CENTER FOR
ECONOMIC
BEHAVIOR &
INEQUALITY

CEBI WORKING PAPER SERIES

Working Paper 10/24

MEASURING WHAT IS TOP OF MIND

Ingar Haaland

Christopher Roth

Stefanie Stantcheva

Johannes Wohlfart

ISSN 2596-447X

CEBI
Department of Economics
University of Copenhagen
www.cebi.ku.dk

Measuring What Is Top of Mind*

Ingar Haaland Christopher Roth
Stefanie Stantcheva Johannes Wohlfart

May 4, 2024

Abstract

We survey the recent literature in economics measuring what is on top of people's minds using open-ended questions. We first provide an overview of studies in political economy, macroeconomics, finance, labor economics, and behavioral economics that have employed such measurement. We next describe different ways of measuring the considerations that are on top of people's minds. We also provide an overview of methods to annotate and analyze such data. Next, we discuss different types of applications, including the measurement of motives, mental models, narratives, attention, information transmission, and recall. Our review highlights the potential of using open-ended questions to gain a deeper understanding of mechanisms underlying observed choices and expectations.

Keywords: Thoughts, Open-ended Questions, Text Data, Methodology, Surveys, Qualitative Research.
JEL Classification: C90, D83, D91

*Ingar Haaland, NHH Norwegian School of Economics and FAIR, email: Ingar.Haaland@nhh.no; Christopher Roth, University of Cologne and ECONtribute, Max Planck Institute for Research on Collective Goods, CEPR; email: roth@wiso.uni-koeln.de; Stefanie Stantcheva, Harvard University, NBER, CEPR, email: sstantcheva@fas.harvard.edu; Johannes Wohlfart, University of Cologne and ECONtribute, Max Planck Institute for Research on Collective Goods, CEBI, CESifo, email: johannes.wohlfart@econ.ku.dk. Maximilian Fell, Paul Grass, Hrishikesh Iyengar, Sarthak Joshi, Apoorv Kanoongo, Julian König and Nicolas Röver provided excellent research assistance. 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.

1 Introduction

Quantitative measures of economic beliefs and choices, included in surveys, have substantially advanced our understanding of human behavior (Fuster and Zafar, 2023; Stantcheva, 2022a). Such measures have been used to study economic decision-making under uncertainty (Tversky and Kahneman, 1974), social preferences (Cappelen et al., 2007; Fehr and Gächter, 2000; Kahneman et al., 1986), policy views (Kuziemko et al., 2015; Stantcheva, 2021), and economic expectations (Coibion et al., 2018; Manski, 2004). Despite their success in characterizing heterogeneity in beliefs and preferences, quantitative measures are sometimes less suited for studying underlying mechanisms, such as respondents' attention allocation, economic reasoning and motives. For example, to understand the origins of economic expectations, it is critical to study attention allocation (Bordalo et al., 2023; Gabaix, 2019), mental models (Andre et al., 2021, 2023b), or the narratives people use to explain the world (Andre et al., 2023a).

In this paper, we review an emerging literature in economics that uses open-ended survey questions to better understand the mechanisms underlying economic beliefs and choices. In such questions, participants are asked to write down the considerations on top of their mind when thinking about a particular issue, decision or prediction problem. Different from quantitative measures, the qualitative text data resulting from open-ended measurement provide a detailed lens into respondents' thoughts and considerations. Unlike structured survey questions, open-ended questions do not prime individuals on any particular aspect through the displayed response options, thereby potentially changing the object researchers aim to measure. These features make open-ended questions suited for studying topics such as attention allocation, reasoning, mental models, or verbal communication, and thereby help to gain a deeper understanding of the mechanisms underlying economic choices and expectations. As shown in Figure 1, the use of open-ended survey questions has become increasingly common in economics.

Open-ended measures allow researchers to test predictions of influential theories of human behavior. For instance, such methods can be used to test whether personal experiences are associated with selective recall (Andre et al., 2021; Graeber et al., 2022) or which features of a decision problem people pay attention to (Bordalo et al., 2023). They also let us paint a more realistic picture of which variables individuals consider relevant when making choices or forming beliefs. Moreover, open-ended survey questions enable researchers to measure motives behind particular decisions (Chopra et al., 2024b; Hager et al., 2023b) or the perceived motives driving others' behaviors (Bursztyn et al., 2023b, 2022).

The techniques presented in this review are also important for policy questions, such as understanding which concerns loom largest in voters' minds (Ferrario and Stantcheva, 2022). For example, Stantcheva (2020) and Ferrario and Stantcheva (2022) use open-ended measurement to characterize the first-order considerations people have when thinking about different policies, such as the income tax or the estate tax.

This paper complements recent reviews on the design of surveys and information provision experiments in economics (Bergman et al., 2020; Fuster and Zafar, 2023; Haaland et al., 2023; Stantcheva, 2022a). Different from these papers, our review focuses on open-ended survey questions, providing a detailed discussion of advantages and disadvantages as well as best-practice recommendations for such questions. We build on research in psychology discussing the opportunities offered by verbal reports as well as limitations of such measures (Berger et al., 2016; Ericsson and Simon, 1980, 1993; Nisbett and Wilson, 1977). While Nisbett and Wilson (1977) highlight the limitations of open-ended measures, more recent evidence suggests that people can accurately report key aspects of their choice process in some contexts (Morris et al., 2023). This review also builds on research in survey methodology and public opinion research, where open-ended questions have been discussed as an alternative to closed-end questions (Geer, 1988, 1991; Krosnick, 1999; Lazarsfeld, 1944).

We proceed as follows: In Section 2, we review studies from various fields in economics that have used open-ended measurement techniques. In Section 3, we discuss different methods for measuring what is top of mind. In Section 4, we provide best-practice recommendations and examples for the analysis of data on thoughts. In Section 5, we discuss fruitful applications of open-ended thought measurement. Finally, Section 6 concludes.

2 Literature using open-ended measurement techniques

In this section, we review studies across different fields of economics that have employed the open-ended measurement of thoughts and considerations in applied settings. Tables 1–7 provide overviews across different fields in economics.

Political economy Political economy is one of the key areas in which open-ended measurement techniques have been applied. Such approaches have been used to understand the formation of policy preferences, political persuasion, and political behavior in the field.

In the context of policy preferences, Stantcheva (2020) has pioneered the use of open-ended questions. Using such questions in large-scale online surveys conducted in the US, Stantcheva (2021) examines what considerations people have in mind when prompted to think about income and estate taxes, along with these policies' goals, shortcomings, and anticipated effects. The open-ended responses indicate pronounced partisan differences in how frequently ideas related to distribution, government spending and public goods are expressed. The patterns observed in the open-ended responses are validated using stated policy views as well as structured response data on knowledge about taxes and reasoning about their efficiency, spillovers, and distributional and fairness effects. Using similar methods, Stantcheva (2022b) studies how people reason about trade. Dechezleprêtre et al. (2022) examine people's considerations in the context of climate change and what the government should do about it. König and Schmacker (2022) examine public attitudes toward sin taxes, focusing on taxes for sugar-sweetened beverages (SSBs). Responses to free response questions about first-order considerations of sin taxes unveil that people primarily oppose these taxes because of perceived paternalism and regressivity. Liscow and Fox (2022) study public perceptions of capital gains taxation. The study employs a free response question to uncover the rationale behind participants' stances on taxing consumption vs income. Jessen et al. (2024) investigate perceptions of life expectancy inequality. An open-ended question reveals better healthcare as the dominant preferred policy measures to combat such inequality, alongside education, minimum wages, and housing improvements. Casarico et al. (2024) study how beliefs about gender gaps in earnings and pensions affect the support of policies aimed at reducing gender inequality. A free response question uncovers that respondents attribute these gaps primarily to parenthood, occupation, salary, and working hours.

Andre (2024) studies the attentional foundations of redistributive preferences using open-ended questions. In particular, spectators decide how to distribute money between two workers. Respondents explain in an open-text format which considerations shaped their merit judgements. In the context of fiscal policy, Gründler and Potrafke (2020) use open-ended questions in which experts are asked to write down their main considerations about fiscal rules in free-text entry boxes.

Free response elicitations are also applied in the context of political persuasion. For example, Hüning et al. (2022) conduct an experiment that collects voting intentions and textual data from interactions in chat groups before the ballot and self-reported votes afterwards. Galasso et al. (2024) assess the effectiveness of video ads countering populist narratives during the 2020 Italian referendum on reducing Parliament members. The authors employ an open-ended question that probes respondents' perceptions of the videos.

It is also possible to use open-ended questions to understand how political correctness norms affect the prevalence and interpretation of public political behavior. For instance, Bursztyn et al. (2020) use such questions for studying inferences about the motives of individuals who made a donation to a xenophobic organization. They validate their open-ended measure with a structured belief measure and establish strong correlations. Bursztyn et al. (2023b) examine the impact of justifications on individuals' willingness to express dissenting opinions, particularly on controversial political and social topics. Following their choice, participants are asked a free response question about their motives for selecting a particular tweet.

Open-ended measurement techniques have also been applied to better understand political behavior in the field. For instance, Hager et al. (2023b) shed light on strategic interactions in political behavior by asking party supporters how they would adjust their campaign efforts in response to learning about the higher effort of their peers. They then ask the respondents to describe the reasoning underlying their decision. Nathan et al. (2023) explore household decisions to protest property taxes, leveraging a free response question to understand why people protest or not. Hager et al. (2023a) use open-ended responses to give party supporters more voice in the context of political campaigns.

Open-ended questions can also be useful for elucidating mechanisms underlying the effects of natural experiments. Lang and Schneider (2023) investigate the enduring impact of historical immigration on nationalist voting patterns in Germany, using a natural experiment. Leveraging a geocoded survey with an open-ended question on the perceived implications of the historical inflow of migrants post-WWII reveals a significant dampening effect on nationalist backlash in areas with greater historical exposure to migrants.

Macroeconomics Open-ended measurement techniques allow us to gain insights on important questions in macroeconomics, such as which mental models economic agents employ when forming macroeconomic expectations, how stories and narratives shape these expectations, and how attention is allocated between different economic variables.

To measure subjective models of the macroeconomy, Andre et al. (2021) conduct large-scale surveys with U.S. households and experts. They use a combination of structured and open-ended survey questions to understand which propagation mechanisms come to mind when thinking about the transmission of canonical macroeconomic shocks. They document striking heterogeneity in "what comes to mind" both within and across their samples of households and experts. For example, households are relatively more likely to think of a "cost channel" in the context of monetary policy shocks than experts. Andre et al. (2021) also provide evidence that the considerations on top of people's minds are significantly associated with their forecasts about unemployment and inflation responses to the shocks, and that they account for part of the differences in forecasts between experts and households. Moreover, Andre et al. (2021) combine priming with open-ended measurement of associations to study the causal effects of attention to supply-side and demand-side factors on households' forecasts of the effects of monetary policy shocks. They show that being primed on demand-side factors in the context of an interest rate hike significantly increases respondents' attention to the demand side and has a negative effect on their predicted inflation response to the shock.

To understand the narratives agents invoke to explain macroeconomic phenomena, Andre et al. (2023a) use free response questions in the context of a historic surge of U.S. inflation experienced in late 2021 and 2022. Respondents explain in an open-ended question which factors they think caused the increase in inflation. To quantify the causal narrative that respondents endorse, the authors represent each of these open-text responses by a directed acyclic graph (DAG). A causal DAG is a network of variables in which links between variables indicate causal relationships. Andre et al. (2023a) also conduct various experiments in which they experimentally manipulate which narratives come to respondents' minds. They quantify the first-stage effects of their treatments using a free response question.

In an application to attention allocation in macroeconomic contexts, Link et al. (2023) employ open-ended questions in panels of German firms and households, specifically asking, "What topics come to mind when you think about the economic situation of your company/household?". They provide evidence on attention allocation over the business cycle and the association between attention and macroeconomic expectations.

Several other studies have used free response questions to better understand the formation of macroeconomic expectations. In the context of inflation expectations, Leiser and Drori (2005) explore people's associations with inflation using open-ended text questions across different groups, including high school students, university students, shopkeepers, and teachers. They identify inflation associations by asking participants to specify terms, concepts, or short phrases related to inflation. This is followed by a more structured question where participants are asked to connect up to five out of nine economic terms, including inflation, and then explain their choices. An et al. (2023) study the formation of gas price and inflation expectations. They use open-ended survey questions to understand the reasoning underlying changes in participants' expectations and spending plans in response to information (e.g., about recent gas price changes).

Another set of applications in macroeconomic contexts is understanding people's attitudes and preferences regarding aggregate outcomes. On why people dislike inflation, Stantcheva (2024) uses a combination of open-ended and closed-ended questions to measure people's understanding, as well as their emotions and feelings related to inflation. Binetti et al. (2024) leverage these methods to shed light on people's understanding and preferences related to inflation – i.e., how they trade-off inflation and other economic outcomes such as unemployment and growth.

Colarieti et al. (2024) examine the dynamic adjustments in household spending and debt following unexpected increases or decreases in income (the 'how'), and the motivations behind these decisions (the 'why'). To understand the 'why' behind household financial behaviors, they guide households through a series of targeted survey questions asking them to select relevant reasons for doing something (e.g., spending more) and not doing something (e.g., not borrowing more). Identical spending behaviors can have diverse motivations and be consistent with different theoretical models and structured, detailed, and cohesive survey questions can uncover the thought processes behind financial decisions and help classify households into distinct types based on their decision-making principles.

Finally, to better understand the impacts of central bank communications on public finance perceptions, Hommes et al. (2023) conduct large scale experiments. Employing open-ended questions, they gauge prior public finance opinions and introduce a metric for macroeconomic policy literacy.

Finance Open-ended measurement techniques are also extremely useful for understanding belief formation and decision-making in finance contexts. In such contexts, many factors can be decision-relevant and a vast amount of potentially relevant information is available. Open-ended questions provide a lens into which of these factors individuals attend to when forming beliefs or making decisions.

One key application of open-ended measurements in finance contexts is understanding the motives and mechanisms underlying observed financial behaviors. Chinco et al. (2022) use hypothetical survey questions to study the relevance of the correlation of an asset's return with risk factors in driving individuals' investment decisions. To corroborate their findings, they also ask respondents to explain in open text what factors are most important to them when deciding what fraction of an endowment to invest in stocks. Subsequently, survey respondents self-classify their open-ended responses into a set of structured options.

Bailey et al. (2019) examine the relationship between home price expectations and mortgage leverage choice. To shed light on the mechanism, they conduct a survey in which individuals make hypothetical leverage choices across scenarios with different home price expectations. Subsequently, they ask respondents a free response question on why their choices differ across scenarios. Liu and Palmer (2023) use a variety of different approaches to show that many individuals seem to rely more on perceived past home price growth rather than on expected future home price growth when making a housing investment decision. In an open-ended elicitation of motives, their respondents cite low confidence in other belief factors or (implicit) extrapolation as a motive to rely more on past than on expected future home price growth.

Chopra et al. (2024a) use open-ended survey questions to shed light on the mechanisms underlying the effects of home price expectations on spending. They ask respondents to explain why an increase in their home price expectations would affect their economic circumstances in a particular way and use these data to disentangle mechanisms driving spending responses. Luttmer and Samwick (2018) assess the welfare impact of perceived policy uncertainty regarding social security benefits. They explore the drivers behind their findings through a free response question.

As a recent methodological innovation in this set of applications, Chopra and Haaland (2023) introduce an AI-assisted method for semi-structured interviews to dissect the stock market participation puzzle. They uncover a range of reasons for non-participation, such as financial constraints, risk concerns, and informational barriers, with respondents often attributing their non-engagement to multiple factors. The depth of the interviews further sheds light on subtler aspects that initial responses might obscure, including fears of financial losses, a preference for low-risk investments, a perception of stock markets akin to gambling, and misconceptions that stock ownership necessitates intensive market research and active trading.

Another set of applications in finance contexts uses open-ended questions to shed light on the drivers underlying individuals' beliefs. For instance, Andre et al. (2023b) investigate the mental models of the stock market underlying the return expectations of retail investors, financial professionals, and academic experts. The study employs open-ended questions asking respondents to explain the reasoning behind the return expectations they report in different hypothetical scenarios. These open-ended data reveal that households and financial professionals often neglect equilibrium price adjustments when forming return expectations. Ba et al. (2023) investigate experts' and nonexperts' beliefs about the impact of the 2020 racial justice protests on the stock performance of law enforcement-associated firms, which experienced unexpected stock price gains in response to the protests. Their study includes an analysis of free response questions asking respondents to explain their estimates. Those who underestimate the stock price gains tend to mention lower trust and budget cuts to the police as a result of the protests. Laudenbach et al. (2022) use an open-ended elicitation to study retail investors' selective retrieval of particular historical episodes when estimating the historical autocorrelation of aggregate stock returns.

Filippini et al. (2021) propose a novel application of open-ended measurement in finance contexts. Specifically, they measure people's financial knowledge by asking respondents to describe which characteristics they think distinguish sustainable financial products from conventional investments.

Labor economics Open-ended measurement techniques have also become frequently used in labor economics. The methods reviewed in this paper have been applied to understand worker productivity, job search decisions, and other labor market behaviors.

One set of applications in labor economics is concerned with understanding worker productivity. For instance, Kaur et al. (2021) measure financial worries among workers using an open-ended question, and study how such worries affect productivity. They document that about half of all respondents mention financial concerns when asked "What were you thinking about while working?" with no prompts related to finances. Abeler et al. (2023) use a combination of field and laboratory experiments to study the effects of incentive scheme complexity on workers' perception of dynamic incentives and their productivity. An open-ended question is used to gauge what fraction of workers recognize dynamic incentives.

Another set of applications is concerned with understanding preferences over job attributes and job search decisions. Rodrik and Stantcheva (2021) ask survey respondents free response questions on what is, to them, a "good job" without priming them one way or the other. When performing text analysis on these answers, the terms that come up most frequently are "good salary," "well paid," "a good environment/good feeling," "good work conditions," and terms related to "private life" and "family life," indicating a desire for work-life balance. A "bad job" is associated with almost the exact opposite attributes. They also ask respondents what features of a job they would pay most attention to if they had to look for a new one. Important features appear to be pay, good relations with colleagues and with one's boss, the possibility to leverage one's skills, autonomy and creativity, career progression, interest and passion in the job, and safe work conditions. Miano (2023) uses open-ended survey questions to examine how workers' beliefs about job search costs and external opportunities influence on-the-job-search. The paper askes respondents to describe the challenges they might face while searching for a new job.

Several papers have used open-ended measurement to understand the role of characteristics such as gender and personal background in labor market behavior. Capozza (2024) investigates women's primary concerns regarding the gender gap in salary negotiation through open-ended and structured questions, revealing two main causal narratives: concerns over potential employer backlash and the belief that women are less inclined to negotiate than men. Ayyar et al. (2023) use essays written by girls at age 11 to construct an index of traditional gender attitudes, and study how such attitudes are related to lifetime earnings. Oh (2023) investigates the influence of caste identity on labor supply decisions in rural India, utilizing a field experiment. The experiment includes a free-form question asking participants to explain their reasons for declining specific job offers, which points to important roles of social image concerns and an intrinsic need to maintain caste identity.

Behavioral economics Open-ended responses have also recently been employed in behavioral economics, where they are particularly useful for dissecting the mechanisms underlying observed behaviors and beliefs. In this section, we review applications of open-ended questions in research on attention and memory, social economics, and motivated cognition.

Open-ended measurement techniques have been a productive tool for research in cognitive economics, focusing on attention and memory. Martínez-Marquina et al. (2019) use open-ended data in the context of experiments highlighting failures in contingent reasoning. To shed light on the underlying cognitive processes, they analyze the written advice participants give to another participant. They provide direct evidence that participants facing uncertainty neglect relevant states in their advice and that participants' advice is correlated with their own failures in contingent reasoning.

Arrieta and Nielsen (2024) explore how complexity influences individual choice processes.

In the context of choices under risk using lotteries, individuals face complex and simple decision scenarios and convey their decision-making methods through direct messages to other participants. These other respondents are then incentivized to replicate the choices with or without the messages. Bordalo et al. (2023) introduce and empirically test a theoretical framework for how individuals' attention to salient features in statistical problems influences their reasoning and decision-making, leading to various biases and judgements. For the empirical tests, the authors utilize free form questions explicitly asking participants how they arrived at their answers to statistical problems. Agranov and Ortoleva (2017) use open-ended questions to understand the motives behind people exhibiting a preference for randomization. The open-ended data illustrate that people randomize their choices because of hedging motives.

Kaufmann et al. (2024) investigate the behavior and market effects of "socially responsible consumers" who care about climate change and other externalities associated with their purchases. To measure consumers' mental models of externalities they leverage free response questions, which highlight substantial heterogeneity in the sophistication of consumers.

Measuring what is on top of mind is also helpful in memory research. For example, Graeber et al. (2022) study selective recall of statistics versus stories. They use hand-coded data based on an unstructured open-ended recall task to provide direct evidence on participants' recall of the provided information. Their open-ended measure provides rich insights into the specific associations that come to respondents' minds for different types of information seeded on the day before.

Open-ended questions have been employed in the context of economics research on communication. Graeber et al. (2024a) examine how explanations affect choice accuracy in financial reasoning tasks. In the experiments, participants first solve financial problems and provide verbal explanations for their choices, which are then relayed to other participants who make their own decisions based on these inputs. Exposure to verbal explanations significantly improves choice accuracy compared to just learning about another's choice. This effect is driven entirely by respondents encountering learning opportunities. The authors then characterize differences in the supply of explanations across tasks and by the accuracy of speakers. Graeber et al. (2024b) investigate how verbal transmission distorts economic information. In the study, participants listen to audio recordings containing economic forecasts and then relay this information to others via voice messages. Using both script-based and belief-based measures of information loss they show that information regarding the reliability of a forecast is more likely to be lost in transmission compared to the forecast's content itself. The study also explores the underlying causes of this differential information loss, suggesting information about reliability simply does not come to mind during transmission. Grunewald et al. (2024) investigate the propagation of motivated beliefs through communication, incorporating an experiment where participants play a dictator game and then use free-form chats to discuss their decisions and beliefs about others' decisions. They discover that communication typically reduces biases, except in environments with external plural opinions, where biases are amplified.

Open-ended elicitations have also been employed to study motivated reasoning and self-deception. Chopra et al. (2024b) use these techniques to measure people's motives for subscribing to a newsletter. Castagnetti and Schmacker (2022) also employ free form questions to understand people's preferences for receiving feedback about their IQ. Saccardo and Serra-Garcia (2023) explore preferences for enabling versus limiting belief distortion in moral dilemmas through experiments in an advisor setting with more than 9,000 participants. They ask respondents to explain the reason behind choosing the order of receiving commission and product quality information.

Behavioral economists have employed open-ended measurement in a variety of other applications. Elias et al. (2023) study how individuals perceive and reason about sudden price increases for different products under different policy regimes. Using textual analysis of open-ended responses, they show that price increases are not seen just as signals of scarcity. Instead, they cause widespread opposition and strong and polarized moral reactions.

Bursztyn et al. (2023a) use open-ended questions in the context of social media consumption to understand why active users might prefer a world without the respective platform (TikTok and Instagram). They also employ free form questions to understand how the respondents would feel if only they deactivated their accounts, revealing significant concerns about fear of missing out (FOMO).

Roth et al. (2024b) study how misconceptions about the effectiveness of psychotherapy influence individuals' willingness to utilize and invest in it, using a study of 1,843 depressed participants. They employ open-ended questions to understand the participants' considerations about their willingness to pay for therapy. The open-ended responses reveal that the perceived effectiveness of therapy is the central concern influencing individuals' valuation of therapy. Roth et al. (2024a) employ a similar methodology, employing open-ended questions, to explore the impact of perceived social stigma on therapy demand.

Development economics Research in development economics has long used qualitative research methods. Unlike most of the applications in this section, these questions are often asked in person and recorded by a surveyor or enumerator. They are often part of a multi-pronged strategy that includes focus group discussions, non-structured or semi-structured interviews, and structured surveys. They serve both as a basis for developing interventions and structured questionnaires, as well as to assess outcomes and mechanisms.

Jayachandran et al. (2023) propose a method to use open-ended interview questions to develop a short series of closed-ended questions that best capture the latent attitudes or beliefs. In their application, they first conduct an interview with open-ended questions related to female agency and score each woman's agency based on the interview. This measure is used as the benchmark measure of agency. They then determine which five closed-ended questions (and index measure based on them) are most predictive of the benchmark using machine learning algorithms (based on LASSO and random forest). This method to select survey questions based on how well they predict qualitative interview questions can be applied in other settings.[1]

In development economics, qualitative interviews are often conducted at the end of interventions to assess participants' understanding or specific mechanisms. Baird et al. (2011), for instance, conduct structured in-depth qualitative interviews with a random subsample of participants to assess their understanding of a cash transfer intervention. Dillon et al. (2012) use qualitative interviews at the end of each day during a pilot phase, when households are asked about their feedback on the survey questions and how they chose responses.

At other times, these questions form the bulk of the data, such as in Romero et al. (2022)

[1] A similar approach is used in Parker and Kozel (2007) to assess the complex reasons behind poverty in Uttar Pradesh and Bihar.

who study schools' managerial practices using the DWMS—an adaptation of the World Management Survey (WMS), including open-ended questions recorded in audio files.

3 Methods for the measurement of thoughts

In this section, we discuss design considerations when measuring people's thoughts. We start with a discussion of open-ended survey questions—the most common tool to elicit top-of-mind responses and the main focus of this review. We next discuss how structured survey questions can sometimes be a good alternative or supplement to open-ended survey questions. Lastly, we discuss two recent technological advances in the measurement of thoughts, namely speech recordings and AI-based interviews.

3.1 Measuring thoughts using open-ended survey questions

One way of measuring thoughts is to ask survey participants to write down in an open-text box their key considerations when thinking about a particular issue, when stating a specific belief, or when taking a specific decision.

Design considerations for open-ended thought measurement Stantcheva (2022a) provides detailed guidelines on best practices for writing open-ended questions. Open-ended response formats require more effort from survey participants than structured formats. To maintain high data quality, it is thus critical to lower the mental costs for respondents to the extent possible. Open-ended questions should thus ideally be asked early on in the survey, and not too many should be asked within the same survey. Moreover, depending on the context, it may be desirable to include an open-ended question at the beginning of the survey and to screen out respondents who are generally unwilling to engage with open-ended questions and who fail basic attention checks. Open-ended questions might also be especially taxing for respondents taking surveys on their mobile phones, making it good practice to encourage the use of a computer for surveys with many open-ended questions.

Another key consideration is how to avoid ex-post rationalization to the extent possible. Specifically, participants could come up with reasons for their choices and stated beliefs after these have been made or expressed. One way of mitigating concerns about ex-post rationalization is to elicit thoughts directly on the decision screen, possibly even before respondents make their choice. However, such a prompt may change behavior by increasing deliberation time (Imas et al., 2022).

Moreover, to reduce heterogeneity in response behavior, it is good practice to give respondents an idea about how much written text is expected in their response to a particular question and to ask people to respond in full sentences. Similarly, it is desirable to give respondents an indication of the amount of time they are expected to spend on the question ("Please spend 1 or 2 minutes"). An alternative approach is to use a minimum length validation, which allows respondents to proceed with the survey only if the open-ended response contains at least a certain number of characters.

Finally, the visual format of the responses boxes should be tailored to the types of answers needed. For single answers, it is desirable to provide a single answer box, while for multiple answers, multiple answer entry fields should be provided. If you want longer written texts you should use larger open-text boxes.

Advantages of open-ended thought measurement Compared to structured approaches, open-ended measurement of considerations offers several advantages. First, open-ended measurement techniques allow respondents to freely express their thoughts, not restricting them to a predefined set of structured response options. This is especially important in settings where the researchers want to discover novel factors and in settings where it might be difficult for the researchers to predict in advance what will be on top of people's minds. Second, open-ended questions do not change people's thoughts by informing them about potential lines of reasoning or drawing their attention to particular issues through the displayed response options. This feature should alleviate concerns about potential confounds, such as social desirability bias or ex-post rationalization. For instance, when eliciting memories, it is much more natural to ask people which events they remember than to give a structured list of response options—in which case it is not clear whether they were reminded of the event or actually remembered it. 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 underlying knowledge. Third, open-ended questions can be asked directly on the screen eliciting the prediction or decision of interest, which allows researchers to document the thoughts that are on respondents' minds while they make their decision or prediction. This further mitigates concerns about ex-post rationalization. Fourth, open-ended responses are arguably more natural to respondents and thus better suited to capture typical reasoning in real-world situations. Fifth, open-ended responses may reveal misunderstanding or confusion on the part of participants, and allow for qualitative insights that cannot be achieved with structured measures.

Disadvantages of open-ended thought measurement Open-ended measurement techniques necessarily also have a series of disadvantages: First, as a result of their unstructured nature, there is likely large variation in the way individuals respond to open-ended questions. This variation may affect the content of the answer and its length.[2] Second, open-ended responses are qualitative in nature, which makes it necessary to apply text analysis methods or develop a coding scheme to quantify and compare responses. 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. 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. There are also subjective judgments to be made when coding up the responses according to the coding scheme, potentially introducing additional noise and measurement error. Although large language models can sometimes reduce costs by annotating open-ended text data, validating LLM performance with human coders remains important, particularly for responses that might require nuanced judgments beyond current LLM capabilities. Third, open-ended responses likely contain non-classical measurement error, as respondents may be unwilling to exert effort when describing their thoughts, which may vary systematically across demographic groups.

[2]For instance, consider the setting in Andre et al. (2021) where respondents describe their thoughts while predicting the effects of macroeconomic shocks on inflation and unemployment. In this setting, a respondent may write that she used her knowledge of economics, but may not indicate which specific economic mechanism she had in mind. Other respondents may describe the full propagation channels of the shocks.

3.2 Measuring thoughts using speech recordings

A recent innovation in the open-ended measurement of thoughts is to ask participants to record their considerations instead of writing them down. These measures allow for a richer analysis of thoughts than text and might be particularly adept at capturing immediate top-of-mind responses as they allow for the immediate expression of thoughts without the delays associated with the writing process.

Design considerations for measuring thoughts using speech recordings The general approach to using speech recordings for thought measurement, as illustrated in the studies by Graeber et al. (2024a) and Graeber et al. (2024b), involves a process where participants are prompted to verbalize their thoughts or relay information presented in audio recordings. This method captures real-time thought processes and articulation in a dynamic manner. In Graeber et al. (2024a), the task involves explaining choices made in financial reasoning tasks, whereas Graeber et al. (2024b) focus on the transmission of economic forecasts. The recordings are subsequently analyzed to assess various factors such as the accuracy of information, the reliability of transmission, and the presence of specific language markers such as the usage of modal verbs.

Advantages of measuring thoughts using speech recordings There are several advantages to using speech recordings for measuring thoughts. Next to content features that are also captured by writing in a text box, speech recordings capture the spontaneity and natural flow of thoughts, which is often lost in written communication. Speech recordings capture more features than just text, including information about emotions, tone, emphasis, and natural disfluencies Graeber et al. (2024a). For instance, when eliciting narratives—stories people tell to make sense of the world—documenting the broader thought process and the emotional tone people use to discuss different relevant factors (e.g., the rise in inflation or the past financial crisis) might give nuanced insights into their thinking.

Disadvantages of measuring thoughts using 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 et al. (2024a) suggests that participants feel comfortable recording themselves. Additionally, analyzing speech data can be more complex and time-consuming than written responses due to the need to interpret non-verbal cues. Finally, technical limitations, such as poor audio quality or speech impediments, can pose challenges in ensuring clarity and usability of the recordings, even though this very rarely matters in practice (Graeber et al., 2024a,b).

3.3 Measuring considerations using AI interviews

While most studies reviewed in this paper measure considerations using a single open-ended survey question, a recent methodological advance is to use artificial intelligence (AI) to conduct qualitative interviews in which the initial top-of-mind response is followed by a series of follow-up questions.[3]

Methodological framework for AI-assisted interviews Chopra and Haaland (2023) develop an innovative AI-assisted semi-structured interview method. They leverage the advanced capabilities of transformer-based language models, specifically an API integration with OpenAI's GPT-4, to simulate human-like interviewing processes. The text-based interviews are conducted using a chat interface that mirrors popular text messaging applications and can easily be integrated into standard survey software, such as Qualtrics.

The AI interviewer is programmed to adhere to the methodological best practices inherent in qualitative research, such as using broad, open-ended, and neutral questions. The key advantage of AI-assisted interviews compared to using a series of pre-defined 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.

For instance, when surveying less literate populations or those with lower educational

[3]This complements work using human-led qualitative interviews, which are also increasingly used in economics (Bergman et al., 2024; Bustos et al., 2022; Duflo, 2017, 2020; Duflo et al., 2013).

backgrounds, who may struggle to articulate their thoughts, 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. 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.

Advantages of AI-assisted interview techniques The AI-assisted interview method offers a significant advantage in scalability, as it enables the efficient conduct of a large number of interviews simultaneously, surpassing the logistical and resource limitations often encountered with traditional human-led interviews. The primary advantage of this AI-driven interview method is its ability to mimic the depth and flexibility of human interviews while maintaining consistency and objectivity. By using open-ended questions as a starting point and allowing for in-depth probing, the method captures a rich and nuanced understanding of respondents' thoughts and opinions. Additionally, the AI's ability to adaptively manage the conversation based on previous responses adds a layer of contextual sensitivity often lacking in traditional survey methods. Chopra and Haaland (2023) demonstrate high levels of satisfaction among interviewees, highlighting the effectiveness and acceptability of this approach.

Disadvantages in AI-assisted interviews AI-assisted 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-assisted interviews face unique challenges, such as algorithmic biases and potential concerns about data privacy. Another disadvantage of AI-conducted interviews compared to single open-ended responses is that they increase the size and complexity of the resulting text corpus, making it more costly for the researchers to categorize and analyze the data. Furthermore, if the interest mainly is in collecting spontaneous top-of-mind responses, a full interview might introduce too much deliberate thinking to be revealing of top-of-mind thoughts.

4 Analyzing data on thoughts

In this section, we discuss different ways of analyzing data on thoughts, focusing on both text analysis methods, following the overview in Ferrario and Stantcheva (2022), and human coding of text. On top of this, we also provide a brief overview on how one can use large-language models to classify text data.

4.1 Text analysis methods

Text analysis methods offer an easy-to-implement, cost-effective and scalable way of analyzing open-ended data on thoughts.[4]

Preparing text data To quantitatively analyze text data, one can apply common techniques from computational linguistics to reduce the number of separate word features and to combine words originating from the same root. First, one should delete all punctuation, digits, and words with fewer than three characters. Second, one can stem words using a Porter Stemmer (Porter, 1980), meaning that one reduces words to their roots by cutting off the suffix. Third, it is best practice to filter out stop words.

Word clouds Word clouds are utilized to display the most frequently used words in a dataset, offering a visually appealing representation of data where word sizes are proportional to their frequency. The creation of these displays involves algorithms that dynamically adjust the placement and size of words to optimize both aesthetics and readability. However, despite their visual appeal, word clouds have limitations as they emphasize word frequency without accounting for context or semantic similarity. This can lead to misleading interpretations where common but less informative words appear disproportionately significant. Despite these limitations, word clouds serve as a useful starting point for identifying the main patterns in the data. To enhance readability, one should restrict the number of words in the word cloud to a reasonable value (e.g., 100).[5]

[4] Text analysis methods are widely used in economics (Ash and Hansen, 2023) and sociology (Popping, 2015). For an overview of different methods, see Gentzkow et al. (2019).

[5] Some of the features shown in word clouds are usually 2-grams, i.e., two words combined into a frequently used phrase.

Lasso procedure Lasso procedures are commonly employed to select features from unstructured text that best predict the variation in structured measures of interest, such as structured measures of thoughts. For Lasso procedures to be applied, in addition to the preparations described above, text responses must be converted to a numerical format which can be done using, e.g., bag-of-words, term frequency-inverse document frequency or word embeddings. By imposing a penalty, the Lasso method encourages a sparse solution, effectively minimizing prediction error while promoting model simplicity.[6] This characteristic makes Lasso particularly adept at handling high-dimensional data, like large text datasets. Yet, Lasso methods are more likely to face challenges in the presence of high multi-collinearity.

Keyness and keywords analysis To study the relative frequency of words across groups, it is possible to conduct keyness analysis. This method identifies words that are best at predicting group membership. For example, it can be used to predict differences in words used across demographic groups (see, e.g., Stantcheva 2020). It can also be used to predict differences across different treatment groups. For instance, Chopra et al. (2024b) and Bursztyn et al. (2022) employ the methodology proposed by Gentzkow and Shapiro (2010) to determine the words that are most characteristic of being in different treatment groups. Given two groups A and B of respondents, they calculate Pearson's χ^2 statistic for each word w,

$$\chi^2_{wAB} = \frac{(f_{wA} f_{\sim wB} - f_{wB} f_{\sim wA})^2}{(f_{wA} + f_{wB})(f_{wA} + f_{\sim wA})(f_{wB} + f_{\sim wB})(f_{\sim wA} + f_{\sim wA})} \tag{1}$$

where f_{wA} and f_{wB} denote the total number of times that the word w is mentioned by respondents in group A and B, respectively. Similarly, $f_{\sim wA}$ and $f_{\sim wB}$ refer to the total number of times words *other* than w are mentioned. Their subsequent analysis focuses on the words with the largest χ^2.

Topic analysis Topic analysis is a powerful tool to analyze open-ended text data (Roberts et al., 2014), beneficial for both preliminary exploratory analysis before using human coders and for making well-grounded inferences about the effects of treatments, frames, or covariates on the content of responses. This approach is based on a keywords-count model (Wekhof, 2024).

[6]For an overview of different selection methods and penalty terms, see Freo and Luati (2024).

Topics are defined by sets of keywords. For example, in Stantcheva (2021) and Stantcheva (2022b), topics are defined by keywords that capture particular aspects of a policy. It is then possible to display the distribution of topics that are most commonly mentioned in an open-ended question and to study group differences in the topics that are on top of people's minds. To identify topics and their corresponding keywords, methods range from manual to more automated techniques like semi-supervised or unsupervised approaches; see for an overview of some key methods Ferrario and Stantcheva (2022). For shorter text responses, such as those found in surveys, extracting and analyzing the document-term matrix helps in understanding word usage patterns (Ferrario and Stantcheva, 2022). Sensitivity checks on the identified topics and keywords are crucial, especially when adjusting for differences in answer lengths across respondent groups.

Identifying causal structure in text Ash et al. (2023) provides an unsupervised method to quantify latent narrative structures in text documents. Their software package relatio identifies coherent entity groups and maps explicit relations between them in the text. They apply their method to study political and economic narratives in the US congress.

4.2 Human coding of scripts

An alternative to textual analysis is human coding. Human coding is particularly useful if researchers aim to measure nuanced lines of reasoning (such as thoughts of particular theoretical mechanisms), which automated methods may struggle to capture. Compared to automated methods, human coding is more labor-intensive, more prone to potential human biases, and much more difficult to scale.

Several studies have used hand-coding of open-ended responses to analyze unstructured text data. Andre et al. (2021) elicit respondents' thoughts 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 thoughts related to economic propagation mechanisms of the shocks or whether responses include general political or normative statements. Zooming in on respondents that mention economic propagation mechanisms, they also employ a more fine-grained coding scheme that differentiates between particular economic mechanisms (such as firms passing on higher costs to product prices or demand-side channels such as changes in household spending).

Andre et al. (2023a) use hand-coding of respondents' open-ended explanations for why inflation in the US increased. They represent each respondent's explanation by its Directed Acyclical 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.

Hager et al. (2023b) hand-code motives underlying effort adjustments in response to new information, while Chopra et al. (2024b) hand-code perceived motives underlying a newspaper's reporting strategy. As another example, Bursztyn et al. (2020) use hand-coded data from a question on the purpose of the study. A team of research assistants classify the open-ended responses on the study purpose into different categories. Graeber et al. (2024a) rely on human-coded data for different types of arguments across 15 different financial reasoning tasks. Chinco et al. (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. The structured categories were selected based on open-text responses in pilot studies.

Inductive vs deductive coding schemes 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 are "grounded" in the data, is very 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. Furthermore, some coding schemes might include elements of both. A typical application of open-ended responses in economics is to test whether participants correctly guessed the study hypothesis. While it is natural to always include a code that corresponds to the correct option, a reading of the actual responses might inform a broader coding scheme that includes other beliefs about the study's purpose.

Best-practice recommendations for human coding of scripts To ensure a high quality of manually coded open-ended data, we recommend implementing the following steps. First, coders should be given a written coding scheme and instructed in detail on how to apply the scheme to the data in question. Ideally, all involved coders participate in a joint training session and subsequently take a test to ensure a thorough and common understanding of the coding scheme. Second, making sure that the coders do not know the research hypothesis can reduce the potential for biases in coding. Lastly, double-coding of responses and resolving discrepancies through discussion between coders can reduce measurement error and mitigate the effect of biases of individual coders. Double-coding also allows researchers to calculate the intercoder reliability (ICR), which we discuss at length in the next subsection.

4.3 Intercoder reliability

ICR is a numerical measure of the agreement between different coders regarding how the same data should be coded. This section discusses its uses and advantages, and some practical implementation suggestions. It largely follows the paper by O'Connor and Joffe (2020).

Uses and advantages Advantages of ICR are that it allows for the assessment of the rigor and transparency of the coding scheme and its application to the data and is a signal of quality for the reader (that is otherwise especially difficult to assess with open-ended data). ICR may also have benefits internal to the research process: it can motivate researchers to ensure consistency in coding decisions, which is especially important with multiple researchers coding and cross-linguistic studies. It also fosters iteration within the research team that can improve the coding process.

One pitfall of ICR—perhaps similar to other summary statistics—is that it may convey a sense of false precision, even in a study that is otherwise of poor quality. Another important shortcoming is akin to "multiple hypothesis testing." As Hruschka et al. (2004) warn, repeated ICR testing followed by discussions and changes in the procedure may end up leading to "interpretative convergence" that may or may not be warranted.

Practical considerations ICR requires a minimum of two independent coders. With the rise of AI tools, it is conceivable that one of the "coders" might be an AI tool. More coders are informative and can be beneficial on net depending on the resources and importance of the project. When a single person has performed the coding, it is necessary to recruit an additional coder to code a sample and verify the reliability. There is little consensus on what share of the data should be subject to multiple coders. O'Connor and Joffe (2020) recommend that, depending on the size of the data, 10-25% of it are randomly selected and representative of the entire data.

It is also critical that the interaction between coders is set and documented before commencing the coding (including to alleviate the concern about interpretative convergence raised by Hruschka et al. 2004). Pre-registration of the procedure may be desirable and, if these methods become more widespread, might be recommended by journals, the same way as it is currently done for other types of analysis and experiments. If the main goal is to ensure objectivity, then no to few interactions and the recruiting of external coders is desirable.

If the main goal is instead to enhance the analytical process by encouraging conversations among researchers and pinpointing areas that require further explanation, then it is essential to have discussions between various independent coding sessions. In this case, a valuable strategy can be to agree on a cutoff IRC score (see below) above which the coding will be considered reliable and consistent across coders and from which point on, a single coder can be used.

A decision that needs to be taken is how to segment the data. Using larger units of text (e.g., full survey responses or multiple responses) implies greater validity but also more complexity, which may result in lower ICR scores. Smaller units can be tough to determine while still preserving the context of the whole. O'Connor and Joffe (2020) recommend that the first and more experienced researcher segments the data and then the alternative coder follows the initial segmentation.

ICR scores are likely to be higher if there are fewer codes (i.e., fewer categories to classify answers in) and if the question is more factual rather than conceptually complex. In that sense, paradoxically, the more nuanced and sophisticated the question posed is, the likelier that the ICR score will be low.

Computing and presenting the ICR score The most common method in the literature is a percentage-based approach but it is also a problematic one as it does not account for agreement by chance between coders and is hard to apply when there are more than two coders (Lombard et al., 2002). Statistical tests such as Cohen's kappa, Krippendorff's alpha, Scott's pi, Fleiss' K, Analysis of Variance binary ICC, and the Kuder-Richardson 20 correct for agreement by chance and can be used when there are multiple coders. The most popular and flexible one is Krippendorff's alpha (O'Connor and Joffe, 2020).

Presenting a single ICR measure, as an average value for all codes, is commonly done. A more informative alternative is to present the distribution of the ICR scores for all codes. There are no unanimously accepted cutoffs to represent a "good level" of agreement (O'Connor and Joffe, 2020), so the interpretation depends on the score used and the context.

Nevertheless, Miles and Huberman (1994) suggests that a good reliability means that coders agree to an extent of 80% on 95% of codes that they jointly looked at. Neuendorf (2002) cites as "rules of thumb" ICR scores above 0.9 as "acceptable by all," and scores above 0.8 as "acceptable by many." Below these thresholds, however, there is substantial disagreement below.

What if the ICR scores are low? The approach depends on what the goal of the computation was (from the two main goals described above). If it is a one-time appraisal of the coding process, possibilities include to discard codes below a certain threshold and/or simply report the results. If it is part of an iterative approach to improve the coding, the research group can modify the procedure and repeat the process until the score is sufficiently improved.

4.4 Using Large Language Models to annotate text data

An alternative to human coding of qualitative data is to exploit AI methods. Recent evidence suggests that Large Language Models (LLMs) can annotate text data in a reliable and reproducible manner (Gilardi et al., 2023). Several recent papers in economics have also started to classify text data using LLMs, such as OpenAI's flagship model GPT-4 (Bursztyn et al., 2023a; Graeber et al., 2024a,b; Link et al., 2023). These papers demonstrate that classifications using GPT-4 and human coding in many cases yield very similar results.

While LLMs, such as GPT-4, are a low-cost and viable alternative to human coding of qualitative data in many cases, it is important to emphasize that their performance might depend both on the quality of the prompting, the type of LLM used, and the complexity of the setting (Rathje et al., 2023). It is therefore important, especially in novel or complex settings, to validate the quality of the LLM coding with human coding of the same data. If LLMs and human coding produce similar results on a random subset of the data, one can more safely rely on LLM coding of the full data set.

To classify open-ended responses with an LLM such as GPT-4, it is good practice to first develop a coding scheme as discussed in Section 4.2. While it is possible to use GPT-4 to assist in this process by simply giving it example responses and asking it to suggest a potential coding scheme, it is important to read through some responses and use human judgment when creating the coding scheme. The coding scheme should include a name and description of each category as well as some example responses. A prompt can then include the name and description of the category and the example responses from the coding scheme.[7] For instance, Bursztyn et al. (2023a) 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." Then, a specified hand-coding scheme was given, including category names, definitions, and illustrative examples.[8]

[7]This approach is called *few-shot prompting* because it provides the model with some examples to learn how to apply the coding scheme. A more comprehensive approach in which the LLM is trained on a task-specific data set with more examples that can fit in a single prompt is called fine-tuning.

[8]For instance, in Bursztyn et al. (2023a), the category FOMO includes the following description: "Respondent

5 Major applications

In this section, we discuss different types of applications for the open-ended measurement of thoughts.

5.1 Designing surveys

Open-ended questions can be an intermediate step in the design of large-scale surveys with structured response options. First, a challenge with structured response options is that the researchers have to decide in advance on the relevant factors. One approach to designing structured response options is purely data-driven. In a first step, a pilot study is used to understand what type of responses participants give in an unstructured open-ended question. Subsequently, the researchers design structured response options based on the open-ended pilot data. An advantage of this approach is that the response options closely reflect what is naturally on top of people's minds. It also helps to capture the most relevant options, some of which might not be obvious to the researcher. When using structured response options, it is good practice to include an "other" category with a free response format. This approach is especially useful for cases in which it is not possible to ask open-ended measures, e.g. because of length restrictions on the survey.

Second, in many cases, it is desirable to combine open-ended and structured measures of reasoning. Including both types of questions allows for the validation of the unstructured data with the structured data. For example, Andre et al. (2021) show that word counts in their unstructured data are predictive of their structured measures of thoughts. Stantcheva (2024) also includes a mix of open-ended and closed-ended questions on the same issue, where the closed-ended questions' answer options reflect economic theory. The comparison between the two sets of answers reveals the phenomena that economic theory might be missing, but which are top-of-mind for respondents. To rule out that the response options in the structured question change participants' responses to the open-ended question, structured questions on

mentions fear of missing out, feeling out of the loop, their wish to stay connected, or justifies usage through others' usage." and the 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."

thoughts should ideally be presented on a separate survey screen appearing later than the open-ended question.

5.2 Reasoning and (perceived) motives behind decisions

A key application of open-ended measurement is to study the reasoning and motives underlying specific decisions. First, open-ended questions can be used to understand the considerations behind decisions within experiments. Such decisions could include choices about the consumption of pieces of news offered by the experimenter (Chopra et al., 2024b), the willingness to Tweet particular content (Bursztyn et al., 2023b), or making a particular economic forecast or prediction (Andre et al., 2021). Typically, researchers will first ask respondents to make the decision or the prediction in question. On a subsequent survey screen or further down on the decision screen, respondents are then asked to report the main considerations or reasons underlying their decision in an open-text box. The resulting text data allow researchers to understand how people justify their decisions and which factors they consider important for their choices.

Second, instead of measuring motives behind decisions in experimental settings, open-ended questions can be used to characterize motives underlying decisions in the real world. Applications include protesting property taxes (Nathan et al., 2023), willingness to engage in a political campaign (Hager et al., 2023b), stock market non-participation (Chopra and Haaland, 2023), and the consumption of goods with externalities (Kaufmann et al., 2024). Typically, researchers first elicit the behavior of interest using a structured survey question and subsequently pose an open-ended question on people's motives by asking people why they behave in a specific way.

Third, open-ended responses can also be leveraged for the measurement of inference about others' motives. In particular, it is possible to ask respondents about the motives underlying another respondent's choice (Bursztyn et al., 2023b, 2022, 2020). In practice this involves researchers asking respondents why they think someone else acts in a particular way.

5.3 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 the causal accounts for why a specific event occurred (Akerlof and Snower, 2016; Pennington and Hastie, 1992; Shiller, 2017; Sloman and Lagnado, 2015; Trabasso and van den Broek, 1985). Mental models can be thought of as beliefs about the co-movement between different variables and the underlying mechanisms driving this co-movement. Open-ended measurements can be a useful tool to measure and understand narratives and mental models. The most common applications include: asking respondents about (i) the causes of a given phenomenon (Andre et al., 2023a), (ii) the mechanisms underlying the relationship between different variables (Andre et al., 2021, 2023b) and (iii) the consequences of a given change in a variable without specifying one specific outcome variable (Stantcheva, 2024).

5.4 Attention allocation

Open-ended measurement approaches can be used to measure people's attention allocation. For instance, such measurement can be applied to better understand which concerns loom largest in voters' minds when they think about government policies (Ferrario and Stantcheva, 2022) or households' and firm' managers attention allocation across different economic topics (Link et al., 2023). Different from the measurement of motives behind decisions, these applications do not ask an open-ended question about a previously made decision or stated belief. Instead, survey respondents are confronted with a prompt asking them what issues come to their mind when thinking about a specific topic. Respondents write down their thoughts in an open-text box. The resulting text data provide insights into people's spontaneous considerations and attention allocation in the context of the prompted topic.

5.5 Priming interventions

The mechanisms underlying the effects of priming interventions have been widely criticized for being a black box. Open-ended questions open up the possibility of measuring how priming interventions affect attention allocation (Andre et al., 2021, 2023a). Priming interventions are typically used to exogenously draw respondents' attention to a particular issue or aspect of a decision problem. This allows the researcher to study the causal effect of attention to a particular issue on beliefs, decisions or behaviors elicited later in the survey.

For instance, survey questions can be ordered differentially such as to generate variation in the contextual cues treated and control respondents are exposed to (Alesina et al., 2023; Roth et al., 2024a). An open-ended question can then be used to measure what is on top of respondents' minds, e.g., when taking a specific decision within the experiment. The resulting text data then allows us to estimate the "first-stage" effect of the priming intervention on attention allocation. Structured questions are less suited for this purpose, as the included response options might themselves change respondents' attention, potentially interfering with the treatment variation.

5.6 Recall

To study selective memory, researchers regularly measure free recall with the help of open-ended questions (Graeber et al., 2022; Jiang et al., 2023; Kahana, 2012). Such measures lend themselves to study the recall of information seeded in a baseline experiment. For example, Graeber et al. (2022) 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, i.e., 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 as a structured incentivized task. This suggests that unstructured open-ended elicitations are a reliable measure of recall even in the absence of incentives for accuracy. Moreover, open-ended data reveal additional nuance about the types of information being recalled and also provide the opportunity to detect memory distortions and confusion.

5.7 Information transmission

Given that most communication relies on natural language, open-ended questions also lend themselves to studying information transmission. For example, it is possible to use open-ended questions in which respondents record voice messages as a tool to study the causal impact of verbal explanations on social learning (Graeber et al., 2024a) and information transmission (Graeber et al., 2024b). Similarly, it is possible to study communication through writing in an open-text box (Grunewald et al., 2024). Such measurement is compelling as it mimics key features of communication in the real world and allows individuals to express their thoughts, feelings, and experiences in their own words, without being constrained by predefined options.

5.8 Experimenter demand effects

An important concern for most experimental work are experimenter demand effects (de Quidt et al., 2018; Zizzo, 2010). 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 (Andre et al., 2023a; Chopra et al., 2024b; Jäger et al., 2024; Roth et al., 2024a). 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. Typically, only low fractions of respondents correctly guess the experimental hypothesis in such open-ended questions, while the large majority have only vague beliefs about it.

6 Conclusion and directions for future research

This review provides an overview of techniques for the measurement of thoughts and considerations based on open-ended survey questions. Such techniques can be used to measure participants' attention to different aspects when thinking about a particular issue or when making an economic decision or prediction. Crucially, open-ended survey questions avoid priming respondents on a particular set of answer categories.

Given their wide applicability and specific advantages, we believe that open-ended measurement techniques will continue to grow in popularity. For instance, the increasing interest in better understanding the attentional foundations of belief formation and decision-making will likely spur more widespread use of these methods. We hope that this review lowers the barriers for researchers and practitioners who would like to make use of such methods.

The availability of new generative artificial intelligence technologies, such as GPT-4, offers substantial new opportunities for the analysis of unstructured data collected in open-ended questions: first, GPT-4 can improve the efficiency and accuracy of analyzing textual data by better capturing the context, semantics, and sentiment of open-ended responses than existing textual analysis tools. Second, the automation of text analysis by GPT-4 reduces the need for manual coding of open-ended responses. Third, artificial intelligence methods allow for systematic data-driven approaches to generate classification schemes.

References

Abeler, Johannes, David B. Huffmann, and Collin Raymond, "Incentive Complexity, Bounded Rationality and Effort Provision," *CESifo Working Paper*, 2023, (10541).

Agranov, Marina and Pietro Ortoleva, "Stochastic Choice and Preferences for Randomization," *Journal of Political Economy*, 2017, *125* (1), 40–68.

Akerlof, George A. and Dennis J. Snower, "Bread and Bullets," *Journal of Economic Behavior & Organization*, 2016, *126*, 58–71.

Alesina, Alberto, Armando Miano, and Stefanie Stantcheva, "Immigration and redistribution," *The Review of Economic Studies*, 2023, *90* (1), 1–39.

An, Zidong, Carola Binder, and Xuguang Simon Sheng, "Gas price expectations of Chinese households," *Energy Economics*, 2023, *120*, 106622.

Andre, Peter, "Shallow Meritocracy," *The Review of Economic Studies*, 2024. Forthcoming.

_ , Carlo Pizzinelli, Christopher Roth, and Johannes Wohlfart, "Subjective Models of the Macroeconomy: Evidence from Experts and Representative Samples," *The Review of Economic Studies*, 2021, *89(6)*, 2958–2991.

_ , Ingar Haaland, Christopher Roth, and Johannes Wohlfart, "Narratives about the Macroeconomy," *CESifo Working Paper*, 2023, (10535).

_ , Philipp Schirmer, and Johannes Wohlfart, "Mental Models of the Stock Market," *Working Paper*, 2023.

Arrieta, Gonzalo and Kirby Nielsen, "Procedural Decision-Making In The Face Of Complexity," *Working Paper*, 2024.

Ash, Elliott and Stephen Hansen, "Text Algorithms in Economics," *Annual Review of Economics*, 2023, *15*, 659–688.

_ , Germain Gauthier, and Philine Widmer, "Relatio: Text semantics capture political and economic narratives," *Political Analysis*, 2023.

Ayyar, Sreevidya, Uta Bolt, Eric French, and Cormac O'Dea, "Imagine your Life at 25: Children's Gender Attitudes and Later-Life Outcomes," *Preliminary Draft*, 2023.

Ba, Bocar A, Roman Rivera, and Alexander Whitefield, "Forecasting the Impact of Racial Uprisings, Market versus Stakeholders' Expectations," *NBER Working Paper*, 2023, (31857).

Bailey, Michael, Eduardo Dávila, Theresa Kuchler, and Johannes Stroebel, "House Price Beliefs and Mortgage Leverage Choice," *The Review of Economic Studies*, 2019, *86* (6), 2403–2452.

Baird, Sarah, Craig McIntosh, and Berk Özler, "Cash or Condition? Evidence from a Cash Transfer Experiment," *Quarterly Journal of Economics*, 2011, *126* (4), 1709–1753.

Berger, Christopher C, Tara C Dennehy, John A Bargh, and Ezequiel Morsella, "Nisbett and Wilson (1977) revisited: The little that we can know and can tell," *Social Cognition*, 2016, *34* (3), 167–195.

Bergman, Abigail, Alex Chinco, Samuel M Hartzmark, and Abigail B Sussman, "Survey Curious? Start-up Guide and Best Practices for Running Surveys and Experiments Online," *Start-Up Guide and Best Practices For Running Surveys and Experiments Online (October 05, 2020)*, 2020.

Bergman, Peter, Raj Chetty, Stefanie DeLuca, Nathaniel Hendren, Lawrence F. Katz, and Christopher Palmer, "Creating Moves to Opportunity: Experimental Evidence on Barriers to Neighborhood Choice," *American Economic Review*, 2024. Forthcoming.

Binetti, Alberti, Francesco Nuzzi, and Stefanie Stantcheva, "People's Understanding of Inflation," *Harvard University Working Paper*, 2024.

Bordalo, Pedro, John J Conlon, Nicola Gennaioli, Spencer Yongwook Kwon, and Andrei Shleifer, "How People Use Statistics," *NBER Working Paper*, 2023, (31631).

Breyer, Magdalena, Tabea Palmtag, and Delia Zollinger, "Narratives of Backlash? Perceptions of Changing Status Hierarchies in Open-Ended Survey Responses," *URPP Equality of Opportunity Discussion Paper Series*, 2023, (15).

Burgstaller, Lilith, Annabelle Doerr, and Sarah Necker, "Do Household Tax Credits Increase the Demand for Legally Provided Services?," *CESifo Working Paper*, 2023, (10211).

Bursztyn, Leonardo, Benjamin R Handel, Rafael Jimenez, and Christopher Roth, "When Product Markets Become Collective Traps: The Case of Social Media," *NBER Working Paper*, 2023, (31771).

_, Georgy Egorov, Ingar Haaland, Aakaash Rao, and Christopher Roth, "Justifying dissent," *The Quarterly Journal of Economics*, 2023, *138* (3), 1403–1451.

_, _, _, _, _ et al., "Scapegoating During Crises," *AEA Papers and Proceedings*, 2022, *112*, 151–155.

_, Ingar Haaland, Aakaash Rao, and Christopher Roth, "Disguising Prejudice: Popular Rationales as Excuses for Intolerant Expression," *NBER Working Paper*, 2020, (27288).

Bustos, Sebastián, Dina Pomeranz, Juan Carlos Suárez Serrato, José Vila-Belda, and Gabriel Zucman, "The Race Between Tax Enforcement and Tax Planning: Evidence From a Natural Experiment in Chile," *NBER Working Paper*, 2022, (30114).

Capozza, Francesco, "Beliefs about the Gender Gap in Salary Negotiation," *Working Paper*, 2024.

Cappelen, Alexander W, Astri Drange Hole, Erik Ø Sørensen, and Bertil Tungodden, "The pluralism of fairness ideals: An experimental approach," *American Economic Review*, 2007, *97* (3), 818–827.

Casarico, Alessandra, Jana Schuetz, and Silke Uebelmesser, "Gender Inequality Over the Life Cycle, Information Provision and Policy Preferences," *CESifo Working Paper*, 2024, (10916).

Castagnetti, Alessandro and Renke Schmacker, "Protecting the ego: Motivated information selection and updating," *European Economic Review*, 2022, *142*, 104007.

Chinco, Alex, Samuel M Hartzmark, and Abigail B Sussman, "A New Test of Risk Factor Relevance," *The Journal of Finance*, 2022, *77* (4), 2183–2238.

Chopra, Felix and Ingar Haaland, "Conducting qualitative interviews with AI," *CESifo Working Paper*, 2023, (10666).

_ , Christopher Roth, and Johannes Wohlfart, "Home Price Expectations and Spending: Evidence From a Field Experiment," *Working Paper*, 2024.

_ , Ingar Haaland, and Christopher Roth, "The Demand for News: Accuracy Concerns versus Belief Confirmation Motives," *Economic Journal*, 2024. Forthcoming.

Coibion, Olivier, Yuriy Gorodnichenko, and Saten Kumar, "How Do Firms Form their Expectations? New Survey Evidence," *American Economic Review*, 2018, *108* (9), 2671–2713.

Colarieti, Roberto, Pierfrancesco Mei, and Stefanie Stantcheva, "The How and Why of Household Reactions to Income Shocks," *NBER Working Paper*, 2024, (32191).

de Quidt, Jonathan, Johannes Haushofer, and Christopher Roth, "Measuring and Bounding Experimenter Demand," *American Economic Review*, 2018, *108* (11), 3266–3302.

Dechezleprêtre, Antoine, Adrien Fabre, Tobias Kruse, Bluebery Planterose, Ana Sanchez Chico, and Stefanie Stantcheva, "Fighting climate change: International attitudes toward climate policies," *NBER Working Paper*, 2022, (30265).

Dillon, Andrew, Elena Bardasi, Kathleen Beegle, and Pieter Serneels, "Explaining variation in child labor statistics," *Journal of Development Economics*, 2012, *98* (1), 136–147.

Duflo, Esther, "The Economist as Plumber," *American Economic Review*, 2017, *107* (5), 1–26.

_ , "Field Experiments and the Practice of Policy," *American Economic Review*, 2020, *110* (7), 1952–73.

_ , Michael Greenstone, Rohini Pande, and Nicholas Ryan, "Truth-telling by Third-party Auditors and the Response of Polluting Firms: Experimental Evidence from India," *The Quarterly Journal of Economics*, 2013, *128* (4), 1499–1545.

Elias, Julio J, Nicola Lacetera, and Mario Macis, "Is the Price Right? The Role of Morals, Ideology, and Tradeoff Thinking in Explaining Reactions to Price Surges," *NBER Working Paper*, 2023, (29963).

Ericsson, Anders K. and Herbert A. Simon, "Verbal Reports as Data," *Psychological Review*, 1980, *87* (3), 215–251.

— **and** — , *Protocol Analysis: Verbal Reports as Data*, MIT Press, 1993.

Fehr, Ernst and Simon Gächter, "Cooperation and punishment in public goods experiments," *American Economic Review*, 2000, *90* (4), 980–994.

Ferrario, Beatrice and Stefanie Stantcheva, "Eliciting People's First-Order Concerns: Text Analysis of Open-Ended Survey Questions," *AEA Papers and Proceedings*, 2022, *112*, 163–169.

Filippini, Massimo, Markus Leippold, and Tobias Wekhof, "Sustainable Finance Literacy and the Determinants of Sustainable Investing," *Swiss Finance Institute Research Paper*, 2021, (22-02).

Freo, Marzia and Alessandra Luati, "Lasso-based variable selection methods in text regression: the case of short texts," *AStA Advances in Statistical Analysis*, 2024, *108* (1), 69–99.

Fuster, Andreas and Basit Zafar, "Survey experiments on economic expectations," in Rüdiger Bachmann, Giorgio Topa, and Wilbert van der Klaauw, eds., *Handbook of Economic Expectations*, Academic Press, 2023, pp. 107–130.

Gabaix, Xavier, "Behavioral inattention," in B. Douglas Bernheim, Stefano DellaVigna, and David Laibson, eds., *Handbook of Behavioral Economics: Applications and Foundations 1*, Vol. 2, Elsevier, 2019, pp. 261–343.

Galasso, Vincenzo, Massimo Morelli, Tommaso Nannicini, and Piero Stanig, "The Populist Dynamic: Experimental Evidence on the Effects of Countering Populism," *CESifo Working Paper*, 2024, (10949).

Geer, John G, "What do open-ended questions measure?," *Public Opinion Quarterly*, 1988, *52* (3), 365–371.

— , "Do open-ended questions measure "salient" issues?," *Public Opinion Quarterly*, 1991, *55* (3), 360–370.

Gentzkow, Matthew and Jesse M Shapiro, "What Drives Media Slant? Evidence From U.S. Daily Newspapers," *Econometrica*, 2010, *78* (1), 35–71.

__ , Bryan Kelly, and Matt Taddy, "Text as data," *Journal of Economic Literature*, 2019, *57* (3), 535–74.

Gilardi, Fabrizio, Meysam Alizadeh, and Maël Kubli, "ChatGPT outperforms crowd workers for text-annotation tasks," *Proceedings of the National Academy of Sciences*, 2023, *120* (30), e2305016120.

Graeber, Thomas, Christopher Roth, and Constantin Schesch, "Explanations," *Working Paper*, 2024.

__ , __ , and Florian Zimmermann, "Stories, Statistics, and Memory," *CESifo Working Paper*, 2022, (10107).

__ , Shakked Noy, and Christopher Roth, "Lost in Transmission," *Working Paper*, 2024.

Gründler, Klaus and Niklas Potrafke, "Experts and Epidemics," *CESifo Working Paper*, 2020, (8556).

Grunewald, Andreas, Victor Klockmann, Alicia von Schenk, and Ferdinand von Siemens, "Are biases contagious? The influence of communication on motivated beliefs," *Würzburg Economic Papers*, 2024, (109).

Haaland, Ingar, Christopher Roth, and Johannes Wohlfart, "Designing Information Provision Experiments," *Journal of Economic Literature*, 2023, *61* (1), 3–40.

Hager, Anselm, Lukas Hensel, Christopher Roth, and Andreas Stegmann, "Voice and Political Engagement: Evidence from a Field Experiment," *The Review of Economics and Statistics*, 2023, pp. 1–34.

__ , __ , Johannes Hermle, and Christopher Roth, "Political Activists as Free Riders: Evidence from a Natural Field Experiment," *The Economic Journal*, 2023, *133* (653), 2068–2084.

Hommes, Cars, Julien Pinter, and Isabelle Salle, "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*, 2023, (10574).

Houde, Sébastien and Tobias Wekhof, "Using narratives to infer preferences in understanding the energy efficiency gap," *Nature Energy*, 2023, *8*, 965–977.

Hruschka, D. J., D. Schwartz, D.C. St John, E. Picone-Decaro, R. A. Jenkins, and J. W. Carey, "Reliability in coding open-ended data: Lessons learned from HIV behavioral research," *Field Methods*, 2004, *16*, 307–331.

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

Imas, Alex, Michael A Kuhn, and Vera Mironova, "Waiting to choose: The role of deliberation in intertemporal choice," *American Economic Journal: Microeconomics*, 2022, *14* (3), 414–440.

Jäger, Simon, Christopher Roth, Nina Roussille, and Benjamin Schoefer, "Worker Beliefs About Outside Options," *Quarterly Journal of Economics*, 2024.

Jayachandran, Seema, Monica Biradavolu, and Jan Cooper, "Using machine learning and qualitative interviews to design a five-question survey module for women's agency," *World Development*, 2023, *161*, 106076.

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

Jiang, Zhengyang, Hongqi Liu, Cameron Peng, and Hongjun Yan, "Investor Memory and Biased Beliefs: Evidence from the Field," *Working Paper*, 2023.

Kahana, Michael Jacob, *Foundations of human memory*, OUP USA, 2012.

Kahneman, Daniel, Jack L Knetsch, and Richard Thaler, "Fairness as a constraint on profit seeking: Entitlements in the market," *The American Economic Review*, 1986, pp. 728–741.

Kaufmann, Marc, Peter Andre, and Botond Kőszegi, "Understanding Markets with Socially Responsible Consumers," *The Quarterly Journal of Economics*, 2024. Forthcoming.

Kaur, Supreet, Sendhil Mullainathan, Suanna Oh, and Frank Schilbach, "Do Financial Concerns Make Workers Less Productive?," *NBER Working Paper*, 2021, (28338).

Krosnick, Jon A, "Survey research," *Annual Review of Psychology*, 1999, *50* (1), 537–567.

Kuziemko, Ilyana, Michael I Norton, Emmanuel Saez, and Stefanie Stantcheva, "How Elastic are Preferences for Redistribution? Evidence from Randomized Survey Experiments," *American Economic Review*, 2015, *105* (4), 1478–1508.

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

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

Laudenbach, Christine, Ruediger Weber, Annika Weber, and Johannes Wohlfart, "Beliefs About the Stock Market and Investment Choices: Evidence from a Field Experiment," *CEBI Working Paper*, 2022.

Lazarsfeld, Paul F, "The controversy over detailed interviews—an offer for negotiation," *Public Opinion Quarterly*, 1944, *8* (1), 38–60.

Leiser, David and Shelly Drori, "Naïve understanding of inflation," *Journal of Socio-Economics*, 2005, *34* (2), 179–198.

Link, Sebastian, Andreas Peichl, Christopher Roth, and Johannes Wohlfart, "Attention to the Macroeconomy," *CESifo Working Paper*, 2023, (No. 10858).

Liscow, Zachary and Edward Fox, "The psychology of taxing capital income: Evidence from a survey experiment on the realization rule," *Journal of Public Economics*, 2022, *213*, 104714.

Liu, Haoyang and Christopher Palmer, "Implicit Extrapolation and the Beliefs Channel of Investment Demand," *Working Paper*, 2023.

Lombard, M., J. Snyder-Duch, and C. Bracken, "Content Analysis in Mass Communication: Assessment and Reporting of Intercoder Reliability," *Human Communication Research*, 2002, *28*, 587–604.

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

Manski, Charles F, "Measuring Expectations," *Econometrica*, 2004, *72* (5), 1329–1376.

Martínez-Marquina, Alejandro, Muriel Niederle, and Emanuel Vespa, "Failures in Contingent Reasoning: The Role of Uncertainty," *American Economic Review*, 2019, *109* (10), 3437–74.

Miano, Armando, "Search Costs, Outside Options, and On-the-Job Search," *Working Paper*, 2023.

Miles, B. and A. Huberman, *Qualitative data analysis: An expanded sourcebook*, Sage, 1994.

Morris, Adam, Ryan W Carlson, Hedy Kober, and Molly Crockett, "Introspective access to value-based choice processes," *Working Paper*, 2023.

Nathan, Brad C, Ricardo Perez-Truglia, and Alejandro Zentner, "My Taxes are Too Darn High: Why Do Households Protest their Taxes?," *NBER Working Paper*, 2023, (27816).

Neuendorf, K., *The content analysis guidebook*, Sage, 2002.

Nisbett, Richard E and Timothy D Wilson, "Telling more than we can know: Verbal reports on mental processes.," *Psychological Review*, 1977, *84* (3), 231.

O'Connor, Cliodhna and Helene Joffe, "Intercoder reliability in qualitative research: Debates and practical guidelines," *International Journal of Qualitative Methods*, 2020, *19.*

Oh, Suanna, "Does Identity Affect Labor Supply?," *American Economic Review*, 2023, *113* (8), 2055–83.

Parker, Barbara and Valerie Kozel, "Understanding poverty and vulnerability in India's Uttar Pradesh and Bihar: A Q-squared approach," *World Development*, 2007, *35* (2), 296–311.

Pennington, Nancy and Reid Hastie, "Explaining the Evidence: Tests of the Story Model for Juror Decision Making," *Journal of Personality and Social Psychology*, 1992, *62* (2), 189–206.

Popping, Roel, "Analyzing Open-ended Questions by Means of Text Analysis Procedures," *Bulletin of Sociological Methodology/Bulletin de Méthodologie Sociologique*, 2015, *128* (1), 23–39.

Porter, Martin F, "An algorithm for suffix stripping.," *Program*, 1980, *14* (3), 130–137.

Rathje, Steve, Dan-Mircea Mirea, Ilia Sucholutsky, Raja Marjieh, Claire Robertson, and Jay J. Van Bavel, "GPT Is an Effective Tool for Multilingual Psychological Text Analysis," *PsyArXiv*, May 2023.

Roberts, Margaret E, Brandon M Stewart, Dustin Tingley, Christopher Lucas, Jetson Leder-Luis, Shana Kushner Gadarian, Bethany Albertson, and David G Rand, "Structural Topic Models for Open-Ended Survey Responses," *American Journal of Political Science*, 2014, *58* (4), 1064–1082.

Rodrik, Dani and Stefanie Stantcheva, "A Policy Matrix for Inclusive Prosperity," *NBER Working Paper*, 2021, (28736).

Romero, Mauricio, Juan Bedoya, Monica Yanez-Pagans, Marcela Silveyra, and Rafael De Hoyos, "Direct vs indirect management training: Experimental evidence from schools in Mexico," *Journal of Development Economics*, 2022, *154*, 102779.

Roth, Christopher, Peter Schwardmann, and Egon Tripodi, "Depression Stigma," *Working Paper*, 2024.

_, _, and _, "Misperceived Effectiveness and the Demand for Psychotherapy," *Working Paper*, 2024.

Rothschild, Jacob E, Adam J Howat, Richard M Shafranek, and Ethan C Busby, "Pigeonholing partisans: Stereotypes of party supporters and partisan polarization," *Political Behavior*, 2019, *41*, 423–443.

Saccardo, Silvia and Marta Serra-Garcia, "Enabling or Limiting Cognitive Flexibility? Evidence of Demand for Moral Commitment," *American Economic Review*, 2023, *113* (2), 396–429.

Shiller, Robert J, "Narrative economics," *American economic review*, 2017, *107* (4), 967–1004.

Sloman, Steven A. and David Lagnado, "Causality in Thought," *Annual Review of Psychology*, 2015, *66*, 223–247.

Stantcheva, Stefanie, "Understanding Economic Policies: What do People Know and How Can they Learn?," *Harvard University Working Paper*, 2020.

_ , "Understanding tax policy: How do People Reason?," *Quarterly Journal of Economics*, 2021, *136* (4), 2309–2369.

_ , "How to Run Surveys: A guide to creating your identifying variation and revealing the invisible," *Annual Review of Economics*, 2022, *15.*

_ , "Understanding of Trade," *NBER Working Paper*, 2022, (30040).

_ , "Why Do We Dislike Inflation?," *Brookings Papers on Economic Activity*, 2024. Forthcoming.

Trabasso, Tom and Paul van den Broek, "Causal Thinking and the Representation of Narrative Events," *Journal of Memory and Language*, 1985, *24* (5), 612–630.

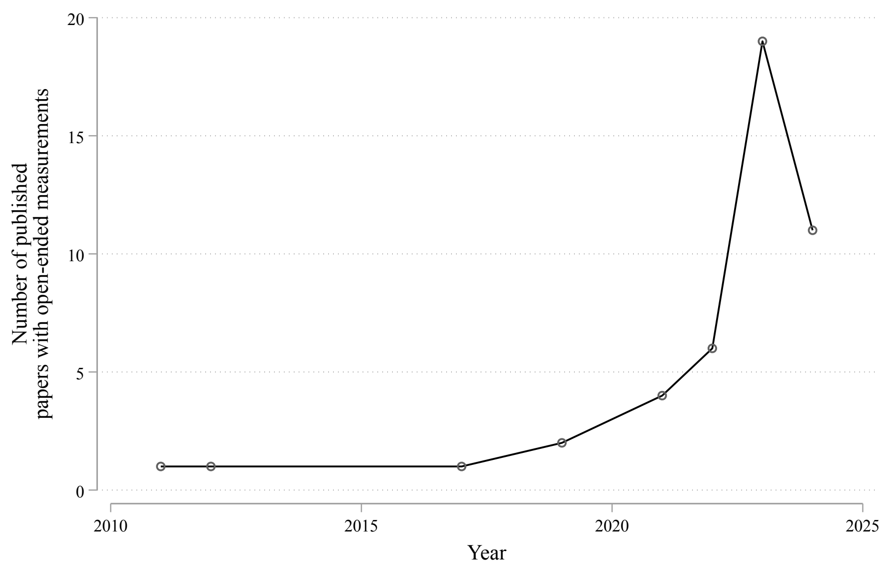
Tversky, Amos and Daniel Kahneman, "Judgment Under Uncertainty: Heuristics and Biases," *Science*, 1974, *185* (4157), 1124–1131.

Wekhof, Tobias, "Conditional Topic Allocations for Open-Ended Survey Responses," *Working Paper*, 2024.

Zizzo, Daniel John, "Experimenter Demand Effects in Economic Experiments," *Experimental Economics*, 2010, *13* (1), 75–98.

Zollinger, Delia, "Cleavage Identities in Voters' Own Words: Harnessing Open-Ended Survey Responses," *American Journal of Political Science*, 2022, pp. 1–48.

Figure 1: Number of studies with open-ended measurements published in leading journals and working paper series since 2014



Notes: This figure shows the number of studies with open-ended measurements published in leading journals since 2010. For 2024, publications and forthcoming papers as of mid-March are included. The figure is based on publications in leading journals (American Economic Review, American Economic Journal: Applied Economics, American Economic Journal: Economic Policy, American Economic Journal: Macroeconomics, Econometrica, 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, the Review of Economic Studies), the AEA Papers and Proceedings, and worker paper series (CEPR, CESifo, and NBER). To identify articles, we used Google Scholar to search for all articles published in these journals since 2010 containing the words "open-ended", "open and then verified which of the search results featured an actual open-ended measurement. We also supplemented with papers covered in our review that were not captured using this search algorithm.

Table 1: Overview of studies: Political economy

Paper name	**Domain**	**Measurement**	**Analysis of text data**
Andre (2024)	Explanation of people's distributive choices	"Please explain why you made your choice the way you did."	Hand-coding of responses.
Bursztyn et al. (2023b)	Motives for choosing Tweet	"Why did you choose this Tweet rather than the other Tweet?"	Word counts and simple machine learning techniques (Gentzkow and Shapiro, 2010).
Bursztyn et al. (2020)	Beliefs about motives underlying xenophobic expression	"Why do you think your matched respondent chose to donate to Fund the Wall?" On top of this, the authors employ structured measures of beliefs about the matched respondent's type.	Pre-specified word counting procedure; Support Vector Machine classifier to predict structured belief measures based on text data.
Dechezleprêtre et al. (2022)	Considerations about climate change	"When thinking about climate change, what are your main considerations? What should [country] government do regarding climate change?"	Text analysis.
Galasso et al. (2024)	Anti-populist video ads regarding a populist referendum in Italy	Respondents were invited to share their thoughts about a video in an open-ended question to compare how two videos are comparatively perceived.	Hand-coding and supervised text analysis (FEEL-IT)
Gründler and Potrafke (2020)	Attitudes towards fiscal rules	"What are your main considerations about fiscal rules?"; "What should be the goal of fiscal rules?"; "What are the main shortcomings of fiscal rules."	Word cloud and ML methods for sentiment analysis.
Hager et al. (2023a)	Voice and political engagement	4 Treatment groups with open-ended questions:"Would you like to tell us more about which issues we should particularly emphasize in the election campaign?(/Would you like to tell us more about which topics are particularly close to your heart?)(/+ After the completion of the survey, we will send you a summary of the results.)"	Hand-coding of responses.
Hager et al. (2023b)	Motives underlying change in effort in response to info	"Why would the results of the survey affect or not affect your decision? Please answer using whole sentences"	Human coding of scripts.
Hüning et al. (2022)	Attitudes towards rent control	Discuss pro and cons of rent control	NLP and text analysis techniques.
Jessen et al. (2024)	Policy demand to reduce socioeconomic inequality in life expectancy	List as many measures as possible that the government could use to improve the life expectancy of the poor.	Word cloud.
König and Schmacker (2022)	Sin taxes, i.e., taxes on sugar-sweetened beverages (SSB)	4 Open-ended questions: 1st the general considerations of SSB taxes; 2nd regarding the goals; 3rd and 4th the benefiting respective losing groups.	Word cloud and keyness analysis.
Lang and Schneider (2023)	Influence of post-WWII German immigrant movement on nationalist sentiment and electoral responses	"What do you think is the significance of the fact that many Germans had experience of expulsion, flight, and immigration?"	Hand-coding of responses.
Liscow and Fox (2022)	Attitudes towards capital tax realization rule	Why preferred to defer taxation until sold respectively why preferred to tax before sold.	Word counting.
Luttmer and Samwick (2018)	Impact of policy uncertainty in social security on individual welfare	"We are interested in better understanding why you chose uncertain benefits around [B]% of the Social Security benefits you are supposed to get under current law over guaranteed benefits equal to [L]% of the Social Security benefits you are supposed to get under current law. Could you tell us the main reason for your choice?"	Hand-coding of responses.
Nathan et al. (2023)	Reasoning for filling property tax complain or not.	"If you can, please explain why you will (or will not) protest in 2020."	Handcoding of responses.
Stantcheva (2020)	Understanding, reason and learning about 4 economic policies: i) income, and ii) estate taxation, iii) health insurance, iv) trade	"What are your main considerations about [policy]...?" and more specific subquestions regarding perceived goals and shortcomings, as well as the anticipated effects (e.g., which group would gain) from the specific policy.	Text analysis techniques (keyness analysis, topic analysis, word clouds).
Stantcheva (2021)	First-order concerns about income and estate tax	"When you think about federal personal income taxation and whether the U.S. should have higher or lower federal personal income taxes (/federal estate tax), what are the main considerations that come to your mind?"	Text analysis techniques (keyness analysis, topic analysis, word clouds).
Stantcheva (2022b)	Attitudes towards trade	"When you think about trade policy and whether the U.S. should put some restrictions on trade with other countries, such as tariffs, what are the main considerations that come to your mind?" "What would be the effects on the U.S. economy if barriers to trade, such as tariffs, were increased?" "Which groups of people do you think would gain if trade barriers such as tariffs were increased?"	Text analysis techniques (keyness analysis, topic analysis, word clouds)

Table 2: Overview of studies: Political science

Paper name	Domain	Measurement	Analysis of text data
Breyer et al. (2023)	Perceived status gains for women or minorities.	"E.g., Now think about the people who have tended to gain recognition compared to the past. How would you describe these people? What kind of characteristics, lifestyles, and opinions do these people have?"	Text analysis with a parsimonious dictionary.
Roberts et al. (2014)	Views on immigration; Intuition versus reflection in Public Goods Game; American National Election Survey	"Of the stories you read, what stories do you remember best? (If you don't remember the names, just describe the stories)."; "Please write a paragraph (approximately 8-10 sentences) describing a time your intuition/first instinct(/time carefully reasoning) led you in the wrong(/right) direction and resulted in a bad(/good) outcome."; "What has been the most important issue to you personally in this election?" and "What do you think is the most important political problem facing the United States today?"	Introduce their Structural Topic Model (STM), which relies on machine learning methods, and apply it to the three examples and compare it to hand-coding.
Rothschild et al. (2019)	Stereotypes about the two American parties	"Please write down four words that typically describe people who support the [Democratic/Republican] Party."	Structural Topic modelling.
Zollinger (2022)	Attitudes towards voter-party links	"If you imagine people with a lifestyle and opinions similar to your own, what kind of people would these be? How would you describe them?", "And someone who is not at all like you? Someone who lives completely differently and who has very different opinions? How would you describe them?"	Text analysis techniques (keyness analysis, latent semantic scaling)

Table 3: Overview of studies: Development economics and other areas

Paper name	Domain	Measurement	Analysis of text data
Baird et al. (2011)	Role of conditionality in cash transfers	Qualitative interview of random subsample.	Hand-coding of responses.
Burgstaller et al. (2023)	Tax credits influence on demand for legally provided services	"What reasons could there be for someone not claiming the government support?"	Hand-coding of topics which are used for keyness analysis.
Dillon et al. (2012)	Child labor	In the pilot phase, qualitative interviews with open-ended questions were conducted to solicit how respondents thought about the survey questions, why they chose the responses they did, and how they thought about concepts such as work, household production, and their primary activities.	Hand-coding of responses.
Houde and Wekhof (2023)	Investment in energy efficiency	"Describe the reasons why you decided (not) to carry out energy efficiency retrofits. Please write a short text of about 4 sentences."	Semi-manual classification validated with hand-coding and machine learning methods.
Jayachandran et al. (2023)	Woman's agency	Semi-structured interview with open-ended questions.	Hand-coding of responses to calculate benchmark score.
Parker and Kozel (2007)	Poverty and vulnerability in India	'Semi-structured interview with open-ended questions.	Qualitative analysis methods are used to inform a quantitative survey.
Romero et al. (2022)	Direct vs. indirect management training	DWMS, an adaptation of the World Management Survey, was used for an interview that included 23 open-ended questions, such as " How do you keep track of what teachers are doing in the classrooms?"	Hand-coding of responses.

Table 4: Overview of studies: Macroeconomics

Paper name	Domain	Measurement	Analysis of text data
An et al. (2023)	Gas price and inflation expectations	Asked respondents to describe the main considerations that come to their mind regarding the impact of the war on China's economy, overall prices in China, and gas prices in China.	Word cloud and hand-coding.
Andre et al. (2021)	Unemployment and inflation predictions	Ask respondents about their "main considerations in making the prediction" and about how they "come up with [their] prediction". On top of this, the authors employ structured measures of the considerations respondents had on their mind.	Human coding of scripts and simple word counting techniques.
Andre et al. (2023a)	Narratives about the rise of inflation	Ask respondents "Which factors caused the rise in inflation?"	Human coding of text into DAGs.
Binetti et al. (2024)	placeholder	placeholder	placeholder
Hommes et al. (2023)	Prior Knowledge of Public Finance	"Which risk(s) do you have in mind?" or "Which advantage(s) do you have in mind?"	Word cloud and classification.
Leiser and Drori (2005)	Inflation expectations	Ask participants to specify terms, concepts, or short phrases related to inflation.	Human coding of text and classification into broader categories.
Link et al. (2023)	Current economic situation	"What topics come to mind when you think about the economic situation of your company/household?"	Human coding of scripts and word counting.
Stantcheva (2024)	Causes and personal impact of inflation	E.g., "What were the most important impacts of inflation on your life?"; "When inflation gets very high, what do you think is the reason?"	Topic analysis and word clouds.

Table 5: Overview of studies: Labor economics

Paper name	Domain	Measurement	Analysis of text data
Abeler et al. (2023)	Incentive complexity and effort provision	"If someone were trying to get the most money, total, from [Period 3 and Period 4], what do you think would be the best approach?"	Hand-coding of responses.
Ayyar et al. (2023)	Gender attitudes influence on lifetime earnings.	"Imagine you are now 25 years old. Write about the life you are leading, your interests, your home life and your work at the age of 25. (You have 30 minutes to do so)."	Word-embedding model to identify gender attitudes in essays.
Capozza (2024)	Concerns of women regarding gender gap in salary negotiations	"Which factors do you think caused the gender gap in salary negotiation?"	Word cloud, keyness analysis and hand-coding.
Casarico et al. (2024)	Causes of gender gap in earnings and pensions in Germany	"What do you think are the causes of the differences between men and women in gross annual earnings and retirement pensions?"	Word cloud and keyness analysis.
Kaur et al. (2021)	Financial worries	"What makes you worry about money issues?"	Word clouds and text-counting.
Miano (2023)	Beliefs about on-the-job search	"Imagine you wanted to look for a new job at a new employer now, while still working at your current employer. Are there any issues that would make looking for a new job difficult for you now? What are the first ones that come to your mind?""	Word cloud.
Oh (2023)	Labor supply decisions related to caste	During the follow-up survey, workers were asked why they turned down specific offers.	Surveyors classified free-form answers based on training.
Rodrik and Stantcheva (2021)	Beliefs about what makes a good job	"What is a good job?"	Text analysis techniques

Table 6: Overview of studies: Finance

Paper name	Domain	Measurement	Analysis of text data
Andre et al. (2023b)	Prediction of Stock Market Returns	Ask respondents for their reasoning for their stock return predictions based on a pair of hypothetical scenarios involving stale news about future company earnings.	Word count and hand-coding of open-ended data.
Ba et al. (2023)	Forecasting the stock market impact of racial uprisings	“Please explain your prediction using 2 to 3 sentences.”	Hand-coding of responses.
Bailey et al. (2019)	Mortgage leverage choice	Ask respondents why their mortgage leverage choice differs across hypothetical scenarios with different projected home price changes.	Overview of representative text responses
Chinco et al. (2022)	Stock investment decisions	Ask respondents what factors are most important to them when deciding what fraction of an endowment to invest in stocks	Self-classification of open-ended responses by survey participants.
Chopra and Haaland (2023)	Stock market non-participation puzzle	Conduct AI-assisted interviews with respondents, exploring their reasons for not investing in the stock market including an "what if" scenario and counterfactual reasoning.	Hand-coding 50 interviews and using these for OpenAI’s API to query GPT-4 for code assignment.
Chopra et al. (2024a)	Home Price Expectations	How would this change in your expectations about future home prices affect your expectations about your household’s future economic situation. Please explain why. Respond in full sentences.	Hand-coding of open-ended data.
Filippini et al. (2021)	Financial literacy	“Describe which characteristics you think distinguish sustainable financial products from conventional investments. Please write a short text of about three sentences”	Text analysis techniques (topic specific word counts)
Jiang et al. (2023)	Selective recall of past returns	“First think about the overall stock market movement since you opened an account. Since you started trading, what is the episode of market movement that first comes to mind? Please enter the starting month and ending month of this episode.”	Hand-coding of dates.
Laudenbach et al. (2022)	Beliefs about the stock market	Ask respondents to describe in open text which specific historical episodes – if any – they had in mind when estimating the historical autocorrelation of aggregate stock returns.	Human coding of text responses into different historical episodes.

Table 7: Overview of studies: Behavioral economics

Paper name	Domain	Measurement	Analysis of text data
Arrieta and Nielsen (2024)	Explanation of people's lottery and charity choices	"Please write a message to another participant describing how you made your last five decisions."	Other respondents replicate choice with or without message, Robustness with GPT-4 which also classified text in procedural categories.
Bordalo et al. (2023)	Explanation of people's solving strategies for statistical problems	"Could you describe to us in your own words how you came up with your answer to the previous question?"	Classification with GPT-3.5 which specific features of the problem were attended to.
Bursztyn et al. (2023a)	Motives for preferring a world without Tiktok or Instagram and feelings about being the sole user to leave the platform	"You mentioned you would prefer to live in a world without [platform]. Why do you still use it?" and "How would you feel if you were the only one who deactivated [platform] and everyone else kept using it?".	Hand-coding of responses and AI based classification using GPT-4.
Castagnetti and Schmacker (2022)	Motivated information selection and updating	"Please explain, in general, how you decided between feedback modes across the five scenarios. For example, why did you choose one feedback mode over another? What specific characteristics of the feedback modes were you looking at?"	Hand-coding of responses.
Chopra et al. (2024b)	Motives for (not) subscribing to newsletter	"Why did you (not) subscribe to the newsletter?"	Word counts and simple machine learning techniques (Gentzkow and Shapiro, 2010).
Agranov and Ortoleva (2017)	Motives for choices between lotteries	"In Part III of the experiment each question was asked to you three times. If you chose different options, could you please tell us why did you do it? (Please elaborate)."	Hand-coding of responses.
Elias et al. (2023)	Attitudes towards sudden price increases and price regulation	'Using the slider below, please rate this scenario as: -10 (completely unfair) to +10 (completely fair)','We now ask you to select, among the two scenarios described above, the one that you would prefer to have in place in your country.','Please briefly describe in the space provided the main reason(s) for your answers and choice above'	Text analysis.
Graeber et al. (2022)	Memories about information provided	"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, i.e., the first thought first, then the next one etc."	Hand-coding of responses.
Graeber et al. (2024a)	Verbal explanations of financial reasoning choices	"We are interested in how you would give advice in an informal conversation: You should share an explanation behind your response. Your recording will be played to a few other participants who will have to respond to the same question."	Transcribe transcripts by Phonic using Amazon Transcribe and GPT-4 for classification of transcribed text.
Graeber et al. (2024b)	Oral transmission of information using speech recordings	"Think about the first(/ or second) opinion you listened to about changes in house price growth in a large US city. We will now ask you to record a voice message summarizing this opinion."	Classify by hand-coding and GPT of responses whether level and reliability are transmitted.
Grunewald et al. (2024)	Potential reinforcement of motivated beliefs through ccommunication.	"[Quotes & Chat] To start the conversation and to give you some food for thought, here are two quotes by famous personalities: I think we are living in selfish times. — Javier Bardem, Hollywood actor and Oscar winner I'm just thankful I'm surrounded by good people. — Jon Pardi, singer and songwriter"	Word lists and bigram and trigram analysis.
Kaufmann et al. (2024)	Explanation of people's impact prediction on externalities	"Please explain why you chose this response.", "Please explain why you gave the same(/different) answer(/s) in the two situations." and "Please explain why you would be willing to pay money in situation 2 where the total impact would be zero."	Hand-coding of responses.
Martínez-Marquina et al. (2019)	Provide incentivized advice to another participant for making a guess	"In the box below you can provide advice on what price you think the advisee should submit and a justification for your recommendation."	Hand-coding of responses.
Roth et al. (2024a)	Social Stigma and demand for psychotherapy	"Imagine a person with depression. What views about depressed people by others does this person worry about most?"	Hand-coding of responses.
Roth et al. (2024b)	Effectiveness and demand for psychotherapy	"What considerations do you have on your mind when choosing how much you would be willing to spend on 4 weeks of online therapy from BetterHelp? Please write 2-3 sentences. You may mention both downsides and benefits of buying therapy (if any were on your mind)."	Hand-coding of responses.
Saccardo and Serra-Garcia (2023)	Enable or limit their capacity to distort beliefs in moral dilemmas	"When you had to decide between learning about your commission Before or After getting information about the quality of Product B [A, if the order was flipped], how did you make this decision?"	Hand-coding of responses.