use LLMs to code open-ended data. While human coders are still preferred in certain cases, the presence of LLMs creates a potential agency problem: Research assistants might use LLMs for classification tasks. This is especially problematic in cases where the purpose of human coding is to create a human benchmark to examine the quality of the LLM coding. Researchers who rely on research assistants to code open-ended data should be aware of this issue. To mitigate this problem, researchers should clearly explain to research assistants why using LLMs defeats the purpose of the task.

4.1.3 *Assessing Inter-Coder Reliability (ICR)*

Another advantage of double-coding is that it allows for the calculation of the inter-coder reliability (ICR), which measures the extent of overlap between different coders. On the one hand, calculating the ICR is useful when developing and refining a coding scheme. On the other hand, the ICR can be reported in the final paper as a measure of the quality of the data and coding procedure. While the ICR is a standard metric in other fields, such as sociology and anthropology, it is rarely reported in economics.

ICR requires at least two coders, with 10–25 percent of the data typically coded by multiple individuals (O'Connor and Joffe 2020). Common measures include Cohen's kappa, Krippendorff's alpha, Scott's pi, and Fleiss's K, with Krippendorff's alpha being the most flexible (O'Connor and Joffe 2020). While a simpler measure like percent agreement might be more intuitive, it has the disadvantage of not taking into account agreement by chance. ICR scores improve with fewer codes and less complex questions, but can decrease with nuanced or sophisticated topics. Miles and Huberman (1994) suggest 80 percent agreement on 95 percent of codes, and Neuendorf (2002) proposes thresholds of 90 percent (widely acceptable) and 80 percent (commonly acceptable), but the optimal cutoff depends on the application and the goal of the exercise.

The ICR can be used to iteratively improve a coding scheme as follows: The ICR is calculated after several coders have coded the same batch of open-ended responses. Subsequently, the coding scheme is refined in discussions between the coders. Then, another batch of responses is coded by all coders and the ICR is recalculated. When the ICR has reached an acceptable level, the scheme is completed and used to code the final data, in many cases based on single-coding. However, as Hruschka et al. (2004) caution, repeated ICR testing can lead to "interpretative convergence," potentially reducing validity through the suppression of ambiguity or loss of nuance. One way of reducing this risk is to predefine and document interaction protocols (Hruschka et al. 2004).

4.1.4 *Example Applications*

Several studies have used hand coding of open-ended responses to analyze unstructured text data. Andre et al. (2022) elicit respondents' reasoning when forecasting changes in unemployment and inflation in response to hypothetical macroeconomic shocks using an open-text question on the forecast survey screen. They classify the open-ended text responses into broad response type categories, such as whether responses mention considerations related to economic propagation mechanisms of the shocks or whether responses include general political or normative statements.

Andre et al. (2024) use hand coding of respondents' open-ended explanations for why inflation in the United States increased. They represent each respondent's explanation by its directed acyclic graph (DAG). Thus, their hand-coding procedure not only identifies the respondents' perceived causal drivers of the inflation rate, but also the causal connections between different variables. For example, their coding scheme allows them to differentiate between perceived root causes and intermediary causes of inflation. Representing open-text data as DAGs brings these data into a quantitative format and allows researchers to analyze open-text data using methods from graph theory and network theory.

Chinco, Hartzmark, and Sussman (2022) pursue a different approach in the context of investment choices. They let survey participants themselves classify the open-text explanations of their considerations into structured categories, significantly reducing the costs of analyzing the data. The structured categories were selected based on open-text responses in pilot studies.

4.2 *Using LLMs to Code Open-Ended Data*

While human coding of open-ended responses allows the researcher to capture rich nuances and context from the data, it is very time and resource intensive, especially for large-scale data collections. Recent evidence suggests that LLMs can code open-ended data in a reliable and reproducible manner, outperforming crowd workers and making them a low-cost and easily scalable alternative to human coding of open-ended data (Gilardi, Alizadeh, and Kubli 2023). Several recent studies in economics leveraging frontier LLMs, such as OpenAI's GPT-4o or Anthropic's Claude 3 Opus, further demonstrate that they provide very similar results to human coding in many cases (Bursztyn et al. forthcoming; Braghieri et al. 2024; Link et al. 2024; Chopra and Haaland 2023).

4.2.1 *Developing a Coding Scheme with LLMs*

To use LLMs to code responses, it is first necessary to develop a coding scheme. As the reasoning abilities of frontier LLMs continue to improve and their context windows expand to allow them to analyze more open-ended responses simultaneously, they are becoming increasingly useful in the coding scheme development phase. While it is possible to rely exclusively on LLMs in this process by giving them access to the open-ended data and asking the model to suggest a classification scheme, it is still considered best practice that researchers familiarize themselves with the nature of the collected open-ended data and develop a coding manual by reading through at least a random subset of responses, as discussed in Section 4.1.[12]

After initially developing a coding scheme by reading through responses, a good next step can be to ask an LLM to generate its own coding scheme. The researcher can then compare and contrast the LLM-generated coding scheme and the human-generated one. The LLM might have

[12] For instance, Chopra and Haaland (2023) use an LLM to create a coding scheme and categorize open-ended responses from AI-conducted interviews. While the coding scheme looked reasonable at first sight and led to some informative insights, a detailed reading of the interview transcripts led to a much richer coding scheme that better captured the richness of the data, as reported in Chopra and Haaland (2023).

identified subtle themes that humans might have missed and vice versa, making a collaboration between humans and LLMs more powerful than either of the two alone. After the LLM has suggested its own scheme, a good follow-up strategy can be to give it access to the human-derived coding scheme and ask it to create suggestions for how to consolidate the two coding schemes. In the final step, the researcher—using their best judgment and understanding from a detailed reading of the open-ended data—should decide which suggestions to implement.[13]

4.2.2 *Can LLMs Completely Replace Human Coding?*

While frontier LLMs can be a low-cost and easily scalable alternative to human coding, their performance can depend on the quality of the prompting, the type of LLM used, and the complexity of the setting (Rathje et al. 2024). Moreover, predicting the performance of an LLM across different contexts is challenging (Vafa, Rambachan, and Mullainathan 2024). It is therefore important, especially in novel or complex settings, to compare the LLM coding with human coding for a random subset of the data (Ludwig, Mullainathan, and Rambachan 2025). If LLM coding and human coding align well on this random subset, it is reasonable to rely on LLM coding for the full dataset.

To assess the degree of alignment with the human benchmark, it is standard practice to calculate F1 scores (Dell 2025), which combine *precision* (true positives divided by true positives plus false positives) and *recall* (true positives divided by true positives plus false negatives) into a single metric ranging from 0 (worst score) to 1 (best score; perfect precision and recall). In addition to calculating an aggregate F1 score for all categories in the coding scheme, calculating individual F1 scores for each category allows researchers to identify specific categories where LLMs may struggle to match human performance. These insights can inform adjustments to the prompts guiding the LLM, such as incorporating additional examples to clarify proper classifications or refining variable definitions to improve the accuracy and reliability of the categorization. Furthermore, to avoid over-fitting the prompt to the idiosyncrasies of the test data, it is good practice to refine prompts using a separate validation set and reserve the test set for final evaluation (Dell 2025).

When working iteratively to improve prompts using F1 scores to assess the quality of AI coding, it is important that the underlying human coding is of high quality, as it serves as the ground truth for classification. Ideally, the human coding should be based on double-coded data, with any discrepancies between coders discussed and resolved. While a ground truth is necessary to evaluate the quality of LLM coding, researchers should recognize that human coding can be biased, even when there is high inter-coder agreement. Furthermore, the LLM may pay attention to relevant factors that humans may overlook. For this reason, it is good practice to instruct the LLM to provide a brief explanation justifying its classification. The explanation can help determine whether the LLM made an error that calls for a refinement of the prompts or correctly classified the response, in which case the human coding protocol should be revised instead. Providing an explanation before the decision can also help enhance the quality of LLM coding through better model reasoning (Wei et al. 2022).

[13] While hallucinations—cases where the LLM "discovered" factors that are not supported by the underlying text data—used to be a common problem, it is becoming increasingly rare with recent frontier models. The researcher should still verify that all LLM suggestions incorporated in the final coding scheme are grounded in the open-ended data.

4.2.3 *Implementing LLM Coding at Scale with an API*

When using LLMs for coding, it is standard practice to leverage an API to automate and document the workflow. The most common one is the OpenAI API, which gives the user access to all of OpenAI's GPT models. Importantly, unlike interactions with ChatGPT, OpenAI API requests are not used for training data, which can be important for privacy reasons. When using the API, researchers must specify which model to use. Frontier models are typically more expensive, but might be needed to achieve acceptable performance. With very large datasets—typically much larger than in most survey research applications—frontier models might be prohibitively costly, making more cost-efficient models preferable.[14] For a detailed discussion of different state-of-the-art LLMs, their sizes, and the trade-offs involved, see the regularly updated online version of Korinek (2023).[15]

Similar to human research assistants, LLMs benefit from a comprehensive coding scheme that provides a clear definition of each category with clear examples of how the coding scheme should be applied. In addition to providing examples, it can be useful to include a justification for why the code should apply or not for a given example. How many examples to give and how detailed the justifications should be depends on the complexity of the coding scheme and how familiar the LLM is with the task at hand. For a very simple classification task, examples might not even be needed.

While it is often possible to provide an LLM with a full coding scheme and ask it to classify multiple codes simultaneously, especially with frontier models, it can sometimes improve performance to make separate API calls for each code. This could be especially useful when working with large and complex coding schemes or large text responses, such as full interview transcripts. While making separate calls is costlier and less time efficient, it allows the LLM to focus on one classification task at a time, improving accuracy and consistency (Link et al. 2024; Bursztyn et al. forthcoming; Chopra and Haaland 2023).

4.2.4 *A Practical Example*

To get started with the API, the first step involves designing a prompt that incorporates the coding scheme and provides clear instructions for the model. The prompt should include a clear task description. For instance, Bursztyn et al. (forthcoming) use the following prompt to categorize text responses about social media platform usage: "You will be supplied with a list of responses. The responses refer to the usage of different platforms, the platform will be indicated in parentheses at the end of the response. Please classify responses based on the coding scheme below. Each open-ended response can fall into multiple categories or none." The prompts also included the full coding scheme, including category names, definitions, and illustrative examples. For instance, the category FOMO includes the following description: "Respondent mentions fear of missing out [FOMO], feeling out of the loop, their wish to stay connected, or justifies usage through others' usage." It also includes the

[14] In such cases, fine-tuning—where the researcher trains the model on an existing dataset hand-coded by humans—might be necessary to achieve acceptable performance. While fine-tuning yields limited improvements for frontier models, it can enable smaller models to approach the performance levels of frontier models at much lower costs (Braghieri et al. 2024). However, fine-tuning is becoming increasingly redundant as frontier models become cheaper and more able.

[15] https://www.aeaweb.org/articles?id=10.1257/jel.20231736

following examples: "I feel compelled to keep 'in touch' with what I perceive as being the culturally relevant 'thing' at the moment. It breeds a sense of FOMO when you don't use it" and "Everyone else uses it so I feel that I will be missing out if I don't." Supplemental Appendix B provides more details with Python code from the actual research process.

4.3 *Traditional Text Analysis Methods*

While LLMs offer a scalable tool for analyzing open-ended data, traditional text analysis methods, such as dictionary-based approaches, topic modeling, keyness analysis, and machine learning classifiers remain widely used in economics, as discussed in recent reviews (Gentzkow, Kelly, and Taddy 2019; Ash and Hansen 2023; Dell 2025; Ferrario and Stantcheva 2022). As documented by Rathje et al. (2024), LLMs have the potential to outperform simple text analysis methods like dictionary-based approaches and match the performance of more advanced machine learning methods without requiring additional training data. Given their ease of use and the lack of existing training data for most open-ended survey applications, we believe LLMs will be the preferred choice over most existing text analysis methods for survey researchers. However, in cases involving large datasets where high-quality training data is available, or when the size of the text corpus makes LLMs prohibitively expensive, machine learning classifiers might be a better option (Dell 2025).

4.4 *Considerations When Choosing between Different Procedures*

Open-ended survey data can be analyzed through human coding, LLM coding, or traditional text analysis methods. The choice of method depends on factors such as the research objective and the nature of the data. In this section, we outline some key considerations to help guide the selection of the most appropriate approach for the question at hand.

4.4.1 *Desired Level of Richness*

The first question one should ask is how important it is to exploit the full richness of the data. If it is important to capture subtle nuances and context from the open-ended responses, human coding or LLM coding might be preferable. For instance, if the research question is to understand how people reason about inflation (Andre et al. 2024; Binetti, Nuzzi, and Stantcheva 2024), traditional text analysis methods are unlikely to reveal the full richness of the responses. In other applications, where the purpose might be to test a very specific mechanism, more traditional text analysis methods, such as keyness analysis or machine learning classifiers, might be sufficient to detect the relevant patterns in the data (Bursztyn et al. 2022).

4.4.2 *Scalability, Costs, Replicability, and Privacy Considerations*

When conducting surveys with relatively small samples, human coding is often a convenient and cost-effective way of analyzing the data. For large-scale surveys with potentially several thousand participants, human coding can become prohibitively costly in terms of both time and resources. While LLM coding is often a viable alternative, there are still some general drawbacks with LLM coding that might make more traditional text analysis methods preferable. First, API requests can become expensive with large sample sizes or very elaborate coding schemes with many factors. Second, while API requests are typically not used for training data, it could still be problematic from a privacy perspective to send survey data to external servers. Third, results from the use of a commercial API might not be reproducible if the model used is discontinued. However, many of these potential issues can be mitigated by using open-source LLMs, such as Meta's LLama. These models can be run locally, allowing the user to run LLM queries without having to send data externally or pay for API costs. The open-source nature of the models also allows for full replicability. However, setting up such a system locally often requires significant computational resources and technical expertise (Dell 2025).

4.5 *Insights from Other Social Sciences*

Economists often approach qualitative data analysis with goals distinct from those of researchers in fields like sociology and anthropology. In this section, we discuss these differences and highlight how economists might benefit from adopting techniques commonly used in other disciplines.

Anthropologists and sociologists use tools such as Dedoose and NVivo to code and interpret qualitative data. Their capabilities—hierarchical coding, pattern visualization, and multimodal data handling—are distinct from typical economic software. Moreover, Dedoose's collaborative features support team-based coding, reflecting anthropological practices where multiple analysts bring diverse perspectives. One potential limitation of these tools is that they might be less suited to analyzing datasets with a large number of observations. Yet, the integration of AI tools might mitigate these scale limitations going forward.

Anthropologists and sociologists often present their findings through detailed narrative accounts designed to immerse readers in the studied environment. These accounts incorporate direct quotes, anecdotes, and personal observations to vividly convey participants' perspectives (Bernard 2018). Rather than prioritizing contextual depth, economic research emphasizes the identification and measurement of variables and patterns (Bardhan and Ray 2006; Starr 2014). A summary of these methodological differences across disciplines in the analysis of unstructured data is provided in Table 3. Incorporating some anthropological techniques, such as collaborative coding and in-depth contextual analysis, could offer an opportunity for economists to interpret qualitative data in a richer and more nuanced way.

4.6 *Reproducibility*

Concerns about low reproducibility of research results are paramount in economics and the social sciences more broadly (Christensen and Miguel 2018; Camerer et al. 2016). Open-ended data poses several new challenges for reproducibility, given large degrees of freedom in the analysis of such data. In particular, researchers have considerable degrees of freedom in devising coding schemes, which can be especially problematic for studies concerned with hypothesis testing.

4.6.1 *Documentation of Coding Schemes*

One potential way to mitigate concerns about researcher degrees of freedom is a transparent, standardized, and detailed documentation of coding schemes. A comprehensive codebook that includes definitions and examples is crucial for ensuring that codes can be independently

TABLE 3
Comparing Qualitative Approaches in Anthropology and Sociology versus Economics

Fields	Anthropology & Sociology	Economics
Data collection strategies	Focus groups, interviews, participant observation (Bernard 2018); could be unstructured and driven by participants (Ashwin et al. 2022)	Surveys and textual analysis; generally top-down approaches (Rao 2023)
Sampling	Purposive; typically smaller samples focusing on specific populations for in-depth insights (Seawright and Gerring 2008; Starr 2014)	Random; large samples for statistical power
Objective of data collection	Achieve saturation: collect data until no new themes emerge (Small 2009)	Establish representativeness: aim for generalizable and replicable trends (Rao 2023)
Role of researcher	Reflexive and often involved in data collection process; emphasizes the process and interpretation (Burawoy 1998)	Objective; typically detached from their research subjects and focused on analysis (Rao 2023)
Potential bias	More susceptible to researcher's influence; smaller sample sizes and thus not generalizable (Small 2011; Starr 2014)	Emphasizes generalizable relationships, abstracting from local context (Bardhan and Ray 2006)
Analysis and inference	Emphasizes contextual complexity; seeks hidden meanings and patterns, richer descriptions of social phenomena (Geertz 1973)	Focuses on variables and patterns that can be measured (Bardhan and Ray 2006; Starr 2014); often uses LLMs for analysis due to large samples (Ashwin, Chhabra, and Rao 2025)

understood and applied by other researchers. Decision logs that provide a record of the rationale behind specific coding choices—especially ambiguous cases—can also be useful, at least for internal purposes. Such documentation not only makes the coding transparent but also provides a basis for subsequent adjustments in the interpretation. Furthermore, in studies using LLMs for coding of open-ended responses, including the full prompts used for coding in an online appendix is good practice to encourage replicability and transparency around the results.

4.6.2 *Preregistration of Coding Schemes and LLM Prompts*

Preregistering coding schemes can mitigate concerns about researcher flexibility by specifying categories, definitions, and coding rules in advance. Preregistration reduces post-hoc adjustments and enhances the credibility of findings but is mainly an option for studies concerned with hypothesis testing. For instance, a study testing whether a priming intervention successfully changes attention to a topic, as measured with an open-ended response, could benefit from a preregistration. The preregistration could include a coding scheme and the LLM prompt used to classify responses. However, a full preregistration is not feasible in many applications where the focus is at least partly on discovery. For studies focused on hypothesis generation, a preregistration could mainly be used to specify the coding procedures rather than the coding scheme.

4.6.3 *Data Anonymization*

Whether to publish the raw data alongside codes depends on the nature of the data and its potential sensitivity. In nonpersonal applications, such as expectations about macroeconomic variables without identifiable or sensitive information, sharing raw text data is often both feasible and desirable. However, when it comes to open-ended data describing personal experiences, publishing the raw data may not be appropriate due to the risk of reidentification, even after anonymization. In such cases, researchers might instead provide detailed metadata, summaries, or illustrative examples that maintain the essence of the findings without exposing the original text. Sharing the analysis codes alone can still promote transparency and reproducibility by allowing others to understand and validate the methods. For sensitive qualitative datasets, access could be restricted through secure data repositories, requiring approval or agreements that prioritize ethical considerations. This approach balances the need for scientific openness with the responsibility to protect participants' privacy.

5. *Conclusion and Avenues for Future Research*

This review provides an overview of how open-ended questions can be used to uncover mechanisms behind economic beliefs and behaviors. Given their wide applicability and advantages, we believe that open-ended survey data will continue to grow in popularity. For instance, the quest to better understand the foundations of belief formation and decision-making will likely spur more widespread use of these methods in economics. We conclude this review with a discussion of avenues for future research in the context of open-ended survey data.

5.1 *New Opportunities through LLMs*

The availability of LLMs, such as OpenAI's GPT-4 or Anthropic's Claude, provides new opportunities and dramatically reduces costs in the collection and analysis of open-ended data: LLMs can be used to conduct qualitative interviews at scale; LLMs can improve the analysis of open-ended data by better capturing the context, semantics, and sentiment of responses than existing tools; LLMs can be used to automatize the classification of open-ended data; and they offer a systematic, data-driven approach to generating (initial) classification schemes. We hope that this review lowers the barriers for researchers and practitioners who would like to make use of open-ended survey data.

5.2 *Incentives*

As discussed in our review, one concern about open-ended questions is that respondents may exert low effort when writing their responses. Another concern is that effort levels may vary widely across respondents depending on their intrinsic motivations, which reduce interpersonal comparability.

A potential solution to these problems is the use of monetary incentives. Incentivizing open-ended responses is a complex challenge because there is typically no objective benchmark against which to evaluate them. More generally, unlike closed questions, where specific answers can be defined as correct or desirable, open-ended responses vary significantly in content, depth, and style, making it difficult to assess quality or value in a consistent, objective way. Without clear benchmarks, it is challenging to design incentive structures that reliably encourage respondents to provide thoughtful, meaningful answers rather than simply writing more or producing responses that may seem impressive but lack genuine insight. This lack of objective standards leaves open the question of how to best motivate quality responses in a way that aligns with research goals. It is conceivable, for example, that providing monetary incentives for more effortful responses distorts responses away from the true reasoning processes guiding behavior.

More generally, it remains an open question whether and how incentives can be used to increase the truthfulness of open-ended responses. While prior work emphasizes the importance of motivational prompts (Smyth et al. 2009), future research should examine whether incentives can be designed to enhance the truthfulness and accuracy of open-ended responses.

5.3 *Voice-Based AI Interviews*

An emerging innovation in the field of interviews is the use of AI-driven systems to conduct semi-structured, voice-based interviews with respondents. In this setup, an AI interviewer interacts directly with respondents through natural conversation, leveraging voice technology to capture not only the content of their answers but also vocal nuances such as tone, pace, and emotion. Voice-based AI interviews can enable richer interactions by probing deeper into areas of interest based on both spoken words and vocal cues. By combining the conversational flexibility of AI with the expressive power of voice, this method opens new possibilities for understanding respondent sentiment and intent, making it particularly valuable for exploring complex topics in a scalable yet empathetic way. Eventually, adding videos to AI-conducted interviews can further increase their potential by allowing the AI interviewer to also pick up on facial expressions.

5.4 *Neuroeconomics*

Combining methods from neuroeconomics with unstructured, open-ended responses presents a unique opportunity to deepen our understanding of belief formation, emotions, and considerations in decision-making contexts. Neuroeconomics provides tools to measure neural and physiological responses—such as brain imaging (fMRI or EEG), eye tracking, and skin conductance (Camerer, Loewenstein, and Prelec 2005)—capturing the immediate, often nonconscious, reactions that might accompany or precede verbalized considerations. Meanwhile, open-ended responses offer an introspective and narrative perspective, revealing subjective interpretations, emotional nuances, and the complexity of considerations that might not be apparent in more structured data. By integrating these approaches, researchers could measure both the internal processing and the explicit expression of considerations. For example, neuroeconomic methods might reveal brain areas activated during moments of conflict or ambivalence, which could then be connected to participants' own descriptions of doubt or conflicting considerations in their responses. This could illuminate how certain neural patterns correlate with specific ways of interpreting experiences or reasoning about decisions.

References

Abeler, Johannes, David B. Huffmann, and Collin Raymond. 2023. "Incentive Complexity, Bounded Rationality and Effort Provision." IZA Discussion Paper 16284.

Agranov, Marina, and Pietro Ortoleva. 2017. "Stochastic Choice and Preferences for Randomization." *Journal of*

Political Economy 125 (1): 40–68.
Akram, Arslan. 2023. "An Empirical Study of AI Generated Text Detection Tools." Preprint, arXiv. https://doi.org/10.48550/arXiv.2310.01423.
Alesina, Alberto, Armando Miano, and Stefanie Stantcheva. 2023. "Immigration and Redistribution." *Review of Economic Studies* 90 (1): 1–39.
Algan, Yann, Eva Davoine, Thomas Renault, and Stefanie Stantcheva. 2025. "Emotions and Policy Views." Unpublished.
Alsan, Marcella, Joshua Schwartzstein, and Stefanie Stantcheva. 2025. "The Universal Pursuit of Safety and the Demand for (Lethal, Non-Lethal or No) Guns." Unpublished.
An, Zidong, Carola Binder, and Xuguang Simon Sheng. 2023. "Gas Price Expectations of Chinese Households." *Energy Economics* 120: 106622.
Andre, Peter. 2025. "Shallow Meritocracy." *Review of Economic Studies* 92 (2): 772–807.
Andre, Peter, Ingar Haaland, Christopher Roth, Mirko Wiederholt, and Johannes Wohlfart. 2024. "Narratives about the Macroeconomy." SAFE Working Paper 426.
Andre, Peter, Carlo Pizzinelli, Christopher Roth, and Johannes Wohlfart. 2022. "Subjective Models of the Macroeconomy: Evidence from Experts and Representative Samples." *Review of Economic Studies* 89 (6): 2958–91.
Andre, Peter, Philipp Schirmer, and Johannes Wohlfart. 2024. "Mental Models of the Stock Market." CESifo Working Paper 10691.
Arrieta, Gonzalo, and Kirby Nielsen. 2024. "Procedural Decision-Making in the Face of Complexity." Unpublished.
Ash, Elliott, and Stephen Hansen. 2023. "Text Algorithms in Economics." *Annual Review of Economics* 15: 659–88.
Ashwin, Julian, Aditya Chhabra, and Vijayendra Rao. 2025. "Using Large Language Models for Qualitative Analysis Can Introduce Serious Bias." *Sociological Methods and Research*. https://doi.org/10.1177/00491241251338246.
Ashwin, Julian, Vijayendra Rao, Monica Biradavolu, Aditya Chhabra, Arshia Haque, Afsana Khan, and Nandini Krishnan. 2022. "A Method to Scale-Up Interpretative Qualitative Analysis, with an Application to Aspirations in Cox's Bazaar, Bangladesh." World Bank Policy Research Working Paper 10046.
Ayyar, Sreevidya, Uta Bolt, Eric French, and Cormac O'Dea. 2024. "Imagine Your Life at 25: Children's Gender Attitudes and Later-Life Outcomes." NBER Working Paper 32789.
Ba, Bocar A., Roman Rivera, and Alexander Whitefield. 2023. "Forecasting the Impact of Racial Uprisings, Market versus Stakeholders' Expectations." NBER Working Paper 31857.
Bailey, Michael, Eduardo Dávila, Theresa Kuchler, and Johannes Stroebel. 2019. "House Price Beliefs and Mortgage Leverage Choice." *Review of Economic Studies* 86 (6): 2403–52.
Baird, Sarah, Craig McIntosh, and Berk Özler. 2011. "Cash or Condition? Evidence from a Cash Transfer Experiment." *Quarterly Journal of Economics* 126 (4): 1709–53.
Bardhan, Pranab, and Isha Ray. 2006. "Methodological Approaches to the Question of the Commons." *Economic Development and Cultural Change* 54 (3): 655–76.
Bauer, Rob, Katrin Gödker, Paul Smeets, and Florian Zimmermann. 2024. "Mental Models in Financial Markets: How Do Experts Reason about the Pricing of Climate Risk?" IZA Discussion Paper 17030.
Becker, Howard S. 1998. *Tricks of the Trade: How to Think about Your Research While You're Doing It*. University of Chicago Press.
Berger, Christopher C., Tara C. Dennehy, John A. Bargh, and Ezequiel Morsella. 2016. "Nisbett and Wilson (1977) Revisited: The Little That We Can Know and Can Tell." *Social Cognition* 34 (3): 167–95.
Bergman, Peter, Raj Chetty, Stefanie DeLuca, Nathaniel Hendren, Lawrence F. Katz, and Christopher Palmer. 2024. "Creating Moves to Opportunity: Experimental Evidence on Barriers to Neighborhood Choice." *American Economic Review* 114 (5): 1281–337.
Bernard, H. Russell. 2018. *Research Methods in Anthropology: Qualitative and Quantitative Approaches*. Rowman and Littlefield.
Bewley, Truman F. 1995. "A Depressed Labor Market as Explained by Participants." *American Economic Review* 85 (2): 250–54.
Bewley, Truman F. 1999. *Why Wages Don't Fall during a Recession*. Harvard University Press.
Bewley, Truman. 2002. "Interviews as a Valid Empirical Tool in Economics." *Journal of Socio-Economics* 31 (4): 343–53.
Binetti, Alberti, Francesco Nuzzi, and Stefanie Stantcheva. 2024. "People's Understanding of Inflation." *Journal of Monetary Economics* 148: 103652.
Blinder, Alan S., Elie R. D. Canetti, David E. Lebow, and Jeremy B. Rudd. 1998. *Asking about Prices: A New Approach to Understanding Price Stickiness*. Russell Sage Foundation.

Bordalo, Pedro, Giovanni Burro, Katherine Coffman, Nicola Gennaioli, and Andrei Shleifer. 2024. "Imagining the Future: Memory, Simulation, and Beliefs." *Review of Economic Studies* 92 (3): 1532–63.
Bordalo, Pedro, John J. Conlon, Nicola Gennaioli, Spencer Y. Kwon, and Andrei Shleifer. 2023. "Memory and Probability." *Quarterly Journal of Economics* 138 (1): 265–311.
Bordalo, Pedro, John Conlon, Nicola Gennaioli, Spencer Kwon, and Andrei Shleifer. 2025. "How People Use Statistics." *Review of Economic Studies*. https://doi.org/10.1093/restud/rdaf022.
Bordalo, Pedro, Nicola Gennaioli, and Andrei Shleifer. 2022. "Salience." *Annual Review of Economics* 14: 521–44.
Bourdieu, Pierre. 1990. *The Logic of Practice*. Translated by Richard Nice. Stanford University Press.
Braghieri, Luca, Sarah Eichmeyer, Ro'ee Levy, Markus Mobius, Jacob Steinhardt, and Ruiqi Zhong. 2024. "Article-Level Slant and Polarization of News Consumption on Social Media." CEPR Discussion Paper 19807.
Braghieri, Luca, Peter Schwardmann, and Egon Tripodi. 2024. "Talking across the Aisle." CEPR Discussion Paper 19753.
Breyer, Magdalena, Tabea Palmtag, and Delia Zollinger. 2023. "Narratives of Backlash? Perceptions of Changing Status Hierarchies in Open-Ended Survey Responses." URPP Equality of Opportunity Discussion Paper Series 15.
Brosius, Anna, Michael Hameleers, and Toni G. L. A. van der Meer. 2022. "Can We Trust Measures of Trust? A Comparison of Results from Open and Closed Questions." *Quality and Quantity* 56 (5): 2907–24.
Bruner, Jerome. 1991. "The Narrative Construction of Reality." *Critical Inquiry* 18 (1): 1–21.
Burawoy, Michael. 1998. "The Extended Case Method." *Sociological Theory* 16 (1): 4–33.
Burgstaller, Lilith, Annabelle Doerr, and Sarah Necker. 2023. "Do Household Tax Credits Increase the Demand for Legally Provided Services?" CESifo Working Paper 10211.
Bursztyn, Leonardo, Georgy Egorov, Ingar Haaland, Aakaash Rao, and Christopher Roth. 2022. "Scapegoating during Crises." *AEA Papers and Proceedings* 112: 151–55.
Bursztyn, Leonardo, Georgy Egorov, Ingar Haaland, Aakaash Rao, and Christopher Roth. 2023. "Justifying Dissent." *Quarterly Journal of Economics* 138 (3): 1403–51.
Bursztyn, Leonardo, Ingar Haaland, Aakaash Rao, and Christopher P. Roth. 2020. "Disguising Prejudice: Popular Rationales as Excuses for Intolerant Expression." NBER Working Paper 27288.
Bursztyn, Leonardo, Benjamin R. Handel, Rafael Jimenez, and Christopher Roth. Forthcoming. "When Product Markets Become Collective Traps: The Case of Social Media." *American Economic Review*.
Bustos, Sebastián, Dina Pomeranz, Juan Carlos Suárez Serrato, José Vila-Belda, and Gabriel Zucman. 2022. "The Race between Tax Enforcement and Tax Planning: Evidence from a Natural Experiment in Chile." NBER Working Paper 30114.
Camerer, Colin F., Anna Dreber, Eskil Forsell, Teck-Hua Ho, Jürgen Huber, Magnus Johannesson, Michael Kirchler, et al. 2016. "Evaluating Replicability of Laboratory Experiments in Economics." *Science* 351 (6280): 1433–36.
Camerer, Colin, George Loewenstein, and Drazen Prelec. 2005. "Neuroeconomics: How Neuroscience Can Inform Economics." *Journal of Economic Literature* 43 (1): 9–64.
Capozza, Francesco. 2024. "Beliefs about the Gender Gap in Salary Negotiation." CESifo Working Paper 11228.
Cappelen, Alexander W., Ranveig Falch, and Bertil Tungodden. 2025. "Experimental Evidence on the Acceptance of Males Falling Behind." *Journal of the European Economic Association*. https://doi.org/10.1093/jeea/jvaf016.
Casarico, Alessandra, Jana Schuetz, and Silke Uebelmesser. 2024. "Gender Inequality over the Life Cycle, Information Provision and Policy Preferences." CESifo Working Paper 10916.
Case, Karl E., and Robert J. Shiller. 2003. "Is There a Bubble in the Housing Market?" *Brookings Papers on Economic Activity* 34 (2): 299–362.
Case, Karl E., Robert J. Shiller, and Anne Thompson. 2012. "What Have They Been Thinking? Home Buyer Behavior in Hot and Cold Markets." *Brookings Papers on Economic Activity* 43 (2): 265–315.
Castagnetti, Alessandro, and Renke Schmacker. 2022. "Protecting the Ego: Motivated Information Selection and Updating." *European Economic Review* 142: 104007.
Charness, Gary, and Martin Dufwenberg. 2006. "Promises and Partnership." *Econometrica* 74 (6): 1579–601.
Chinco, Alex, Samuel M. Hartzmark, and Abigail B. Sussman. 2022. "A New Test of Risk Factor Relevance." *Journal of Finance* 77 (4): 2183–238.
Chopra, Felix, and Ingar Haaland. 2023. "Conducting Qualitative Interviews with AI." CESifo Working Paper 10666.
Chopra, Felix, Ingar Haaland, and Christopher Roth. 2024. "The Demand for News: Accuracy Concerns versus Belief Confirmation Motives." *Economic Journal* 134 (661): 1806–34.
Chopra, Felix, Christopher Roth, and Johannes Wohlfart. 2025. "Home Price Expectations and Spending: Evidence From a Field Experiment." *American Economic Review* 115 (7): 2267–305.
Christensen, Garret, and Edward Miguel. 2018. "Transparency, Reproducibility, and the Credibility of Economics

Research." *Journal of Economic Literature* 56 (3): 920–80.
Cohn, Alain, and Michel André Maréchal. 2016. "Priming in Economics." *Current Opinion in Psychology* 12: 17–21.
Colarieti, Roberto, Pierfrancesco Mei, and Stefanie Stantcheva. 2024. "The How and Why of Household Reactions to Income Shocks." NBER Working Paper 32191.
Conlon, John J. 2024. "Memory Rehearsal and Belief Biases." Unpublished.
Connor Desai, Saoirse, and Stian Reimers. 2019. "Comparing the Use of Open and Closed Questions for Web-Based Measures of the Continued-Influence Effect." *Behavior Research Methods* 51 (3): 1426–40.
Corbin, Juliet, and Anselm Strauss. 2015. *Basics of Qualitative Research: Techniques and Procedures for Developing Grounded Theory*. 4th ed. SAGE Publications, Inc.
Dechezleprêtre, Antoine, Adrien Fabre, Tobias Kruse, Bluebery Planterose, Ana Sanchez Chico, and Stefanie Stantcheva. 2025. "Fighting Climate Change: International Attitudes toward Climate Policies." *American Economic Review* 115 (4): 1258–300.
Dell, Melissa. 2025. "Deep Learning for Economists." *Journal of Economic Literature* 63 (1): 5–58.
Denzin, Norman K., and Yvonna S. Lincoln. 2017. *The SAGE Handbook of Qualitative Research*. 5th ed. SAGE Publications, Inc.
de Quidt, Jonathan, Johannes Haushofer, and Christopher Roth. 2018. "Measuring and Bounding Experimenter Demand." *American Economic Review* 108 (11): 3266–302.
Dillman, Don A. 2007. *Mail and Internet Surveys: The Tailored Design Method*. John Wiley and Sons.
Dillman, Don A., Jolene D. Smyth, and Leah Melani Christian. 2014. *Internet, Phone, Mail, and Mixed-Mode Surveys: The Tailored Design Method*. John Wiley and Sons.
Dillon, Andrew, Elena Bardasi, Kathleen Beegle, and Pieter Serneels. 2012. "Explaining Variation in Child Labor Statistics." *Journal of Development Economics* 98 (1): 136–47.
Dube, Arindrajit, Suresh Naidu, and Adam D. Reich. 2022. "Power and Dignity in the Low-Wage Labor Market: Theory and Evidence from Wal-Mart Workers." NBER Working Paper 30441.
Dunn, Amina, and Vianney Gómez. 2023. "Nonresponse Rates on Open-Ended Survey Questions Vary by Demographic Group, Other Factors." Pew Research Center, March 7. https://www.pewresearch.org/decoded/2023/03/07/nonresponse-rates-on-open-ended-survey-questions-vary-by-demographic-group-other-factors/.
Duraj, Kamila, Daniela Grunow, Michael Haliassos, Christine Laudenbach, and Stephan Siegel. 2024. "Rethinking the Stock Market Participation Puzzle: A Qualitative Approach." IMFS Working Paper 210.
Dutta, Subhadra, and Eric M. O'Rourke. 2020. "Open-Ended Questions: The Role of Natural Language Processing and Text Analytics." In *Employee Surveys and Sensing: Challenges and Opportunities*, edited by William H. Macey and Alexis A. Fink, 202–18. Oxford University Press.
Dutz, Deniz, Ingrid Huitfeldt, Santiago Lacouture, Magne Mogstad, Alexander Torgovitsky, and Winnie van Dijk. 2022. "Selection in Surveys: Using Randomized Incentives to Detect and Account for Nonresponse Bias." NBER Working Paper 29549.
Elias, Julio J., Nicola Lacetera, and Mario Macis. 2023. "Is the Price Right? The Role of Morals, Ideology, and Tradeoff Thinking in Explaining Reactions to Price Surges." NBER Working Paper 29963.
Emerson, Robert M., Rachel I. Fretz, and Linda L. Shaw. 1995. *Writing Ethnographic Fieldnotes*. University of Chicago Press.
Enke, Benjamin. 2024. "The Cognitive Turn in Behavioral Economics." Unpublished.
Ericsson, K. Anders, and Herbert A. Simon. 1980. "Verbal Reports as Data." *Psychological Review* 87 (3): 215–51.
Ericsson, K. Anders, and Herbert A. Simon. 1993. *Protocol Analysis: Verbal Reports as Data*. MIT Press.
Erkal, Nisvan, Lata Gangadharan, and Erte Xiao. 2022. "Leadership Selection: Can Changing the Default Break the Glass Ceiling?" *Leadership Quarterly* 33 (2): 101563.
Ferrario, Beatrice, and Stefanie Stantcheva. 2022. "Eliciting People's First-Order Concerns: Text Analysis of Open-Ended Survey Questions." *AEA Papers and Proceedings* 112: 163–69.
Filippini, Massimo, Markus Leippold, and Tobias Wekhof. 2024. "Sustainable Finance Literacy and the Determinants of Sustainable Investing." *Journal of Banking and Finance* 163: 107167.
Fowler, Floyd J., Jr. 1995. *Improving Survey Questions: Design and Evaluation*. SAGE Publications, Inc.
Fuster, Andreas, and Basit Zafar. 2023. "Survey Experiments on Economic Expectations." In *Handbook of Economic Expectations*, edited by Rüdiger Bachmann, Giorgio Topa, and Wilbert van der Klaauw, 107–30. Academic Press.
Gabaix, Xavier. 2019. "Behavioral Inattention." In *Handbook of Behavioral Economics: Applications and Foundations 1*, Vol. 2, edited by B. Douglas Bernheim, Stefano DellaVigna, and David Laibson, 261–343. North-Holland.
Galasso, Vincenzo, Massimo Morelli, Tommaso Nannicini, and Piero Stanig. 2024. "The Populist Dynamic: Experimental Evidence on the Effects of Countering Populism." CESifo Working Paper 10949.

Galasso, Vincenzo, Tommaso Nannicini, and Debora Nozza. 2024. "We Need to Talk: Audio Surveys and Information Extraction." CEPR Discussion Paper 19749.
Galesic, Mirta, and Michael Bosnjak. 2009. "Effects of Questionnaire Length on Participation and Indicators of Response Quality in a Web Survey." *Public Opinion Quarterly* 73 (2): 349–60.
Geer, John G. 1988. "What Do Open-Ended Questions Measure?" *Public Opinion Quarterly* 52 (3): 365–67.
Geer, John G. 1991. "Do Open-Ended Questions Measure 'Salient' Issues?" *Public Opinion Quarterly* 55 (3): 360–70.
Geertz, Clifford. 1973. *The Interpretation of Cultures*. Basic Books.
Geiecke, Friedrich, and Xavier Jaravel. 2024. "Conversations at Scale: Robust AI-Led Interviews with a Simple Open-Source Platform." CEPR Discussion Paper 19705.
Gentzkow, Matthew, Bryan Kelly, and Matt Taddy. 2019. "Text as Data." *Journal of Economic Literature* 57 (3): 535–74.
Gibson, Kerry. 2022. "Bridging the Digital Divide: Reflections on Using WhatsApp Instant Messenger Interviews in Youth Research." *Qualitative Research in Psychology* 19 (3): 611–31.
Gilardi, Fabrizio, Meysam Alizadeh, and Maël Kubli. 2023. "ChatGPT Outperforms Crowd Workers for Text-Annotation Tasks." *Proceedings of the National Academy of Sciences* 120 (30): e2305016120.
Graeber, Thomas, Shakked Noy, and Christopher Roth. 2024. "Lost in Transmission." CESifo Working Paper 10903.
Graeber, Thomas, Christopher Roth, and Constantin Schesch. 2024. "Explanations." CESifo Working Paper 11131.
Graeber, Thomas, Christopher Roth, and Florian Zimmermann. 2024. "Stories, Statistics, and Memory." *Quarterly Journal of Economics* 139 (4): 2181–225.
Gründler, Klaus, and Niklas Potrafke. 2020. "Experts and Epidemics." CESifo Working Paper 8556.
Grunewald, Andreas, Victor Klockmann, Alicia von Schenk, and Ferdinand von Siemens. 2024. "Are Biases Contagious? The Influence of Communication on Motivated Beliefs." Würzburg Economic Paper 109.
Haaland, Ingar, Christopher Roth, Stefanie Stantcheva, and Johannes Wohlfart. 2024. "Understanding Economic Behavior Using Open-Ended Survey Data." NBER Working Paper 32421.
Haaland, Ingar, Christopher Roth, and Johannes Wohlfart. 2023. "Designing Information Provision Experiments." *Journal of Economic Literature* 61 (1): 3–40.
Hager, Anselm, Lukas Hensel, Johannes Hermle, and Christopher Roth. 2023. "Political Activists as Free Riders: Evidence from a Natural Field Experiment." *Economic Journal* 133 (653): 2068–84.
Hager, Anselm, Lukas Hensel, Christopher Roth, and Andreas Stegmann. 2025. "Voice and Political Engagement: Evidence from a Field Experiment." *Review of Economics and Statistics* 107 (4): 1149–58.
Handel, Benjamin, and Joshua Schwartzstein. 2018. "Frictions or Mental Gaps: What's Behind the Information We (Don't) Use and When Do We Care?" *Journal of Economic Perspectives* 32 (1): 155–78.
Hanna, Rema, Sendhil Mullainathan, and Joshua Schwartzstein. 2014. "Learning through Noticing: Theory and Evidence from a Field Experiment." *Quarterly Journal of Economics* 129 (3): 1311–53.
Hansen, Karolina, and Aleksandra Świderska. 2023. "Integrating Open- and Closed-Ended Questions on Attitudes towards Outgroups with Different Methods of Text Analysis." *Behavior Research Methods* 56 (5): 4802–22.
Henkel, Luca, Christoph Oslislo, and Frederik Schwerter. 2024. "Priming, Interference and Economic Choice." Unpublished.
Himelein, Kristen. 2016. "Interviewer Effects in Subjective Survey Questions: Evidence from Timor-Leste." *International Journal of Public Opinion Research* 28 (4): 511–33.
Ho, Karen. 2009. "Disciplining Investment Bankers, Disciplining the Economy: Wall Street's Institutional Culture of Crisis and the Downsizing of 'Corporate America.'" *American Anthropologist* 111 (2): 177–89.
Höhne, Jan Karem, Joshua Claassen, Saijal Shahania, and David Broneske. 2024. "Bots in Web Survey Interviews: A Showcase." *International Journal of Market Research* 67 (1): 3–12.
Holland, Dorothy, and Naomi Quinn. 1987. *Cultural Models in Language and Thought*. Cambridge University Press.
Hommes, Cars, Julien Pinter, and Isabelle Salle. 2023. "What People Believe about Monetary Finance and What We Can('t) Do about It: Evidence from a Large-Scale, Multi-country Survey Experiment." CESifo Working Paper 10574.
Houser, Daniel, and Erte Xiao. 2011. "Classification of Natural Language Messages Using a Coordination Game." *Experimental Economics* 14 (1): 1–14.
Hruschka, Daniel J., Deborah Schwartz, Daphne Cobb St John, Erin Picone-Decaro, Richard A. Jenkins, and James W. Carey. 2004. "Reliability in Coding Open-Ended Data: Lessons Learned from HIV Behavioral Research." *Field Methods* 16 (3): 307–31.

Hüning, Hendrik, Lydia Mechtenberg, and Stephanie W. Wang. 2022. "Exiting the Echo Chamber: Can Discussions in Randomly Formed Groups Change Opinions and Votes?" Unpublished.

Imas, Alex, Michael A. Kuhn, and Vera Mironova. 2022. "Waiting to Choose: The Role of Deliberation in Intertemporal Choice." *American Economic Journal: Microeconomics* 14 (3): 414–40.

Israel, Glenn D. 2010. "Effects of Answer Space Size on Responses to Open-Ended Questions in Mail Surveys." *Journal of Official Statistics* 26 (2): 271–85.

Jabarian, Brian, and Elia Sartori. 2024. "Critical Thinking and Storytelling Contexts." CESifo Working Paper 11282.

Jayachandran, Seema, Monica Biradavolu, and Jan Cooper. 2023. "Using Machine Learning and Qualitative Interviews to Design a Five-Question Survey Module for Women's Agency." *World Development* 161: 106076.

Jerolmack, Colin, and Shamus Khan. 2014. "Talk Is Cheap: Ethnography and the Attitudinal Fallacy." *Sociological Methods and Research* 43 (2): 178–209.

Jessen, Lasse J., Sebastian Köhne, Patrick Nüß, and Jens Ruhose. 2024. "Socioeconomic Inequality in Life Expectancy: Perception and Policy Demand." CESifo Working Paper 10940.

Jiang, Zhengyang, Hongqi Liu, Cameron Peng, and Hongjun Yan. 2025. "Investor Memory and Biased Beliefs: Evidence from the Field." *Quarterly Journal of Economics* 140 (4): 2749–804.

Kaufmann, Marc, Peter Andre, and Botond Kőszegi. 2024. "Understanding Markets with Socially Responsible Consumers." *Quarterly Journal of Economics* 139 (3): 1989–2035.

Kaur, Supreet, Sendhil Mullainathan, Suanna Oh, and Frank Schilbach. 2025. "Do Financial Concerns Make Workers Less Productive?" *Quarterly Journal of Economics* 140 (1): 635–89.

Kazdin, Alan E. 2022. *Research Design in Clinical Psychology*. 5th ed. Cambridge University Press.

Kelley, Stanley, Jr. 1983. *Interpreting Elections*. Princeton University Press.

Knott, Eleanor, Aliya Hamid Rao, Kate Summers, and Chana Teeger. 2022. "Interviews in the Social Sciences." *Nature Reviews Methods Primers* 2 (1): 73.

König, Tobias, and Renke Schmacker. 2022. "Preferences for Sin Taxes." CESifo Working Paper 10046.

Korinek, Anton. 2023. "Generative AI for Economic Research: Use Cases and Implications for Economists." *Journal of Economic Literature* 61 (4): 1281–317.

Krosnick, Jon A. 1999. "Survey Research." *Annual Review of Psychology* 50: 537–67.

Krosnick, Jon A. 2018. "Questionnaire Design." In *The Palgrave Handbook of Survey Research*, edited by David L. Vannette and Jon A. Krosnick, 439–55. Palgrave Macmillan.

Kvale, Steinar. 1996. *InterViews: An Introduction to Qualitative Research Interviewing*. SAGE Publications, Inc.

Kvale, Steinar, and Svend Brinkmann. 2007. *InterViews: Learning the Craft of Qualitative Research Interviewing*. 2nd ed. Sage Publications, Inc.

Lang, Valentin, and Stephan A. Schneider. 2023. "Immigration and Nationalism in the Long Run." CESifo Working Paper 10621.

Laudenbach, Christine, Annika Weber, Rüdiger Weber, and Johannes Wohlfart. 2024. "Beliefs about the Stock Market and Investment Choices: Evidence from a Field Experiment." *Review of Financial Studies*. https://doi.org/10.1093/rfs/hhae063.

Lazarsfeld, Paul F. 1944. "The Controversy over Detailed Interviews—An Offer for Negotiation." *Public Opinion Quarterly* 8 (1): 38–60.

Leiser, David, and Shelly Drori. 2005. "Naïve Understanding of Inflation." *Journal of Socio-Economics* 34 (2): 179–98.

Levitt, Steven D., and Sudhir Alladi Venkatesh. 2000. "An Economic Analysis of a Drug-Selling Gang's Finances." *Quarterly Journal of Economics* 115 (3): 755–89.

Link, Sebastian, Andreas Peichl, Christopher Roth, and Johannes Wohlfart. 2024. "Attention to the Macroeconomy." CESifo Working Paper 10858.

Liscow, Zachary, and Edward Fox. 2022. "The Psychology of Taxing Capital Income: Evidence from a Survey Experiment on the Realization Rule." *Journal of Public Economics* 213: 104714.

Loewenstein, George, and Zachary Wojtowicz. 2025. "The Economics of Attention." *Journal of Economic Literature* 63 (3): 1038–89.

Ludwig, Jens, Sendhil Mullainathan, and Ashesh Rambachan. 2025. "Large Language Models: An Applied Econometric Framework." NBER Working Paper 33344.

Luttmer, Erzo F. P., and Andrew A. Samwick. 2018. "The Welfare Cost of Perceived Policy Uncertainty: Evidence from Social Security." *American Economic Review* 108 (2): 275–307.

Machlup, Fritz. 1946. "Marginal Analysis and Empirical Research." *American Economic Review* 36 (4): 519–54.

Martínez-Marquina, Alejandro, Muriel Niederle, and Emanuel Vespa. 2019. "Failures in Contingent Reasoning:

The Role of Uncertainty." *American Economic Review* 109 (10): 3437–74.
Mavletova, Aigul. 2013. "Data Quality in PC and Mobile Web Surveys." *Social Science Computer Review* 31 (6): 725–43.
Miano, Armando. 2023. "Search Costs, Outside Options, and On-the-Job Search." Unpublished.
Miles, Matthew B., and A. Michael Huberman. 1994. *Qualitative Data Analysis: An Expanded Sourcebook*. SAGE Publications, Inc.
Millar, Morgan, and Don Dillman. 2012. "Do Mail and Internet Surveys Produce Different Item Nonresponse Rates? An Experiment Using Random Mode Assignment." *Survey Practice* 5 (2).
Miller, Angie L., and Amber D. Lambert. 2014. "Open-Ended Survey Questions: Item Nonresponse Nightmare or Qualitative Data Dream?" *Survey Practice* 7 (5).
Morduch, Jonathan, and Rachel Schneider. 2017. *The Financial Diaries: How American Families Cope in a World of Uncertainty*. Princeton University Press.
Morris, Adam, Ryan W. Carlson, Hedy Kober, and Molly Crockett. 2025. "Introspective Access to Value-Based Multi-attribute Choice Processes." *Nature Communications* 16: 3733.
Namey, Emily, Greg Guest, Amy O'Regan, Christine L. Godwin, Jamilah Taylor, and Andres Martinez. 2020. "How Does Mode of Qualitative Data Collection Affect Data and Cost? Findings from a Quasi-experimental Study." *Field Methods* 32 (1): 58–74.
Nathan, Brad, Ricardo Perez-Truglia, and Alejandro Zentner. 2025. "My Taxes Are Too Darn High: Why Do Households Protest Their Taxes?" *American Economic Journal: Economic Policy* 17 (1): 273–310.
Neuendorf, Kimberly A. 2002. *The Content Analysis Guidebook*. SAGE Publications, Inc.
Nisbett, Richard E., and Timothy D. Wilson. 1977. "Telling More than We Can Know: Verbal Reports on Mental Processes." *Psychological Review* 84 (3): 231–59.
O'Connor, Cliodhna, and Helene Joffe. 2020. "Intercoder Reliability in Qualitative Research: Debates and Practical Guidelines." *International Journal of Qualitative Methods* 19: 1609406919899220.
Oh, Suanna. 2023. "Does Identity Affect Labor Supply?" *American Economic Review* 113 (8): 2055–83.
Parker, Barbara, and Valerie Kozel. 2007. "Understanding Poverty and Vulnerability in India's Uttar Pradesh and Bihar: A *Q*-Squared Approach." *World Development* 35 (2): 296–311.
Patton, Michael Quinn. 2002. *Qualitative Research and Evaluation Methods*. 3rd ed. SAGE Publications, Inc.
Piore, Michael J. 2006. "Qualitative Research: Does It Fit in Economics?" *European Management Review* 3 (1): 17–23.
Rao, Vijayendra. 2023. "Can Economics Become More Reflexive? Exploring the Potential of Mixed-Methods." In *Handbook on Economics of Discrimination and Affirmative Action*, edited by Ashwini Deshpande, 323–49. Springer.
Rathje, Steve, Dan-Mircea Mirea, Ilia Sucholutsky, Raja Marjieh, Claire E. Robertson, and Jay J. Van Bavel. 2024. "GPT Is an Effective Tool for Multilingual Psychological Text Analysis." *Proceedings of the National Academy of Sciences* 121 (34): e2308950121.
Reja, Urša, Katja Lozar Manfreda, Valentina Hlebec, and Vasja Vehovar. 2003. "Open-Ended vs. Close-Ended Questions in Web Questionnaires." *Developments in Applied Statistics* 19 (1): 159–77.
RePass, David E. 1971. "Issue Salience and Party Choice." *American Political Science Review* 65 (2): 389–400.
Roberts, Margaret E., Brandon M. Stewart, Dustin Tingley, Christopher Lucas, Jetson Leder-Luis, Shana Kushner Gadarian, Bethany Albertson, and David G Rand. 2014. "Structural Topic Models for Open-Ended Survey Responses." *American Journal of Political Science* 58 (4): 1064–82.
Rodrik, Dani, and Stefanie Stantcheva. 2021. "A Policy Matrix for Inclusive Prosperity." NBER Working Paper 28736.
Romero, Mauricio, Juan Bedoya, Monica Yanez-Pagans, Marcela Silveyra, and Rafael de Hoyos. 2022. "Direct vs Indirect Management Training: Experimental Evidence from Schools in Mexico." *Journal of Development Economics* 154: 102779.
Roth, Christopher, Peter Schwardmann, and Egon Tripodi. 2024a. "Depression Stigma." CESifo Working Paper 11012.
Roth, Christopher, Peter Schwardmann, and Egon Tripodi. 2024b. "Misperceived Effectiveness and the Demand for Psychotherapy." *Journal of Public Economics* 240: 105254.
Rothschild, Jacob E., Adam J. Howat, Richard M. Shafranek, and Ethan C. Busby. 2019. "Pigeonholing Partisans: Stereotypes of Party Supporters and Partisan Polarization." *Political Behavior* 41 (2): 423–43.
Rozado, David. 2024. "The Political Preferences of LLMs." *PLoS ONE* 19 (7): e0306621.
Rubin, Herbert J., and Irene S. Rubin. 2011. *Qualitative Interviewing: The Art of Hearing Data*. SAGE Publishing, Inc.
Saccardo, Silvia, and Marta Serra-Garcia. 2023. "Enabling or Limiting Cognitive Flexibility? Evidence of Demand for Moral Commitment." *American Economic Review* 113 (2): 396–429.

Saldaña, Johnny. 2021. *The Coding Manual for Qualitative Researchers*.4 ed. SAGE Publications, Inc.
Seawright, Jason, and John Gerring. 2008. "Case Selection Techniques in Case Study Research: A Menu of Qualitative and Quantitative Options." *Political Research Quarterly* 61 (2): 294–308.
Shiller, Robert J. 1997. "Why Do People Dislike Inflation?" In *Reducing Inflation: Motivation and Strategy*, edited by Christina D. Romer and David H. Romer, 13–70. University of Chicago Press.
Shiller, Robert J. 2017. "Narrative Economics." *American Economic Review* 107 (4): 967–1004.
Singer, Eleanor, and Mick P. Couper. 2017. "Some Methodological Uses of Responses to Open Questions and Other Verbatim Comments in Quantitative Surveys." *Methods, Data, Analyses* 11 (2): 115–34.
Sinkowitz-Cochran, Ronda L. 2013. "Survey Design: To Ask or Not to Ask? That Is the Question..." *Clinical Infectious Diseases* 56 (8): 1159–64.
Sloman, Steven. 2009. *Causal Models: How People Think about the World and Its Alternatives*. Oxford University Press.
Small, Mario Luis. 2009. "'How Many Cases Do I Need?': On Science and the Logic of Case Selection in Field-Based Research." *Ethnography* 10 (1): 5–38.
Small, Mario Luis. 2011. "How to Conduct a Mixed Methods Study: Recent Trends in a Rapidly Growing Literature." *Annual Review of Sociology* 37: 57–86.
Smyth, Jolene D., Don A. Dillman, Leah Melani Christian, and Mallory McBride. 2009. "Open-Ended Questions in Web Surveys: Can Increasing the Size of Answer Boxes and Providing Extra Verbal Instructions Improve Response Quality?" *Public Opinion Quarterly* 73 (2): 325–37.
Spradley, James P. 1979. *The Ethnographic Interview*. Holt, Rinehart and Winston.
Stantcheva, Stefanie. 2020. "Understanding Economic Policies: What Do People Know and Learn?" Unpublished.
Stantcheva, Stefanie. 2021. "Understanding Tax Policy: How Do People Reason?" *Quarterly Journal of Economics* 136 (4): 2309–69.
Stantcheva, Stefanie. 2022. "Understanding of Trade." NBER Working Paper 30040.
Stantcheva, Stefanie. 2023. "How to Run Surveys: A Guide to Creating Your Identifying Variation and Revealing the Invisible." *Annual Review of Economics* 15: 205–34.
Stantcheva, Stefanie. 2024. "Why Do We Dislike Inflation?" *Brookings Papers on Economic Activity* 55 (1): 1–46.
Starr, Martha A. 2014. "Qualitative and Mixed-Methods Research in Economics: Surprising Growth, Promising Future." *Journal of Economic Surveys* 28 (2): 238–64.
Stefkovics, Ádám, and Endre Sik. 2022. "What Drives Happiness? The Interviewer's Happiness." *Journal of Happiness Studies* 23 (6): 2745–62.
Tiezzi, Silvia, and Erte Xiao. 2016. "Time Delay, Complexity and Support for Taxation." *Journal of Environmental Economics and Management* 77: 117–41.
Törnberg, Petter. 2024. "Large Language Models Outperform Expert Coders and Supervised Classifiers at Annotating Political Social Media Messages." *Social Science Computer Review*. https://doi.org/10.1177/08944393241286471.
Vafa, Keyon, Ashesh Rambachan, and Sendhil Mullainathan. 2024. "Do Large Language Models Perform the Way People Expect? Measuring the Human Generalization Function." Preprint, arXiv. https://doi.org/10.48550/arXiv.2406.01382.
Venkatesh, Sudhir. 2008. *Gang Leader for a Day: A Rogue Sociologist Takes to the Streets*. Penguin Press.
Wagner, Michael, and Duane G. Watson. 2010. "Experimental and Theoretical Advances in Prosody: A Review." *Language and Cognitive Processes* 25 (7-9): 905–45.
Wei, Jason, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2022. "Chain-of-Thought Prompting Elicits Reasoning in Large Language Models." Preprint, arXiv. https://doi.org/10.48550/arXiv.2201.11903.
Weiss, Robert S. 1994. *Learning from Strangers: The Art and Method of Qualitative Interview Studies*. Free Press.
Wekhof, Tobias. 2024. "Explaining the Attitude-Behavior Gap for Sustainable Investors: Open vs. Closed-Ended Questions." http://dx.doi.org/10.2139/ssrn.4838449.
Wekhof, Tobias, and Sébastien Houde. 2023. "Using Narratives to Infer Preferences in Understanding the Energy Efficiency Gap." *Nature Energy* 8 (9): 965–77.
Ziegler, Jeffrey. 2022. "A Text-As-Data Approach for Using Open-Ended Responses as Manipulation Checks." *Political Analysis* 30 (2): 289–97.
Zollinger, Delia. 2022. "Cleavage Identities in Voters' Own Words: Harnessing Open-Ended Survey Responses." *American Journal of Political Science* 68 (1): 1–48.
Züll, Cornelia. 2016. *Open-Ended Questions (Version 2.0)*. GESIS Survey Guidelines. GESIS–Leibniz-Institut für Sozialwissenschaften. https://doi.org/10.15465/gesis-sg_en_002.