Title
Essays on Economics of Beliefs

Author
Galashin, Mikhail

Publication Date
2023

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Essays on Economics of Beliefs

A dissertation submitted in partial satisfaction

of the requirements for the degree

Doctor of Philosophy in Management

by

Mikhail Galashin

2023

ABSTRACT OF THE DISSERTATION

Essays on Economics of Beliefs

by

Mikhail Galashin
Doctor of Philosophy in Management
University of California, Los Angeles, 2023
Professor Melanie Sharon Wasserman, Co-Chair
Professor Ricardo Perez-Truglia, Co-Chair

This dissertation studies the effects and elicitation of beliefs. In the first chapter, we estimate the effects of macroeconomic expectations on consumer decisions. We examine this question using an experiment with 2,872 credit card customers at a large commercial bank. We provide participants with expert forecasts of inflation and the nominal exchange rate and measure the consumption response to this information using detailed data on individual credit card transactions. We find that forecasts strongly affect inflation and exchange rate expectations, but do not change spending or self-reported consumption plans as predicted by standard models of intertemporal choice. Results from a supplementary survey experiment suggest that consumers are sophisticated enough to anticipate nominal rigidities that lower expected real income and reduce spending on durables for precautionary reasons, counteracting the effects predicted by standard models of intertemporal optimization. The absence of a link between consumer expectations and behavior has potentially important implications for macroeconomic policies such as forward guidance.

The counter-intuitive results motivate the development of more flexible methods of belief

elicitation, which could facilitate understanding of the subjects' decision-making process. While the elicitation of numerical variables, such as inflation, is well understood, certain beliefs, such as action plans during an inflation hike, cannot be represented numerically and require verbal elicitation. Verbal elicitation requires the researcher either to use open-ended questions or to know the most important answers in advance. We propose a method to crowdsource potential answers to open-ended questions. This ensures that the survey is adaptive and does not miss important answers while maintaining a low-cost, multiple-choice format. We propose two measures of information loss to evaluate the quality of multiple-choice questions: the semantic similarity of selected answers to the open-ended answers to the same question and the probability of selecting any answer from the list. We conduct an experiment to examine the impact of monetary incentives and characteristics of respondents on the quality of crowdsourced answers. Our findings indicate that incentives can increase quality and effort, but the effects are relatively small compared to the variation across respondents. We observe that option authors' similarity to the respondents in beliefs, political views, and demographics, explains a large fraction of the variation in quality. These results imply that sample selection is likely to be more important for crowdsourcing hypotheses than incentive design.

The third chapter outlines a research agenda extending the work on the verbal elicitation techniques. In particular, we focus on the potential of large language models (LLMs) to improve belief elicitation techniques with natural language. We review the current state of belief elicitation in economics. Next, we introduce the main architectures and relevant fine-tuning techniques for LLMs. Lastly, we discuss the potential applications of LLMs to belief elicitation and the utilization of beliefs in empirical work. This includes using LLMs for representing beliefs numerically, interacting with subjects during belief elicitation, and generating hypotheses regarding how beliefs affect actions.

The dissertation of Mikhail Galashin is approved.

Paola Giuliano

Nico Voigtländer

Romain T. Wacziarg

Ricardo Perez-Truglia, Committee Co-Chair

Melanie Sharon Wasserman, Committee Co-Chair

University of California, Los Angeles

2023

To my parents

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

It is hard to express in words my gratitude to my advisors: Paola Giuliano, Ricardo Perez-Truglia, Nico Voigtländer, Romain Wacziarg, and Melanie Wasserman. Their invaluable guidance, unwavering support, and consistent encouragement have been indispensable.

I have also immensely benefited from advice from other faculty members including Kathleen Ngangoué, Clemence Tricaud, Charlie Sprenger, Kirby Nielsen, Isabelle Brocas, Graeme Blair, and Jerry Nickelsburg. My early work and later conversations with Leonardo Bursztyn were particularly significant. I am thankful to my earlier mentors, Sergey V. Popov, Ruben Enikolopov, and Maria Petrova, for their lasting support and advice.

I am deeply grateful to my co-authors, Alena Buinskaya, Yury Olshanskiy, and Martin Kanz, from whom I learned so much. Chapter 1 of this dissertation is based on our joint work with Martin Kanz and Ricardo Perez-Truglia. Chapter 2 is derived from our work with Alena Buinskaya.

I am profoundly thankful to my peers and friends including Vasily Korovkin, Olga Novikova, Aleksandr Gevorkian, Zach Sauers, Matias Giaccobasso, Asya Tsaturyan, Sebastian Ottinger, Yangkeun Yun, Ekaterina Gurkova, Arseniy Samsonov, Vitaly Titov, Bruno Pellegrino, Lidia Kosenkova, Ksenia Shakhgildyan, Stefano Fiorin, Mikhail Poyker, Logan Spencer, Elisabetta Campagna, Anh Nguyen, Artem Timoshenko, Anton Sobolev, Maxim Ananyev, Ekaterina Volkova, and all participants of the experimental meetings at Caltech. I am especially grateful to Natalia Lamberova.

Finally, I gratefully acknowledge financial support from Harold and Pauline Price Center for Entrepreneurship & Innovation. I am particularly grateful to Jennifer McMillen for her assistance in navigating all the administrative issues with the Center.

VITA

2009–2013	Bachelor of Science in Economics International College of Economics and Finance, National Research University Higher School of Economics, Moscow
2013–2015	Master of Arts in Economics New Economic School, Moscow
2015–2023	Ph.D. Candidate in Management UCLA Anderson School of Management
2016-2019	Research Assistant to Prof. Ricardo Perez-Truglia Global Economics and Management Department, UCLA Anderson School of Management.
2017–2023	Teaching Assistant, Global Economics and Management Department UCLA Anderson School of Management

CHAPTER 1

Macroeconomic Expectations and Credit Card Spending

1.1 Introduction

Models of macroeconomics and household finance that use the Euler equation to microfound individual behavior typically assume that the savings and consumption choices of households respond directly to changes in macroeconomic expectations. This notion is so deeply ingrained in economic thought that it is often taken for granted. However, surprisingly little causal evidence exists on the effect of macroeconomic expectations on consumer behavior.

Recent insights from behavioral economics provide several reasons why the link between macroeconomic expectations and household behavior could be more tenuous than is generally assumed. The typical household may, for example, not be sophisticated enough to understand how to optimally revise consumption plans in response to changed macroeconomic expectations, make inference mistakes that prevent optimal intertemporal substitution, or be uncertain about how to interpret information about macroeconomic events (Andre et al., 2022; Gabaix and Laibson, 2006). Macroeconomic expectations may also not be sufficiently salient, especially when consumers make spending decisions involving small transaction amounts. The resulting failure of consumers to factor macroeconomic expectations into their decisions could have far-reaching consequences for the impact of various macroeconomic policies, such as forward guidance, that are explicitly based on the premise that changes in expectations will affect consumer behavior and real economic activity (Bernanke, 2007; Coibion et al., 2023).[1]

In this paper, we provide novel evidence on the causal effect of macroeconomic expectations on consumption decisions. We conduct an information-provision experiment with credit card customers of a large commercial bank in an emerging market and focus on two of the macroeconomic variables that arguably receive the most attention in models of household finance and macroeconomics: the inflation rate and the foreign exchange rate. In the experiment, we provide credit card customers with randomized expert forecasts of inflation and the nominal exchange rate and examine how this information affects macroeconomic expectations, spending plans, and actual consumption decisions, observed in detailed transaction-level data on credit card spending among experimental participants.

We collaborate with the consumer finance division of a large Malaysian bank and integrate a randomized experiment into the bank's standard customer communications. The experiment was implemented as part of a phone survey with the partner bank's credit card customers and proceeded in four steps. The survey first elicited respondents' exchange rate and inflation expectations. Second, randomly chosen subsets of the sample population were provided with expert forecasts of inflation, the exchange rate, or both. Third, we elicited participants' posterior beliefs and self-reported spending plans, using questions similar to those employed in the elicitation of prior beliefs. Finally, we merged results of the survey experiment with comprehensive transaction-level data on credit card spending provided by the partner bank. This allows us to examine the impact of macroeconomic expectations on actual consumption behavior, observed free of measurement error in credit card transaction data.

The setting and customer population we use in our experiment have several advantages

[1] The Federal Reserve, for example, explains on its website that "when central banks provide forward guidance, individuals and businesses will use this information in making decisions about spending and investments. Thus, forward guidance about future policy can influence financial and economic conditions today." See `www.federalreserve.gov/faqs`

that help us explore the link between macroeconomic expectations and economic decisions. Macroeconomic trends in Malaysia are quite representative of many small open economies and, as such, provide an interesting contrast between the evolution of inflation and exchange rates. On the one hand, the inflation rate has been stable at low levels over the past decades: since 2005, the inflation rate has hovered between 1% and 3% per year. The nominal exchange rate, on the other hand, has been highly volatile, with two-digit depreciation swings over the same period: since 2005, the exchange rate has fluctuated between 3.08 and 4.45 Malaysian Ringgit (MYR) per US Dollar (US$). This volatility, combined with a high share of imported goods in the average household's consumption basket[2] means that, at least according to models of inattention, it may be comparatively more important for consumers to keep up to date with the exchange rate and factor exchange rate expectations into their spending decisions. This feature of the study setting allows us to benchmark the effect of information on macroeconomic indicators with different levels of salience, which can help us shed light on the effects of policy changes in different macroeconomic environments, such as high- versus low-inflation regimes. Another advantage of our setting lies in the characteristics of the sample population. Individuals in our sample are among the most educated and financially experienced and thus should be one of the populations that is most likely to revise consumption decisions in line with macroeconomic expectations. As the consumers in our study are relatively affluent, their consumption bundle contains significant shares of durable and tradable goods, which correspond to the categories one would expect to be most affected by changes in exchange rate and inflation expectations.

We present two main results. The first result is that the information provided in the experiment has a strong effect on the formation of consumer expectations. The vast majority of individuals whose inflation and exchange rate expectations are not in line with expert

[2] While comprehensive data on the import content of different consumption categories is not available for Malaysia, data for the limited number of categories where such information is available point to a high import content and thus a potentially important link between the exchange rate and consumer prices. The largest category in the Malaysian consumption basket, for example, is food and non-alcoholic beverages, which accounts for 30.2% of total consumption for the average household and has an import content of 60%.

forecasts update their expectations in response to our information treatments. Specifically, a 1 percentage point (ppt) increase in the information shock about future inflation increases inflation expectations by 0.237 ppt ($p < 0.001$). Similarly, a 1 pp increase in the information shock about the future nominal exchange rate increases exchange rate expectations by 0.065 ppt ($p = 0.036$). The finding that expectations are more responsive to information about inflation than to information about the nominal exchange rate is consistent with consumers in our setting having greater incentives to be informed about the exchange rate and thus having stronger prior beliefs.

The second result is that changes in macroeconomic expectations induced by our experiment do not translate into changes in consumption behavior in the direction predicted by a standard model of intertemporal consumption choice. Specifically, we test three basic predictions from a standard model of optimal consumption: (i) higher inflation expectations should increase spending on durables; (ii) higher expected exchange rate depreciation should increase spending on tradable durable goods; and (iii) conditional on the nominal interest rate, higher inflation expectations should increase credit card borrowing. We do not find empirical support for any of these predictions. Instead, the effects of information shocks induced by the experiment on spending are close to zero and statistically insignificant. For specific spending categories, such as durables consumption, the point estimates suggest a reduction in spending in response to an anticipated increase in inflation and exchange rate devaluations, contrary to the predictions of a standard model of intertemporal choice. While we cannot rule out small effects on any specific outcome, we have sufficient statistical power to rule out moderate or large effects. Moreover, shocks to expectations do not affect self-reported spending plans, which we elicit immediately after the information-provision part of the survey experiment.

To better understand why changes in macroeconomic expectations do not affect spending decisions as predicted by standard models of consumer choice, we conduct an additional "mental model experiment." In this follow-up experiment, participants were presented with randomly assigned inflation and exchange rate scenarios and asked about their spending plans in each scenario. As part of this exercise, we additionally elicited measures of financial literacy and demand for inflation-indexed securities, which allow us to further narrow down which consumers respond to macroeconomic information. We explore several mechanisms that could explain the absence of a spending response by testing how the response to randomly assigned macroeconomic scenarios varies with consumer characteristics and complement this evidence with results from the main experiment.

Our preferred interpretation, based on the results of this exercise, is that consumers reduce spending in anticipation of nominal rigidities. That is, consumers appear to be sophisticated enough to understand that their income is not indexed to inflation or the exchange rate and reduce spending, especially on durable goods, for precautionary reasons because they correctly anticipate the purchasing power of their income to decline as inflation rises or the value of the local currency depreciates. This counteracts the effects of exchange rate devaluations and heightened inflation expectations on spending predicted by a standard model of intertemporal choice and explains the lack of a spending response to changed macroeconomic expectations. Moreover, our results are not qualitatively different between the consumer response to changes in inflation or exchange rate beliefs, which suggests that the mechanism we illustrate is not limited to one specific macroeconomic variable.

We discuss, and provide evidence against, a number of alternative mechanisms that could explain the absence of a spending response. One alternative explanation is that consumers are unable to interpret the information that is provided to them, or might not be sophisticated enough to re-optimize their spending plans based on revised macroeconomic expectations.[3] Indeed, substantial evidence suggests that consumers fail to optimize in many simpler economic decisions, either due to behavioral biases or lack of knowledge (see Campbell et al.,

[3]News about macroeconomic events may, for example, not affect consumer expectations in the first place, if they are not sufficiently salient, or if consumers are unable to interpret them. Coibion et al. (2021) show evidence of this in the case of the Federal Reserve's announcement of its new average inflation targeting policy.

2011; Beshears et al., 2018). In the context of credit card spending, Ponce et al. (2017) and Gathergood et al. (2019) show that consumers do not borrow using the lowest interest rate card and do not prioritize repayment of the card with the highest interest rate.[4] We present several tests of this hypothesis and find that lack of financial knowledge is unlikely to explain our results. We first show that respondents in our setting update their expectations substantially in response to the information that is provided to them, which rule out the possibility that consumers are entirely unable to interpret the information given to them in the experiment. Moreover, we show that consumers with high (above median) financial literacy, as measured using a standard test, do not respond differently to inflation and exchange rate information than consumers with low (below median) financial literacy. In addition, both groups show substantial demand for indexed securities when presented with a high inflation or high exchange rate depreciation scenario, which suggests a relatively high degree of consumer sophistication.

Second, we test whether the absence of a spending response can be explained by time inconsistency and commitment problems. Specifically, it is possible that consumers revise their spending plans in response to updated macroeconomic expectations but are unable to follow through on these plans due to self-control problems. We provide a direct test of this hypothesis. If self-control problems were responsible, we would expect the information treatments to affect spending plans but not actual spending. Instead, we use self-reported spending plans, elicited immediately after the information provision stage of the experiment to show that providing information about inflation or the exchange rate has no impact on self-reported spending plans.

Third, we can also rule out the possibility that consumers reduce spending because they associate higher inflation or anticipated exchange rate depreciations with worsening overall economic conditions. Results from the main experiment show that our information treat-

[4] See also Chetty et al. 2020, who show that credit card spending fails to react to anticipated income shocks.

ments have no effect on participants' expectations about their personal financial situation or the overall state of the economy.

Finally, it is possible that the effects of information shocks on expectations are not sufficiently long-lasting to affect consumer decisions. We provide some evidence against this interpretation. First, we find that our results are robust if, instead of looking at spending behavior in the subsequent three months, we look at shorter time horizons.[5] Second, we show that information shocks do not affect spending plans, which are self-reported immediately after the information-provision experiment when one would expect the information to still be fresh and salient. Third, evidence from several other studies suggests that providing information through an experiment tends to have long-lasting effects on expectations. For instance, the effects of information shocks on inflation expectations last for at least a few months (Cavallo et al., 2017). Shocks to other economic beliefs have been found to last from months up to one year (Bottan and Perez-Truglia, 2020b; Fehr et al., 2019). Fourth, information experiments similar to the one in this study have been shown to affect high-stakes decisions (see **?**, for a review): employees work harder after increasing their expectations of future salary increases, and home sellers are less likely to sell their homes when their home price expectations increase (Cullen and Perez-Truglia, 2022; Bottan and Perez-Truglia, 2020a). Information interventions have also been shown to have meaningful effects on personal financial decisions. Bursztyn et al. 2019 show that information about the moral and material consequences of delinquency improves credit card repayment and savings and investment decisions in response to feedback about the choices of their peers, and Beshears et al. 2015 and Bursztyn et al. (2014) show that people revise their savings and investment decisions in response to information about the choices of their peers.

Our study of macroeconomic expectations and consumer behavior relates to several strands of the literature. First, and most directly, our paper relates to research on the role of subjective expectations in macroeconomics and household finance (Roth and Wohl-

[5]These results are available upon request

fart, 2020; Beshears et al., 2018). Macroeconomic expectations play a central role in models of household finance and macroeconomics that rely on the consumption Euler equation to microfound individual saving and consumption behavior. While traditional economic theory assumes that individuals form statistically optimal expectations based on all available information, the available evidence shows that there are large information frictions and wide disagreement in the interpretation of macroeconomic information (Mankiw and Reis, 2002; Mankiw et al., 2003; Armantier et al., 2016; Cavallo et al., 2016, 2017; **?**; Giglio et al., 2021; Roth and Wohlfart, 2020). Growing evidence shows how macroeconomic expectations are formed, and which deviations from optimality are common. Andre et al. (2021) examine how individuals rationalize changes in macroeconomic conditions using narratives that place different weight on alternative propagation mechanisms and show that this heterogeneity may explain disagreement in beliefs among agents who observe the same macroeonomic shock and have access to the same information set. Similarly, D'Acunto et al. (2023) show that cognitive ability correlates with the ability to form accurate macroeconomic expectations in the cross-section. Less is known, however about how individuals factor macroeconomic expectations into their economic decisions (Armantier et al., 2015).

We make two contributions to this line of research. First, our paper is the first to examine the impact of economic expectations using macroeconomic variables that have plausibly different degrees of salience to consumers. Specifically, we focus on the response to inflation and exchange rate expectations, the two macroeconomic variables that have arguably received the most attention in the policy debate and in academic research. While inflation has been low and stable in recent decades and are thus arguably less salient for most households, large fluctuations in the nominal exchange continue to be commonplace around the world and have meaningful economic consequences for many households (Gouvea, 2020; Cravino and Levchenko, 2017). By comparing the response to exchange rate and inflation information, we can shed light on whether belief formation and behavior differ based on the salience of the macroeconomic indicator in question. This has policy implications, for example for the optimal communication of monetary policy in high- versus low-inflation environments.

Second, along with contemporaneous work by Coibion, Gorodnichenko, and Weber (2022), our paper is the first to measure the impact of macroeconomic expectations on actual spending, observed free of measurement error, rather than self-reported survey measures of consumption. We observe consumption decisions using detailed transaction-level data on credit card spending in a setting where credit card transactions account for a meaningful share of total consumption.[6] This overcomes several limitations of existing studies that rely on survey data to measure impacts on behavior, and are therefore susceptible to misreporting, measurement error, and experimenter demand effects. Indeed, we provide evidence that concerns about survey data should be taken seriously, as we document a surprisingly weak correlation between self-reported spending plans and actual spending decisions.

Finally, this study also speaks to a literature on the impact of expectations on personal financial decisions more generally. Giglio et al. (2021) and Giglio et al. (2020) use survey data to test how macroeconomic beliefs affect the decisions of retail investors. They show that beliefs are reflected in asset allocation and change in response to discrete macroeconomic events, such as a stock market crash. Aaronson et al. (2012), Agarwal et al. (2007), and Agarwal and Qian (2014) use credit card data to test consumer responses to wage increases, tax changes, and unanticipated income shocks and find effects inconsistent with fully rational expectations. Beshears et al. (2015) survey the literature on behavioral household finance and highlight the important role of subjective expectations. We contribute to this literature by examining how macroeconomic expectations influence the fundamental intertemporal savings and consumption decisions that lies at the heart of household finance.

The remainder of the paper is structured as follows. Section 1.2 provides a stylized theoretical framework to motivate our experimental design and develop our hypotheses.

[6]Based on income data provided by our partner bank, the estimated ratio of credit card spending to monthly income is approximately 35% in our sample. This is comparable to the level of credit card spending in most high-income economies.

In Section 1.3, we summarize the institutional setting. Section 1.4 presents the research design and implementation of the experiment. In Section 1.5, we describe additional data sources and provide descriptive statistics. Section 1.6 reports the results, and the last section concludes.

1.2 Theoretical Framework and Hypotheses

We use a standard model of intertemporal consumer choice to motivate our experimental design. Letting subscript t denote the time period, we assume that the consumer faces an exogenous stream of nominal income Y_t and can have positive or negative holdings of an asset A_t that pays an exogenous nominal interest rate R_t. There are four types of consumption goods, which we can classify according to their durability and tradability: durable tradables (denoted X_t^T), durable nontradables (X_t^N), nondurable tradables (C_t^T), and nondurable nontradables (C_t^N). We assume that durable goods depreciate at a rate of δ, tradable goods (both durable and nondurable) have an exogenous price P_t^T, and nontradable goods (both durable and nondurable) have price P_t^N. The consumer gets utility $U(C_t^N, X_t^N, C_t^T, X_t^T)$ from a given combination of goods, which is concave in each of its arguments and has a discount factor β.

The consumer's optimization problem can thus be summarized as follows:

$$\max_{\{C_t^N, X_t^N, C_t^T, X_t^T, A_t\}_t} \quad \sum_{t=1}^{T} \beta^t U(C_t^N, X_t^N, C_t^T, X_t^T) \tag{1.1}$$

subject to

$$\begin{aligned} P_t^N(C_t^N + X_t^N - X_{t-1}^N + \delta X_{t-1}^N) + P_t^T(C_t^T + X_t^T - X_{t-1}^T + \delta X_{t-1}^T) + A_{t+1} \\ \leq P_t^N Y_t + R_t A_t \end{aligned}$$

We denote the exogenously given rate of inflation from period t to $t+1$ as π_{t+1}, which is defined as follows:

$$\pi_{t+1} = \frac{P^N_{t+1}(\bar{C}^N_t + \Delta\bar{X}^N_t) + P^T_{t+1}(\bar{C}^T_t + \Delta\bar{X}^T_t)}{P^N_t(\bar{C}^N_t + \Delta\bar{X}^N_t) + P^T_t(\bar{C}^T_t + \Delta\bar{X}^T_t)} = w_t\pi^N_{t+1} + (1-w_t)\pi^T_{t+1}$$

$$\text{where} \quad w_t \equiv \frac{(\bar{C}^N_t + \Delta\bar{X}^N_t)}{(\bar{C}^N_t + \Delta\bar{X}^N_t) + P^T_t/P^N_t(\bar{C}^T_t + \Delta\bar{X}^T_t)}$$

$\bar{Z}$ is the average value of variable Z in the economy.

We make the following simplifying assumptions. First, an increase in inflation cannot be accompanied by a decrease in inflation in any specific category of goods:

Assumption 1. $\frac{d\pi^N_{t+1}}{d\pi_{t+1}} \geq 0, \quad \frac{d\pi^T_{t+1}}{d\pi_{t+1}} \geq 0$

Second, defining $d_{t+1} = \frac{E_{t+1}-E_t}{E_t}$ as the exchange rate depreciation between period t and t+1 (E_t denotes the spot exchange rate of the Malaysian Ringgit to the US Dollar at time t), we assume non-zero pass-through of exchange rate depreciation to tradables:

Assumption 2. $\frac{d\pi^T_{t+1}}{dd_{t+1}} \geq 0$

Third, we assume Cobb-Douglas instantaneous utility:

Assumption 3. *Let consumption utility be Cobb-Douglas with parameters α and θ corresponding to the weights of non-durables and non-tradables, respectively:*

$$U(C^N_t, X^N_t, C^T_t, X^T_t) =$$
$$\alpha\theta\log C^N_t + \alpha(1-\theta)\log C^T_t + (1-\alpha)\theta\log X^N_t + (1-\alpha)(1-\theta)\log X^T_t$$

This model yields the following three predictions (for proofs of each proposition, see Appendix 1.A), which motivate the design of our field experiment:

Proposition 1. *Spending on durables (tradable and non-tradable) $P^N_t\Delta X^N_t + P^T_t\Delta X^T_t$ increases with expected inflation π_{t+1}.*

The intuition for this standard result (see, for example, Bachmann et al., 2015) is that one can buy durables to shield against the inflation tax.

The second proposition describes the effect of nominal exchange rate depreciation:

Proposition 2. *Spending on tradable durables* $P_t^T \Delta X_t^T$ *increases with future exchange rate depreciation* E_{t+1}.

The intuition is equivalent to that of the previous proposition: consumers want to consume durable tradables to shield against depreciation. If consumers expect the exchange rate to depreciate, they might be more likely to buy durable tradables such as electronics (as in one of our survey questions) now, because doing so in the future will be more expensive.

The last result describes the relationship between inflation expectations and debt:

Proposition 3. *Net borrowing* $A_t - A_{t+1}$ *increases with inflation* π_{t+1}.

This proposition states that when deciding how much debt or savings to accumulate, individuals care about the real interest rate. Holding constant the nominal interest rate, an increase in inflation will reduce the real interest rate.[7] As a result, an increase in the expected rate of inflation will make it attractive for consumers to borrow more (or save less).

1.3 Background and Setting

1.3.1 Macroeconomic and Institutional Context

We conduct a natural field experiment with credit card customers from a large commercial bank in Malaysia. Malaysia is representative of many small, open economies in that inflation has been stable and low over the past two decades, whereas the exchange rate has been

[7]The assumption of a fixed nominal interest rate matches our setting where, as in many emerging markets, the regulator enforces an interest rate cap on credit card advances that is only slightly above the rates currently charged in the market.

volatile at times. Malaysia's central bank, the Bank Negara Malaysia, follows a mandated 2% inflation target, and its main policy instrument are changes in the overnight policy rate.[8] The central bank communicates the rationale behind changes in the policy rate as well as its future outlook through public 'monetary policy statements', which are released every two months. Thus, our study takes place in a setting where, similar to the United States and the Eurozone, forward guidance and the management of market expectations is an important goal of central bank communication.

Figure 1.2 presents historical data on inflation and the exchange rate. The figure first shows the evolution of the consumer price index over the last four decades. Over this period, Malaysia experienced a short period of high inflation in the early 1980s. However, since then, inflation has been generally moderate and stable at around 2% to 5% annually. Figure 1.2 also shows the evolution of the nominal exchange rate with respect to the U.S. Dollar. The exchange rate has been markedly more volatile than the rate of inflation. Indeed, the difference in volatility would be even more pronounced if we looked at weekly rather than yearly data, as inflation is stable over the year, while the exchange rate is characterized by sharp changes at shorter time intervals. Exchange rate volatility was most pronounced in the Asian financial crisis that began in 1997. During this period, the Malaysian Ringgit depreciated by more than 50% against the U.S. Dollar. Following this experience, Malaysia pegged its currency to the U.S. Dollar at a rate of 3.20 MYR/US$ between 1998 and 2005. Since the end of the currency peg, the exchange rate has fluctuated between 3.08 and 4.45 MYR/US$. Several large exchange rate swings have occurred in recent years, triggered by external and domestic events, such as oil price shocks and political instability surrounding national elections.

Because of its greater volatility and the relatively high share of imported goods in total consumption, it seems plausible that in our setting and time period, the foreign exchange

[8] See https://www.bnm.gov.my for the statutes of Bank Negara Malaysia and recent monetary policy statements.

rate is comparatively more salient and plays a larger role in consumer decisions than the rate of inflation. As our theoretical framework in Section 1.2 highlights, consumers can partially offset the effect of higher inflation by shifting their consumption towards durable goods. We therefore predict that a consumer who expects higher inflation will increase the relative share of durable goods in credit card spending to insure against inflation risk. However, given that inflation in our setting is typically between 2 and 5 percent, this is unlikely to be very meaningful for the consumer's overall finances. In contrast, when considering the purchase of tradable goods, such as consumer electronics or a car, the same consumer might be substantially affected by exchange rate fluctuations due to the magnitude in our setting. If the consumer, for example, expects a 20-25% exchange rate depreciation, the timing of such a purchase could amount to substantial savings.

To examine whether, in our context, changes in the exchange rate are in fact more salient to households than changes in the inflation rate, we exploit data on online searches and newspaper articles. Figure 1.3 plots data from Google Trends that tracks the frequency of online searches. These data have been used in several previous studies to measure public interest in specific topics (see, for example, Perez-Truglia, 2020). Figure 1.3.a plots the frequency of online searches related to the terms "inflation" and "exchange rate", where dark bars correspond to keywords related to inflation and lighter-colored bars correspond to keywords related to the exchange rate. Google reports online searches only in relative terms. We therefore normalize the series, with the nominal exchange rate taking the value 100 in the first period. The figure shows that consumers seek information about the exchange rate more frequently than information about the inflation rate: in an average week of 2019, there were approximately 18 times more searches about the exchange rate than searches related to inflation.[9]

This pattern is supported by a comparison of newspaper articles mentioning either inflation or the exchange rate. Figure 1.3.b plots the frequency of articles in Malaysia's most

[9]The 2019 data cover the period when our experiment was conducted and are similar to other years.

widely read English language newspaper in which either term is mentioned.[10] As in Figure 1.3.a, both series are normalized so that the nominal exchange rate takes the value 100 in the first period. In an average week of 2019, there were approximately twice as many newspaper articles that referred to the exchange rate than newspaper articles mentioning the inflation rate. The two different data sources indicate both higher demand and higher supply for news about the exchange rate, rather than news about the inflation rate, in our setting.

1.3.2 Partner Bank

Our partner institution is one of the largest commercial banks in Asia and has more than a million individual customers in Malaysia. Nearly all of our partner bank's retail banking customers have debit cards, and a significant share additionally have credit cards linked to their account. Although our partner bank covers a broad, socially and geographically diverse customer base, its clients are naturally not a fully representative sample of the population. They are on average younger, more educated, more likely to reside in urban areas, and wealthier (see Section 1.5). Credit card usage in this setting is high and in fact comparable to high-income economies. Using the bank's administrative data, we estimate that monthly credit card spending in our data accounts for approximately 35% of consumers' estimated monthly income which is comparable to many advanced economies. In comparison, Ganong and Noel (2019) use data from the JP Morgan Chase Institute and find that average credit and debit card spending accounted for 51% of monthly income in the United States.[11]

We worked with our partner bank to integrate an information experiment into their regular customer communications and merge the experimental data with the bank's administrative records on credit card transactions. Our partner bank records credit card spending at the transaction level for all customers. The data are obtained directly from the bank

[10] *The Star*, whose archive is available at www.thestar.com.my.

[11] Based on results from Table 1 of Ganong and Noel (2019), who use a sample of credit and debit card customers in the three months prior to becoming unemployed.

and include date, time, and amount of the transaction, along with a location and type of transaction code. This allows us to time each transaction and assign it to a specific spending category and trace the impact of randomized information.

1.4 Research Design

1.4.1 Overview

Figure 1.1 provides a graphical summary of the research design. Our intervention is designed as an information provision experiment and administered through a phone survey with our partner bank's credit card customers. In addition to survey data, we observe pre-treatment and post-treatment administrative data covering the universe of credit card transactions for all participants of our study.

The intervention proceeds in the following four steps. First, all respondents are asked a set of standard questions on demographics and their general economic situation. Second, a survey module elicits macroeconomic expectations and provides randomly selected subsets of respondents with information about inflation, the exchange rate, or both. Third, we measure posterior beliefs to assess whether our information treatments affect respondents' expectations. Finally, we combine survey responses with administrative data on credit card spending to test whether the information provided affects subsequent consumption in the manner predicted by economic theory. We provide additional details on each of these steps in the following sections.

1.4.2 Sample Population

To construct the sample frame for our experiment, we first requested a list of credit card customers from our partner bank. We specified that this list should be restricted to customers who opened their accounts within the previous three years. We received a random sample of

33,000 credit card customers and invited these customers to participate in a phone survey, which included our information experiment.

The survey was conducted by a team of 11 call center operators who were trained to administer a short phone survey and supervised in person by a member of the research team. At the beginning of each workday, the operators were provided with a randomly selected list of credit card customers to call. The operators introduced themselves as surveyors working on behalf of researchers from UCLA and asked participants if they were willing to participate in a short survey about their economic expectations. If operators were unable to reach a respondent on the first attempt, they were instructed to make at least one further attempt at a later time.

1.4.3 Information Experiment

We integrated our experiment into this credit card customer survey administered by our partner bank. The survey instrument, which is available in Appendix 1.D, can be divided into five parts: (i) collecting baseline information, (ii) eliciting prior beliefs, (iii) providing information to a random subset of respondents, (iv) eliciting posterior beliefs, and (v) self-reporting of consumption plans for all respondents. In this section, we describe each component of the intervention in turn.

1.4.3.1 Baseline Information

We begin with a set of standard questions on the respondent's socio-economic background, including employment status, highest level of education attained, marital status, and dependents. We do not ask about gender, age, or income as this information is available in the administrative records obtained from the partner bank. We also include one question about the expected economic conditions in the country over the next 12 months, for which the possible responses are "better off," "about the same," and "worse off." The language in this question, and all other questions about expectations, closely follows the wording used in the most widely used surveys of consumer expectations, such as the University of Michigan's Survey of Consumers and the Federal Reserve Bank of New York Survey of Consumer Expectations (see, for example, Bachmann et al., 2015; Fuster et al., 2020).

1.4.3.2 Elicitation of Prior Beliefs

Next, we elicit participants' inflation and exchange rate expectations at two points in time: immediately before the treated individuals receives information from the experimenter (prior beliefs) and after a randomly chosen subset of respondents is provided with an inflation or exchange rate forecast (posterior beliefs). The wording in both rounds is closely modeled on that used in standard surveys of consumer expectations, and was adjusted to our study context through qualitative interviews and an online pilot.

We elicit beliefs about the future inflation rate and exchange rate. To avoid artificially making one belief more salient than the other, we randomized the order of these two questions. To elicit inflation expectations, the surveyors first provide a definition of inflation by explaining that "[...] inflation is the measure of how prices in Malaysia change in general" and then elicit the respondent's expected inflation rate over the following 12 months.[12] Our wording is similar to that used in one of the most widely used surveys of inflation expectations, the Federal Reserve Bank of New York Survey of Consumer Expectations, which asks about the inflation rate directly.[13] Participants are asked to give their response in percentage

[12] Providing a definition is standard practice in surveys about inflation. The Michigan Survey of Consumers, for instance, avoids the term inflation and asks instead "do you think prices in general will go up, go down, or stay the same".

[13] Another widely used source of data is the Michigan Survey of Consumers, conducted by the Survey Research Center at the University of Michigan, which asks about *prices in general* instead of asking about inflation directly. See Armantier et al. (2016) for a discussion on how these differences in wording might affect responses. Given the similarity of the questions used to elicit expectations, we can benchmark our results to those of related studies that have used the Survey of Consumer Expectations data and, with some caveats that have been highlighted by Armantier et al. (2016) among others, to studies that used the Michigan Consumer Survey data.

points.

As documented in Section 1.3.1, the nominal exchange rate is already salient in news media and online searches in the country of study. This makes it more straightforward to elicit exchange rate expectations. To elicit respondents' nominal exchange rate expectations, the surveyor informs the respondent of the current nominal exchange rate ("as of April 2019, 1 U.S. Dollar is worth around 4.05 Malaysian Ringgit") and then asks what, in their opinion, "[...] the exchange rate will be 12 months from now, in April 2020". This way of eliciting beliefs is consistent with previous work by Cavallo et al. (2017) and was adapted to our research setting through a series of pilot tests and consumer interviews.

1.4.3.3 Information Provision

In the information provision stage of the experiment, all respondents are first read the following message: "In this stage, we randomly select respondents to receive some feedback about the previous questions." Each respondent is then randomly assigned to one of the following three treatment groups with equal probability:

(a) Treatment *inflation*: In the first treatment condition, respondents receive a signal about the future inflation rate: "The consensus among experts from the government and private sector is that inflation in Malaysia will be 2.3% over the next 12 months."

(b) Treatment *exchange rate*: In our second treatment, respondents receive a signal about the future nominal exchange rate: "The consensus among experts from the government and private sector is that 1 U.S. Dollar in Malaysia will be worth 4.10 Malaysian Ringgit 12 months from now".

(c) Treatment *exchange rate and inflation*: In our final treatment condition, respondents receive two signals. The first one relates to the inflation rate: "The consensus among experts from the government and private sector is that inflation in Malaysia will be 2.3% over the next 12 months". The second one relates to the exchange rate: "The consensus among experts from the government and private sector is that 1 U.S. Dollar will be worth 4.10 Malaysian Ringgit 12 months from now".[14]

Expert forecasts of inflation and the exchange rate used in the experiment were obtained from widely used forecast websites and updated once over the course of the experiment to reflect a quarterly forecast revision.[15] Figure 1.8 in the Supplementary Appendix reports information about the historical accuracy of forecasts from the sources used in the experiment. The figure shows that, historical inflation and exchange rate forecasts at the same time horizon used in the experiment are fairly accurate with a mean prediction error of 1.28 percentage points for inflation (0.97 percentage points when excluding an outlier during the pandemic) and 0.39 percentage points for the exchange rate.

1.4.3.4 Elicitation of Posterior Beliefs

The second round of belief elicitation takes place immediately after respondents are provided with information about inflation, the exchange rate, or both. To ensure that responses are comparable to the elicitation of prior beliefs, the second round of belief elicitation uses the exact same wording as the first. The goal of this second round of belief elicitation is to understand whether individuals incorporate the information provided to them through the information treatments into their expectations.

1.4.3.5 Elicitation of Consumption Plans

While the main goal of our experiment is to estimate the effect of information on actual consumption, as measured objectively in administrative data, we also asked a series of ques-

[14] The order of these two pieces of information was consistent with the (randomized) order of the questions on prior beliefs: i.e., if the prior inflation expectations was elicited before the prior exchange rate expectations, then feedback about inflation would come before the feedback about the exchange rate.

[15] Inflation forecasts are taken from Statista (www.statista.com), forecasts of the exchange rate come from Trading Economics (www.tradingeconomics.com).

tions on respondents' self-reported consumption plans. These responses were collected after the elicitation of posterior beliefs and serve two purposes. First, they allow us to test whether our information treatments affect intended behaviors. With that goal in mind, we measure expected future spending in the main consumption categories highlighted by the theoretical framework (durable goods, tradable goods, and credit card debt), as well as other categories that act as useful proxies or benchmarks. Second, these questions allow us to confirm that survey responses have predictive content by testing whether predicted consumption correlates with actual future consumption.[16]

The first of these questions elicits respondents' expected change in total credit card expenditures (which corresponds to the total expenditures we observe in administrative data). Specifically, respondents were asked: "Do you expect your credit card spending to go up, stay the same, or go down during the next 3 months?" We code this and other similar questions on a simple three-step scale. the variable takes the value -1 if the individual responded "go down," 0 if the individual responded "stay the same," and +1 if the individual responded "go up." Another pair of questions uses similar language to elicit total spending, not limited to spending on credit cards, and spending on groceries for comparison.

The next set of questions on expected spending asks about spending on durable goods and uses wording that closely follows the Michigan Survey of Consumers. We first ask respondents: "Do you think now is a good time, a bad time, or neither a good nor a bad time to buy household items, such as furniture or a refrigerator?" We code responses using the same $\{-1, 0, +1\}$ scale as before. The variable takes the value -1 if the individual responded "No, it's a bad time," 0 if the individual responded "It's neither a good nor a bad time," and +1 if the individual responded "Yes, it's a good time." We include three additional questions using this same language, but instead of asking about durable expenditures, we

[16]We measure consumption plans only once, after the intervention, so as to not confound the prior and posterior elicitation exercise on inflation and exchange rate beliefs. Under the assumption that priors are balanced across treatment and control groups due to randomization, we interpret any differences in consumption plans as causal.

ask about electronics, vehicles, and credit card borrowing, respectively.

One potential concern with our design is that the information treatments could affect behavior through a mechanism other than intertemporal optimization of consumption. Intuitively, information about inflation and the exchange rate could affect spending by changing respondents' general optimism or pessimism about the economy. For example, individuals who learn that there will be inflation or depreciation in the future may infer that these are symptoms of an economic downturn and interpret this as bad news for their personal economic situation. We include two questions to shed light on this potential mechanism in the questionnaire. The first asks respondents about their expectations for the economy overall, the second asks about the individual's own financial outlook: "Looking ahead, would you say that you and your family living with you will be better off or worse off financially than you are now?" We code both outcomes using the same $\{-1, 0, +1\}$ scale: -1 if the individual responds "Worse off," 0 if the individual responds "About the same," and +1 if the individual responds "Better off."

1.5 Data and Descriptive Statistics

1.5.1 Sample and Survey Implementation

We implemented the experiment over a four-month period between April and July 2019. During this time, members of the survey team attempted to reach 28,958 credit card clients and completed 2,872 phone surveys, implying a 10% response rate.[17] The survey team used the internal records of the partner bank to ensure that respondent phone numbers could

[17] This final sample excludes individuals who started the survey but did not make it to the end. There are only 174 partially complete surveys and we cannot reject the hypothesis that respondents are missing at random after the information-provision stage of the survey. Our final sample excludes 274 individuals who reported extreme prior beliefs about the nominal exchange rate (above 4.65 or below 3.7 Ringgit per U.S. dollar) because the system used by the surveyors prevents us from knowing the exact expectations of those respondents. We would have excluded those extreme priors anyway to avoid sensitivity to outliers, as is standard in studies using expectations data (Fuster et al., 2020).

be matched to account data. Before commencing the survey, members of the survey team additionally verified the name of the respondent and confirmed that they were the holder of a debit or credit card from our partner bank. Surveys were offered in English and Malay, and 47% of respondents chose to complete the survey in English, while the remaining 53% responded in Malay. Because our survey experiment was integrated in the bank's regular customer outreach program which has no provision for incentivizing respondents, participants were not compensated for their time. Respondents were asked if they wished to participate at the beginning of the interview and could opt out of the survey at any point.[18] The partner bank shared anonymized administrative records for all participants for all survey respondents, as well as for a representative sample of clients who were invited to the survey but did not respond.

Table 1.2 provides descriptive statistics based on administrative data. Column (1) reports data for a random sample from the universe of the bank's credit card clients, which includes both survey respondents and non-respondents. In this sample, 62% of clients are male, they are on average 33.6 years old, have an average monthly income of $3,087 and monthly credit card expenditures of $1,106. As one might expect, the summary statistics shown in column (1) of Table 1.2 indicate that clients of the partner bank are not fully representative of the Malaysian population. On the one hand, the age and gender composition of our sample is not substantially different from the country average: data from the Malaysian Department of Statistics for 2020 indicate that 51.4% of the Malaysian population is male (compared to 62% in our sample of bank customers) with a mean age of 31.4 years (compared to 33.6 years in our sample). On the other hand, we find substantial income differences. According to data from the Salaries and Wages Survey Report,[19] the average Malaysian household earned $1,767 (compared to $3,087 in our sample).

[18] Because the bank released only anonymized data and the research team had no access to personally identified information, informed consent was not obtained from respondents in writing.

[19] Source: Malaysian Department of Statistics, 2017.

However, Table 1.2 also shows that there is no indication of selection into treatment. Customers who participated in our experiment are similar to non-respondents. Columns (2) and (3) compare the characteristics of the 2,872 clients who answered our survey (reported in column (2)) to the sample of 3,126 clients who were invited to the survey but did not respond (column (3)). Comparing columns (2) and (3) indicates that although there are some statistically significant differences in average characteristics, none of these differences are meaningful in magnitude. For example, the average age is 33.28 years among survey respondents as compared to 33.88 among non-respondents. The average monthly income is \$3,128 among survey respondents as compared to \$3,049 among non-respondents. The average credit card expenditures are \$1,095 among survey respondents and \$1,045 among non-respondents. The one possible exception is gender, where we find that men are more represented among survey respondents (67%) than among non-respondents (57%).

Table 1.3 provides additional descriptive statistics, based on survey and administrative data, and presents a test of randomization balance. Column (1) is based on the sample of all 2,872 survey respondents. The summary statistics show that the respondents in this sample are highly educated (87% have a college degree), around half (54%) are married, and around 10% are self-employed. In columns (2) through (4) of Table 1.3, we compare the baseline characteristics and expenditures of the three treatment groups. Column (5) reports p-values for the null hypothesis that these characteristics are equal across all three treatment groups. The results indicate that, consistent with successful random assignment, pre-treatment observables are balanced across treatment groups. As expected, all differences across treatment groups are economically small. The difference is statistically significant ($p = 0.09$) for only one of the 11 characteristics reported in the table: the number of dependents. This result is within expectations, given that 1 out of every 10 differences are expected to be statistically significant at the 10% level simply by chance. We nonetheless follow standard practice and include the number of dependents as a control variable in all regressions.

1.5.2 Credit Card Data

Our partner bank shared administrative data on credit card transactions for all customers in the sample. These data allow us to measure spending and borrowing behavior of all customers in our sample for 12 months prior to the intervention and 3 months after the intervention. The dataset contains detailed records of all credit card transactions that occurred during this time period, which include the transaction amount, description, vendor name, and spending category code.[20] The credit card data also include information about outstanding balances and repayment, which we use to measure consumers' willingness to take on debt. Importantly for our purposes, each transaction in the data contains the standardized Merchant Category Code (MCC), a 4-digit identifier that classifies a business by the types of goods or services it sells. The MCC makes it possible to assign each transaction to a specific spending category. Importantly, for the goal of our analysis, the MCC allows us to distinguish between spending on durable, nondurable, tradable, and nontradable goods. To classify spending into durable versus nondurable goods, we follow the standard categorization used in the literature (see Aaronson et al., 2012; Agarwal and Qian, 2014; Ganong and Noel, 2019; Chetty et al., 2020). For example, some durable spending items include apparel, consumer electronics, and furniture. To the best of our knowledge, no other paper has used MCCs to classify credit card spending into tradable and non-tradable expenditures. We therefore created our own categorization by manually inspecting each individual MCC and classifying it as tradable or nontradable. We follow the standard definition, which identifies a tradable good as a good that can be sold and consumed in a location other than the place where it was produced. In our classification, all codes for services are assigned to the nontradable category, whereas codes for goods are assigned to the tradable category for goods that can be imported or exported. For example, some tradable spending items include apparel and consumer electronics.

[20] We do not obtain data on debit card transactions because that they account for a negligible fraction of spending according to pre-intervention summary data (debit cards are used primarily for cash withdrawals).

We summarize both MCC categorizations in Table 1.1. As there are thousands of individual MCCs, we report summary statistics using standard groupings of MCCs that are commonly used by financial institutions and in the academic literature. Column (1) reports the average spending for each MCC group in our sample. Column (2) indicates the fraction of spending within that MCC group that is classified as durable (the remainder is classified as nondurable by construction). Column (3) indicates the fraction of spending within the MCC group that is classified as tradable (with the remainder classified as nontradable). For example, the third row corresponds to the MCC group "automotive expenditures", in which 100% of codes are classified as durables and 0% as non-durables, and 64% of the spending in this MCC group is classified as tradable versus 36% as nontradable (primarily codes corresponding to services). The last row of the table summarizes codes that we group as "uncategorized." These miscellaneous MCCs do not contain enough information to categorize them as durable, as opposed to nondurable, or tradable, as opposed to nontradable, expenditures.

Figure 1.4 summarizes the breakdown of spending between durable and tradable categories. Each rectangle corresponds to one unit of spending. The blue rectangles towards the right, denoted as uncategorized, correspond to the 10% of spending that cannot be categorized. Among the transactions that can be categorized (90% of all spending), 31% are categorized as tradable and the remaining 69% as nontradable. Among the transactions that can be categorized, 32% are durable and 68% nondurable. Figure 1.4 also shows substantial orthogonal variation between the two categorizations. That is, not all tradables are durables and vice versa.

Table 1.2 also shows average spending statistics for each key spending category used in our analysis. Specifically, column (2) shows that customers who participated in the experiment used their credit cards to spend average monthly amounts of $364 (33% of the $1,095 total credit card spending) on durables and $272 (25% of total spending) on tradable durables. On average, subjects had $1,801 in outstanding credit card debt, equivalent to 1.5 months of spending.

1.6 Main Results

1.6.1 Spending: Survey versus Administrative Data

Existing research has generally studied the impact of economic expectations on consumption using survey data, which may suffer from a number of well-known limitations, such as measurement error, selection problems, and surveyor demand effects. To assess whether this is a source of concern in our study, we explicitly test the relationship between self-reported consumption plans and actual future consumption in our data. If survey measures of consumption track actual consumption closely, measuring consumption in administrative data has few benefits. If, however, there is a disconnect between survey responses and actual spending, this would suggest that using administrative data could be crucially important to avoid spurious results.

Our survey elicited expectations about future credit card spending by asking respondents whether they expect their credit card spending to increase, decrease, or remain the same. Comparing these self-reported consumption plans to actual spending can reveal the extent to which survey measures predict actual consumption. Figure 1.5 presents the results. The x-axis corresponds to the actual change in monthly credit card spending in the 3 months after the survey completion. The y-axis corresponds to self-reported consumption plans on a 3-point scale from -1 (“go down”) to +1 (“go up”).

We find a marginally statistically significant ($p = 0.060$) relationship between the expected change in credit card expenditures and the actual change in spending, indicating that self-reported consumption plans have some information content. This relationship is, however, weak. The estimated slope (0.040) implies that a one-standard-deviation increase in actual future expenditures is associated with an increase in expected future expenditures of only 0.03 standard deviations.[21] This is also highlighted by a low R-squared of 0.021. These results indicate that survey predictions are a useful but very weak indicator of actual future spending.

There are several possible explanations for this finding. One possibility is that individuals may make consumption decisions spontaneously and are therefore not very good at predicting their spending over longer time horizons. Alternatively, individuals may have a clear idea of their future spending but may fail to follow through on their plans, for example due to a lack of self-control or financial constraints (although the latter is unlikely given that all participants of our experiment have access to credit card borrowing by definition). Another explanation could be measurement error and different types of response bias. Consumers may have a clear idea of their future spending but fail to communicate this accurately in their survey responses. This interpretation is somewhat unlikely to apply in our population, given that the participants of our experiment are highly educated, financially experienced (87% have a College degree) and that we elicited survey expectations following standard questionnaires that were adapted for this specific population.

Taken together, the evidence suggests that using survey data to measure treatment effects may be misleading, and provides a strong rationale for using administrative data to measure actual rather than planned consumption.

1.6.2 Prior Beliefs

Figure 1.6 shows the distribution of inflation and exchange rate expectations at baseline. In Figure 1.6.a, we plot the distribution of prior beliefs about the future inflation rate. Mean (3.39 pp) and median (3 pp) inflation expectations at baseline are fairly close to the expert forecast (2.3 pp) and higher than the most recent observed inflation rate at the time of

[21] The standard deviation of the variable shown in the x-axis is $570, and the standard deviation of the variable shown in the y-axis is 0.665.

the experiment (1.4 pp).[22] There is, however, significant dispersion in predictions across individuals, with individuals in the bottom decile of the distribution predicting an inflation rate of 0 pp and individuals in the top decile of the distribution predicting an inflation rate of 10 pp. In Figure 1.6.b, we plot the distribution of exchange rate expectations. The figure shows that beliefs about the future exchange rate follow a similar pattern as those for the future rate of inflation: prior expectations about the nominal exchange rate (4.13 MYR per US$) are centered close to the expert forecast with a mean and median of 4.10 MYR per US$, but there is significant dispersion in individual predictions, with some individuals (bottom 10%) expecting the exchange rate to rise to 3.90 Ringgit per US Dollar and others (top 10%) expecting it to decline to 4.40 Ringgit per U.S. Dollar.

The finding that expectations are centered around the professional forecast but dispersed has been documented widely in the literature on inflation and exchange rate expectations (Armantier et al., 2016; Cavallo et al., 2017), as well as in other contexts (see, for example, Fuster et al., 2020). Our information-provision experiment leverages this dispersion in prior beliefs.

1.6.3 Effect of Information on Posterior Beliefs

We next examine how the information feedback provided through our treatment conditions affects macroeconomic expectations. To do so, we use the standard econometric approach that has been used in information-provision experiments on a wide range of topics, such as inflation (Armantier et al., 2016; Cavallo et al., 2017), cost of living (Bottan and Perez-Truglia, 2020a), and housing prices (Fuster et al., 2020). We show that agents update their beliefs in response to the information provided in the experiment and find learning rates comparable to those in similar information experiments (see Roth and Wohlfart 2020 for a

[22] The 1.4 pp annual rate of inflation corresponds to the estimate for July 2019, and is obtained from the Malaysian Department of Statistics.

review).[23]

Let subscript i index the participants of our experiment and denote $\pi_{i,t}^{prior}$ as individual i's prior belief about the inflation rate, where t denotes the point in time when the belief is elicited and π the expected inflation between time t and t + 12 months. This is the belief about the inflation rate right before the individual reaches the information-provision stage of the experiment. Let $\pi_{i,t}^{signal}$ be the value of the signal that we may or may not show to the individual (the expert forecast at time t of the inflation rate in 12 months). Let $T_{i,t}^{\pi}$ be a binary variable that takes the value 1 if individual i is shown the signal and 0 otherwise. We denote the corresponding posterior belief as $\pi_{i,t}^{post}$. That is, the expected inflation rate after the individual sees, or does not see, the signal.

When priors and signals are distributed normally, Bayesian learning implies that after the individual sees the signal, the mean of the posterior belief should be a weighted average between the signal and the mean of the prior belief, $\pi_{i,t}^{post} = \alpha \cdot \pi_{i,t}^{signal} + (1-\alpha) \cdot \pi_{i,t}^{prior}$, where the parameter α depends on the relative precision of the prior belief and the signal (Hoff, 2009). The parameter α, the learning rate, ranges from 0 (individuals ignore the signal) to 1 (individuals fully adjust to the signal). We can rearrange this identity as follows:

$$\pi_{i,t}^{post} - \pi_{i,t}^{prior} = \alpha \cdot \left(\pi_{i,t}^{signal} - \pi_{i,t}^{prior}\right) \tag{1.2}$$

In other words, the Bayesian model predicts that the belief updates ($\pi_{i,t}^{post} - \pi_{i,t}^{prior}$) should be a linear function of the gap between the signal and the prior belief ($\pi_{i,t}^{signal} - \pi_{i,t}^{prior}$). That is, respondents who overestimate the inflation rate will revise their expectations downward when shown the signal, and those who underestimate the inflation rate will revise their beliefs upward when shown the signal. The model also predicts that the slope of that relationship should be equal to the learning rate, α.

[23]We find learning rates of .47 and .32 for inflation and exchange rate expectations, respectively. These are close to the median learning rate of information experiments surveyed in Roth and Wohlfart (2020), which range from .08 to .88.

In practice, several spurious reasons may explain why individuals revise their beliefs in the direction of the feedback, even if they received no signal. For example, some may take additional time to think when asked a question a second time and may get closer to the truth as a result. This may be particularly true in phone surveys where participants interact with a caller and may feel social pressure to report different beliefs when asked about their expectations again, even if they were not given new information. To allay concerns of such potentially spurious updating, we exploit the randomized assignment from the information provision experiment, following standard specifications in the literature (see Armantier et al., 2016; Cavallo et al., 2016):

$$\pi_{i,t}^{post} - \pi_{i,t}^{prior} = \alpha \cdot \left(\pi_{i,t}^{signal} - \pi_{i,t}^{prior}\right) \cdot T_{i,t}^{\pi} + \beta \cdot \left(\pi_{i,t}^{signal} - \pi_{i,t}^{prior}\right) + \epsilon_i \tag{1.3}$$

In this specification, the parameter β picks up spurious reversion towards the signal and α picks up true learning (i.e., changes in beliefs caused by the information provision) above any spurious revisions. Note that we do not expect subjects to fully update to the signal we provided ($\alpha = 1$) because the signal is an expert forecast that most respondents will correctly interpret as uncertain. Moreover, some individuals may not fully trust the source of the forecast and therefore place lower weight on the forecast. Nevertheless, we should expect α to be significantly greater than zero.

The same logic applies to expectations about the nominal exchange rate. Let $d_{i,t}^{prior}$ denote participant i's prior belief about the depreciation rate (i.e., the growth rate of the nominal exchange rate) before the individual reaches the information-provision experiment. Let $d_{i,t}^{signal}$ be the value of the signal that we may or may not show to the individual (i.e., the forecast). Let $T_{i,t}^{d}$ be a binary variable that takes the value 1 if we showed that signal to individual i and 0 if not. Denote $d_{i,t}^{post}$ as the corresponding posterior belief, that is, the expected depreciation rate after the individual sees, or does not see, the signal.

Our experiment provides respondents with information about inflation and the nominal exchange rate. Thus, it is possible that individuals use feedback about the inflation rate to update beliefs about the exchange rate and vice versa. Indeed, we might expect this type of cross-learning based on macroeconomic evidence. For example, after a devaluation of the local currency, there is partial pass-through to inflation (Dornbusch, 1987). We therefore expand the learning model to accommodate the possibility of cross-learning and estimate the following set of equations:

$$\begin{aligned}\pi_{i,t}^{post} - \pi_{i,t}^{prior} &= \alpha_1 \cdot \left(\pi_{i,t}^{signal} - \pi_{i,t}^{prior}\right) \cdot T_{i,t}^{\pi} + \alpha_2 \cdot \left(d_{i,t}^{signal} - d_{i,t}^{prior}\right) \cdot T_{i,t}^{d} + \\ &\quad \beta_1 \cdot \left(d_{i,t}^{signal} - d_{i,t}^{prior}\right) + \beta_2 \cdot \left(\pi_{i,t}^{signal} - \pi_{i,t}^{prior}\right) + X_{i,t}\gamma_1 + \epsilon_i\end{aligned} \tag{1.4}$$

$$\begin{aligned}d_{i,t}^{post} - d_{i,t}^{prior} &= \alpha_3 \cdot \left(\pi_{i,t}^{signal} - \pi_{i,t}^{prior}\right) \cdot T_{i,t}^{\pi} + \alpha_4 \cdot \left(d_{i,t}^{signal} - d_{i,t}^{prior}\right) \cdot T_{i,t}^{d} + \\ &\quad \beta_1 \cdot \left(d_{i,t}^{signal} - d_{i,t}^{prior}\right) + \beta_2 \cdot \left(\pi_{i,t}^{signal} - \pi_{i,t}^{prior}\right) + X_{i,t}\gamma_2 + \epsilon_i\end{aligned} \tag{1.5}$$

Note that this equation also includes a vector of control variables denoted $X_{i,t}$. Given random assignment, this vector of control variables should not change the point estimates but it can help absorb the variance of the error term and improve statistical power. We use the exact same set of control variables in all regressions presented in this paper: a set of 10 surveyor dummies, four dummies for the week of the year when the respondent completed the survey, the number of dependents, and 20 variables to control flexibly for the pre-treatment spending patterns.[24]

The main parameters of interest are α_1, measuring how individuals incorporate feedback about inflation into their inflation expectations, and α_4, measuring how individuals incorporate feedback about the exchange rate into their exchange rate expectations. The parameters α_2 and α_3 measure cross-learning by capturing how individuals incorporate exchange rate

[24]More specifically, we include a set of four variables with the average monthly spending over each of the last four quarters before the survey date, as well as the corresponding set of variables for each of the following spending categories: durable, tradable durable, and nondurable.

feedback into their inflation expectations and inflation feedback into their exchange rate expectations.

Before presenting the regression results, Figure 1.7 provides a graphical summary of the impact of our information treatments on macroeconomic expectations. Figure 1.7.a shows a binned scatterplot corresponding to the effects of the inflation feedback. The x-axis corresponds to the potential update in response to the provision of feedback (i.e., the difference between the feedback on inflation expectations and the corresponding prior belief). The y-axis shows the actual belief update (i.e., the difference between the posterior belief and the prior belief). The gray circles correspond to the control group (i.e., individuals who do not receive inflation feedback). The slope of this linear relationship (the gray line) corresponds to the coefficient β in the learning equation (3.3), which measures "spurious" learning. We find significant spurious learning, which is consistent with findings from related studies (see, for example, Fuster et al., 2020; Cullen and Perez-Truglia, 2022).[25] In turn, the red squares correspond to the treatment group (i.e., individuals who receive the inflation feedback). Most importantly, the slope of the relationship is significantly larger ($p < 0.001$) in the treatment group (0.472) than in the control group (0.247). This difference in slopes corresponds to the coefficient α from the learning equation (3.3) (i.e., the true rate of learning that can be attributed to the information provision). Figure 1.7.b is similar to Figure 1.7.a, but reports updating on exchange rate expectations instead of inflation expectations. Again, consistent with genuine learning from the feedback, we find that the slope is stronger in the treatment group than in the control group, although the difference is smaller in magnitude (0.317 vs 0.255) and statistical significance ($p = 0.048$).

We next turn to the regression results, presented in Table 1.4. The first two columns of this table correspond to the regression specifications given by equations (1.4) and (1.5),

[25] In terms of magnitude, however, the degree of spurious learning seems larger in our data. Our preferred interpretation for this difference is that, unlike other surveys experiments that are conducted online, our survey was conducted via phone. As a result, some individuals may have felt pressured to revise their posterior beliefs even if they did not receive any feedback.

respectively. In column (1), the dependent variable is the updating on inflation expectations. In column (2), the dependent variable is the updating on the expected exchange rate depreciation. These results differ from the simpler binned scatterplots in Figure 1.7 in that they include additional control variables and allow for cross-learning. Table 1.4 reports the coefficients of the two key independent variables, corresponding to the interactions between the treatment assignments and the size of the information shock. For simplicity, we refer to these variables as information shocks.

The first coefficient from column (1) of Table 1.4 indicates that information about inflation has a significant effect on inflation expectations: a 1 pp increase in inflation shock increases inflation expectations by 0.236 pp ($p < 0.001$). The second coefficient from column (1) of Table 1.4 is close to zero (-0.030) and statistically insignificant ($p = 0.189$), indicating that the information shock about the exchange rate does not have a significant effect on inflation expectations. In other words, individuals use the feedback in a compartmentalized manner.

The magnitude of the pass-through from the inflation feedback to the inflation expectations is in the same order of magnitude as the pass-through estimated in other information experiments. For example, Bottan and Perez-Truglia (2020a) shows that a 1 pp increase in feedback about future home prices increases the home price expectations by 0.205 pp. However, the degree to which subjects incorporate the information is lower than that reported in other studies. For example, Cavallo et al. (2017) find that, when forming inflation expectations, the average Argentine respondent assigns a weight of 0.432 to the feedback and the remaining 0.568 to their prior beliefs (coefficient α-statistics reported in Panel B, column (1) of Table 1). The fact that individuals are less prone to incorporating information in our context may reflect a more educated and financially savvy population that has more confidence in their prior beliefs. However, this difference in rates of learning could be attributed to differences in the survey methods. For example, other studies provide information and elicit beliefs on a computer screen, whereas our study uses phone surveys, which could arguably make the information less salient. Also, other studies where subjects are paid to fill out the survey could generate experimenter demand effects. Subjects in our survey were not paid for their participation.

The second coefficient in Table 1.4, column (2), indicates that information about the exchange rate has a significant effect on exchange rate expectations: a 1 pp information shock increases expectations of a nominal exchange rate depreciation by about 0.064 pp ($p = 0.038$). Again, we find compartmentalized learning about the exchange rate: the first coefficient in column (1) is close to zero (0.032) and statistically insignificant ($p = 0.214$), indicating that information about the inflation rate does not have a significant effect on participants' exchange rate expectations. We find that the magnitude of the learning effects for the exchange rate (coefficient of 0.064) is lower than the magnitude of learning effects for inflation (0.236), and the difference between the two is statistically significant ($p < 0.001$). Following a Bayesian learning approach, we offer two potential interpretations for this difference. First, individuals may have stronger prior beliefs about the exchange rate than about the inflation rate. This interpretation is consistent with the evidence documented in Section 1.3.1 showing substantially more interest in learning about the exchange rate than the rate of inflation, presumably because it is more consequential for everyday economic decisions. An alternative interpretation is that individuals do not trust the precision of the signal. That is, they are less likely to trust expert forecasts about inflation than about the exchange rate. However, as we do not provide specific information about the sources of inflation and exchange rate forecasts used in our experiment, this interpretation seems less likely.

1.6.4 Effect of Information on Spending

Having shown that our information treatments are effective at shifting beliefs, we turn to their impacts on consumption. The main goal of our experiment is to test whether changes in macroeconomic expectations affect actual spending, as measured in administrative data covering the universe of credit card transactions for bank customers in our sample. To examine this question, we estimate the following regression equation:

$$\begin{aligned} Y_{i,t+1} &= \alpha_Y^{\pi} \cdot \left(\pi_{i,t}^{signal} - \pi_{i,t}^{prior}\right) \cdot T_{i,t}^{\pi} + \alpha_Y^{d} \cdot \left(d_{i,t}^{signal} - d_{i,t}^{prior}\right) \cdot T_{i,t}^{d} + \\ &\quad \beta_Y^{\pi} \cdot \left(d_{i,t}^{signal} - d_{i,t}^{prior}\right) + \beta_Y^{d} \cdot \left(\pi_{i,t}^{signal} - \pi_{i,t}^{prior}\right) + X_{i,t}\gamma_Y + \epsilon_i \end{aligned} \tag{1.6}$$

Note that the right-hand-side of equation (1.6) is identical to the learning equations in (1.4) and (1.5). The only difference is that the dependent variable is now a generic outcome $Y_{i,t+1}$. For example, this dependent variable may be the average monthly spending in the 3 months post-treatment. Recall that in the set of control variables ($X_{i,t}$), we include the pre-treatment spending, which exploits the persistence in spending patterns to help reduce the variance of the error term and improve statistical power (see McKenzie, 2012).

Table 1.4 reports the results. In columns (1) and (2), we estimate the relationship between the information shock and the resulting change in self-reported macroeconomic expectations. The results confirm that participants update their macroeconomic expectations in response to the information provided to them through our experiment. In Table 1.4, columns (3) through (6), we use the same empirical specification to examine whether the changes in macroeconomic expectations documented in columns (1) and (2) translate into changes in consumption behavior. To do so, we estimate equation (1.6), with spending on durables, tradable durables, credit card debt, and total spending as the respective outcomes. Each outcome is measured in the administrative data obtained from our partner bank. We observe credit card spending for 3 months after the intervention and average the monthly spending over the entire period to mitigate concerns about outliers or seasonality of expenditures. Total credit card debt is the amount of debt outstanding after the monthly repayment due date.

The results in columns (3) through (5) of Table 1.4 test the key predictions of the theoretical framework presented in Section 1.2. The first coefficient in column (3) measures the effect of the inflation shock on durables consumption. This is a direct test of Proposition 1, which states that spending on durables should increase with expected inflation. We do not find support for this prediction in the data. While the point estimate has the correct sign, the coefficient is small in magnitude (1.646) and not statistically different from 0 ($p = 0.556$). The information shock delivered by our experiment moves inflation expectations by an average of 0.088 standard deviations, but our estimate implies that it increases average monthly spending on durables by only 0.005 standard deviations (or less than \$2).

Similarly, the second coefficient estimate in column (4) of Table 1.4 provides a test of Proposition 2, which states that a decrease in the expected exchange rate (an increase in the expected rate of depreciation) should increase spending on tradable durables. We also do not find evidence consistent with this prediction. The point estimate is negative, small in magnitude (-2.514), and not statistically significant ($p = 0.196$). While our intervention moves exchange rate expectations by an average of 0.023 standard deviations, this coefficient estimate implies a negligible impact on spending on tradable durables, shifting expenditures in this category by only 0.01 standard deviations.

Finally, column (5) of Table 1.4 shows that, consistent with Proposition 3, an increase in the expected inflation rate leads to a slight increase in total credit card debt. However, this coefficient is statistically insignificant and small in magnitude: it implies that for each 1 pp increase in the inflation shock, individual credit card debt increases by just 0.013 standard deviations.

In column (6), we consider an additional hypothesis that is not motivated by intertemporal consumption models. As argued by Coibion et al. (2019), individuals may see future inflation and exchange rate depreciation as signs of a weak economy. According to that view, an increase in expected inflation and depreciation may discourage the individual from spending in general, for precautionary reasons. To explore this additional hypothesis, column (6) of Table 1.4 uses total spending as the dependent variable. We do not find any evidence that inflation and exchange expectation shocks have significant effects on total spending. The coefficients are negative but economically small and statistically insignificant. For example, a 1 pp inflation shock reduces total spending by just 0.002 standard deviations ($p = 0.776$), while a 1 pp depreciation shock reduces total spending by just 0.005 standard deviations ($p = 0.480$).

As an additional robustness test to check the regression specification, we leverage data on pre-treatment spending, which allows us to conduct a falsification test in the spirit of an event-study analysis. We estimate the same regression but use pre-treatment instead of post-treatment spending as the dependent variables. The outcomes are measured before participants receive information and thus should not show effects of the information on pre-treatment spending. Appendix 1.C.1 presents the results. As expected, we find no effects of the information shocks on pre-treatment spending outcomes.

1.6.5 Magnitude of Coefficients

In the previous section, we show that our information treatments shift expectations but do not have a statistically significant effect on consumption. However, this does not necessarily mean that the effects are precisely zero. To get a better quantitative sense of the effect sizes, we take a hypothetical information shock of 1 pp and estimate its impact on the outcomes of interest in terms of dollars and standard deviations. We first consider the effect of an inflation shock on durable consumption and find that a 1 pp information shock is predicted to increase durable spending by a statistically insignificant \$1.646, equivalent to less than 0.005 standard deviations of the corresponding outcome. To examine the possibility of an undetected increase in durable spending, we inspect the confidence interval of our estimate. The upper bound of the 95% confidence interval is approximately 7.12, which rules out positive effects larger than \$7.12. Relative to the standard deviation of the outcome variable, we can rule out effects above 0.021 standard deviations, which is very close to zero.

Note that our estimates are intention-to-treat (ITT) effects, because the information shock given to the subjects only partially translates into changes in their posterior beliefs. For this reason, we refer to equation (1.6) as the reduced-form effects of the information experiment. For a more direct measure of the effect of expectations on behavior, we can use an instrumental variables version of the reduced-form equation but with two endogenous variables corresponding to the belief updates for inflation and exchange rate expectations.

We report the results from the instrumental variables regressions in Table 1.5. The first prediction of interest is that an increase in inflation expectations should increase durable expending. The coefficient on inflation expectations in Table 1.5, column (1), indicates that a 1 pp increase in inflation expectations causes an increase in durable spending of $13.4, or only 0.039 standard deviations for this outcome. Looking at the upper bound of the 95% confidence interval, we rule out an increase in this outcome above $38.5, or 0.11 standard deviations. This suggests that, while we cannot rule out that inflation expectations have small effects on spending behavior, we can rule out moderate to large effects.

The results are similar for other hypotheses that we tested. The second prediction of our theoretical framework is that an increase in expected depreciation should increase spending on tradable durables. Contrary to this prediction, the coefficient from column (2) indicates that a 1 pp increase in the expected devaluation reduces rather than increases spending on tradable durables by $34.4, which is equivalent to a reduction of 0.138 standard deviations. Inspecting the bounds of the 95% confidence interval, we rule out an increase of more than 0.116 standard deviations for this outcome. The third prediction of the theoretical framework is that an increase in inflation expectations should increase credit card borrowing. The coefficient from column (3) indicates that a 1 pp increase in expected inflation increases credit card debt by only $31.7, which is equivalent to 0.042 standard deviations. Moreover, the 95% confidence interval rules out positive effects of more than 0.122 standard deviations. Last, column (4) shows the effects of expectations on total spending. The results indicate that a 1 pp increase in inflation expectations increases total spending by $1.54, or 0.002 standard deviations. In turn, a 1 pp increase in depreciation expectations decreases total spending by \$64.3, or 0.071 standard deviations.

These results involve tests of multiple related predictions, involving multiple combinations of expectations and spending margins. The small and statistically insignificant coefficients across the board suggest that while expectations may have some effect on spending behavior, those effects appear to be very small and therefore difficult to detect. As an additional test to rule out the presence of economically meaningful effects, we estimate the relationship between expectations and credit card spending using the full variation in expectations, rather than restricting our attention to the exogenous variation generated by our experiment. Table 1.6 presents the results. The results reported in the table correspond to the ordinary least squares equivalent of the instrumental variable regressions reported in Table 1.5. There is a simple trade-off between these two approaches. On the one hand, the experimental estimates provide better identification of the causal relationship between expectations and consumption. On the other hand, the OLS estimates exploit all available variation in expectations and thus lead to substantially more precisely estimated coefficients.

The results from the two approaches are qualitatively consistent: the estimated effects of expectations on behavior are close to zero and statistically insignificant. However, the OLS estimates from Table 1.6 are substantially more precisely estimated than the corresponding IV estimates from Table 1.5. As a result, the non-experimental estimates can rule out even smaller effects. Take for example the effect of inflation expectations on durable consumption. According to the coefficient from column (1) of Table 1.6, a 1 pp increase in inflation expectations is associated with a reduction in durable expenditures of less than \$2. If we take the upper bound of the 95% confidence interval, we can rule out increases in durable expenditures above \$2.01, which is equivalent to 0.002 standard deviations of that outcome and thus an arguably negligible effect. In summary, both experimental and non-experimental data support the conclusion that our estimates provide evidence of null or small effects of macroeconomic expectations on spending behavior.

1.6.6 Mechanisms: What Explains the Absence of a Spending Response?

There are several potential mechanisms that might explain why the significant changes in macroeconomic expectations, induced by the information experiment, do not translate into changes in credit card spending matching those of a standard model of intertemporal consumer choice. In this section, we present additional tests and results from a follow-up survey experiment to explore which of these mechanisms are most consistent with the pattern of our results.

The "mental model" experiment. We conduct an additional mental model survey experiment to test several candidate mechanisms that might explain the absence of a spending response in our main experiment. This follow-up experiment is similar to our main information intervention, but designed to elicit responses to different exchange rate and inflation scenarios while also randomizing whether expected nominal income is held constant and eliciting several measures of respondents' financial sophistication. In the first step of the experiment, we elicit each participant i's prior beliefs about inflation θ_i^π or the exchange rate θ_i^d, using the same approach as in our main experiment. In the second step, we randomly assigned respondents to one of four inflation or exchange rate depreciation scenarios. The scenarios presented to respondents are defined in reference to their prior beliefs. Specifically, respondents are asked to consider a scenario in which the realized inflation or exchange rate depreciation is $\theta_i^{\pi,d} + \Delta x_i^k$ and the randomly assigned percentage change takes one of the values $\Delta x_i^k \in \{-10, -3, 3, 10\}$. We ask participants about their spending response and elicit demand for inflation and exchange rate indexed securities in each scenario. Finally, we included a brief survey measuring basic indicators of financial literacy.[26] We combine evidence from the mental model exercise with results from our main experiment to test alternative mechanisms that might explain the absence of a spending response to changes in

[26] We find that, overall, financial literacy in our study population is relatively high. For example, 58.5% percent of respondents are able to answer at least two of the standard "big three" financial literacy questions correctly.

macroeconomic expectations.

Time inconsistency. Before turning to the results of the mental model exercise, we explore two mechanisms that can be tested with data from the main experiment. First, one possible explanation for the absence of a spending response is that individuals update their intended behavior but cannot follow through on their consumption plans, for example due to self-control problems or liquidity constraints. Since our survey collected data on spending plans, we can test this hypothesis directly. We estimate equation (1.6) using self-reported spending plans measured post-treatment as the dependent variable, rather than actual spending observed in the credit card data.

The results show that individuals do not change their self-reported spending plans in response to new information about inflation or the exchange rate, as shown in Table 1.7. For reference, columns (1) and (2) reproduce the treatment effects of information shocks on the inflation and exchange rate expectations. In columns (3) through (6), we report the results of estimating equation (1.6) using self-reported spending plans for the four specific subcategories of spending. Each of these outcomes is measured on a three-point scale that takes the values -1 (if the respondent anticipates spending less in the future), 0 (if they anticipate spending about the same), or +1 (if they anticipate spending more). Each outcome in columns (3) to (6) of Table 1.7 corresponds to the survey equivalents of the consumption behavior measured with administrative data in columns (3) to (6) of Table 1.4. In column (3), the dependent variable is the stated intention to increase or decrease spending on durable goods. The prediction from the macroeconomic model is that higher inflation expectations should increase intended spending on durables.

We find no evidence of such an effect. The impact of increased inflation expectations on total consumption is close to zero (-0.006), statistically insignificant and small in magnitude: a 1 pp inflation shock is associated with a reduction in expected durable spending of only 0.001 standard deviations. In column (4), the dependent variable is intended future spending on electronic goods, which is our survey proxy for durable tradables. These estimates test the prediction that expected depreciation should result in an increase in planned spending on durable tradables. Contrary to this prediction, the coefficient on the exchange rate shock is close to zero (0.008) and statistically insignificant. In column (5), the dependent variable is the expected change in borrowing. The theoretical prediction is that an increase in inflation expectations will lead to higher expected borrowing. We do not find support for this hypothesis. The coefficient on inflation shock is close to zero (-0.003) and statistically insignificant. In column (6), the dependent variable is the intention to increase total spending. This regression tests the hypothesis that expectations of inflation or exchange rate depreciation may be interpreted as a sign of an overall economic slowdown and should therefore lead to a decline in spending. Instead, we find that the coefficients on the inflation shock and the exchange rate shock are both close to zero (-0.006 and 0.003) and statistically insignificant.

Transaction amounts. It is also possible that changes in macroeconomic expectations are not reflected in credit card spending because macroeconomic considerations are not sufficiently salient when consumers make small purchases and only enter consumers' decision making in the case of large purchases. We test this mechanism using data from the main experiment and disaggregating consumers' credit card transactions by amount and type of transaction and estimating treatment effects separately for large and small purchases.

To implement this, we first use administrative data on credit card purchases in the pre-treatment period and construct an indicator variable for transactions in the top quartile (large transactions) and the bottom quartile (small transactions) of the pre-treatment distribution of credit card purchases. We then apply the cutoffs from the pre-treatment distribution to identify large and small purchases in the post-treatment data and estimate separate treatment effects for large and small transactions, using our standard specification. In all regressions, we normalize the spending variables to avoid bias due to the inherent difference in transaction sizes between the top and bottom quartiles of the distribution. Intuitively, if the lack of a spending response to updated macroeconomic expectations is due to insufficient salience of macroeconomic considerations in small transactions, one would expect anticipated inflation or currency depreciation to have no effect on small purchases but affect spending for larger purchases. Table 1.8 presents the results. We find that the treatment effects are overall close to zero and not statistically different between the small and large purchase categories.

As an additional test, Table 1.12 in the Supplementary Appendix uses a different classification of large and small purchases based on the credit card spending categories in the administrative data that have the largest average transaction sizes. In Table 1.12, we compare the treatment effect for the three categories with the largest median transaction size to the treatment effect for the three categories with the smallest median transaction sizes. We find that, again there is no economically or statistically significant difference in the treatment effect between categories with large and small transaction amounts. Taken together, this suggests that insufficient salience of macroeconomic considerations in small transactions is not the mechanism that can explain the absence of a spending effect in response to revised macroeconomic expectations.

Impact of information over time. In addition to transaction amounts, another possibility that could explain the absence of a spending response could be that the impacts of the information treatment are not sufficiently long lasting to affect consumption choices. While we cannot test this possibility directly, growing evidence in similar settings indicates that participants retain information provided in the context of an information treatment for months after the experiment (see Cavallo et al., 2017; Bottan and Perez-Truglia, 2020a) or even a year later. This suggests that it is unlikely that participants in our experiment simply discard the acquired information when making consumption decisions.Another result that is inconsistent with the information treatments not being long-lasting enough is that the treatment effects we find in the main experiment are not zero across the board, we do find significant reductions in expenditures on durable goods, for example. While the sign of the effects is inconsistent with a standard model of intertemporal consumer choice, the presence of an effect suggests that the information provided in our experiment did affect some dimensions of consumer choice at a longer time horizon.

Consumer sophistication. Another mechanism that could explain the absence of a spending response is the possibility that agents fail to optimize due to limited financial sophistication or financial literacy, a pattern that has been documented in many similar household finance settings (Campbell et al. 2011; Beshears et al. 2018; Ponce et al. 2017). Specifically, in the context of our experiment it is possible that consumers change their expectations in response to expert forecasts, but simply do not know how to respond optimally to changes in inflation or the exchange rate.

We examine this mechanism using a series of tests based on standard measures of financial literacy we collected as part of the mental model experiment. In the first test, we elicit a simple measure of financial literacy using the standard "big three" financial literacy questions and examine whether responses to the inflation and exchange rate scenarios presented in the mental model experiment differ systematically for more financially literate respondents. Specifically, we split the sample of the mental model experiment into high financial literacy (above median) and low financial literacy (below median) respondents and compare their responses to the randomly assigned inflation and exchange rate scenarios using within-person variation.

Table 1.9, Panel A, first confirms that the reaction to anticipated changes in inflation and the exchange rate matches that of our main experiment. Table 1.9, Panel B, reports the results separately for respondents with high and low financial literacy. Interestingly, we find that the negative effects of the high inflation and exchange rate depreciation scenarios on durable consumption, which point toward the anticipation of nominal rigidities, appear to be driven entirely by the more financially literate respondents in our sample. This indicates that, rather than limited financial literacy explaining the lack of transmission from expectations to actual spending, it is the most financially literate consumers who reduce durable spending in a way that counteracts the predictions of a standard model of intertemporal choice. This is arguably because the more sophisticated consumers in our sample understand that their income is not indexed to inflation or the exchange rate and they correctly anticipate their real income to decline in the high inflation or exchange rate depreciation scenario.

Nominal rigidities. To provide a more direct test of the hypothesis that consumers are sophisticated enough to anticipate nominal rigidities, we examine whether customers with large updates in inflation and exchange rate beliefs have demand for inflation indexed or exchange rate indexed securities. In one arm of the mental model experiment, participants were presented with an hypothetical asset indexed to either inflation or the exchange rate and asked whether they would be interested in buying this asset. The results, reported in Table 1.10, show that consumers with large updates in their inflation and exchange rate expectations have substantially higher demand for the indexed asset, both in the simple updating experiment and when they are additionally assigned to receive the fixed income script. The fact that demand for indexed securities is not reduced by the fixed income condition provides additional suggestive evidence in favor of the hypothesis that consumers reduce spending in anticipation of nominal rigidities. This result resonates with the hypothesis, proposed in Christiano et al. (1999), that when making consumption decisions, consumers care about the wage Philipps curve rather than the price Phillips Curve. It is also consistent with recent work that finds a low passthrough from inflation expectations to income growth expectations (see Hajdini et al., 2022).

Overall, the results suggest that the lack of a spending response to information shock cannot be explained by a lack of consumer sophistication. On the contrary, consumers are sophisticated enough to associate exchange rate depreciations and inflation with a decline in real income and reduce spending for precautionary reasons, which attenuates the possibility of a positive spending response to higher expected inflation and exchange rate depreciations.

In summary, the pattern of results suggests that the absence of a spending response to changed macroeconomic expectations is not the result of an anticipated worsening in the overall economic situation, time inconsistency, lack of financial literacy, insufficient salience of macroeconomic considerations. Instead, the results suggest that consumers are sophisticated enough to anticipate nominal rigidities that will erode the purchasing power of their non-indexed income and reduce their consumption of durable goods for precautionary reasons. This counteracts the spending effects predicted by standard models of intertemporal optimization and contributes to the absence of an overall credit card spending response to revised macroeconomic expectations that matches the predictions of a standard model of intertemporal consumer choice.

1.7 Conclusion

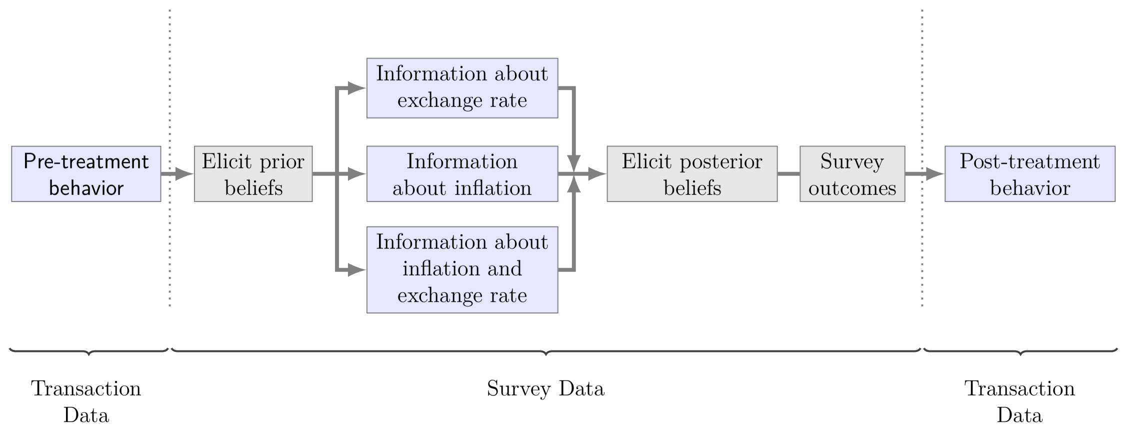
How do macroeconomic expectations affect individual consumption decisions? To explore this question, we conducted a field experiment with 2,872 credit card customers of a large commercial bank. We created exogenous variation in macroeconomic expectations through an information-provision experiment in which participants were provided with expert forecasts of inflation and the exchange rate. We then measure the effects of these information shocks on consumers' subsequent macroeconomic expectations, self-reported spending plans (measured in survey data), and actual spending (measured in administrative data). We test several predictions from a standard model of intertemporal consumer choice, such as whether an increase in inflation expectations increases spending on durables. We find that information provision shifts beliefs but does not change consumers' actual spending behavior as predicted by these models.

We test several mechanisms that could explain the absence of a spending response to the significant changes in macroeconomic expectations induced by our experiment. We show that consumers do not fail to optimize because of limited financial sophistication or behavioral factors, such as time inconsistency, commitment problems, or mistaken beliefs about the link between inflation, exchange rates and the state of the overall economy. Instead, the interpretation most consistent with our findings is that consumers correctly anticipate nominal rigidities and reduce expenditures, especially on durable goods, for precautionary reasons. This counteracts the effects predicted by standard models of intertemporal optimization and accounts for the absence of a spending response.

These results have direct implications for the transmission of macroeconomic policy. Many macroeconomic policies are explicitly based on the premise that changes in economic expectations will affect households' consumption choices. For example, central banks may try to engineer higher inflation expectations to stimulate spending (Bachmann et al., 2015), or they may try to manipulate expectations about the exchange rate to affect the consumption of foreign goods. Our results suggest that such policies might be ineffective, or at least less effective than previously believed, because consumers do not factor macroeconomic expectations into their consumption decisions in the manner predicted by standard economic models.

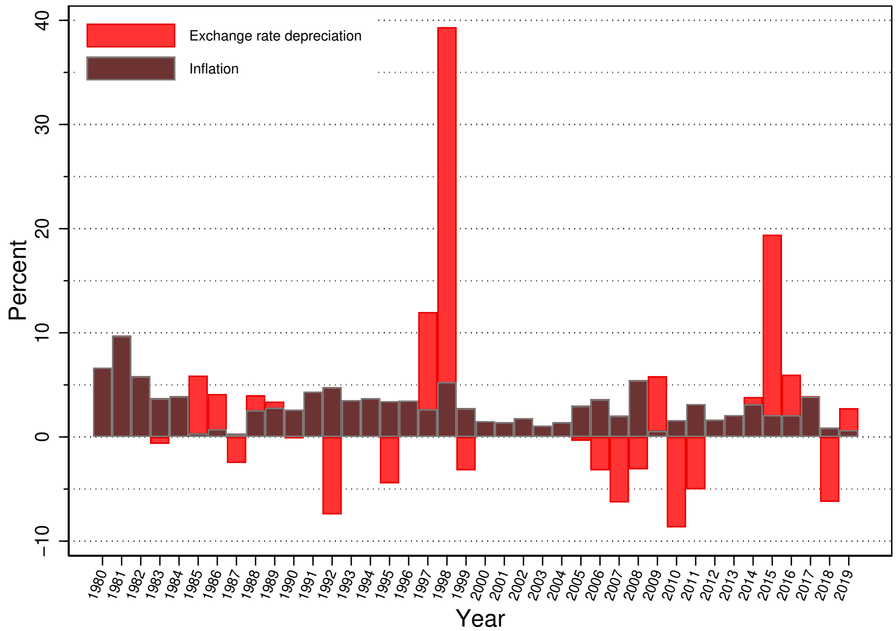
Our results also highlight the important role of consumer heterogeneity in the transmission of economic expectations to the real economy. We find that even within the relatively homogeneous population of our experiment, precautionary consumption reductions in response to updated inflation or exchange rate expectations are concentrated among more financially literate respondents. This is in line with the widely documented disagreement on macroeconomic expectations among households (Andre et al., 2022) and further complicates the task of predicting the aggregate effects of macroeconomic polices such as forward guidance on the real economy.

Figure 1.1: Experimental Design



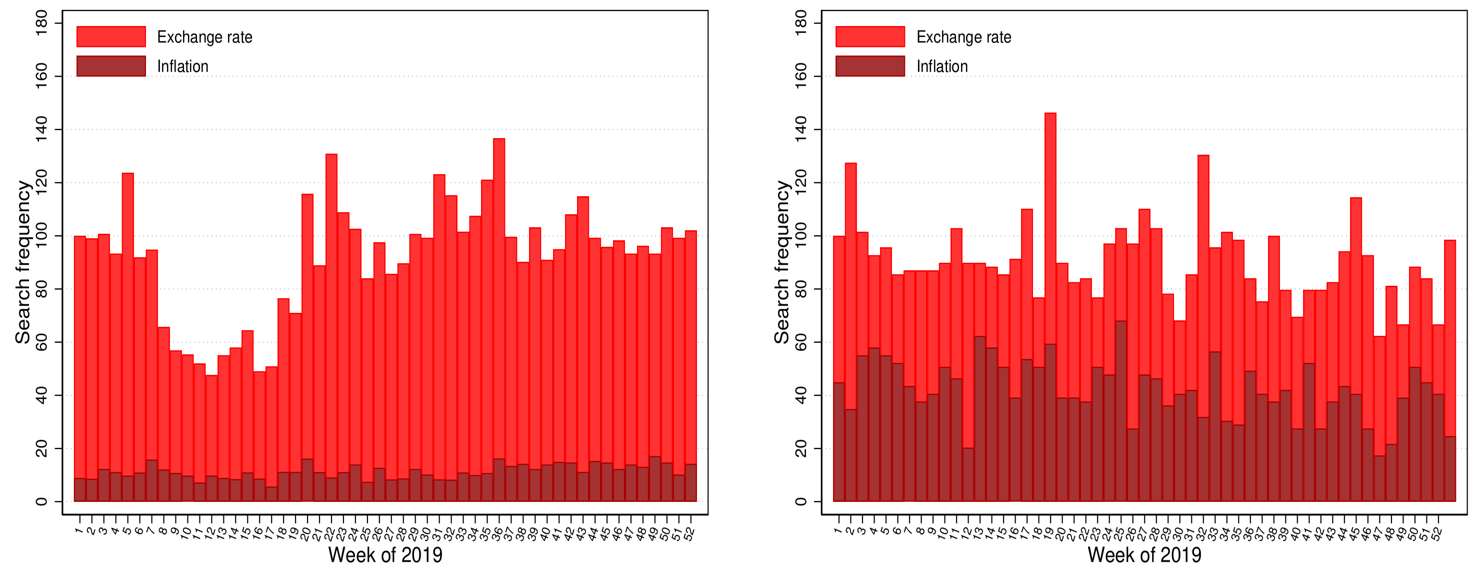
Notes: The figure summarizes the treatment conditions and timeline.

Figure 1.2: Inflation and Nominal Exchange Rate 1980-2019



Notes: The figure shows the time series of the annual inflation rate and the time series of changes in the nominal exchange rate of the Malaysian Ringgit against the U.S. Dollar for the period 1980-2019. Source: Federal Reserve Bank of St. Louis.

Figure 1.3: Public Interest in Inflation and the Exchange Rate

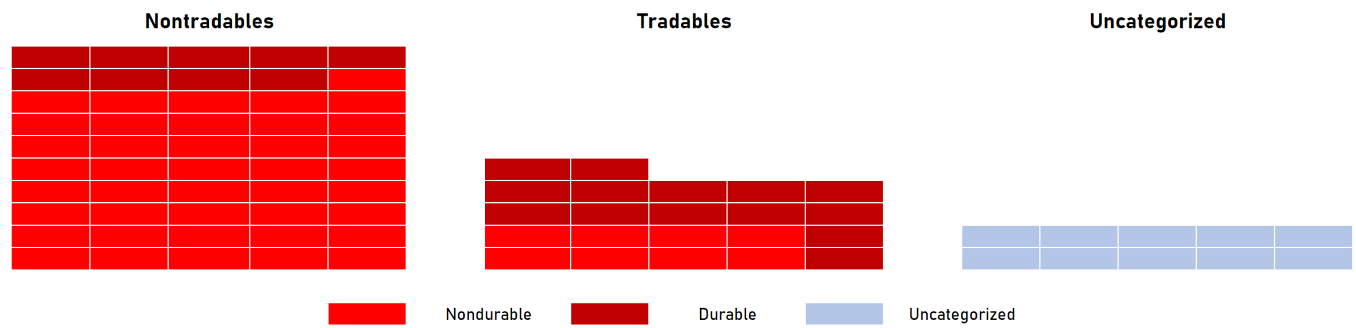


(a) Google Searches

(b) Newspaper Articles

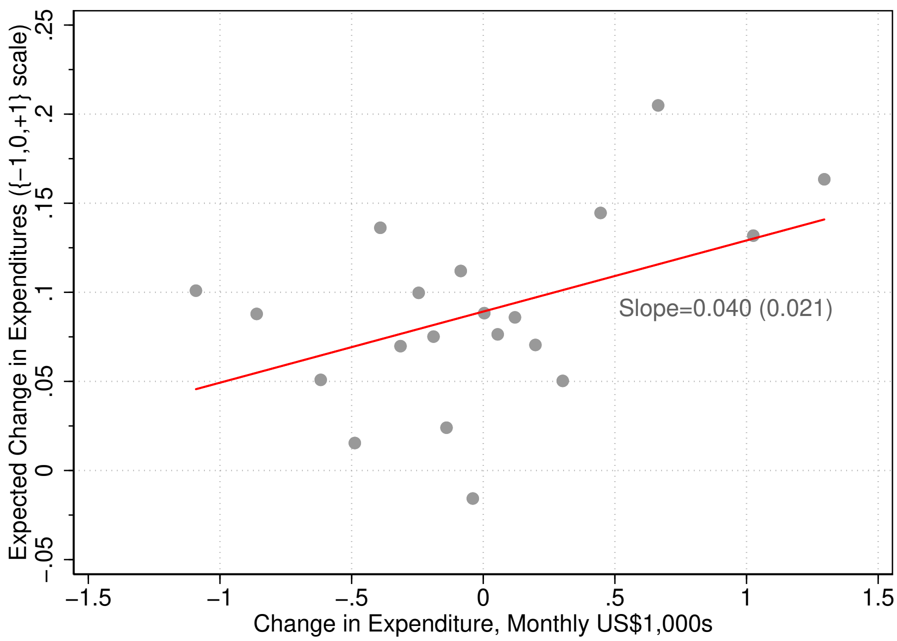
Notes: The figure shows descriptive statistics on public interest in inflation and the nominal exchange rate. Panel (a) shows the frequency of Google searches for the terms "inflation" and "dollar" in English and Malay between January and December 2019. Data on Google searches is reported only in relative terms with reference to a numeraire category. We therefore normalize the series so that exchange rate searches in the first week of 2019 are equal to 100. Panel (b) shows the frequency of articles containing the terms "inflation" and "dollar" in the country's most widely read English language newspaper between January and December 2019 (100=70 articles).

Figure 1.4: Expenditures by Category



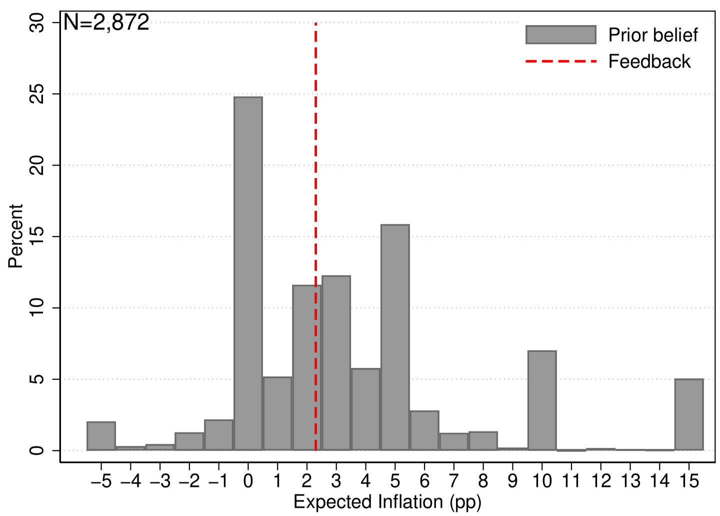
Notes: Each of the 82 squares in the figure represent $\frac{1}{82}$ of the total spending in the credit card data. The leftmost group of squares corresponds to spending on nontradable goods, the middle group corresponds to spending on tradables, the rightmost group corresponds to spending that cannot be categorized. The leftmost and middle groups are subdivided into nondurable spending and durable spending. All expenditures were categorized based on MCCs. For additional details, see Table 1.1.

Figure 1.5: Self-Reported Spending Plans versus Actual Spending

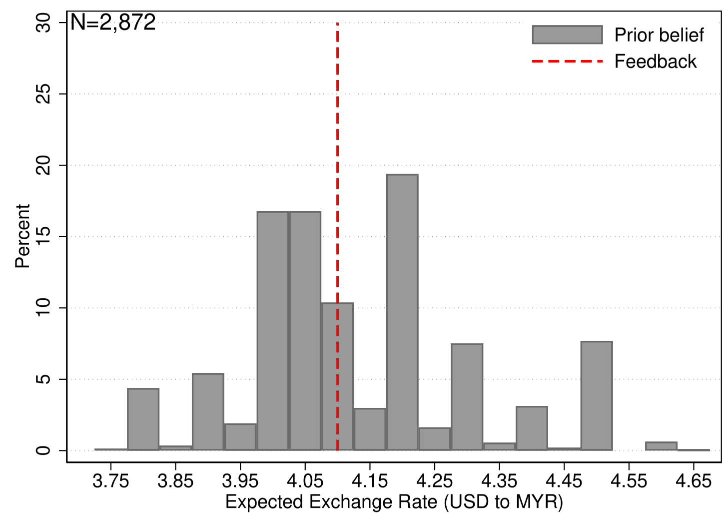


Notes: The figure shows the relationship between the actual change in credit card expenditures, measured in administrative data, and self-reported spending plans, based on survey data. The regression controls for surveyor and week fixed effects. Expenditure in administrative data is measured as the difference in average monthly expenditure across three months post-treatment (the post-survey period for which data is available) and average monthly expenditure for the twelve months pre-treatment (the pre-survey period for which data is available). The predicted change in expenditure corresponds to survey responses on planned credit card expenditure, recorded as 1 if a respondent expects to spend more, 0 if they expect to spend about the same, and -1 if they expect to spend less. 'Slope' is the OLS coefficient of the relationship, with robust standard errors in parentheses.

Figure 1.6: Distribution of Prior Expectations



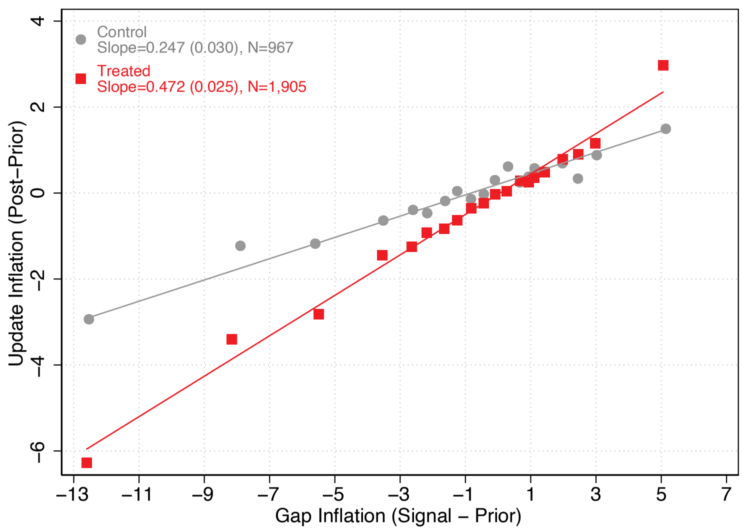
(a) Inflation Priors



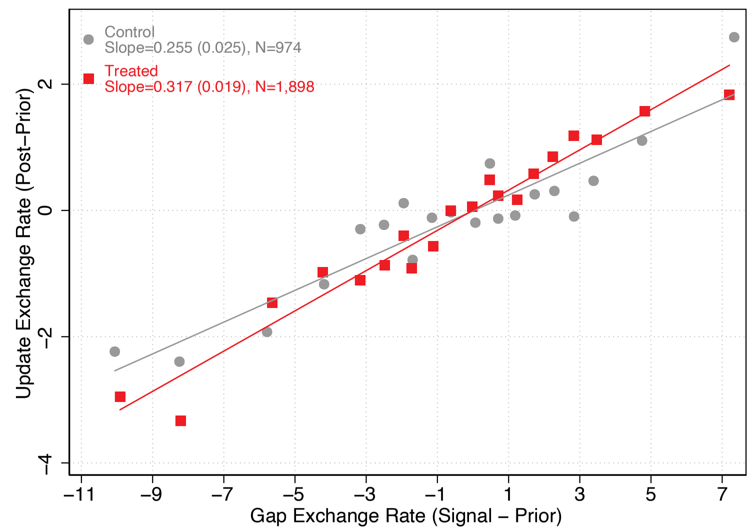
(b) Exchange Rate Priors

Notes: The figure shows the distribution of prior beliefs about future inflation in panel (a) and the future nominal exchange rate in panel (b), elicited prior to the information experiment for all survey respondents. Dashed vertical lines correspond to the feedback on the future inflation and exchange rate subsequently provided through our intervention. The mean (median) of inflation expectations is 3.39 pp (3 pp). The mean (median) of exchange rate expectations is 4.1 MYR/US$ (4.1 MYR/US$).

Figure 1.7: Belief Updating



(a) Inflation Expectations



(b) Exchange Rate Expectations

Notes: The figure shows the relationship between information shocks provided and changes in inflation expectations in panel (a), and information shocks provided and exchange rate expectations in panel (b). The x-axis in panel (a) plots the gap between the inflation signal shown to respondents and their prior inflation expectations $\pi_{i,t}^{signal} - \pi_{i,t}^{prior}$, while the y-axis plots the difference between prior and posterior inflation expectations $\pi_{i,t}^{post} - \pi_{i,t}^{prior}$. The x-axis in panel (b) plots the gap between the exchange rate signal shown to respondents and their prior exchange rate expectations $d_{i,t}^{signal} - d_{i,t}^{prior}$, while the y-axis plots the difference between prior and posterior exchange rate expectations $d_{i,t}^{post} - d_{i,t}^{prior}$. In panel (a), treatment and control groups denote whether the subject was chosen to receive feedback about the inflation rate or not. In panel (b), treatment and control groups denote whether the subject was chosen to receive feedback about the exchange rate or not. The analysis controls for number of dependents, week fixed effects, surveyors fixed effects, and 20 additional variables controlling for spending patterns during the four pre-treatment quarters.

Table 1.1: Durable and Tradable Expenditures

	Average monthly expenditure, USD (1)	Durables (%) (2)	Tradables (%) (3)
Airline and Travel	78	0	0
Apparel	30	100	100
Automotive	43	100	64
Books and Stationery	5	100	100
Business Service	47	0	0
Camera and Photo	2	33	33
Car Rental	2	0	0
Computer Equipment	13	100	100
Department Store	40	100	100
Dept Store	34	100	100
Dining	66	0	0
Direct marketing	39	0	0
Education	9	100	0
Electronics	28	100	66
Entertainment	6	0	0
Financial services	21	0	0
Food and beverage	61	0	0
Furniture	21	100	66
Government	21	0	0
Groceries	58	0	0
Health and beauty	30	0	50
Home improvement	17	100	60
Hotel	2	0	0
Insurance	20	0	0
Jewellery and watches	17	100	100
Medical and optical	38	100	16
Music store	3	0	100
Others	13	0	7
Petrol	95	0	100
Retail	28	0	0
Sporting store	8	100	100
Telecommunications	52	100	33
Toys	3	100	100
Utilities	23	0	0
Uncategorized	137	–	–

Notes: The table shows average monthly credit card spending by Merchant Category Code (MCC) groups, and the classification of MCC groups according to whether they are tradable or durable. Column (1) shows monthly spending by category. Columns (2) and (3) report the share of purchases in each category that are classified as durable and tradable goods, respectively.

Table 1.2: Summary Statistics for Participants and Non-Participants

	All	Responded to survey: Yes	Responded to survey: No	p-value
	(1)	(2)	(3)	(4)
Panel A: demographics				
Male	0.62	0.67	0.57	0.000
	(0.01)	(0.01)	(0.01)	
Age	33.59	33.28	33.88	0.001
	(0.09)	(0.13)	(0.13)	
Monthly income	3,087	3,128	3,049	0.113
	(24.97)	(34.28)	(36.09)	
Panel B: monthly expenditures, pre–treatment				
Total	1,069.44	1,095.09	1,045.89	0.265
	(22.27)	(28.06)	(34.08)	
Durables	343.20	364.20	323.91	0.015
	(8.19)	(13.47)	(9.69)	
Tradable durables	259.34	271.88	247.83	0.098
	(7.16)	(11.87)	(8.35)	
Debt balance	1,805.97	1,800.84	1,810.67	0.895
	(37.71)	(47.49)	(57.72)	
Observations	6,000	2,872	3,128	

Notes: The table reports summary statistics on survey respondents and non-respondents. Panel A reports demographic characteristics, based on the bank's administrative data. Panel B reports summary statistics on pre-treatment spending, based on average monthly credit card spending in the 12 months prior to the experiment. Column (1) reports summary statistics for the full sample, column (2) reports summary statistics for credit card customers who participated in the experiment and column (3) reports statistics for customers that we attempted to contact, but who did not participate in the experiment. Column (4) reports p-values for a test for equality of means between the group of survey respondents and non-respondents. Robust standard errors of the mean in parentheses.

Table 1.3: Test of Randomization Balance

	All	Treatment			
		Exchange Rate	Inflation Rate	Both	p-value
	(1)	(2)	(3)	(4)	(5)
Panel A: demographics					
College	0.87	0.86	0.87	0.86	0.36
	(0.01)	(0.01)	(0.01)	(0.01)	
Married	0.54	0.53	0.54	0.52	0.19
	(0.01)	(0.01)	(0.01)	(0.02)	
Number of dependents	0.86	0.82	0.87	0.82	0.09
	(0.02)	(0.03)	(0.03)	(0.04)	
Self-employed	0.10	0.09	0.09	0.09	0.61
	(0.01)	(0.01)	(0.01)	(0.01)	
Monthly income	3,128	3,132	3,128	3,136	0.99
	(34.28)	(42.53)	(41.81)	(60.45)	
Panel B: monthly expenditures, pre-treatment					
Total	1,128.55	1,135.64	1,122.15	1,129.91	0.94
	(32.97)	(44.16)	(39.64)	(65.89)	
Durables	371.97	387.27	353.48	365.34	0.29
	(16.73)	(24.02)	(12.90)	(20.78)	
Tradable durables	274.10	286.23	259.47	268.88	0.37
	(15.32)	(22.30)	(10.26)	(16.59)	
Debt	1,909.97	1,883.86	1,934.88	1,907.70	0.67
	(47.53)	(57.22)	(61.10)	(87.98)	
Panel C: prior beliefs					
Prior exchange rate	-0.29	-0.37	-0.19	-0.27	0.16
	(0.08)	(0.10)	(0.10)	(0.14)	
Prior inflation	3.39	3.47	3.28	3.32	0.13
	(0.08)	(0.10)	(0.09)	(0.13)	
Observations	2,872	967	974	931	

Notes: The table reports pre-treatment characteristics and a test of randomization balance. Panel A reports demographic characteristics, based on the bank's administrative data. Panel B reports summary statistics on average monthly credit card spending in the 12 months prior to the intervention by category. The experiment was conducted over 3 months, therefore the expenditures in this panel are not perfectly aligned with Panel B of Table 1.2. Panel C reports data on prior beliefs elicited before respondents reached the information provision stage of the experiment. Column (1) reports pre-treatment characteristics for all survey respondents, columns (2) to (4) report the same characteristics for each of the three treatment conditions, that is, for respondents assigned to receive information about the exchange rate, the inflation rate, or both. Column (5) reports p-values of a test for the null hypothesis that the average pre-treatment characteristics are equal between the three treatment groups. Robust standard errors of the mean in parentheses.

Table 1.4: Effects of Information on Expectations and Behavior: Reduced Form Estimates

	Survey Data		Transaction Data			
	(1)	(2)	(3)	(4)	(5)	(6)
	Δ Inflation	Δ Depreciation	Durables	Trad. Dur.	Debt	Total
$\left(\pi_{i,t}^{signal} - \pi_{i,t}^{prior}\right) \cdot T_{i,t}^{\pi}$	0.236***	0.032	1.646	1.497	10.036	-1.718
	(0.037)	(0.026)	(2.793)	(2.062)	(6.472)	(6.028)
$\left(d_{i,t}^{signal} - d_{i,t}^{prior}\right) \cdot T_{i,t}^{d}$	-0.030	0.064**	-3.402	-2.552	4.120	-4.195
	(0.023)	(0.031)	(2.595)	(1.916)	(6.930)	(5.944)
Observations	2,872	2,872	2,872	2,872	2,872	2,872
R-squared	0.393	0.236	0.249	0.199	0.052	0.371
Outcome mean	-0.369	-0.212	255.641	176.445	99.096	947.399
Outcome SD	2.695	2.837	339.340	250.156	758.368	902.024

Notes: Each column corresponds to a separate OLS regression with the same independent variables but different dependent variables. These regressions present the reduced-form effects of the information provision experiment. Column (1) corresponds to equation (1.4), column (2) to equation (1.5) and columns (3) through (6) correspond to equation (1.6). $\left(\pi_{i,t}^{signal} - \pi_{i,t}^{prior}\right)$ is the gap between the feedback about inflation rate that could be shown to the individual and the individual's prior belief about the inflation rate. $T_{i,t}^{\pi}$ is an indicator variable that takes the value 1 if the feedback was shown to the subject and 0 otherwise. $\left(d_{i,t}^{signal} - d_{i,t}^{prior}\right)$ and $T_{i,t}^{d}$ are the corresponding variables for the exchange rate instead of the inflation rate. All regressions control for $\left(\pi_{i,t}^{signal} - \pi_{i,t}^{prior}\right)$ and $\left(d_{i,t}^{signal} - d_{i,t}^{prior}\right)$ as well as the usual set of additional controls: number of dependents, week fixed effects, surveyors fixed effects, and 20 variables on the spending patterns during the four pre-treatment quarters. The dependent variables are listed as follows. Δ Inflation is the difference between the posterior and prior beliefs on the inflation rate (i.e., $\pi_{i,t}^{post} - \pi_{i,t}^{prior}$). Δ Depreciation is the difference between the posterior and prior beliefs on the exchange rate (i.e., $d_{i,t}^{post} - d_{i,t}^{prior}$). Durables is the monthly average expenditure across 3 months post-treatment in the durables category. Trad. Dur. is the monthly average expenditure across 3 months post-treatment in the tradable durables category. Debt is the monthly credit card debt accrued in the 3 months post-treatment. Total is the total average expenditure across 3 months post-treatment. Robust standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 1.5: Effects of Expectations on Behavior: Instrumental Variables Estimates

	Transaction Data			
	(1) Dur.	(2) Trad. Dur.	(3) Debt	(4) Total
Δ Inflation	13.353	11.062	31.749	1.539
	(12.827)	(9.322)	(31.186)	(26.105)
	[-112.349, 139.055]	[-80.2917, 102.416]	[-273.87, 337.369]	[-254.286, 257.363]
Δ Exchange rate	-46.533	-34.420	78.673	-64.333
	(44.068)	(32.407)	(109.620)	(93.003)
	[-478.393, 385.327]	[-351.998, 283.159]	[-995.586, 1152.93]	[-975.749, 847.083]
Observations	2,872	2,872	2,872	2,872
Outcome mean	255.641	176.445	99.096	947.399
Outcome SD	339.340	250.156	758.368	902.024
Kleiberg-Paap F-statistics	2.394	2.394	2.394	2.394

Notes: Each column corresponds to a separate Instrumental Variables regression. The endogenous variables are: Δ Inflation is the difference between the posterior and prior beliefs on the inflation rate (i.e., $\pi_{i,t}^{post} - \pi_{i,t}^{prior}$); Δ Exchange Rate is the difference between the posterior and prior beliefs on the exchange rate (i.e., $d_{i,t}^{post} - d_{i,t}^{prior}$). The excluded instruments are $\left(\pi_{i,t}^{signal} - \pi_{i,t}^{prior}\right) \cdot T_{i,t}^{\pi}$ and $\left(d_{i,t}^{signal} - d_{i,t}^{prior}\right) \cdot T_{i,t}^{d}$. All regressions control for $\left(\pi_{i,t}^{signal} - \pi_{i,t}^{prior}\right)$ and $\left(d_{i,t}^{signal} - d_{i,t}^{prior}\right)$ as well as the usual set of additional controls: number of dependents, week fixed effects, surveyors fixed effects, and 20 variables on spending patterns during the four pre-treatment quarters. The dependent variables are listed as follows. Durables is the monthly average expenditure across 3 months post-treatment in the durables category. Trad. Dur. is the monthly average expenditure across 3 months post-treatment in the tradable durables category. Debt is the monthly credit card debt accrued in the 3 months post-treatment. Total is the total average expenditure across 3 months post-treatment. Robust standard errors in parentheses. Weak instruments Anderson-Rubin confidence interval at 95% level in brackets. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 1.6: Effects of Expectations on Behavior: OLS Estimates

	Transaction Data			
	(1) Durables	(2) Trad. Dur.	(3) Debt	(4) Total
Δ Inflation	-1.849	-0.358	-7.032	5.145
	(1.966)	(1.383)	(4.825)	(5.032)
Δ Exchange rate	-0.532	-0.662	-6.483	-3.519
	(2.000)	(1.523)	(4.798)	(4.798)
Observations	2,872	2,872	2,872	2,872
R-squared	0.249	0.198	0.052	0.371
Outcome mean	255.641	176.445	99.096	947.399
Outcome SD	339.340	250.156	758.368	902.024

Notes: Each column corresponds to a separate OLS regression. Δ Inflation is the difference between posterior and prior beliefs on inflation (i.e., $\pi_{i,t}^{post} - \pi_{i,t}^{prior}$). Δ Exchange Rate is the difference between the posterior and prior beliefs on the exchange rate (i.e., $d_{i,t}^{post} - d_{i,t}^{prior}$). All regressions include the usual set of additional controls: number of dependents, week fixed effects, surveyors fixed effects, and 20 variables on spending patterns during the four pre-treatment quarters. The dependent variables are listed as follows. Durables is the monthly average expenditure across 3 months post-treatment in the durables category. Trad. Dur. is the monthly average expenditure across 3 months post-treatment in the tradable durables category. Debt is the monthly credit card debt accrued in the 3 months post-treatment. Total is the total average expenditure across 3 months post-treatment. Robust standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 1.7: Effects of Information on Expectations and Survey Outcomes: Reduced Form Estimates

	Survey Data					
	(1)	(2)	(3)	(4)	(5)	(6)
	Δ Inflation	Δ Depreciation	Dur.	Trad. Dur.	Debt	Total
$\left(\pi_{i,t}^{signal} - \pi_{i,t}^{prior}\right) \cdot T_{i,t}^{\pi}$	0.236***	0.032	-0.001	-0.002	-0.003	-0.006
	(0.037)	(0.026)	(0.007)	(0.007)	(0.008)	(0.006)
$\left(d_{i,t}^{signal} - d_{i,t}^{prior}\right) \cdot T_{i,t}^{d}$	-0.030	0.064**	0.004	0.002	0.009	0.003
	(0.023)	(0.031)	(0.007)	(0.007)	(0.007)	(0.006)
Observations	2,872	2,872	2,872	2,872	2,872	2,872
R-squared	0.393	0.236	0.030	0.034	0.073	0.037
Outcome Mean	-0.369	-0.212	-0.055	0.005	-0.055	0.088
Outcome SD	2.695	2.837	0.857	0.775	0.857	0.665

Notes: The table reports reduced-form effects of the information provision experiment. Column (1) corresponds to equation (1.4), column (2) to equation (1.5) and columns (3) through (6) correspond to equation (1.6). $\left(\pi_{i,t}^{signal} - \pi_{i,t}^{prior}\right)$ is the gap between the feedback about inflation rate that could be shown to the individual and the individual's prior belief about the inflation rate, while $T_{i,t}^{\pi}$ is an indicator variable that takes the value 1 if the feedback was shown to the subject and 0 otherwise. $\left(d_{i,t}^{signal} - d_{i,t}^{prior}\right)$ and $T_{i,t}^{d}$ are the corresponding variables for the exchange rate instead of the inflation rate. All regressions control for $\left(\pi_{i,t}^{signal} - \pi_{i,t}^{prior}\right)$ and $\left(d_{i,t}^{signal} - d_{i,t}^{prior}\right)$ as well as the usual set of additional controls: number of dependents, week fixed effects, surveyors fixed effects, and 20 variables on spending patterns during the four pre-treatment quarters. The dependent variables are listed as follows. Δ Inflation is the difference between the posterior and prior beliefs on the inflation rate (i.e., $\pi_{i,t}^{post} - \pi_{i,t}^{prior}$). Δ Depreciation is the difference between the posterior and prior beliefs on the exchange rate (i.e., $d_{i,t}^{post} - d_{i,t}^{prior}$). The dependent variables in columns (3) through (6) correspond to the stated future consumption as measured in the survey, and they can take values +1 (if participants say they are going to spend more or think it is a good time to buy goods in the category), 0 (if they say that they are going to spend about the same or think it's neither good nor bad time to buy the goods) or -1 (if they are going to spend less or think it is a bad time to buy the goods). Durables corresponds to the future spending in durables, Trad. Dur. correspond to the future spending in electronics, Debt corresponds to future credit card borrowing, and Total corresponds to total future spending. Robust standard errors in parentheses. $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 1.8: Effects of Information on Expectations and Behavior by Size of Purchases: Reduced Form Estimates

	Durables			Tradable Durables			All		
	(1) p25	(2) p75	(3) Diff	(4) p25	(5) p75	(6) Diff	(7) p25	(8) p75	(9) Diff
$(\pi_{i,t}^{signal} - \pi_{i,t}^{prior}) \cdot T_{i,t}^{\pi}$	-0.004	0.009	0.013	-0.007	0.008	0.015	-0.003	-0.003	0.000
	(0.007)	(0.009)	(0.011)	(0.007)	(0.009)	(0.011)	(0.005)	(0.007)	(0.009)
$(d_{i,t}^{signal} - d_{i,t}^{prior}) \cdot T_{i,t}^{d}$	-0.005	-0.011	-0.006	0.000	-0.010	-0.010	0.005	-0.006	-0.011
	(0.006)	(0.008)	(0.010)	(0.007)	(0.008)	(0.011)	(0.006)	(0.007)	(0.009)
Observations	2872	2872	2872	2872	2872	2872	2872	2872	2872
R-squared	0.406	0.240		0.389	0.196		0.599	0.379	
Outcome Mean	0.000	0.000		0.000	0.000		0.000	0.000	
Outcome SD	1.000	1.000		1.000	1.000		1.000	1.000	

Notes: Notes: Columns (1), (2), (4), (5), (7) and (9) correspond to a separate OLS regression with the same independent variables but different dependent variables. $\left(\pi_{i,t}^{signal} - \pi_{i,t}^{prior}\right)$ is the gap between the feedback about inflation rate that could be shown to the individual and the individual's prior belief about the inflation rate. $T_{i,t}^{\pi}$ is an indicator variable that takes the value 1 if the feedback was shown to the subject and 0 otherwise. $\left(d_{i,t}^{signal} - d_{i,t}^{prior}\right)$ and $T_{i,t}^{d}$ are the corresponding variables for the exchange rate instead of the inflation rate. All regressions control for $\left(\pi_{i,t}^{signal} - \pi_{i,t}^{prior}\right)$ and $\left(d_{i,t}^{signal} - d_{i,t}^{prior}\right)$ as well as the usual set of additional controls: number of dependents, week fixed effects, surveyors fixed effects, and 24 variables on the spending patterns during the four pre-treatment quarters. The dependent variables are listed as follows. Durables is the monthly average expenditure across 3 months post-treatment in the durables category. Tradable Durables is the monthly average expenditure across 3 months post-treatment in the tradable durables category. All is the monthly average expenditure across 3 months post-treatment in the tradable durables category in all categories. p25 and p75 are the monthly average expenditure across 3 months post-treatment in purchases whose price correspond to the lowest and top quartile in the distribution of the individual prices of purchases of their corresponding category, respectively. Columns (3), (6) and (9) correspond to the difference between the estimated coefficients by price of purchases in each category. Each dependent variable is standardized for comparability. Robust standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 1.9: Effects of Hypothetical Shocks on Planned Expenditures

Scenario:	Inflation				Depreciation			
	(1) Dur.	(2) Trad. Dur.	(3) Debt	(4) Total	(5) Dur.	(6) Trad. Dur.	(7) Debt	(8) Total
Panel A: Baseline								
Δ Belief	-0.019***	-0.021***	0.001	0.003	-0.025***	-0.027***	0.000	0.002
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
Observations	2302	2302	2302	2302	2302	2302	2302	2302
R-squared	0.187	0.202	0.150	0.202	0.230	0.202	0.139	0.172
Outcome mean	0.033	0.038	-0.126	-0.225	-0.096	-0.180	0.160	0.177
Outcome SD	0.743	0.775	0.745	0.742	0.757	0.752	0.748	0.784
Panel B: By financial literacy								
Δ Belief	-0.008**	-0.010***	0.007*	0.003	-0.012***	-0.013***	-0.001	0.001
	(0.003)	(0.003)	(0.004)	(0.003)	(0.003)	(0.004)	(0.003)	(0.003)
Δ Belief $\cdot L_i$	-0.017***	-0.017***	-0.007	-0.001	-0.020***	-0.023***	-0.002	0.004
	(0.004)	(0.004)	(0.005)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)
Observations	1910	1910	1910	1910	1910	1910	1910	1910
R-squared	0.219	0.226	0.158	0.224	0.237	0.205	0.144	0.176
Outcome mean	0.033	0.038	-0.126	-0.225	-0.096	-0.180	0.160	0.177
Outcome SD	0.743	0.775	0.745	0.742	0.757	0.752	0.748	0.784

Notes: Each column corresponds to a separate OLS regression, reporting the reduced-form effects of the subjective model experiment. Panel A shows the effect of increase in expected inflation and depreciation rates in a hypothetical scenarios relative to the respondents prior beliefs. Panel B shows the effect for high and low financial literacy groups. Columns (1) through (4) correspond to the estimates for the inflation scenario and columns (5) through (8) correspond to the depreciation scenario. For the inflation scenario, Δ Belief$_i = \left(\pi_i^{scenario} - \pi_i^{prior}\right)$, which denotes the gap between the hypothetical inflation rate shown to the individual and the individual's prior belief. For the depreciation scenario, Δ Belief$_i = \left(d_i^{scenario} - d_i^{prior}\right)$, which is the analogous gap for the depreciation rate. L_i is an indicator variable that takes the value 1 if the respondent has the big-3 financial literacy score higher or equal to the median across waves and 0 otherwise. Each regression controls for the corresponding "status-quo" hypothetical outcome. Panel B regressions also control for the literacy group indicator L_i and the interaction L_i with the "status quo" hypothetical outcome. Durables corresponds to the future spending in durables, Trad. Dur. correspond to the future spending in electronics, Debt corresponds to future credit card borrowing, and Total corresponds to total future spending. The dependent variables in columns (3) through (6) correspond to the stated future consumption as measured in the survey, and they can take values +1 (if participants say they are going to spend more or think it is a good time to buy goods in the category), 0 (if they say that they are going to spend about the same or think it's neither good nor bad time to buy the goods) or -1 (if they are going to spend less or think it is a bad time to buy the goods). Robust standard errors in parentheses. $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 1.10: Effects of Hypothetical Shocks on Demand for Inflation Indexed Security

Dep. var.:	Demand for inflation indexed asset			
Scenario:	Inflation		Depreciation	
	(1)	(2)	(3)	(4)
Δ Belief	0.015***	-0.010	0.020***	0.000
	(0.004)	(0.009)	(0.004)	(0.009)
Δ Belief $\cdot L_i$		0.048***		0.028**
		(0.012)		(0.012)
Observations	789	397	789	397
R-squared	0.047	0.079	0.107	0.112
Outcome mean	-0.185	-0.109	-0.165	-0.132
Outcome SD	0.914	0.927	0.908	0.912

Notes: Each column corresponds to a separate OLS regression. These regressions present the reduced-form effects of the subjective model experiment. Panel A shows the effect of increase in expected inflation and depreciation rates in a given hypothetical scenario relative to the respondents prior beliefs. Panel B shows the effect for high and low financial literacy groups. Columns (1) through (3) correspond to the estimates for the inflation scenario and columns (4) through (6) correspond to the depreciation scenario. For the inflation scenario, $\Delta\ \text{Belief}_i = \left(\pi_i^{scenario} - \pi_i^{prior}\right)$, which denotes the gap between the hypothetical inflation rate shown to the individual and the individual's prior belief. For the depreciation scenario, $\Delta\ \text{Belief}_i = \left(d_i^{scenario} - d_i^{prior}\right)$, which is the analogous gap for the depreciation rate. L_i is an indicator variable that takes the value 1 if the respondent has the big-3 financial literacy score higher or equal to the median across waves and 0 otherwise. Each regression controls for the corresponding "status-quo" hypothetical outcome. Panel B regressions also control for the literacy group indicator L_i and the interaction L_i with the "status quo" hypothetical outcome. Durables corresponds to the future spending in durables, Trad. Dur. correspond to future spending on electronics, Debt corresponds to future credit card borrowing, and Total corresponds to total future spending. The dependent variables in columns (3) through (6) correspond to self-reported consumption plans as measured in the survey, and can take values +1 (if participants say they are going to spend more or think it is a good time to buy goods in the category), 0 (if they say that they are going to spend about the same or think it's neither good nor bad time to buy goods) or -1 (if they are going to spend less or think it is a bad time to buy goods). Robust standard errors in parentheses. $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

1.A Proof of Propositions

1.A.1 Lemma 1

The three statements are true:

(1) C_t^N and C_t^T are non-decreasing in π_{t+1}^N and π_{t+1}^T.

(2) X_t^N is increasing in π_{t+1}^N. X_t^T is increasing in π_{t+1}^T.

(3) X_t^N is non-decreasing in π_{t+1}^T. X_t^T is non-decreasing in π_{t+1}^N.

Proof:

Observe that Cobb-Douglas structure of preferences allows us to write a closed form solution for C_t^N:

$$C_t^N = \frac{\alpha\theta}{\sum_{k=0}^{\infty}\beta^k}\frac{\sum_{k=0}^{\infty}P_{t+k}^N Y_{t+k} / \prod_{i=1}^{k}R_{t+i}}{P_t^N} =$$

$$\frac{\alpha\theta}{\sum_{k=0}^{\infty}\beta^k}\sum_{k=0}^{\infty}\frac{\prod_{i=1}^{k}\pi_{t+i}^N Y_{t+k}}{\prod_{i=1}^{k}R_{t+i}}$$

Hence:

$$\frac{dC_t^N}{d\pi_{t+1}^N} = \frac{\alpha\theta}{\sum_{k=0}^{\infty}\beta^k}\sum_{k=0}^{\infty}\frac{\prod_{i=2}^{k}\pi_{t+i}^N Y_{t+k}}{\prod_{i=1}^{k}R_{t+i}} > 0$$

$$\frac{dC_t^N}{d\pi_{t+1}^T} = 0$$

From the first order conditions, one can obtain:

$$\frac{\theta}{1-\theta}\frac{C_t^T}{C_t^N} = \frac{P_t^N}{P_t^T}$$

This implies similar conditions for C_t^T:

$$\frac{dC_t^T}{d\pi_{t+1}^N} > 0 \quad \text{and} \quad \frac{dC_t^T}{d\pi_{t+1}^T} = 0$$

, which concludes the proof of statement (1).

From the first order conditions one gets:

$$\frac{(1-\alpha)C_t^N}{\alpha X_t^N} = \left(1 - (1-\delta)\frac{\pi_{t+1}^N}{R_{t+1}}\right)$$

$$\frac{(1-\alpha)C_t^T}{\alpha X_t^T} = \left(1 - (1-\delta)\frac{\pi_{t+1}^T}{R_{t+1}}\right)$$

This implies that for $I \in \{N, T\}$, $\frac{X_t^I}{C_t^I}$ increases in π_{t+1}^I. Since by statement (1), C_t^I does not decrease in π_{t+1}^I, X_t^I has to increase. This proves statement (2).
Finally, since i) the conditions pin down the $\frac{X_t^I}{C_t^I}$ for fixed π_{t+1}^I, and ii) C_t^I is non-decreasing in π_{t+1}^{-I}, X_t^I does not decrease in π_{t+1}^{-I} ($-I$ denotes $\{N, T\} \setminus I$), which proves statement (3).

1.A.2 Proof of Proposition 1

By Assumption 1, an increase in π_{t+1} does not decrease π_{t+1}^I, $I \in \{N, T\}$ and has to strictly increase at least one of them. By Lemma 1, this implies that X_t^I do not decrease and at least one of them increases. Since X_{t-1}^I and P_t^I are fixed, the same is true for $P_t^I \Delta X_t^I$. Hence, $P_t^N \Delta X_t^N + P_t^T \Delta X_t^T$ increases in π_{t+1}.

1.A.3 Proof of Proposition 2

By Assumption 2, π_{t+1}^T increases in E_{t+1}. By Lemma 1, this implies that X_t^T increases in E_{t+1}. Hence $P_t^T \Delta X_t^T$ increases in E_{t+1}.

1.A.4 Proof of Proposition 3

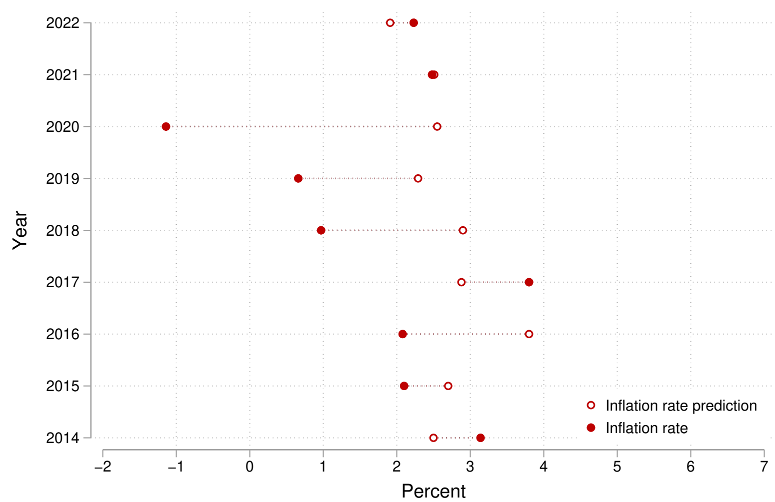
$$A_{t+1} = P_t Y_t + R_t A_t - P_t^N (C_t^N + X_t^N - X_{t-1}^N + \delta X_t^N) -$$

$$-P_t^T (C_t^T + X_t^T - X_{t-1}^T + \delta X_t^T)$$

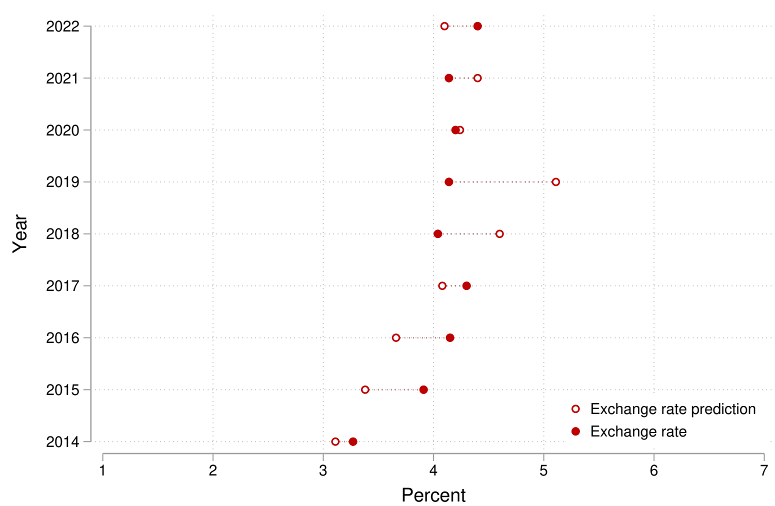
By Assumption 1, neither π_{t+1}^T nor π_{t+1}^T decrease, and at least one of them increases with π_{t+1}. This and Lemma 1 imply that neither of C_t^I and X_t^I, $I \in \{N, T\}$ decreases, and at least one of X_t^I increases. Hence, A_{t+1} decreases and $A_t - A_{t+1}$ increases.

1.B Appendix Figures

Figure 1.8: Accuracy of Expert Forecasts



(a) Inflation Expectations



(b) Exchange Rate Expectations

Notes: The figure shows the difference between predicted and realized inflation and exchange rates in historical data, taken from the same sources used to provide information in the experiment. Forecasts are recovered for the same time horizon (12 months) at which we provide information in the experiment. Inflation forecasts are taken from Statista (www.statista.com), exchange rate forecasts are taken from Trading Economics (www.tradingeconomics.com).

1.C Additional Analysis and Robustness Checks

1.C.1 Event-Study Falsification Tests

we leverage data on pre-treatment spending, which allows us to conduct a falsification test in the spirit of an event-study analysis. We estimate a similar regression as in equation (1.6) but using pre-treatment instead of post-treatment spending as the dependent variables:

$$\begin{aligned} Y_{i,t-1} &= \alpha_Y^{\pi} \cdot \left(\pi_{i,t}^{signal} - \pi_{i,t}^{prior}\right) \cdot T_{i,t}^{\pi} + \alpha_Y^{d} \cdot \left(d_{i,t}^{signal} - d_{i,t}^{prior}\right) \cdot T_{i,t}^{d} + \\ &\quad \beta_Y^{\pi} \cdot \left(d_{i,t}^{signal} - d_{i,t}^{prior}\right) + \beta_Y^{d} \cdot \left(\pi_{i,t}^{signal} - \pi_{i,t}^{prior}\right) + X_{i,t}\gamma_Y + \epsilon_i \end{aligned} \tag{1.7}$$

The dependent variable $Y_{i,t+1}$ refers to the average monthly spending in the 3 months pre-treatment, and the set of control variables ($X_{i,t}$) include just the number of dependents, week fixed effects, surveyors fixed effects.

The results are presented in Table 1.11. Since the outcomes are measured at a point in time when participants had not yet been provided with information, there should be no effects of the information on pre-treatment spending. As expected, we find no "effects" of the information shocks on the pre-treatment spending outcomes. For example, the first coefficient from column (1) indicates that a 1 pp increase in the inflation shock had an "effect" on pre-treatment spending on durables that is close to zero ($0.185, or ¡0.001 standard deviations) and statistically insignificant. Likewise, the rest of the coefficients from Table 1.11 are close to zero and statistically insignificant.

Table 1.11: Effects of Information on Behavior: Event-Study Falsification Tests

	Transaction Data			
	(1)	(2)	(3)	(4)
	Durables	Trad. Dur.	Debt	Total
$\left(\pi_{i,t}^{signal} - \pi_{i,t}^{prior}\right) \cdot T_{i,t}^{\pi}$	-0.154	-1.826	3.161	0.356
	(2.708)	(1.963)	(6.608)	(5.779)
$\left(d_{i,t}^{signal} - d_{i,t}^{prior}\right) \cdot T_{i,t}^{d}$	0.209	-2.506	-10.146	4.119
	(2.706)	(2.080)	(6.841)	(5.912)
Observations	2,872	2,872	2,872	2,872
R-squared	0.221	0.165	0.025	0.346
Outcome Mean	284.843	196.553	157.014	958.194
Outcome SD	352.911	260.234	794.367	887.150

Notes: Each column corresponds to a separate OLS regression with the same independent variables but different dependent variables. All regression corresponds to equation (1.7). $\left(\pi_{i,t}^{signal} - \pi_{i,t}^{prior}\right)$ is the gap between the feedback about inflation rate that could be shown to the individual and the individual's prior belief about the inflation rate. $T_{i,t}^{\pi}$ is an indicator variable that takes the value 1 if the feedback was shown to the subject and 0 otherwise. $\left(d_{i,t}^{signal} - d_{i,t}^{prior}\right)$ and $T_{i,t}^{d}$ are the corresponding variables for the exchange rate instead of the inflation rate. All regressions control for $\left(\pi_{i,t}^{signal} - \pi_{i,t}^{prior}\right)$ and $\left(d_{i,t}^{signal} - d_{i,t}^{prior}\right)$ as well as the following set of additional controls: number of dependents, week fixed effects, and surveyors fixed effects. The dependent variables are listed as follows. Dur. is the monthly average expenditure across 3 months pre-treatment in the durables category. Trad. Dur. is the monthly average expenditure across 3 months pre-treatment in the tradable durables category. Debt is the monthly credit card debt accrued in the 3 months pre-treatment. Total is the total average expenditure across 3 months pre-treatment. Robust standard errors in parentheses. $p < 0.10$, $**p < 0.05$, $***p < 0.01$.

Table 1.12: Effects of Information on Expectations and Behavior by Size of Purchases of Categories: Reduced Form Estimates

	Durables			Tradable Durables			All		
	(1) Bottom 3 Cat.	(2) Top 3 Cat.	(3) Diff	(4) Bottom 3 Cat.	(5) Top 3 Cat.	(6) Diff	(7) Bottom 3 Cat.	(8) Top 3 Cat.	(9) Diff
$(\pi_{i,t}^{signal} - \pi_{i,t}^{prior}) \cdot T_{i,t}^{\pi}$	-0.006	0.000	0.006	-0.006	0.003	0.009	0.004	0.000	-0.005
	(0.008)	(0.009)	(0.012)	(0.008)	(0.009)	(0.012)	(0.007)	(0.009)	(0.012)
$(d_{i,t}^{signal} - d_{i,t}^{prior}) \cdot T_{i,t}^{d}$	0.005	0.004	-0.001	0.005	0.004	-0.001	-0.004	0.004	0.008
	(0.008)	(0.008)	(0.011)	(0.008)	(0.008)	(0.012)	(0.008)	(0.008)	(0.011)
Observations	2872	2872	2872	2872	2872	2872	2872	2872	2872
R-squared	0.089	0.155		0.089	0.104		0.176	0.155	
Outcome Mean	0.000	0.000		0.000	0.000		0.000	0.000	
Outcome SD	1.000	1.000		1.000	1.000		1.000	1.000	

Notes: Columns (1), (2), (4), (5), (7) and (9) report separate OLS regressions with the same independent variables but different dependent variables. $\left(\pi_{i,t}^{signal} - \pi_{i,t}^{prior}\right)$ is the gap between the feedback about inflation that could be shown to the individual and their prior belief about the inflation rate. $T_{i,t}^{\pi}$ is an indicator variable that takes the value 1 if the feedback was shown to the subject and 0 otherwise. $\left(d_{i,t}^{signal} - d_{i,t}^{prior}\right)$ and $T_{i,t}^{d}$ are the corresponding variables for the exchange rate instead of the inflation rate. All regressions control for $\left(\pi_{i,t}^{signal} - \pi_{i,t}^{prior}\right)$ and $\left(d_{i,t}^{signal} - d_{i,t}^{prior}\right)$ as well as the usual set of additional controls: number of dependents, week fixed effects, surveyors fixed effects, and 24 variables on the spending patterns during the four pre-treatment quarters. The dependent variables are listed as follows. Durables is the monthly average expenditure across 3 months post-treatment in the durables category. Tradable Durables is the monthly average expenditure across 3 months post-treatment in the tradable durables category. All is the monthly average expenditure across 3 months post-treatment in the tradable durables category in all categories. Bottom 3 Cat. and Top 3 Cat. are the monthly average expenditure across 3 months post-treatment in the categories with the lowest and highest median prince of individual purchases, respectively. The categories with the lowest median price are Direct Marketing, Groceries and Books and Stationery. The categories with the highest median price are Automotive, Jewellery and Watches and Insurance Columns (3), (6) and (9) correspond to the difference between the estimated coefficients by price of purchases in each category. Robust standard errors in parentheses.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

1.D Survey Instrument

Hello! My name is [surveyor name]. I am working for researchers at the University of California, Los Angeles, currently working in Malaysia. We are conducting a short survey to know Malaysians overall economical situation. Do you have five minutes to respond to the survey?

- Yes
- No

[If the answer to the previous question was "yes":] Great, thank you so much. By the way, if you'd prefer to do the survey in Alternative Language, let me know. I will start asking a few questions about your background.

What is your current employment situation?

- Full-time employee
- Part-time employee
- Self-employed
- Not working

What is your highest education level?

- No school
- High school
- College or some college

- After bachelor degree

Are you married or single?

- Married
- Single
- Divorced

Do you have any children or other dependents that you look after?

- Yes
- No

[If the answer to the previous question was "yes":] How many?

- 1
- 2
- 3
- 4
- 5 or more

Regarding business conditions in the country as a whole, do you think that during the next 12 months the Malaysian economy will be better off, about the same, or worse off?

- Better off

- About the same
- Worse off

Now we want to ask you about the annual inflation rate, which is a measure of how prices in Malaysia change in general. In your opinion, what will be the inflation rate over the next 12 months?

- [ ] %

Now we want to ask you about the exchange rate. As of April 2019, 1 U.S. Dollar is worth around 4.05 Ringgit Malaysia. In your opinion, what will the exchange rate be 12 months from now, in April 2020?

- [ ] Ringgit Malaysia

In this stage, we randomly select respondents to receive some feedback about the previous questions. [***Subjects are randomly assigned to one of the following three treatments.***]

Treatment Exchange Rate: **The consensus among economic experts both from the government and the private sectors is that 1 U.S. Dollar will be worth 4.10 Ringgit Malaysia one year from now.**

Treatment Inflation: **The consensus among economic experts both from the government and the private sectors is that the inflation in Malaysia will be 2.3% over the next 12 months.**

Treatment Both: **The consensus among economic experts both from the government and the private sectors is that the inflation in Malaysia will be 2.3% over**

the next 12 months and 1 U.S. Dollar will be worth 4.10 Ringgit Malaysia one year from now.

What will the inflation rate be over the next 12 months?

- [ ] %

What will be the exchange rate from U.S. Dollar to Ringgit 12 months from now, in April 2020?

- [ ] Ringgit Malaysia

Regarding business conditions in the country as a whole, do you think that during the next 12 months the Malaysian economy will be better off, about the same, or worse off?

- Better off
- The same
- Worse off

Looking forward, would you say that you and your family living with you will be better off or worse off financially than you are now?

- Better off
- About the same
- Worse off

Do you expect your credit card spending to go up, stay the same, or go down during the next 3 months?

- Go up
- Stay the same
- Go down

Do you expect that your spending on groceries to go up, stay the same, or go down during the next 3 months?

- Go up
- Stay the same
- Go down

Do you expect your total spending to go up, stay the same, or go down during the next 3 months?

- Go up
- Stay the same
- Go down

Do you think now is a good time, a bad time, or neither a good nor a bad time to buy household items, such as furniture or a refrigerator? More examples: television, stove or others

- Yes, it's a good time
- It's neither a good nor a bad time
- No, it's a bad time

Do you think now is a good time, a bad time, or neither a good nor a bad time to buy electronic items, such as a computer, TV, phone, washing machine and so on?

- Yes, it's a good time
- It's neither a good nor a bad time
- No, it's a bad time

Do you think now is a good time, a bad time, or neither good nor a bad time to buy a vehicle, car or motorbike?

- Yes, it's a good time
- It's neither a good nor a bad time
- No, it's a bad time

Do you think now is a good time, a bad time, or neither good or bad time to buy big items on an installment basis? [If asked, provide the following examples: installments such as AEON Credit, Courts Mammoth; items such as a car, motorbike, television set, washing machine and so on.]

- Yes, it's a good time
- It's neither a good nor a bad time
- No, it's a bad time

1.E Complementary Survey Experiment, Questionnaire

Hello. We are conducting a survey about the economic outlook for Malaysia. This survey consists of 42 questions and takes approximately 10 minutes. The questions in this survey have no right or wrong answers — we are interested in your views and opinions. Your responses are 100% confidential. At the end of the survey you will have a box where you can let us know if there are any problems with the survey.

- YES, I would like to participate in this survey
- NO, I don't want to participate in this survey

Which of these words is the most associated with the word "paint"?

- draw
- run
- sports
- loud

Which of these words is the most associated with the word "cucumber"?

- video
- trigger
- vegetable
- heel

Recent research on decision making shows that choices are affected by the context in which they are made. Differences in how people feel, in their previous knowledge and experience, and in their environment can influence the choices they make. To help us understand how people make decisions, we are interested in information about you, specifically whether you actually take the time to read the instructions; if you don't, some results may fail to tell us very much about decision making in the real world. To help us confirm that you have read these instructions, please select the "none of the above" option below. Thank you very much.

- Interested
- ...
- None of the above

To get a general picture of the people answering this survey, we would like to know a few things about your background. What is your current employment situation?

- Working full time for someone
- Working part time for someone
- Self-employed
- Not working

Do you have any of the following types of financial products or accounts?

- Bank account — Yes/No
- Debit card — Yes/No

- Charge (prepaid) card — Yes/No
- Credit card — Yes/No

What is your approximate monthly income?

- Less than MYR 1000
- Between MYR 1000 and 2000
- Between MYR 2000 and 3000
- Between MYR 3000 and 5000
- More than MYR 5000

What is your highest education level?

- No school
- High school
- College or some college
- After bachelor degree

Are you married or single?

- Married
- Single
- Divorced

Do you have any children or other dependents that you look after?

- Yes
- No
- Display This Question:
- If Do you have any children or other dependents that you look after? = Yes

[If answered 'Yes' to previous question]

How many?

- 1
- 2
- 3
- 4
- 5 or more

Please confirm that that you are not a robot.

[Captcha box]

Now we want to ask you about the exchange rate. As of [Current month, year], 1 U.S. Dollar is worth around [Current exchange rate] Ringgit Malaysia. In your opinion, what will the exchange rate be 12 months from now, in [Current month, next year]?

- 3.70 RM
- ...

- 4.70 RM

Now, we will give you a couple of hypothetical scenarios, and we want to ask about your expectations and spending plans in each scenario.

In this section we sequentially ask about three scenarios

[Here is the first scenario we want you to consider:/ Now, we give you a second, different scenario: / Now, we give you a third, different scenario:]

Prior: suppose the inflation rate will be [Prior inflation] and the exchange rate will go from [Current forex] RM to [Prior forex] RM per 1 U.S. Dollar (a X% depreciation) over the next 12 months.

Inflation: Suppose the inflation rate will be [Prior inflation + random inflation change] over the next 12 months.

Forex: Suppose the exchange rate will go from [Current forex] RM to [Prior forex · (1 + random forex change)] RM per 1 U.S. Dollar (a X% depreciation) over the next 12 months.

Each random change is sampled from $\{-10, -3, 3, 10\}$ *percentage points.*

The fixed income treatment add "Assume that, other than this, business conditions, interest rates, and your personal financial situation remain the same." to each scenario.

Here is the first scenario we want you to consider: **[Present scenario]**

[Present scenario]

Imagine that you are offered to choose between two savings accounts. The first pays an interest rate of 5% per year. The second pays 1% per year after correcting for the inflation rate. Which account would you choose?

- First account: 5% per year

- Second account: 1% per year + inflation
- Don't know

[Present scenario]
In this scenario, do you expect your credit card spending to go up, stay the same, or go down during the next 3 months?

- Go up
- Stay the same
- Go down

[Present scenario]

In this scenario, do you expect your spending on groceries to go up, stay the same, or go down during the next 3 months?

- Go up
- Stay the same
- Go down

[Present scenario] In this scenario, do you expect your total spending to go up, stay the same, or go down during the next 3 months?

- Go up
- Stay the same
- Go down

[Present scenario] In this scenario, do you think it would be a good time, a bad time, or neither a good nor a bad time to buy household items, such as furniture, television, stove, or a refrigerator?

- Yes, it would be a good time
- It'd be neither a good nor a bad time
- No, it would be a bad time

[Present scenario] In this scenario, do you think it would be a good time, a bad time, or neither a good nor a bad time to buy electronic items, such as a computer, handphone, and so on?

- Yes, it would be a good time
- It'd be neither a good nor a bad time
- No, it would be a bad time

[Present scenario]

In this scenario, do you think it would be a good time, a bad time, or neither good nor a bad time to buy a vehicle, car, or motorbike?

- Yes, it would be a good time
- It'd be neither a good nor a bad time
- No, it would be a bad time

[Present scenario]

In this scenario, do you think it would be a good time, a bad time, or neither good or bad time to buy big items on an installment basis?

- Yes, it would be a good time
- It'd be neither a good nor a bad time
- No, it would be a bad time

[Present scenario]

In this scenario, would you say that you and your family living with you will be better off or worse off financially than you are now?

- Better off
- About the same
- Worse off

Imagine that you are offered to choose between two savings accounts. The first pays an interest rate of 5% per year. The second pays 1% per year after correcting for the inflation rate. You expect that the inflation rate is going to be 7%.Which account would you choose?

- First account: 5% per year
- Second account: 1% per year + inflation
- Don't know

Suppose that the inflation rate increases from [Prior inflation belief] to [Prior inflation belief + 10 pp.]. Assume that, other than this, business conditions, interest rates, and your personal financial situation remain the same. Which of the following is the best thing to do in this situation?

- It is best to spend more with a credit card
- It is best to spend less with a credit card
- I don't know if it is better to spend more or less with a credit card

Suppose that you expect the value of the Ringgit to decrease from [Prior forex belief] US dollars per 1 RM to [Prior forex belief · 1.1] US dollars per 1 RM one year from now and you are planning to buy some electronics. Assume that, other than this, business conditions, interest rates, and your personal financial situation remain the same. Which of the following is the best thing to do in this situation?

- It is best to buy electronics now
- It is best to buy electronics later
- I don't know if it's better to buy electronics now or later in this scenario.

Suppose that you expect the value of the Ringgit to decrease from [Prior forex belief] US dollars per 1 RM to [Prior forex belief · 1.1] US dollars per 1 RM and you are planning to buy a car or a motorbike. Assume that, other than this, business conditions, interest rates, and your personal financial situation remain the same. Which of the following is the best thing to do in this situation?

- It is best to buy a car or a motorbike now

- It is best to buy a car or a motorbike later
- I don't know if it's better to buy a car or a motorbike now or later in this scenario

Suppose you had $100 in a savings account and the interest rate was 2% per year. After 5 years, how much do you think you would have in the account if you left the money to grow?

- More than $102
- Exactly $102
- Less than $102
- Do not know

Imagine that the interest rate on your savings account was 1% per year and inflation was 2% per year. After 1 year, how much would you be able to buy with the money in this account?

- More than today
- Exactly the same
- Less than today
- Do not know

Please tell me whether this statement is true or false. "Buying a single company's stock usually provides a safer return than a mutual fund that invests in the stocks of multiple companies."

- True

- False
- Do not know

How would you rate your understanding of the questions included in this survey?

- I understood all the questions
- There were a few questions I did not understand
- There were several questions I did not understand
- There were many questions I did not understand

Thank you for your participation in this survey!

CHAPTER 2

Survey Crowdsourcing: Integrating Community Information in Research Design

2.1 Introduction

Surveys and opinion polls are the universal tool when it comes to obtaining private subjective information about attitudes and preferences, whether it is measuring a community's demand for a policy and evaluating its success, uncovering social norms, or eliciting people's perceptions about the current state of the economy and its prospects. Social sciences have progressed considerably in developing the best practices and establishing standards for survey design and execution (Manski, 2004; Stantcheva, 2022); still, some questions are yet to be settled. Whether open or closed question format is preferable is among those controversial questions (Lazarsfeld, 1944; Converse, 1984). Arguably, the lack of standards for evaluating the quality of information obtained through various methods has contributed to the longevity of this debate. In the absence of such standards, we can neither compare the tradeoffs between the two methods, nor measure the success of the recent survey design methods, such as crowdsourcing, attempting to reconcile the two sides of the debate Salganik and Levy, 2015.

Despite strong arguments in favor of open questions (Ferrario and Stantcheva, 2022), the closed-questions approach to surveys came to dominate the field. Arguably, its success stems from two indisputable practical advantages. The first advantage is the relative simplicity of analyzing collected data, even with a sizable sample. The second one is the low cognitive effort required from the respondents, which translates to lower cost of data collection. While the former advantage has been losing its importance due to technological advancements, the second consideration, combined with growing concerns about access to a representative sample (Keeter et al., 2006), is still pertinent. Reliance on expert-constructed closed-question surveys may result in potentially significant loss of information and affect the data quality. The three major concerns that crowdsourcing can mitigate are the limitation of expert knowledge, the demand effects, and the observer effect. Crowdsourcing has the potential to alleviate these concerns. Moreover, while average predictions of an expert "crowd" do better, they do not necessarily outperform non-expert "superforecasters," who can be identified via objective measures (DellaVigna and Pope, 2018a).

In certain contexts, non-experts may have an advantage due to their superior local knowledge or, ironically, their ignorance, as expert knowledge relies on a body of literature potentially affected by a publication bias (Franco et al., 2014; Camerer et al., 2018). Two recent public health studies illustrate the efficacy of crowdsourced interventions. In their megastudy of the flu vaccination uptake, Milkman et al. (2022) note that the scientists who designed the megastudy interventions were unable to predict their average or relative performance, while lay survey respondents recruited via Prolific made fairly accurate predictions of both. Otis (2022) takes this approach a step further and shows that non-experts can successfully generate effective behavioral nudges to increase the opt-in rate for COVID-19-related notifications.

Schuman and Presser (1979) is a striking demonstration of both the power and the limitations of a closed-question survey. When asked what people value in jobs, 99% of the respondents who answered the closed form of the question ($n = 460$) chose one of the answers provided by the researchers, with almost 60% selecting "the work is important and gives a feeling of accomplishment."However, when asked in open form ($n = 436$), most respondents provided a new answer that was not among the five closed-question choices, and only 21% of respondents provided an answer that falls in the "stimulating work" category. Even when provided with an option to write in their response, survey respondents tend to confine their responses to the choices offered (Krosnick, 1999) due to the relatively high cognitive cost of constructing a new response versus selecting a preexisting one, being prompted to think about one of the available options (priming effect), or the desire to be compliant and select one of the offered options (demand effects). The crowdsourcing stage of the survey design exposes researchers to a variety of potential answers, hence allowing to mitigate these effects. Finally, by offering *reasonable* options for a closed-form question, the "knowledge-cursed" expert may introduce the respondents to new information and thus change their perceptions. For example, research indicates that people suffer from innumeracy when it comes to estimating the prevalence of minority groups (Nadeau et al., 1993; Sigelman and Niemi, 2001; Chiricos et al., 1997; Wong, 2007), with misperceptions strongly associated with political attitudes (Wong, 2007). Introducing options with *reasonable* brackets in such cases may alleviate respondents' innumeracy and obscure crucial information about the population they represent. Crowdsourcing answer options from the target population ensures that questions contain only information already available to the group to an extent.

In this paper, we propose a crowdsourcing method of constructing a closed-question survey that avoids certain limitations of an expert-constructed survey through what is known as the *wisdom of crowds*. We propose and validate two measures of information loss for closed questions: (1) the probability of an option being selected from the list and (2) sentence similarity to the open-ended answer. As a proof of concept, we conduct an online experiment and construct a survey consisting of four questions corresponding to four different topics. We evaluate our approach using the proposed measures. Finally, we examine the role incentives play in extracting information. In this paragraph, we sketch the theoretical framework used in the paper. It consists of two type of actors: first-stage respondent (or forecaster) and second-stage respondent (or evaluator). The forecasters supply options that can be used in the survey. Evaluators use the supplied options to convey their beliefs.

In our model we assume that the second-stage respondents derive expressive utility from answering the option that best represents her opinion, which has a deterministic component and a random shock of extreme value type 1 distribution. The respondent chooses to opt-out (select "Other / none of the above" option) if the option is too far from her beliefs. That gives rise to a standard logit demand model (e.g. Aguirregabiria 2012). We use the framework to motivate the two measures of option quality. First, a variable measures the distance of an option to true beliefs if and only if it predicts the choices of the respondents. That motivates our test of cosine similarity of the option to the respondent's open-ended answer, which is discussed later. Second, by logit properties, the expected similarity of the chosen option to the true beliefs of the respondent is inversely proportional to the share of the reserve option. That motivates using the share of respondents choosing any option from the list (as opposed to the reserve, "Other" option) as not only a measure of extensive margin of informativeness but also expected distance to true beliefs in general.

The first-stage respondents choose how many and which options to supply. They derive utility from monetary payments as well as non-monetary utility (which we need to rationalize any effort spent without the incentive scheme) and the dislike effort spent on writing the options. We derive intuitive comparative statics results: the monetary payment increases the expected similarity of the chosen option, the probability that any option is chosen as well as its expectation by the respondent, as well as the number of options supplied and effort spent per option.

We propose two measures of information loss for multiple-choice questions: the probability of choosing any answer from the list and sentence similarity to the open-ended answer. Implementing the first measure is straightforward. To implement the second measure, we need to quantify the semantic similarity of the options and open-ended answers. To do so, we use a state-of-the-art sentence embedding model by Reimers and Gurevych (2019). It represents each answer or option as a vector of 768 "characteristics". We use the cosine similarity of the embeddings as a measure of the distance between the answers.

In the first stage of the experiment, we asked participants to create answer options for two open-ended questions. These questions were randomly selected out of the set of four questions adapted from previous surveys targeting U.S. adults[1]. Each question corresponds to one of the four topics—taxation, inflation narratives, purchase of durable goods, and charitable giving. We incentivized participants in one of the two tasks (questions) with a bonus determined by how their options performed in the second stage. The bonus was 20 cents per each of 10 second-stage respondents selecting one of the provided options over option "other / my option is not listed," and could be as low as 0 and as big as 2 U.S. dollars. Half of the participants were offered the opportunity to earn the bonus in the first task, which was followed by an unincentivized task, and vice versa. We did not restrict the participants in terms of the number of options provided total number of options. After completing the main task, the participants answered the same questions from their perspective. Finally, they predicted how well another set of options will perform with the second-stage respondents.

The main task of the second stage also consisted of two questions. For each of the questions, the respondents were asked to write their own one-sentence answer to the question on the screen, after which they answered the same question 12 more times using the multiple-choice option lists we provided. Each list consisted of randomly-ordered options created by a first-stage participant and the option "other." These answers determined the bonus of the first-stage participants.

In each stage of the experiment, the main task was preceded by a set of economic literacy questions and attention checks. After completing the main part, participants also answered questions about their sociopolitical and demographic backgrounds. We ran all parts of the experiment on CloudResearch panel of mTurk, Amazon's online recruitment platform, in October and November 2022 using the survey software Qualtrics. Participation eligibility

[1]That includes Andre et al. (2021); Stantcheva (2021) and Michigan Survey of Consumers.

was restricted to U.S. based adults who had not participated in our pilot study. Additionally, participating in the first stage of the experiment excluded the subject from the pool for the second stage, and vice versa. The target sample size for this study consisted of 788 first-stage participants, 398 of whom were offered a bonus incentive on the first question, and 2409 second-stage participants.

To assess the quality of our similarity measure for options and open-ended answers, we fit a multinomial logistic regression, predicting the choices of the evaluator respondents. To get the appropriate benchmark, in every case, we use a similar model omitting the similarity regressor. We compare the out-of-sample accuracies of the similarity-based and benchmark models.

We also propose a methodological improvement for describing the heterogeneity across subjects. Since each forecast is evaluated on average 30 times, it is likely that the naively estimated distribution will be over-dispersed due to the measurement error: forecasts with high average outcomes might end up there both due to high expected scores and by luck. To circumvent the problem, we follow a three-fold cross-fitting procedure inspired by Abadie et al. (2018) and Chernozhukov et al. (2018). For each forecast, we split the evaluations into three folds of approximately 10 observations. We estimate the quintile of each outcome using two folds. Then we use the remaining fold to estimate the quintile average outcome. The cross-fitting procedure allows us to break the mechanical relationship between the quintile and the measurement error. The average outcomes are unbiased conditional on the quintile. However, quintiles might get misclassified, hence the estimates of the inter-quintile range are conservative.

We find that the quality of options varies significantly between respondents. The average probability of an option being chosen from the list increases two-folds between the bottom and the top 20% (35.3% to 80.6%), while the average cosine similarity increases three-folds (0.107 in the bottom 20% to 0.369 in top 20%). Monetary incentives significantly increase the time participants spend on writing the option lists, as well as the number of options in the list and the number of characters per option, indicating a considerable effect of monetary incentives on participants' effort. The quality of the lists also improves by both measures.

The remainder of this paper is organized as follows. Section 2 builds the theoretical foundations for the experiment and derives this paper's core testable prediction. Section 3 presents the experimental design in detail and describes the data we collect. Section 4 outlines the empirical framework. Section 5 presents the results, followed by the interpretation, a battery of robustness checks, and a discussion of the limitations. Section 6 concludes with practical recommendations and suggestions for future research.

2.2 Experimental design

Our experiment to estimate the effect of incentives and forecaster characteristics on the quality of the responses consists of two stages. In the first stage, we crowdsource sets of potential responses to questions of interest. In the second stage, we evaluate the responses with an actual survey of respondents from the same population. Each of the stages proceeds in three parts: background knowledge questions, the main task and stage-specific questions, and demographic questions. The first and third parts are common across stages.

This section begins with a description of the main tasks in each stage. Next, we describe the treatments and their implementation. Then we describe the characteristics we collect to explain the individual heterogeneity in performance. We conclude the section with a description of the outcomes: response set quality measures and measures of effort.

2.2.1 Main tasks

The goal of the main part of our experiment is crowdsourcing and evaluating survey response sets for a question of interest.

Questions We compose a list of four questions to serve as examples for our method. We select the questions that are initially open-ended either to be representative of the current literature and/or have a proximate action that depends on them. The full list of questions can be found in Appendix 2.A.

The two pre-existing questions are selected to be representative of the recent work in economics. They are taken directly from studies that are either published or under revision in top-5 economics journals. One question asks on perceptions of tax policy (Stantcheva, 2021) and another asks about perceived causes of increased inflation (Andre et al., 2021). The question on the causes of inflation uses the estimate of inflation from September 2022 (8.2%), the latest available at the start of the experiment.

Two other questions are based on established surveys and ask about determinants of some concrete action or judgement. This will allow us to compare the informational content of the answers: intuitively, if respondents differ in the factors they perceive as most important and we have more information on the most important consideration for a respondent, we should have higher predictive accuracy for the decision.[2]

The first question elicits rationales for buying durable goods. The determinants in purchases of durables have received considerable attention in the recent literature on macroeconomic expectations and forward guidance. (cite a bunch) Consumer expectation surveys often ask if it is a good or bad time to buy various items such as durable goods or a house. Our question is based on one of such surveys (Michigan Survey of Consumer Attitudes and Behaviors, 2022) and asks to give the most important factors when deciding if it is a good time or a bad time to buy durable goods. The associated judgment is whether the respondent thinks that it is a good time to buy durables at the time of the survey.

The second question asks the respondents how they decide to donate to a charitable organization. The literature on charitable giving is vast and has pointed out a number of

[2]We use 12-month selections of the Consumer Price Index published monthly by the Bureau of Labor Statistics (BLS).

reasons how people decide on the donation. We base our question on one of the Gallup (2016) polls. The associated action is the amount donated to charitable organizations over the last year.

Stage 1: forecasting In the first stage of the experiment, we asked participants (from now on, we refer to them as "forecasters") to create sets of potential answers for two randomly selected questions from our list. This allows us to use within-subject comparisons while keeping the overall survey relatively short. The explanations of the task are presented over 3 screens. First, we present the question for writing the sets. Second, we explain that we will use their answers with 10 members of the same population. We explicitly tell the subjects that their goal is to write a set of answers that will be popular with the respondents, that is chosen over the "Other" option.[3] Third, we present the incentive scheme.

The fourth screen briefly summarizes the instructions and offers the forecaster a textbox to write in the options. The respondents are not limited with respect to the length of answers per text box and are allowed to add any number of text boxes.[4] We disable copy-pasting text to the text boxes to limit the returns to searching for answers on the web. Although it does not rule out the searches, it removes the opportunity to use the answers from the web without typing.

After the forecaster submits the responses, the process repeats for the second question. The order in which the two selected questions appear is random. We repeat the instructions as is before the second question to minimize the order effect. For example, if the respondents have full access to the information: if the respondent didn't pay attention to the instructions on the first question, she still has the opportunity to read it before answering the second. For both questions, we record the final answer, the total time to answer, as well as 5-second snapshots of the responses to have a detailed picture of the time allocation.

---

[3] We also say that we might remove answers that we deem unethical. This never occurred in practice.

[4] For technical reasons we had a limit of 20 text boxes, but it was never binding.

Stage 2: evaluation In the second stage, we evaluate the forecasted response sets with an actual survey of respondents drawn from the same population. The main task of the second stage consists of two steps. At the first step, the respondent is shown the question of interest and asked to answer it in one full sentence. In the second step, the respondent is asked the same question 12 more times using multiple-choice answer sets. Each set consisted of options created by a first-stage participant. We add the option "Other" to each set and explicitly say that he should use "Other" if none of the answers matches his opinion. We don't edit or rearrange the sets. After the respondent picks answers from the 12 multiple-choice sets, the steps are repeated for the second question. The two questions are randomly selected from the list of four questions and the order is random.

The main task is designed to mimic the usual survey practices, yet we have introduced several notable differences. We take several measures to make sure it does decrease the generality of our findings. The first difference is that we ask the questions repeatedly. On the one hand, it dramatically increases the number of observations and allows us within-subject comparisons. On the other hand, it might cause interference: the first sets can affect the answers to the later ones. Luckily, we can address this issue simply by checking the robustness to exclusion of late answers. The second difference is that the responses are crowdsourced and are out of our control. This introduces a concern that the respondents might stop taking the survey seriously if they see some forecasted sets as unprofessional. For example, the forecasts can potentially contain orthographic mistakes or politically charged language. We check the robustness to this by introducing a note that the responses are crowdsourced from other people on the same platform. We present the note to a random half of the respondents. Third, we ask an open-ended version of the question before the multiple choice one. This is key to our design, as it allows us to have a measure of the respondent's beliefs that is independent of the forecasted set.

One of the concerns comes from similarity to most opinion surveys: truth-telling is not incentivized in the second stage. This makes it subject to the usual criticisms that such questions attract. Our approach is flexible enough to accommodate incentivized versions of the second stage (e.g. using Bayesian Truth Serum, Prelec (2004); Witkowski and Parkes (2012); Hussam et al. (2022)), but it is out of the scope of this paper.

2.2.2 Incentive treatments

The first goal of our experiment is to estimate the effect of incentives on the quality of sets created in the forecasting stage. To do so, we offer the forecasters a piece-rate bonus for their performance in one of the two questions. The bonus was 20 cents per each of the 10 second-stage respondents selecting one of the provided options over option "other / my option is not listed".[5] The incentives for each of the questions are explained gradually over 3 screens of instructions. First, we notify if the respondent that the question is incentivized or not. We hypothesize that, as part of the mechanism for the incentive treatment, incentivized forecasters might spend greater effort when reading the instructions. Hence we start the explanations with the total amount that could be earned:

> *In this section, we will [ask you to write/you can earn up to $2 in bonuses by writing] potential answers to the following question:*
>
> *[Question]*

Second, we proceed with the instructions on the goal of the task described above. Third, we explain the incentives in detail:

> *Please do your best and write the answers you think are most likely. You can write as many answers as you see fit. We will compute the number of people who*

[5] We used the 10 respondents for the simplicity of exposition. In principle, the number of second stage respondents does matter for risk-averse forecasters: the larger is the sample size, the less stochastic is the outcome, conditional on the true distribution of respondent types and the forecasted answers. One can theoretically remove this feature by binarizing the payment similarly to Hossain and Okui (2013), but it is not certain that it will improve elicitation (Selten et al., 1999; Danz et al., 2022).

choose one of your answers and not "other."

[Your payment will not depend on the number of such people./ You will get a 20 cent bonus payment for each such person (up to $2).]

Half of the forecasters were offered the opportunity to earn the bonus in the first question, which was followed by an unincentivized question, and vice versa. The instructions are approximately the same in both cases, allowing for minor alterations..

Our incentive treatment has three important features. First, we aim to test the *behavioral* incentive compatibility of our incentive scheme, hence we choose to transparently communicate the scheme to the participants. We also don't claim that truth-telling is in their best interest. Danz et al. (2022) show that some of the prominent incentives schemes perform worse if explained transparently. As Charness et al. (2021) points out, eliciting the beliefs with intransparent schemes rely on the respondent's trust to the researcher and not the incentive compatibility itself. Since we want to establish incentive compatibility independently of trust, we opt for the transparent option.

Second, our incentive scheme is relatively simple. It essentially asks the respondent to describe the highest modal responses and is theoretically analogous to other frequency or interval-based mode-elicitation mechanisms. Charness et al. (2021) classify frequency and interval-based mechanisms as simple and hypothesize that simple mechanisms would perform robustly better than unincentivized elicitation. They also call for more testing of simple mechanisms as most of the literature is focused on complex ones (such as the binarized scoring rule, Hossain and Okui, 2013). We aim to offer one such test.

Third, we follow DellaVigna and Pope (2018b) in using a similar sentence structure across treatment arms to ensure the effects are driven by incentives only.

2.2.3 Individual characteristics

The second goal of the experiment is to estimate which respondent characteristics lead to better forecasts. To do so we ask a series of questions that we expect to be relevant to the task. This includes beliefs on the subject of our main questions and general background questions. We keep most of the questions identical across stages in content and order. However, we change the order or omit some questions for the evaluation stage, either due to concerns of contamination or survey fatigue.

Beliefs We collect the subjects' own beliefs about the question of interest as well as second-order beliefs.

Own beliefs: We ask subjects in both stages to write their opinion about the questions from the main task as an open-ended response. On the one hand, it allows us to compare the beliefs elicited with the crowdsourced sets to the beliefs elicited in an open-ended way. On the other hand, we hypothesize that the crowdsourced responses are likely to be related to the forecaster's own beliefs since the correlation of own and second-order beliefs is a robust empirical regularity (Bursztyn and Yang, 2022).

Beliefs about other sets: We hypothesize that the ability to write answers should be directly related to the ability to predict which answers would be popular when given a pre-existing set. To test that we elicit forecasters' beliefs about answer sets that we have collected in the first wave of the study. The first wave consisted of 74 forecasters resulting in 35 to 39 answer sets per each of our 4 questions. We evaluate the answer sets in with 30 evaluators each and select 3 sets for each of the 4 questions. The selection is random, but we stratify on the realized quality (1 from the best, 1 from around the median, and 1 from the bottom) and keep the total number of options for predictions to be 9 across all sets (including "Other").

Since our main quantity of interest is the prediction error, it is important precisely to measure the true fraction of responses for the prediction sets. We include the sets in all subsequent evaluation stages: 2 out of 12 sets for each question for each subsequent evaluator are the prediction sets. This results in around 750 evaluator responses per answer set. Hence, our estimates of the true fractions are bounded by approximately 1.8 percentage points.

We incentivize the forecasters for the precision of their answers to this question. We explain that we will ask the question to 100 respondents and pay for guessing the correct fractions within 3 percentage points of the true value. As Danz et al. (2022) and Abeler et al. (2019) argue, this incentive is simple, does not induce central tendency, and the elicited quantity (the mode) is plausibly close to the expected value. Since we elicit multiple fractions that are jointly distributed, we use randomization to avoid hedging across questions: we randomly choose one of the 9 predicted fractions and pay \$2 if the guess for this fraction is within 3 percentage points of the truth (Azrieli et al., 2018).

Beliefs about own sets: In a similar fashion, we ask the forecasters to predict the popularity of the sets that they have just written. This allows us to test for misperceptions in the popularity of the written sets. Since incentives for this question would introduce regret[6], we do not incentivize this elicitation.

Specific knowledge and opinions: We ask one question to capture background knowledge on the topic of the main question. Specifically, we ask for the share of personal income paid by people in the top federal personal income tax bracket for the taxation question; the current level of inflation for the inflation question; the fraction of people in the U.S. who donated for charity this year for the charity question; if it is a good time to buy durable goods for the durables question. To minimize fatigue, we ask only the questions that match the topic of the main questions. Finally, we ask all the subjects for the approximate quantity donated to charities over the last 12 months.

Understanding the crowdsourcing: In order to identify forecasters who understood our incentive scheme, we ask them to estimate their bonus for the crowdsourcing task if 6 people

[6] for example, one might be better off writing no options in the crowdsourcing part as she would be certain to win the prize by guessing 100% "other" in this prediction exercise

used their options. We only remind them of the total possible bonus for the question, not of the piece-rate. We also ask for their self-reported understanding of the scheme. Finally, we ask them to write their opinion on the hardest part of this task to gauge the mechanisms behind the results.

General information Our main questions are either political or economic in nature. Hence we ask a set of questions on political views and financial literacy. Political views include party affiliation, as well as voting decision and preferred candidate in 2020 presidential election. Financial literacy is measured with standard "big three" financial literacy questions. We also include two attention checks and a set of demographic questions including gender, age, and income bracket.

Differences in forecaster and evaluator stages Here we list the differences in characteristics elicited during the forecaster and evaluator stages.

Own beliefs: The forecasters are asked for their opinions after all questions in the main task; evaluators are asked right before each question in the main task. The reason is that our main concern at the forecasting stage is survey fatigue: if we ask the forecasters to write before the main task, it is likely that the forecasters will start the main task more tired and both the effects of treatments and characteristics will be attenuated. Our main concern for the evaluator survey is contamination: the respondents will be primed by the answers in the forecasted sets.

Beliefs about others and own sets: We don't ask these questions to the evaluators.

2.2.4 Outcomes

In this subsection, we describe the outcomes that we collect as a result of our experiment. The main set of outcomes measures the quality of survey answer sets. The secondary set of outcomes measures effort and respondents' beliefs about the quality: the outcomes allow us to better understand the mechanisms behind the changes in the quality.

Measuring answer quality We define the quality of an answer set as the quantity of information it allows to transmit by the respondent. To quantify this, we propose two measures. First, we look at the probability that a random respondent picks an answer from the list, as opposed to picking the "Other" option. Second, we look at the semantic distance between the open-ended response and the answer chosen from the set.

Probability of "Other" option: Our first measure service quality is straightforward. We look if the respondent selects the "Other" option as opposed to one of the options in the list. This measure assesses the extent to which respondents feel that their views are not captured by the provided options. The use of the "Other" option in survey questions is common practice and a natural way to measure information lost in surveys. It is related to the concept of saturation which is the extent to which survey response options fully capture the spectrum of possible answers (e.g. Saunders et al., 2018; Guest et al., 2020; Hennink and Kaiser, 2022). We assume that the number of potential answers is large enough for the other option to be uninformative.

Semantic distance to the open-ended answer: Our second measure of quality is the semantic similarity between the respondent's open-ended answers and the answer selected in the multiple-choice question. We use Sentence-BERT (Reimers and Gurevych, 2019), a state-of-the-art language model fine-tuned to predict sentence similarity.[7] Specifically, we compute the cosine similarity between the embedding vectors of the open-ended and multiple-choice answers.[8] This allows us to capture the intensive margin differences between the answer sets. For example, allows us to discriminate between the answer sets that have the same

[7]In particular, we use all-mpnet-base-v2 implementation as the most precise general-purpose model in the library according to benchmarks. We later plan to fine-tune the model for our particular application.

[8]Automated text analysis has been viewed as a tool that works only with large corpora of texts. However, the development of large pre-trained language models (Devlin et al., 2019) has allowed for analysis of smaller data sets, such as survey responses (e.g. Bursztyn et al. (2022))

span but differ in the granularity of options. In the later section, we discuss the theoretical justification of the measure and provide a test of its validity in our context.

Secondary outcomes To complement the results on outcomes, we measure the effort exerted by the forecasters and their beliefs about quality. We observe the time spent on the task, the total number of answers written in the set, and the number of characters used to type the answers. As DellaVigna et al. (2022) point out, some dimensions of output and effort are more elastic than others: most of the adjustment is likely to come from the extra time spent per task. We intend to check if this is indeed the mechanism in our case. As for the beliefs, we use the number of respondents they expect to use one of their options elicited after they have written the sets.

2.2.5 Sample and recruitment

The sample size for this study consists of 788 first-stage participants, 398 of whom were offered a bonus incentive on the first question, and 2409 second-stage participants.

We ran all parts of the experiment on CloudResearch panel of mTurk, Amazon's online recruitment platform, in October and November 2022 using the survey software Qualtrics. Participation eligibility was restricted to U.S.-based adults who had not participated in our pilot study. Additionally, participating in the first stage of the experiment excluded the subject from the pool for the second stage, and vice versa.

2.3 Model

This section has two goals. First, we built a simple theory of survey responses and derive the multinomial logistic demand model for answer options. We use the model provide a validation scheme for our semantic similarity measure based on the demand estimation exercise. Second, we build an illustrative model of forecaster behavior based on monopolist location choice under horizontal differentiation (e.g. Anderson et al. (1992)). We use the model to obtain testable predictions for the reactions of forecasters to the incentives and the relationship between the forecaster's beliefs and answer quality.

2.3.1 Respondent's behavior

In this section, we describe a simple model of a survey respondent behavior when facing a multiple-choice question. We assume that the respondent derives expressive utility from sharing her opinion. More restrictively, we assume that the respondent values only the expressive value of the survey and hence only maximizes the similarity of her chosen response to her true beliefs. Further, we assume that the utility can be decomposed into two components: mean utility δ_{ijk} and random utility shock ε_{ijk} distributed i.i.d. with extreme value type 1 distribution.[9]

$$u(\theta_j, x_{ijk}) = \delta(\theta_j, x_{ijk}) + \varepsilon_{ijk} \tag{2.1}$$

where $\theta_j \in \Theta$ denotes the respondent j's true beliefs. We assume that the options and beliefs can be represented as real vectors, $\Theta \subset \mathbb{R}^N$ and Θ is closed. This gives rise to a standard multinomial logit model. Further, assume that the mean utility of option x_k is linear in distance between the option and the true belief θ:

$$\delta(\theta_j, x_{ijk}) = \alpha \left(\kappa - \frac{1}{2} \|\theta_j - x_{ijk}\|^2 \right) \tag{2.2}$$

α is the weight the respondents puts on the similarity of the chosen option to the true belief. κ can be interpreted as the reserve distance: if all options are further from the true opinion θ_j than $\sqrt{2\kappa}$, then the respondent is on average better off choosing the reserve ("Other") option 0. When we move to estimation, it will be convenient to assume that all the belief

[9]Depending on the way one conceptualizes the utility shock, it might or might not be included into the similarity measure. We choose to assume that the utility shocks are not related to the similarity.

vectors x_{ijk} and θ_j are normalized.

$$\|x_{ijk}\| = \sum_l x_{ijkl}^2 = 1, \ \|\theta_j\| = \sum_l \theta_{jl}^2 = 1 \tag{2.3}$$

This allows us to represent the distances in terms of the cosine similarities:

$$m_{ijk} = \theta_j \cdot x_{ijk} = 1 - \frac{1}{2}\|x_{ijk} - \theta_j\|^2 \tag{2.4}$$

and hence we can write the mean utilities as:

$$\delta(\theta_j, x_{ijk}) = \alpha\left(\kappa - \frac{1}{2}\|\theta - x_{ijk}\|^2\right) = \alpha\left(\theta_j \cdot x_{ijk} - \beta\right) \tag{2.5}$$

where $\beta = 1 - \kappa$. Denote the chosen option by $x^*(\theta, X, \varepsilon_{ij})$, that is:

$$x^*(\theta, X, \varepsilon_{ij}) = \arg\max_{x_{ijk} \in X} \{u(\theta, x_{ijk})\} \tag{2.6}$$

Next, define the expected similarity of the chosen option:

Definition 4. *We call $M(\theta, X) = \mathbb{E}[\delta(\theta, x^*(\theta, X, \varepsilon_{ij}))|\theta]$ the expected similarity of the opinion of respondent of type θ to the chosen option from the set X.*

We will also denote the share of option $k \in \{0, ..., K\}$ for the respondent of type θ for the set X as $s_k(\theta, X)$.

2.3.2 Demand estimation and validation of the similarity measure

We use the model to validate our main measure of information loss: the similarity of the chosen response to the open-ended answer. Validating measures extracted from texts is a key issue in all research using text as data (Ash and Hansen, 2023). Our methodological innovation is to derive the validation procedure directly from our discrete choice model.

Validation procedure In Section 2.3.1 we have introduced a model of demand for response options. In the model, the demand depends only on the semantic similarity between the answer options in a given set and the true beliefs about the issue, which is precisely the variable that we intend to measure.

Our feasible measure differs from the variable in two ways. First, we use the open-ended answer as a proxy for the true belief about the issue. Second, we approximate the similarity as perceived by the respondent by the similarity predicted by the Sentence-BERT language model. It is reasonable to assume that the noise will appear at both steps, therefore it is important to check if the measure is at all correlated with the true similarity. To do so, we estimate the model of demand for options using the feasible measure instead of the true similarity. Assuming that the measure is linked to the choices only through the similarity, the measure will be predictive of the choices only if the measure predicts the true similarity. Recall from the Section 2.3.1 the demand model and plug in the feasible similarity measure $\hat{m}_{ijk}$:

$$u_{ijk} = b + a\hat{m}_{ijk} + \varepsilon_{ijk} \tag{2.7}$$

$$\hat{m}_{ijk} = m_{ijk} + \nu_{ijk},\ \mathbb{E}[\nu|m_{ijk}] = 0 \tag{2.8}$$

Hence, we can estimate the probabilities of a respondent j choosing option k from a set i using the following multinomial logit regression:

$$\mathbb{P}(C_{ij} = k|\delta) = \frac{\exp(\delta_k)}{\sum_l \exp(\delta_l)} \tag{2.9}$$

$$\delta_k = b + a\hat{m}_{ijk}, \delta_0 = 0 \tag{2.10}$$

where C_{ij} is the option chosen by the respondent j from set i. The estimated parameter $\hat{\alpha}$ is attenuated due to the measurement error ($plim\ \hat{\alpha} < \alpha$). However, it is positive if and only if the parameter is itself positive. ($plim\ \hat{\alpha} > 0 \Leftrightarrow \alpha > 0$). This brings us to the following test:

Hypothesis 1. *The estimated parameter $\hat{\alpha}$ is positive, $\hat{\alpha} > 0$, and hence the measure is positively correlated with the true similarity, $cov(\hat{m}_{ijk}, m_{ijk}) > 0$.*

We also consider accuracy, the probability that the model guesses the choice correctly. We compare the accuracies of the full model to the short model that only discriminates between the given options and the "Other" option (setting $\hat{\alpha} = 0$). This allows for an intuitive measure of the precision of the model. Similarly to the test before, the full model should outperform the short model if and only if the feasible similarity measure is valid.

2.3.3 Forecaster's behavior

In this subsection, we formulate the forecasters' problem and derive comparative statics. A forecaster chooses number K and location of options x_k in the answer set writes, $X = \{x_k\}_{k=1}^{K}, x_k \in \Theta$. The forecaster has to spend the cost of effort $c(K)$ that depends on the number of written options. The effort cost $c(K)$ is strictly convex. For simplicity, we will analyze the case with one respondent.[10] The beliefs of the forecaster are distributed according to the probability density function $\hat{f}(\theta)$. Following DellaVigna and Pope (2018b) we allow for non-monetary utility λ that rationalizes non-zero output without monetary incentives. This gives the following expected utility function:

$$\max_X u^F(X) = \mathbb{E}_\theta \left[1 - s_0(X,\theta)\right](p+\lambda) - c\left(K\right)$$

$$s_0(X,\theta) = \frac{1}{1 + \sum_{k=1}^{K} \exp(\delta(\theta, x_k))}$$

[10]Another equivalent approach is to assume risk-neutrality or binarized payment scheme. Allowing for multiple respondents and risk aversion is important quantitatively, but would yield qualitatively similar results.

Now we are ready to obtain the first set of comparative statics results.

Proposition 5. *The following statements are true:*

1. *At least one solution X^* exists.*

2. *The optimal number of answers $K^* = |X^*|$ is unique (up to non-divisibility) and weakly increase in p.*

3. *The total cost of effort $c(K^*)$ is weakly increasing in p.*

4. *The forecaster's expectation of the share of the outside option $\mathbb{E}s_0(X^*, \theta)$ weakly decreases in p.*

Proof. Sketch: For (1): $\hat{f}(\theta) \longrightarrow 0$ as $\theta \longrightarrow \infty$ in a any direction, hence we can cut Θ to be compact. This ensures solution exists for every K. Denote $X^*(K)$ the solution to the problem for fixed K. $\mathbb{E}(1 - s_0(X^*(K), \theta))$ has decreasing differences in K that converge to 0 as $K \longrightarrow \infty$ and $c(K)$ is convex and increasing, hence optimal number of options K^* exists and is unique. For (2-4): $\mathbb{E}(1 - s_0(X^*(K), \theta))(p + \lambda) - c(K)$ is supermodular in (K, p), hence K^* increases in p. $\square$

Proposition 5 states that the incentive treatment should increase effort, number of written options, and quality according to our two measures: the probability that one of the options ends up being chosen and the expected similarity of true beliefs to the chosen option. As K is discrete, it is possible to have two neighboring optimal quantities K^* and $K^* + 1$. This is a razor-sharp case, so we just take the smaller set as the solution.

Next, we analyze the location of the options. To do so, we make a further simplifying assumption. In our main model, we have allowed for Θ to have a high dimension. This is important, as many issues are hard to shoehorn into a single dimension. For example Ahler and Broockman (2018) shows that representing issues on a one-dimensional partisan spectrum can mask a lot of belief heterogeneity and change conclusions. However, this makes the analysis of locations complex, as one has to look at directional changes along the belief space. To simplify the analysis we assume that the belief space is one-dimensional, $\Theta = \mathbb{R}$, and that its probability density function exists, is symmetric, and log-concave. We will also allow the density function to depend on a shape parameter t.

Assumption 4. *The belief set $\Theta = \mathbb{R}$. The probability density function of θ $f(\theta - \hat{\theta}_m, t)$ is symmetric around the perceived mean $\hat{\theta}_m$ and log-concave.*

This ensures there exists a solution with a simple symmetric structure.

Proposition 6. *Suppose that Assumption 4 holds. Then there exists a symmetric solution:*

$$X^* = \begin{cases} \{\hat{\theta}_m - \tilde{x}_{\bar{K}}, ..., \hat{\theta}_m - \tilde{x}_1, \hat{\theta}_m + \tilde{x}_1, ..., \hat{\theta}_m + \tilde{x}_{\bar{K}}\}, K^* - even; \\ \{\hat{\theta}_m - \tilde{x}_{\bar{K}}, ..., \hat{\theta}_m - \tilde{x}_1, \hat{\theta}_m, \hat{\theta}_m + \tilde{x}_1, ..., \hat{\theta}_m + \tilde{x}_{\bar{K}}\}, K^* - odd. \end{cases} \tag{2.11}$$

where $\bar{K} = \lfloor K/2 \rfloor, \tilde{x}_k \geq 0$. Moreover, if $f(\theta - \hat{\theta}_m, t)$ is strictly log-concave, then the solution is unique.

Proof. $1 - s_0(x, \theta))$ is log-concave for $x_1 \leq x_2 \leq ... \leq x_K$. Hence, $\mathbb{E}[1 - s_0(x, \theta))] = \int_{-\infty}^{\infty}[1 - s_0(x, \theta))]f(\theta - \hat{\theta}_m, t)d\theta$ is also log-concave in x for every K. Hence, if x^* satisfies the first order condition, it solves the problem. Since FOC is symmetric around $\hat{\theta}_m$, symmetric solution exists. □

We will analyze the solution to derive the comparative statics.

Finally, in order to obtain comparative statics for own opinions of the forecasters we need to relate the opinions to the beliefs about the opinions of others. Let's denote the own opinion of the forecaster as θ^o. We assume that the beliefs about the opinions of others are biased towards the forecaster's own opinion.

Assumption 5. *The perception of the mean $\hat{\theta}_m$ increases with own opinion θ^o and the forecaster with mean belief is unbiased, $\hat{\theta}_m(\theta^o)\Big|_{\theta^o = \theta_m} = \theta_m$.*

This is sometimes called “self-predicting prior” assumption (Radanovic and Faltings, 2013). The assumption is plausible: Bursztyn and Yang (2022) find that biases in beliefs towards own opinions is one of the robust empirical regularities in the literature.

This allows us to state the second comparative statics result:

Proposition 7. *Suppose that Assumptions 4 and 5 hold. Then, each $x_k^* \in X^*$ is increasing in forecaster's mean belief $\hat{\theta}_m$ and forecaster's own belief θ^o.*

Assumption 6. *The weight of the option characteristics dominates the noise, $\alpha \to \infty$.*

Under the Assumption 6, the model becomes the usual Hotelling model. This drastically simplifies analysis and allows us to have comparative statics for the answer shares and surpluses under the objective distribution and present a closed-form solution for the model. The assymption looks strong, so we defend it in two ways. First, the reason for introducing the extreme value type-1 ε_{ijk} noise in the model is often computational: some of the consequences of the noise are counterintuitive and the pure characteristics models are generally assumed to be a closer representation of reality (Lu and Saito, 2022). Second, as we show in the following sections, our estimate of α is relatively large compared to the average similarity of beliefs. Since the estimate is likely to be drastically attenuated due to measurement error, it is reasonable to expect the EVT1 noise to be relatively unimportant empirically.

Proposition 8. *Suppose that Assumptions 4, 5, and 6 hold. Then the optimal forecaster's expected share of respondents served is:*

$$\mathbb{E}s(X^*, \theta) = 2\hat{F}\left(\frac{bK}{2}, 0, t\right) - 1 \tag{2.12}$$

and there is a solution to the forecaster's problem of the form:

$$X^* = \begin{cases} \{\hat{\theta}_m - b\bar{K}, ..., \hat{\theta}_m - b, \hat{\theta}_m + b, ..., \hat{\theta}_m + b\bar{K}\}, K^* - even; \\ \{\hat{\theta}_m - b\bar{K}, ..., \hat{\theta}_m - b, \hat{\theta}_m, \hat{\theta}_m + b, ..., \hat{\theta}_m + b\bar{K}\}, K^* - odd. \end{cases} \tag{2.13}$$

where $b = 2\sqrt{2\kappa}$, $\bar{K} = \lfloor K/2 \rfloor$. If $f(\theta, \hat{\theta}_m, t)$ is strictly log-concave, then the solution is unique.

Proof. A respondent of type θ is served if and only if the distance to her from the closest option is smaller than $\sqrt{2\kappa}$ as $\alpha\left(\kappa - \frac{1}{2}\|\theta_j - x_{ijk}\|^2\right) \geq 0$. Since the utility depends only on the served share, the optimal distance between the points is at least $2\sqrt{2\kappa}$. Since the density is log-concave and symmetric, the optimal solution is such that a symmetric interval around the mode is served. □

We use this simplified structure to relate the number of options written to the shares and expected similarities according to the objective distribution $f(\theta, \theta_m)$.

Proposition 9. *Suppose that Assumptions 4, 5, and 6 hold. For any objective distribution $f(\theta, \theta_m)$. Then the expected share of respondents served $\mathbb{E}s(X^*, \theta)$ and the expected similarity $\mathbb{E}M(X^*, \theta)$:*

1. *weakly increase in piece-rate monetary incentive p.*
2. *decrease in the difference in the belief mean to the objective distribution mean, $|\hat{\theta}_m - \theta_m|$.*
3. *increase in forecaster's own belief θ^o for $\theta^o < \theta_m$ and decrease in θ^o for $\theta^o \geq \theta_m$.*

Proof. Since the forecaster serves the central segment $[\hat{\theta}_m - \frac{bK}{2}, \hat{\theta}_m + \frac{bK}{2}]$, it is straightforward to show that for any symmetric log-concave objective distribution of respondent types the expected shares and similarities increase with the incentive p and decrease with the bias in beliefs $|\hat{\theta}_m - \theta_m|$. Driven by the result for the bias in beliefs is the result for the own type: suppose that there is some type $\bar{\theta}_o$ whose own beliefs correspond to an unbiased second-order belief distribution. Then, the expected shares and similarities will be decreasing in the distance to this type as by Assumption 5 second-order beliefs will become more and more biased as one moves from the unbiased type. □

2.3.4 Testable implications

In this subsection, we relate the theoretical results to the empirical application. First, from Propositions 9 we obtain the following hypothesis:

Hypothesis 2. *The probability of respondents choosing an option from the list, and the average semantic similarity to the of the chosen option to the open-ended answer increases with the incentive treatment.*

The hypothesis simply relates the piece-rate monetary payment to the model. Since our treatment increases p from 0 to 20 cents, one should expect an increase in the quality measures consistent with the model.

The next hypothesis is concerned with the secondary variables affected by the incentives:

Hypothesis 3. *The total time spent, the total number of characters, the number of written options, the number of characters per option, and the forecaster's expectation of the share of respondents choosing an option from the list, increase with the incentive treatment.*

In our model, the increase in quality is driven by the increase in the number of options written by the forecaster. Proposition 5 shows the result. For simplicity, we assume that effort is only spent on the extensive margin i.e. depends on the number of options written. It is reasonable to expect though that the forecaster might also have to spend effort to make each particular option clearer. One could model this either by assuming that the locations of the options are perceived with precision proportional to the effort spent or that the options are included in consideration set with probability proportional to the effort (as in Goeree, 2008). We choose to leave the intensive margins of effort out of the model for tractability, but still include them into our hypotheses.

Finally, by Propositions 8 and 9 we expect the quality to depend on the forecaster's beliefs about the opinions of the respondents. We formulate the hypothesis in terms of distances between subjective and true expected values.

Hypothesis 4. *The probability of respondents choosing an option from the list, and the average semantic similarity to the of the chosen option to the open-ended answer decreases with the distance between objective mean θ_m and the perceived mean $\hat{\theta}_m$ and with the distance between the respondent's opinion θ_m and forecaster's opinion θ^o.*

One could also expect that some of the observable characteristics can be correlated with distances in opinions. Hence, we also will test the effect of demographic distance.

2.4 Results

2.4.1 Validating the similarity measure

We start by validating our measure of semantic similarity. To do so, we estimate the multinomial logit model as specified in Subsection 2.3.2. Our estimator is (quasi-)maximum likelihood. Table 2.2 presents the evidence for the Hypothesis 1. The estimated weight $\hat{\alpha}$ of the similarity measure is positive and highly statistically significant (3.143, $p < 0.001$) for the full sample. The estimates are similar in magnitude across the four questions we have asked: the coefficients vary from 2.840 ($p < 0.001$) for the inflation narratives question to 3.551 ($p < 0.001$) for the durable goods question. The estimates are consistent with Hypothesis 1. We also find the value of the reserve similarity to be quite high: the estimate for the full sample is 0.399 ($p < 0.001$) relative to the average value of similarity *hey Lena where is the table?* These patterns appear consistent with the attenuation effects due to noise in the measure: the reserve option utility seems to be inflated for the topics where the measure performs worse.

Table 2.3 shows the accuracies of the multinomial logit model. The model performs surprisingly well: the full model has 8.7 percentage points (pp) higher accuracy compared to 39% percent of the short model, and the difference is highly statistically significant ($p < 0.001$). This appears to be mostly driven by the intensive margin: when we split the prediction task into predicting which option is selected conditional on selecting one of the options and predicting if any of the options is selected, the gain in the first task is 16.7 percentage points relative to 30% accuracy ($p < 0.001$), while the improvement for the binary classification is only 3.3 pp relative to 66%, albeit statistically significant ($p < 0.001$). The gain is positive for all topics but varies from 3.1 pp relative to 45.6% for inflation ($p < 0.001$) to 11.6 pp relative to 38.7% for taxation ($p < 0.001$).[11]

These results combined imply that our similarity measure is indeed capturing some of the variation in the true similarity of beliefs. Hence, we proceed to estimate the effects of the treatments and characteristics on the measure.

2.4.2 Effect of incentives.

In this section, we show the effects of the incentive treatment on the quality of answers as well as the effort spent on answers.

Regression specification. For our main results we estimate the forecast-level regressions of the form:

$$Y_{it} = X_i\beta + I_{it}^{incentives}\tau + e_{it} \tag{2.14}$$

where Y_{it} is the outcome associated with the forecast done by forecasters i on the elicitation time $t \in \{1, 2\}$, X_i is the vector of controls including forecaster fixed effect, question topic fixed effect, and elicitation time fixed effect. Outcomes include the forecast characteristics and the measures of information loss. The information loss outcomes are observed on the evaluation level. We choose to collapse the observations by forecast to simplify including the null forecasts in the analysis. For such forecasts, we impute 0 for all information loss measures.

[11] This can either reflect the differences in the fit of the embeddings or in the innate predictability of answers within the topics. In the later stages of the project, we will check the robustness of these findings with human-coded similarity and with fine-tuning the embeddings to our dataset.

Incentives increase the quality of answers. Table 2.4 shows the effects of the incentive treatment. Columns (1) and (2) of Panel A provide evidence for Hypothesis 2. It shows the treatment effects on the option list quality measures. The incentives increase the probability that one of the options gets chosen by 2.7 percentage points relative to the baseline of 60.3% and the average similarity of the chosen option to the open-ended answer by 0.017 relative to the baseline of 0.234. Both estimates are statistically significant ($p < 0.001$) and are aligned with each other: a 0.017 increase in similarity for a set of 1 option with an average non-zero similarity of 0.38 would result in about 1.3 percentage point increase in probability of this option being chosen according to our logit model. The difference is likely to be due to the measurement error in the semantic similarity.

For the probability of choosing an option from the list, the correct reference point is probably the fraction of respondents who end up using one of the options due to treatment but wouldn't use one otherwise. This "response discovery rate"[12] is about 7.1% for our incentive treatment.

Column (3) shows the effects for the intensive margin of the similarity. The incentive treatment increased the similarity of chosen option conditional on any option being chosen by 0.011 relative to the baseline of 0.380. The effect is statistically significant ($p = 0.012$). The estimate, combined with columns (1) and (2), suggests that the quality is increasing with the treatment both through the extensive and intensive margin: if the forecaster is incentivized, the respondents are more likely to choose one of the options and the options they choose are closer to their true beliefs.

Incentives increase effort and expected quality. Next, we test the effects of incentives on the secondary variables and provide evidence for Hypothesis 3. First, Column (4) in of Panel A of Table 2.4 shows the effects of the effect of incentives on the forecaster's subjective

[12]This is analogous to the persuasion rate (DellaVigna and Gentzkow, 2010; Jun and Lee, 2018) which computes the fraction of persuaded individuals relative to the number of individuals who wouldn't take the target action if untreated (i.e. those who can potentially be persuaded).

beliefs about the share of respondents who will use one of the options from the list they have written. Consistently with the model, the average expected share increases by 3.774 percentage points ($p < 0.001$). Interestingly, the changes in the subjective expectations are aligned with the actual shift in shares: the point estimates in Columns (1) and (4) differ only by 1.1 percentage points which is statistically indistinguishable from 0 ($p = 0.247$). At the same time, the levels of subjective expectations are significantly biased: the subjective expectations are 18.8 percentage points higher than the actual shares ($p < 0.001$). This might have interesting implications for the optimal payment schemes for the forecasters: while the piece-rate incentives are likely to be perceived correctly, non-linear schemes might be less effective due to misperceptions. For example, if one promises a forecaster a large bonus payment for writing an option list that scores more than 80% share, the incentives are likely to be ineffective if the forecaster wrongly expects the list to hit the target share even without the additional effort.

Panel B further investigates the effects of the incentive treatment on the forecaster's behavior. Columns (1) through (4) present the evidence on the four effort measures: total time spent writing the options list, number of characters used to write the lists, number of options written in the list, and number of characters used per option, respectively. All four measures increase sizably and significantly with the incentive treatment ($p < 0.001$). As DellaVigna et al. (2022) point out, some dimensions of effort might be more elastic than others. We find that the total time spent increases the most in relative terms: treatment has increased the total time spent on the question by 44.9 seconds on average relative to the mean of 102.6 seconds in the control group (43.8% increase). The effect for the number of characters has a similar magnitude: a 53.8 increase over the 146.7 average (36.7%). Next, we find that the total increase in characters can be split approximately equally between the extensive and intensive margin: the number of options increases by 0.5 relative to the mean of 2.3 (21.9% increase) and the number of characters per option increases by 17.6 relative to the mean of 83.6 (21.1% increase).

Taking the estimates from Panels A and B together, it seems that the pass-through of effort into quality is rather low. If we take the "response discovery rate" as a reference, the elasticity of the quality to the effort would be about 0.153. This suggests that the main constraint on the quality of answers might be information and not incentives for writing the options. We will explore this possibility in the next two subsections.

2.4.3 Individual heterogeneity.

In this subsection, we describe the distribution of the quality of answers elicited from the forecasters. Since each forecast is evaluated on average 30 times, it is likely that the naively estimated distribution will be over-dispersed due to measurement error: forecasts with high average outcomes might end up there both due to high expected scores and by luck. The standard approach in economics and psychology literatures is to report cross-correlations between measurements (e.g. Engelmann et al. (2019)). However, this approach only provides information on the fraction of variation explained by the expected outcomes by subject relative to the total variation in the outcome but does not allow to make inferences about the distribution of the expected outcomes. To circumvent the problem, we propose a 3-fold cross-fitting procedure inspired by Abadie et al. (2018) and Chernozhukov et al. (2018). For each forecast, we split the evaluations into 3 folds of approximately 10 observations. We estimate the quintile of each outcome using 2 folds. Then we use the left-out outcome to estimate the quintile average outcome. Formally, we estimate the regressions:

$$\bar{Y}_{it}^{m} = \sum_{g=1}^{5} I\{\bar{Y}_{it}^{-m} \in G_g\}\gamma_g + \epsilon_{it} \tag{2.15}$$

where Y_{it}^{m} is average outcome for option list elicited from forecaster i at time t computed with the fold $m \in \{1, 2, 3\}$, $I\{\bar{Y}_{it}^{-m} \in G_g\}$ is the indicator variable that is equal to 1 if $\bar{Y}_{it}^{-m}$, the average outcome computed with all folds except m, belongs to the quintile G_g, $g \in \{1, 2, 3, 4, 5\}$, and ϵ_{it} is the error term.

The cross-fitting procedure allows us to break the mechanical relationship between the quintile and measurement error. The average outcomes are unbiased conditional on the quintile. However quintiles might get misclassified, hence the estimates of the average inter-quintile differences are conservative.

Table 2.5 documents a large dispersion of qualities of forecasts according to our measures. The average probability of choosing any option from the list ranges from 35.3% for the lowest 20% of forecasts to 80.6% for the top 20%. The difference of 45.2 is statistically significant ($p < 0.001$). The average semantic similarity of the chosen option to the open-ended answer ranges from 0.107 in the bottom 20% to 0.369 in top 20%. The difference of 0.261 is also statistically significant ($p < 0.001$).

The degree of heterogeniety of outcomes by forecaster further puts our estimates of incentive treatments into perspective. For example, 2.7 percentage point increase in the fraction of respondents picking one of the options appears small relative to the difference between the top and bottom quintile (45.2 percentage points). This suggests that selection of respondents might be a more important design feature than the incentive scheme. In the next subsection, we explore the characteristics of the forecasters that explain the heterogeniety.

Effect of forecaster characteristics. This subsection estimates the effects of forecaster characteristics. The characteristics include average similarity of forecaster's beliefs to the beliefs in the target group, precision of guesses about other answer lists, as well as index of political views, financial literacy, and gender. The effect for similarity of beliefs and precision of guesses directly test Hypothesis 4. The regression equation this exercise is:

$$Y_{it} = X_{it}\beta + Z_{itc}\gamma_c + \epsilon^c_{it} \tag{2.16}$$

where Y_{it} is the outcome, X_i is a vector of controls that includes topic and question order fixed effects, Z_{itc} is the c component of the characteristics vector Z_{it} and ϵ^c_{it} is the error term.

The estimand here should be thought of as the causal effect of changing the forecaster with characteristic bundle Z_{it}^1 to one with characteristic bundle Z_{it}^2. Since we don't aim at holding other characteristics constant, we include the regressors one by one. We do not include forecaster fixed effects for this reason.

Table 2.6 shows the results of the estimation. We standardize the characteristics for comparability. Panels A, B, and C show the effects for our three main outcomes: the probability that an option is chosen from the list, the semantic similarity of the chosen response to the open-ended response, and the semantic similarity conditional on one of the options being chosen, respectively. Column (1) shows that a one standard deviation increase in semantic similarity of the forecaster's and target group opinion sizably increases the fraction of respondents choosing one of the options by 7.0 pp ($p < 0.001$), the semantic similarity of the chosen option by 4.4 points ($p < 0.001$), and the semantic similarity conditional on choosing one of the options by 4.3 points ($p < 0.001$). This is consistent with Hypothesis 4: if the forecasters overestimate the popularity of their own views, the ones with more popular views should perform better. Column (2) shows that the effects for the accuracy of guesses for pre-existing options are also significant in line with the Hypothesis 4: a 1 standard deviation increase in the accuracy increases the probability of an option from the list being chosen, the semantic similarity, and the intensive margin of semantic similarity by 2.5 pp ($p < 0.001$), 1.5 points ($p < 0.001$), and 0.7 points ($p < 0.001$) respectively. The effects are smaller for the prediction accuracies, which is likely to be explained by higher measurement error: the prediction accuracy is a more direct measure of the precision of beliefs, but our measurement relies on other crowdsourced responses, which might be less complete than the open-ended answers. Column (3) shows the estimates for the financial literacy index. The effects are large: a 1 standard deviation increases three outcomes by 5.8 pp ($p < 0.001$), 2.7 points ($p < 0.001$), and 1.7 points ($p < 0.001$). Columns (4) and (5) show that the estimates for the political index and gender are also sizable and mostly significant. For example, a 1 standard deviation increase in the political views (i.e. having more Republican political views) decreases the probability of the respondent choosing one of the options by 1.8 pp ($p = 0.004$) and a 1 standard deviation increase in the male dummy variable decreases the probability by 1.7 pp ($p = 0.006$). Since our respondent sample skews Democratic and female, it is reasonable to expect that the result is driven by characteristics distances between forecasters and respondents.

Taken together, these estimates underscore the importance of the similarity of the forecasters to the target respondent group. Although the most predictive characteristics, own opinions on the target question, cannot be easily observed, the results also suggest that they might be partially controlled using demographic controls and other opinions that could be easily elicited.

2.5 Conclusions

We proposed a low-cost, scalable method of crowdsourcing popular answers for "closing" open-ended questions and two theoretically motivated measures of answer set quality. We validated the method with an experiment where we first crowdsourced, and then evaluated the answer options. To validate the quality measures, we developed a structural model of demand for answer options and estimated it using experimental data. Finally, we presented evidence that both monetary incentives and characteristics of the respondents significantly affect the quality of crowdsourced responses.

We hope that our work will result in a useful, practical methodology for the adaptation of questions to new setups. There are, however, two missing pieces that we intend to address in future work: optimal combination of answers into the final response set and benchmarking the set against the current alternatives. First, we need to understand how to combine the crowdsourced answers optimally. Our approach would be to use the estimated demand model to compile the set of answers that minimizes the counterfactual shares and expected semantic distances to open-ended answers. A brute force approach, however, could be computationally demanding: for example, picking a 5-option set from 100 possible options would require evaluating around 75 million combinations, a number growing at a factorial rate. We intend to propose a faster algorithm in future work. The second step is benchmarking the crowdsourced answers against a simple transfer of response options from established surveys, which is standard in the literature. For example, Katz and Krueger (2019) use the set of questions from the Current Population Survey (CPS) for a study of gig workers on Amazon Mechanical Turk.[13] We intend to test how well would our method perform in adapting the CPS questions on part-time work to the mTurk population.

We hope that our findings might also inspire more work in the new literature on the origination and evaluation of hypotheses. For example, unlike Otis (2022), we find that incentives affect the quality of crowd-sourced texts. The difference can be explained in two ways. First, it is possible that the intrinsic motivation of the respondents to provide vaccination nudges is due to prosocial motives. Second, there may be some differences in the task or elicitation method that make effort less of an issue. We view the resolution of this contradiction as a fascinating avenue for future work.

More broadly, we hope our methodology could be useful in bridging the theoretical and empirical literatures on communication and information design. We show how the novel text-as-data methods allow for structural estimation in experimental studies of communication with natural language. We hope further work will illuminate the constraints and costs that agents face when sending messages in real-world setups.

[13] In general, the mTurk worker population is often used in the literature to understand the economics of the new working arrangements, e.g. in Adams-Prassl and Berg (2017); Adams-Prassl (2021).

Table 2.1: Example of bundles

	Question: What do you think is the main issue with or shortcoming of the U.S. federal income tax system?			
	Open-ended: Companies paying little to no taxes.			
	Option text	Cos. sim.	Option text	Cos. sim.
	Set 1:		*Set 2:*	
1	**It is not sufficiently progressive**	**0.041**	Tax rate is too high.	0.438
2	It is politicized too much	0.036	Higher earners donâĂŹt pay enough in taxes.	0.459
3	–	–	**Big companies donâĂŹt pay enough in taxes.**	**0.645**
4	–	–	Too many tax loopholes.	0.460
5	–	–	Too many tax cheats.	0.462
99	Other / my preferred answer is not listed	–	Other / my preferred answer is not listed	–

Notes: The table shows the examples of two sets elicited from the forecasters evaluated by one of the evaluators. The line *Question* shows the question asked. *Open-ended* shows the open-ended answer that was received from the evaluator in response to the question. Columns *Option text* display the answers from the crowdsourced sets and columns *Cos. sim.* shows the cosine similarities of the answers to the open-ended answer.

Table 2.2: Logistic regression of choice on cosine similarity

Sample:	All	Charity	Durables	Inflation	Taxation
	(1)	(2)	(3)	(4)	(5)
Similarity ($\hat{\alpha}$)	3.143***	3.067***	3.551***	2.840***	3.471***
	(0.042)	(0.080)	(0.092)	(0.086)	(0.085)
Outside option ($\hat{b}$)	1.253***	1.102***	1.147***	1.416***	1.472***
	(0.017)	(0.034)	(0.032)	(0.035)	(0.038)
Reserve similarity	0.399***	0.359***	0.323***	0.499***	0.424***
($\hat{\beta} = \hat{b}/\hat{\alpha}$)	(0.011)	(0.018)	(0.032)	(0.017)	(0.011)
Observations	909 120	204 190	233 984	205 605	180 752

Notes: The parameter estimates for the multinomial logistic regression (equation 2.9). Columns (1) reports the estimates for the whole sample. Columns (2) through (5) report the estimates by the main question topic: charitable giving, durable goods, inflation narratives, and taxation respectively. Standard errors are clustered by forecaster. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 2.3: Accuracy with cosine similarity predictors

	Accuracy	Base acc.	p-value
	(1)	(2)	(3)
Panel A: full sample			
	0.477	0.390	¡ 0.001
Panel B: split by margin			
Extensive	0.660	0.627	¡ 0.001
Intensive	0.467	0.300	¡ 0.001
Panel C: split by topic			
Charity	0.457	0.367	¡ 0.001
Durables	0.463	0.356	¡ 0.001
Inflation	0.487	0.456	¡ 0.001
Taxation	0.503	0.387	¡ 0.001

Notes: The table shows the out-of-sample accuracies of the multinomial logistic regression. Column (1) reports the estimates of accuracies with cosine similarity as the sole regressor (equation 2.9). Column (2) shows the accuracies of a model omitting the similarity regressors. Column (3) reports the p-value from a t-test for equality of accuracies. Panel A shows the results for the full sample. For Panel B, line *Extensive* reports results for predicting the value of indicator variable that takes the value of 1 if any response will be chosen from the set, line *Intensive* shows the results conditional on any response being chosen from the set. Panel C reports the estimates by main question topic: charitable giving, durable goods, inflation narratives, and taxation respectively. Standard errors for the test are three-way clustered by evaluation, forecaster, and evaluator.

Table 2.4: Effect of incentives

	(1)	(2)	(3)	(4)
Panel A: Quality outcomes				
Dependent variable:	Option chosen	Similarity	Similarity (int. margin)	Self-predicted
Incentivized	2.655***	1.651***	1.099**	3.774***
	(0.789)	(0.445)	(0.439)	(0.807)
Observations	1576	1576	1563	1576
R^2	0.731	0.697	0.740	0.774
Control mean	60.29	23.40	38.01	79.09
Outcome SD	21.25	11.30	11.96	23.76
Panel B: Effort outcomes				
Dependent variable:	Time (sec)	# characters	# options	Char. per option
Incentivized	44.911***	53.846***	0.509***	17.649***
	(4.849)	(6.358)	(0.059)	(3.220)
Observations	1576	1576	1576	1563
R^2	0.748	0.806	0.853	0.825
Control mean	102.57	146.65	2.32	83.64
Outcome SD	136.33	201.53	2.14	107.93
Forecaster FE	X	X	X	X
Forc. question pos. FE	X	X	X	X
Topic FE	X	X	X	X

Notes: Each column of each panel corresponds to a separate OLS regression. Panel A shows the estimates for quality outcomes: the outcome variables are the fraction of respondents who chose any option from the list for Column (1), the average cosine similarity of open-ended and chosen responses for Column (2), the average cosine similarity of open-ended and chosen responses conditional on choosing any response for Column (3), and the forecaster's prediction of the fraction of respondents who chose any option from the list for Column (4). All outcomes in Panel A are scaled by 100. Panel B shows estimates for effort outcomes: total time spent on the task in seconds for Column (1), the total number of characters used for the answer list for Column (2), the number of options in the answer list for Column (3), and the number of characters used in the answer list divided by the number of options used in the list for Column (4). The explanatory variable is the indicator that takes the value of 1 if the task was incentivized. All regressions control for forecater, question order, and question topic fixed effects. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 2.5: Quintiles of answer quality

Quintile:	1	2	3	4	5	5 - 1
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: probability of an option chosen from the list						
Quintile mean	0.353	0.558	0.653	0.739	0.806	0.452
	(0.008)	(0.007)	(0.006)	(0.006)	(0.005)	(0.010)
Panel B: cosine similarity of the chosen option to open-ended answer						
Quintile mean	0.107	0.196	0.242	0.304	0.369	0.261
	(0.003)	(0.003)	(0.003)	(0.003)	(0.004)	(0.005)

Notes: The table presents quintiles of the distribution of outcome variables by forecast. Columns (1) through (5) show average outcomes in quintiles 1 through 5. Column (6) shows the difference between quintiles 5 and 1. Panel A shows the distribution of the probability. The estimates are 3-fold cross-fitted. Standard errors are clustered by forecast. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 2.6: Effect of forecaster characteristics

Characteristic:	Mean similarity	Pred. accuracy	Fin. literacy	Politics	Male
	(1)	(2)	(3)	(4)	(5)
Panel A: Any option chosen					
Coefficient	7.034***	2.473***	5.848***	-1.764***	-1.712***
	(0.668)	(0.578)	(0.659)	(0.606)	(0.627)
Observations	1576	1428	1576	1576	1576
Within R^2	0.109	0.015	0.080	0.007	0.007
Outcome mean	61.54	62.15	61.54	61.54	61.54
Outcome SD	21.25	20.65	21.25	21.25	21.25
Panel B: Semantic similarity					
Coefficient	4.395***	1.454***	2.686***	-0.678**	-0.989***
	(0.288)	(0.293)	(0.325)	(0.304)	(0.322)
Observations	1576	1428	1576	1576	1576
Within R^2	0.155	0.018	0.061	0.004	0.008
Outcome mean	24.10	24.47	24.10	24.10	24.10
Outcome SD	11.30	11.07	11.30	11.30	11.30
Panel C: Semantic similarity (int. margin)					
Coefficient	4.320***	0.705**	1.690***	0.054	-1.113***
	(0.412)	(0.324)	(0.398)	(0.321)	(0.355)
Observations	1563	1419	1563	1563	1563
Within R^2	0.124	0.004	0.022	0.000	0.010
Outcome mean	38.44	38.84	38.44	38.44	38.44
Outcome SD	11.96	11.70	11.96	11.96	11.96
Forc. question pos. FE	X	X	X	X	X
Topic FE	X	X	X	X	X

* p < 0.1, ** p < 0.05, *** p < 0.01

Notes: Each column of each panel corresponds to a separate OLS regression. The outcome variables are the fraction of respondents who chose any option from the list for Panel A, the average cosine similarity of open-ended and chosen responses in Panel B, and the average cosine similarity of open-ended and chosen responses conditional on choosing any response in Panel C. Each outcome variable is scaled by 100. The explanatory variables are the mean cosine similarity between the forecaster's answer and evaluators' answers in the target group in Column (1), the negative logarithm of mean squared error of the forecaster prediction relative to the actual response shares in Column (2), number of correct answers to "Big three" financial literacy questions in Column (3), the index of political views (higher means more Republican as opposed to Democratic) in Column (4), and an indicator variable taking value of 1 if the forecaster identified as male. All explanatory variables are standardized for comparability. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

2.A Main task questions

Taxation (Stantcheva, 2021). What do you think is the main issue with or shortcoming of the U.S. federal income tax system?

Inflation narratives (Andre et al., 2021). In previous years, the U.S. inflation rate has mostly varied between 1.5% and 2.5%. That means that a bundle of goods and services that costs $1,000 in one year will cost between $1,015 and $1,025 in the next year.

Recently, however, the inflation rate has increased to [current rate]. That means that a bundle of goods and services that costs $1,000 in one year will cost $[estimated cost] in the next year.

What is the main factor that you think caused the increase in inflation?

Durables (based on Survey of Consumer Attitudes and Behavior, 2022) Now we want to ask you about the big things people buy for their homes—such as furniture, a refrigerator, stove, television, and things like that. Suppose you are deciding whether now is a good time or a bad time to buy major household items.

What is the main factor or consideration affecting your decision?

Charitable giving (based on Gallup, 2016) What is the most important factor or reason you consider when deciding whether to donate to a charitable organization?

2.B Stage 1 Main Task

Every Stage 1 participant saw Screens 1–4 twice with no differences unless noted otherwise.

Screen 1

<table>
<tr><td>Iter.</td><td>Unincentivized</td><td>Incentivized</td></tr>
<tr><td>1</td><td>In this section, we will ask you to write potential answers to the following question:</td><td>In this section, you can earn up to $[X[14]] in bonuses by writing potential answers to the following question:</td></tr>
<tr><td rowspan="2">2</td><td>In this section, you can earn no bonuses for your answers.</td><td>In this section, you can earn up to $[X] in bonuses for your answers.</td></tr>
<tr><td colspan="2">As before, we will ask you to write potential answers to the following question:</td></tr>
<tr><td>1&2</td><td colspan="2">[Randomly selected question appears here]</td></tr>
</table>

Screen 2

We will use your answers to poll a group of [K] U.S.-based mTurkers from the CloudResearch panel.

If a respondent doesn't like any of the answers you provided, they can select the option "other." The more popular your answers are, the fewer people choose "other."

Your task is to write a set of answers that you think will be most popular with this group.

Note: we will have to remove answers that don't meet the platform standards (for example, contain obscene language).

Screen 3

Please do your best and write the answers you think are most likely. You can write as many answers as you see fit. We will compute the number of people who choose one of your answers and not "other."

Unincentivized	Incentivized
Your payment will not depend on the number of such people.	You will get a $[Y] bonus payment for each such person (up to $[Z]).

Screen 4

Please do your best and write the answers you think are most likely. You can write as many answers as you see fit. Click "Add field" to write additional answers. [Question restated] [text field] [ADD FIELD button]

2.C Additional Stage 1 tasks

Section 2: Own opinions

Now we want to ask your own opinion on the following question: [*relevant question from Appendix 2.A*]. Please write one full sentence that summarizes your opinion below.

Section 3: Peer prediction

In this section, we will ask you to predict the popularity of answers to the questions you have just considered. We will ask the questions to [W] U.S.-based mTurkers from the CloudResearch® panel and give them different lists of possible answers. Your task is to predict the fractions of respondents who will choose each answer from each list. The program will randomly select one of your predictions for payment.We will pay you a $[Z] bonus if your prediction is accurate, with an error of at most +/- 3 percentage points.

Please predict what fraction of respondents will choose each option. The program will randomly select one of your predictions for payment. We will pay you a $[Z] bonus if your prediction is accurate, with an error of at most +/- 3 percentage points. The sum of the fractions must be 100%.

Section 3: Self-evaluation

We will also ask the questions to [V] U.S.-based mTurkers from the CloudResearch® panel and give them the lists of possible answers that you have written in the previous sections. Please predict what fraction of respondents will choose each option.We cannot offer bonuses for the answers in this section, but we ask you to try your best nevertheless.

2.D Additional Stage 1 questions

Section 2

Understanding check. In one of the previous questions, we said the following: "In this section, you can earn up to $[X] in bonuses." Suppose $[0.6 \times K]$ out of $[K]$ respondents selected one of your answers, and the remaining $[0.4 \times K]$ chose "other." How much (in cents) can you expect to be paid in bonuses?

Section 5

1. In one of the previous questions, we said the following:"Your task is to write a set of answers that you think will be most popular with this group."What do you think was the hardest thing about this task?

2. Do you think you understood how the bonuses are awarded?

   - I didn't understand at all how to receive the bonus payment
   - I somewhat understood how to receive the bonus payment
   - I understood completely how to receive the bonus payment

2.E Stage 2 Main Task

Screen 1

The set of screens below is repeated (w/o alterations unless noted otherwise) twice.

Now we want to ask you the following question:
[Question goes here]
Please write one full sentence.
[text box here]

Screen 2

Next [As before], you will see 12 sets of possible answers to the question you just considered.
Please choose the answer that matches your opinion.
If no answer describes your opinion well, please select "other."

Screens 3 – 14

Please choose the answer that matches your opinion.
If no answer describes your opinion well, please select "other."
[question reiterated]
[list of options from the forecaster]
Other / my preferred answer is not listed

2.F Pre-survey on perceptions and economic literacy

2.F.1 Common questions

1. On which device are you taking this survey?On which device are you taking this survey?

    - Mobile phone
    - Tablet
    - Computer (desktop or laptop)

    First, we want to ask you a few general questions about the U.S. economy. Although you might find some of the questions difficult, we ask you to respond to the best of your ability.

2. Suppose you had $100 in a savings account and the interest rate was 2% per year. After 5 years, how much do you think you would have in the account if you left the money to grow?

    - More than $102
    - Exactly $102
    - Less than $102
    - Do not know

3. Imagine that the interest rate on your savings account was 1% per year and inflation was 2% per year. After one year, how much would you be able to buy with the money in this account?

    - More than today
    - Exactly the same
    - Less than today

- Do not know

4. Is this statement true or false? “Buying a single company’s stock usually provides a safer return than a mutual fund that invests in the stocks of multiple companies.”

   - True
   - False
   - Do not know

5. *(Attention check)* Which of these words is the most associated with the word “paint”?

   - draw
   - run
   - sports
   - loud

2.F.2 Topical questions: shown depending on the topic of the main questions.

1. *(inflation)* Now we want to ask you about the inflation rate, which is a measure of how prices in the US change in general. During the last 12 months, do you think that prices in general have increased or decreased?

   - Prices decreased in general
   - Prices increased in general

   By about what percent have prices [decreased/increased] on average, during the past 12 months?

   List from “less than 1%” to “15% or more”

2. *(taxation)* What share of their total income do people in the top federal personal income tax bracket pay in taxes?

scroller from 0 to 100%

3. *(durables)* Now we want to ask you about the big things people buy for their homes—such as furniture, a refrigerator, stove, television, and things like that. Generally speaking, do you think now is a good or a bad time for people to buy major household items?

    - Yes, it's a good time
    - It's neither a good nor a bad time
    - No, it's a bad time

4. *(charity)* What fraction of people in the U.S. do you think have donated to any charitable organization over the last year?

    scroller from 0 to 100%

2.G Post-survey

To get a general picture of the people answering this survey, we need to know a few things about your background.

1. Please indicate your gender:
    - Female
    - Male
    - Other
    - Prefer not to answer
2. *(Attention check)* Recent research on decision-making shows that choices are affected by the context in which they are made. Differences in how people feel, in their previous knowledge and experience, and in their environment can influence the choices they make. To help us understand how people make decisions, we are interested in information about you, specifically whether you actually take the time to read the instructions; if you don't, some results may fail to tell us very much about decision-making in the real world. To help us confirm that you have read these instructions, please select the "none of the above" option below. Thank you very much.
    - Interested
    - Distressed
    - Excited
    - Upset
    - Strong
    - Scared
    - Hostile

- Enthusiastic
- Proud
- Irritable
- Alert
- Inspired
- Nervous
- Determined
- Attentive
- Jittery
- Active
- None of the above

3. How old are you?

   dropdown menu, ages between 18 and "above 80"

4. What was your TOTAL household income, before taxes, last year?

   - $0 – $9,999
   - $10,000 – $14,999
   - $15,000 – $19,999
   - $20,000 – $29,999
   - $30,000 – $39,999
   - $40,000 – $49,999
   - $50,000 – $69,999
   - $70,000 – $89,999

- $90,000 – $109,999
- $110,000 – $149,999
- $150,000 – $199,999
- $200,000 or higher

5. Approximately, how much money have you donated to charitable organizations over the last 12 months?
   - I did not donate to charity over the last 12 months
   - 0 to 99 dollars
   - 100 to 499 dollars
   - 500 to 999 dollars
   - 1000 dollars or more
6. What is your political affiliation if any?
   - Republican
   - Democrat
   - Independent
   - Other
   - Non-Affiliated
7. Did you vote in the last presidential election?
   - Yes
   - No
8. *(If "Yes")* In the last presidential election, you supported:
   - Joseph Biden

- Donald Trump
- Jo Jorgensen
- Howie Hawkins
- Other

9. *(If "No")* Even if you did NOT vote, please indicate the candidate that you were most likely to have voted for or who represents your views more closely:

   - Joseph Biden
   - Donald Trump
   - Jo Jorgensen
   - Howie Hawkins
   - Other

10. How would you rate your understanding of the questions included in this survey?

    - I understood all the questions
    - There were a few questions I did not understand
    - There were several questions I did not understand

11. Could you please give examples of the questions that you had trouble with?

    textbox

12. *(Stage 1 only)* There was a task where we have asked you to write popular answers to a question.

    What do you think was the hardest thing about this task?

    textbox

13. Do you feel that this survey was biased?

    - Yes, left-wing bias
    - Yes, right-wing bias
    - No, it did not feel biased

CHAPTER 3

Belief Elicitation and Large Language Models: a Review

3.1 Introduction

People typically express their beliefs through natural language. However, most of the belief elicitation procedures in economics are either numeric or give the respondent some restrictive set of answer options. This might limit the scope of messages about beliefs that the respondents could potentially transmit.

The preference for structured elicitation methods is likely influenced by the intent to incorporate beliefs into empirical models (Manski, 2004). Since the models can be fitted only if belief data is encoded numerically, most studies used structured methods from the beginning. Recently, however, two trends emerged in the literature. First, recent studies in economics show that verbal elicitation might be complementary to structured ones. For example Andre et al. (2021), use an open-ended question to ask for the perceived causes of the 2021-22 increase in inflation in the U.S. and then validate this data by testing if the perceived causes are indeed important when subjects update beliefs about the variables elicited with the usual structured methods.

The second trend is the rapid progress of natural language processing (NLP) techniques. This technological progress lowers the cost of applying the structure post-elicitation, rather than prior to it. Continuing our example, Andre et al. (2021) use manual coding to encode the causal relationships stated in the texts. Contemporary NLP methods might be able to find the relationships without human assistance (although benchmarking for this particular task is an empirical question). Even if the task is currently not feasible, it is likely that the NLP methods will catch up with human coders within a few years, given the pace of progress in the field.

This review has four goals: document the current state of belief elicitation in economics, introduce the audience to the recent advances in NLP, discuss the current applications of NLP in belief elicitation, and draft a research program extending the applications.

Our description of belief elicitation methods is driven by the applicability of NLP methods. For a more general review of empirical exercises, one might consult the respective section of Haaland et al. (2023). Most reviews of belief elicitation methodologies are concerned with the incentive schemes for structured elicitations, e.g. Schotter and Trevino (2014); Charness et al. (2021). Our review to a large extent omits the discussion of incentives in order to focus on the natural language processing aspect of the procedures. Instead, we first discuss the goals of belief elicitation and then propose a taxonomy of belief elicitation procedures that easily maps into the recommended LLM-assisted procedures.

The main change in NLP is transfer learning and the advent of the pre-trained large language models (LLM). Historically, automated text analysis has mostly appeared in observational studies: the two recent reviews (Gentzkow et al., 2019; Ash and Hansen, 2023) of the methodology mention virtually no experimental or survey studies. However, since pre-training allows to obtain of meaningful results even with relatively few observations in the final dataset, smaller-scale studies start using text-based variables. We intend to describe the current possibilities of the approach.

Although the large language models are a very recent development, there are some studies that use them for belief elicitation or analysis of communication more broadly. We highlight these studies and propose extensions when possible.

The rest of the review proceeds as follows. Section 2 reviews the state of belief elicitation in economics. Section 3 introduces the main architectures of pre-trained large language models and main fine-tuning techniques. Section 4 discusses the applications of the LLMs for belief elicitation. Section 5 gives a simple empirical example using a generative LLM for belief elicitation. Section 6 concludes.

3.2 Belief elicitation in economics

3.2.1 Goals of belief elicitation

There are two main goals of belief elicitation in economics: collecting data on beliefs to test theory and estimate models and using elicited beliefs as a signal about the true state of the world that can later be used either in research or policy design. The first goal is rather established in the literature, while the second has not been rigorously studied in economics until recently.

Modeling behavior The empirical work in economics was mostly focused on revealed preference analysis and the data on perceptions was seldom collected before the 1990s. However, since then the empirical literature on beliefs has flourished. Due to the volume of the literature using beliefs data for behavior modeling, it seems reasonable to refer to discussing only other literature reviews. Manski (2004) discusses the early work on measuring expectations, including the expectations in games, macroeconomic expectations, and expected returns to education. Brunnermeier et al. (2021) discuss the flourishing literature on beliefs in asset pricing. Weber et al. (2022) review the literature on measurement and effects of inflation expectations. Bursztyn and Yang (2022) provide a meta-analysis and show stylized facts on misperceptions about others. Haaland et al. (2023) review the literature and give practical recommendations on information provision experiments.

Signal about the world Some recent studies have been using elicited beliefs as signals about the objective state of the world, as opposed to subjective expectations to be used in models. In this case, one aggregates the beliefs and uses them for policy or research decision-making. An early example of such an approach is Wolfers and Rothchild (2011). They show that asking for expectations about the election results is more informative than asking about individual voting intentions. The intuition is that voters aggregate more than their own voting intentions when forming the expectations, which results in a more informative signal. Hussam et al. (2022) use an incentive-compatible elicitation scheme to collect beliefs about the entrepreneurial ability of peers, which as they show could later be used in financing decisions. Buinskaya and Galashin (2023) crowdsource potential answers to opinion survey questions. Otis (2022) crowdsources text messages to promote Covid vaccination in Kenya. A large body of literature uses expert and lay people forecasts to understand and select best treatments (Otis, 2021) and guide scientific exploration (DellaVigna et al., 2019). Bergman et al. (2019) and Hampole et al. (2021) use an elaborate belief elicitation procedure to crowdsource hypotheses about mechanisms behind "moves to opportunity" and female MBA career decisions.

3.2.2 Theoretical framework

Social science has multiple ways of eliciting beliefs, including direct questions, incentivized procedures, and interviews. Since different fields have different approaches, there is no single criterion. Our approach to belief elicitation is pragmatic: we aim to elicit beliefs that would explain actions. Formally, we hypothesize that the data-generating process has the form

$$Y_i = f(\delta_i, \varepsilon_i) \tag{3.1}$$

where $\delta_i \in \Delta\Theta$[1] is an element of the belief space about the state $\theta_i \in \Theta$, Θ is the state

[1] $\Delta\Theta$ denotes the set of probability measures on Θ.

space. We assume that it has a metric: one can say that some states are more similar to each other. Y_i is the outcome of interest, which can be either an action or a treatment effect (change in the actions). We call a function $g(\delta_i, \nu_i)$ a belief elicitation procedure that provides the researcher with a feasible belief measure $\hat{\delta}_i \in \Delta\mathbb{R}^n$. We assume that the measured beliefs have to support in the real numbers, as the goal of the exercise is to estimate a model of behavior. The goal of the elicitation procedure is to provide a feasible measure $\hat{\delta}_i$ that would maximize the fit of the estimated model, $\hat{f}(\hat{\delta}_i, \varepsilon_i)$. Since the belief space $\Delta\Theta$ might have an arbitrarily high dimension, any reasonable elicitation procedure reduces dimensionality. Even when one assumes that the state space is a scalar, e.g. if one elicits beliefs about the inflation rate next year, the set of beliefs $\Delta\Theta$ is infinite-dimensional unless we impose some additional structure.

3.2.3 Ex-ante vs ex-post structure

As large language models are increasingly good at detecting structure, one can commit to more flexible data analysis procedures. For the purpose of this review, we will classify the elicitation procedures based on the strength of the ex-ante structure they impose, from most to least.

Structured elicitation Most of the belief elicitation procedures in the literature would belong to this class: for example, Manski (2004) discusses probabilistic expectations, Schotter and Trevino (2014) discuss incentivized elicitation procedures for various summary statistics of belief distributions Hussam et al. (2022) organize the entrepreneurs into pairs and ask their peers to pick the more able one.

Unstructured answers In this class, the question is fixed, but the answers can be communicated with natural language. For example, Bursztyn et al. (2022) elicit beliefs about a vignette character with an open-ended question. Stantcheva (2021) and Ferrario and

Stantcheva (2022) use open-ended questions to obtain topics of first-order concerns about tax policy. Andre et al. (2021) use an open-ended question to elicit the subjective causal models of the recent inflation increase. Buinskaya and Galashin (2023) ask subjects to write survey answers that are likely to be used by other respondents.

Unstructured questions and answers In the least structured class, both the answers and questions are not fixed but are rather asked adaptively depending on the conversation. The recent examples include Bergman et al. (2019) and Hampole et al. (2021). A classical example is Lintner (1956).

3.3 Main architectures of large language models

Large language models and transfer learning has dramatically changed the capabilities of natural language processing (NLP) within the last 5 years. Instead of starting the training process from scratch, transfer learning allows us to benefit from the knowledge learned by models on massive datasets. In NLP, transfer learning has become a common approach due to the scarcity of labeled data for specific tasks and the computational resources required to train models from scratch.

The process of transfer learning in NLP typically involves two steps: pre-training a general language model (also called foundation models) and fine-tuning the model to a particular task. Pre-training is usually performed in a self-supervised manner: the model learns to perform some general language understanding tasks, such as predicting the next word in a sentence or filling in masked words, which helps it capture a broad understanding of language. After pre-training, the model is fine-tuned, which means further training on a task-specific dataset with labeled examples. This task could be sentiment analysis, text classification, named entity recognition, question answering, or any other NLP task. The model is fine-tuned by adjusting its parameters to learn task-specific patterns and features from the labeled data. Sometimes, fine-tuning can be done sequentially as one proceeds to more and more specific tasks. For example, Reimers and Gurevych (2019) fine-tune BERT model (Devlin et al., 2019) for semantic similarity tasks, which can later be further tuned for each particular task at hand.

In this section, we review the two main foundational model architectures. Next, we describe the fine-tuning techniques that are particularly relevant to eliciting beliefs.

3.3.1 Main foundation models

There are two main tasks that foundational models are trained to perform: encoding and decoding. Encoding in this context means representing natural language as embeddings - a series of vectors that capture the meaning of the word or subword tokens in the text. Decoding means generating text that allows it to go from the embedding space to human-readable texts. Sometimes people would also separate (change word) encoder-decoder or sequence-to-sequence models, which are a combination of the two: they take text as an input and produce new text as an output. Each of the two tasks as of now has one dominating architecture type, which we review next. A common feature of the two architectures is that they are both Transformers, which is a class of models that uses self-attention. The self-attention is a mechanism that predicts the degree of relationships of words in the text. For example, if one wants to make sense of the words "student" and "retriever" in a "An MBA student had a golden retriever", one should pay attention to different words to understand the context: "MBA" is likely to be more informative for the first, while "golden" is likely to be more informative for the second. The self-attention mechanism is a block within the neural network that predicts the intensity of such relationships.

Encoders The purpose of encoder models is to represent texts as sequences of real vectors that can be used in computations. As of now, most applications would use some version of the Bidirectional Encoder Representations from Transformers (BERT) model (Devlin et al., 2019). The distinctive feature of its architecture is that it is bidirectional, which means that when it represents words, it takes into account both words to the left and to the right of the target word. This is important when the main task is the correct representation of the semantics of the words for downstream tasks such as regression or classification. This is achieved by using masked word prediction as the main task: the model predicts a fraction of words in the sentence using other words and their positions. For example,

Decoders The purpose of decoder models is to generate human-readable texts. This makes the autoregressive (as opposed to bidirectional) architecture more effective: the model starts with a prompt provided by a user and predicts the next words recursively. One can think of it as a time series regression with word sequences as data. The attention mechanism can be thought of as a flexible way of combining the lagged values to obtain a prediction. As of now, the decoding task is dominated by GPT-family models and analogs.

3.3.2 Fine-tuning

The main advantage of the pre-trained foundational models is general understanding of language. This allows them to be fine-tuned for concrete tasks using relatively small datasets and perform exceptionally well, even in domains with limited labeled data. In this subsection, we review the main fine-tuning techniques used for encoder and decoder models. In some techniques the models are fine-tuned separately, while in others encoders and decoders might be combined, for example, one might make the encoder model estimate the quality of the decoder model output.

Encoders During pre-training, encoder models learn to represent texts as sequences of embeddings, long real vectors that capture the semantic and contextual information about the words. One can think of the vectors as bundles of characteristics. Then, one can use the vectors to perform classification or regression tasks or to aggregate the vectors into a representation of larger text units, such as sentences and documents. The level of classification/regression tasks can vary: for example, sentiment analysis might be performed at the level of the document (such as a tweet), while name-entity recognition classifies words within sentences. In order to fine-tune the encoder model, one has to adjust the architecture of the model: instead of having the last layers predicting the masked tokens, the last layers should be changed to the ones compatible with the target task. For example, if one is interested in classifying which texts contain anti-immigrant sentiment, one can use the last layers of the average word embeddings in the text, collapse the embedding vectors into one number with learned weights and apply the sigmoid function, i.e. run a simple logistic regression with mean word embeddings as regressors.

Decoders After the decoder model has learned the general structure of language during the pre-training, it can be fine-tuned to generate sequences that optimize some particular cost function. The main frameworks relevant to our review are imitation and reinforcement learning from human feedback (RLHF).

Imitation: The basic way of fine-tuning decoder models is to use a particular set of texts to predict the next word in a sequence, just like in the pre-training. This is likely to be a reasonable first step when adapting the model to a new setup or task.

Reinforcement learning with human feedback (RLHF): The RLHF framework usually adds two elements to the decoder network: reward model and human feedback. The reward model in this case means an encoder (classifier or regression) model that is fine-tuned to represent the preferences of the human subjects. The steps for the RLHF framework are:

1. Collect data on human feedback to the task, often performed by other humans;
2. Teach the reward model based on the feedback data;
3. Teach the decoder to generate the sequences that maximize the reward according to the reward model;

4. Evaluate the sequences with human subjects for more actual feedback data.

Steps 2 to 4 are iterated until convergence. It might be useful to think of the exercise as profit maximization with respect to the word sequence. First, we collect data on demand (1). Then we estimate the demand model (2). Next, we optimize the characteristics of the product (text) to maximize the profit given the estimated demand model (3). Finally, we ship the word sequences to the market, observe the actual demand, append it to the dataset, and repeat the steps.

3.3.3 Very large decoder models

One of the recent developments in the field is the advent of very large language models that are able to perform tasks such as classification simply from a verbal prompt. For example Zhong et al. (2023) find that the ChatGPT might have results comparable to standard BERT models on the usual language benchmark tasks. This has given rise to prompt-tuning: it might be more efficient to optimize the output of the very large decoder models by tuning the input prompt, as opposed to the model itself. As of now, this is an area of active research.

3.4 Tasks for large language models

In this section we review the tasks that are likely to be or already are a fruitful application of the large language models in belief elicitation. We categorize the applications in four groups: representing beliefs, interacting with subjects, hypothesis generation, and representing subjects.

3.4.1 Representing beliefs

The most straightforward application of the language models to belief elicitation is the extraction of structure from texts. This can be done in 3 ways. First, the texts might be projected on some variable that is related to the dimension of interest. Second, the texts can be embedded, i.e. represented as vectors in a way that preserves semantic distance. Third, we could use within-sentence prediction to identify structures within beliefs, such as causal relationships.

Projections The first way of using encoder models is to extract how the elicited textual beliefs are related to some treatment or downstream action. Bursztyn et al. (2022) randomize if a vignette character has joined an anti-immigrant organization before or after losing his job during the crisis. Joining after the crisis gives an economic excuse for xenophobic actions. They use a BERT-based classifier to estimate if their information treatment affects the perception of the character elicited through text. This allows them to estimate the effects of the treatment on beliefs without relying on pre-defined scales or dictionaries.

To give an example of a projection of a downstream action, suppose we run a survey of career expectations of MBA students before they start interacting at the program in which we allow for open-ended responses, and wait until the students graduate to observe their career choices. If we use the survey text data to predict characteristics of the chosen job (such as firm sector or position), we obtain a measure of the student's ex-ante career aspirations (this could be also done with a fixed survey of questions, but that would require to commit ex-ante to the characteristics of the job that will be analyzed, while the text allows for more flexible preference discovery). After that one can take a similar, but unrelated group of students (such as the next cohort) and use the predicted job characteristics probabilities to estimate to what extent the career aspirations are contagious. Assuming that there is some randomization in the assignment of the peer group (say, through random study groups assignment) the hypothesis would be that students will end up in sectors to which their peer group aspires.

Another use of projection might be understanding mechanisms behind treatment effects. For example, Danz et al. (2022) use a verbal survey after the experiment to understand the reasons behind the center bias of beliefs elicited with the binarized scoring rule (Hossain and Okui, 2013). This results in texts summarizing the subject's beliefs about the mechanism and the optimal strategies. The paper uses the texts as anecdotal evidence. A more formal and potentially informative exercise would be to predict the treatment effects with the texts.

Embeddings Embeddings are representations of text data that preserve semantic distance. Depending on the application, it can be done at the token, word, sentence, or document level. Usually, one starts with the token-level embeddings and then aggregates it up to the required level. This allows for the representation of texts as bundles of characteristics that lends itself to estimation and even structural modeling. For example, Buinskaya and Galashin (2023) use sentence embeddings to measure the qualities of survey answer sets. They use the similarity of embeddings to estimate the semantic similarity of the open-ended responses and the options chosen from the answer sets. This allows them to estimate the how much information is lost relative to the open-ended question if one uses the given answer set. They also validate the measure by estimating the demand for answer options in a survey, assuming that respondents derive expressive utility that increases with the similarity of the answer option to their true beliefs. One can potentially use the demand model to predict the optimal combination of answers from different lists that minimizes information loss.

Another application would be showing belief updating in a flexible way. Continuing with our MBA career choices example, if one has run the aspirations survey before and after the MBA program, one can test if peers shift aspirations directly, without relying on choice data. This is a complementary approach to the one described before: looking at actual career paths only might mute the peer effects as it is possible that the students won't be able to shift their careers due to some constraints, even though their aspirations have shifted. On the other hand, it is possible that career choices shift through some other mechanism, such as the availability of people who are aspiring to work in the sector and can help with employment. Looking at belief updating equations one can separate the two mechanisms just like it is done in usual learning regressions, e. g. Bottan and Perez-Truglia (2020a):

$$\hat{\theta}_i^{post} - \hat{\theta}_i^{prior} = \alpha_0 + \alpha_1 \cdot \left(\hat{\theta}_i^{signal} - \hat{\theta}_i^{prior}\right) + \epsilon_i \tag{3.2}$$

where $\hat{\theta}_i^{post}$, $\hat{\theta}_i^{prior}$, are the posterior and prior beliefs respectively and $\hat{\theta}_i^{signal}$ is the randomized signal, in our case, a belief of the peer. As in this case, beliefs are high-dimensional, one can replace the gaps between the expectations with the distances between embeddings:

$$\|\hat{\theta}_i^{post} - \hat{\theta}_i^{prior}\| = \alpha_0 + \alpha_1 \cdot \|\hat{\theta}_i^{signal} - \hat{\theta}_i^{prior}\| + \epsilon_i \tag{3.3}$$

Understanding belief structure Finally, one might use encoder networks to extract some finer information from the belief responses. For example, Andre et al. (2021) represent beliefs about the causal relationships between macroeconomic variables as Directed Acyclical Graphs (Pearl, 2009) . This structure can be thought of as an adjacency matrix where the entries $a_{i,j}$ are positive if the respondent believes that the variable i is causally affected by variable j. The authors use human coders to represent these beliefs as a graph. At the same time, the recent work by Ash et al. (2021) allows for finding the causal statements in the data. This involves identifying "*agent* → *verb* → *patient*" relationships and then grouping the agents, verbs, and patients into nodes.[2]

3.4.2 Interacting with subjects

Immediate feedback Attenberg et al. (2015) suggest that immediate feedback from machine learning models might improve effort in some tasks. In their study, they ask Amazon Mechanical Turk workers to give examples of hate speech that would "beat the machine":

[2] In principle, one doesn't have to have discrete groups of narratives and represent narratives as embeddings. For example, the statement "Biden caused inflation" is more similar to "the Democrats caused inflation" than "Covid caused inflation". Depending on the application one might want to have a continuous measure of similarity or aggregate the groups differently.

confidently misclassified as not containing hate speech by their machine learning model. This design might be used for belief elicitation. For example, Buinskaya and Galashin (2023) crowdsource lists of response options and estimate a model predicting popularity of the lists. A possible improvement of their methodology is finding a way to reduce the repetitiveness of the suggested answers: the crowdsourced responses are likely to repeat unless there is some encouragement for the later respondents to provide different answers. One of the ways of increasing the diversity of responses is incentivizing the respondents to give the answers that they think will outperform relative to the current version of the model. This could also help with incentives: their experimental subjects are overconfident, which might make the incentive payments less effective. In this case, using the model to correct misperceptions should improve performance.

Generating answers Buinskaya and Galashin (2023) crowdsource response sets from human subjects. One can also think of an exercise using a generative model to come up with the answers. However, the range of implementation strategies for this task is large: one can try fine-tuning or prompt-tuning an open-source decoder LLM (e.g. Touvron et al., 2023) or use the crowdsourced responses for prompts to large decoder models, for example ChatGPT (Ouyang et al., 2022) or GPT-3 (Brown et al., 2020). Some of the strategies might incorporate the estimated "demand for options" as a reward model for the RLHF framework. It is ultimately an empirical question of which strategy will be feasible and effective for the task.

Conversations Some belief elicitation procedures rely not only on unstructured answers but on unstructured questions. This approach includes qualitative interviews, cognitive interviews, and focus groups. In this case, the respondent is asked questions adaptively depending on her answers. This might allow for more precise elicitation, as it, for example, gives the researcher an opportunity to clarify the meaning a respondent assigns to words in her response or zoom in to some aspect of beliefs that appears to be important based on past answers. Such elicitations have appeared as a source of hypotheses or evidence on mechanisms in Bergman et al. (2019); Hampole et al. (2021). Jayachandran et al. (2021) uses interviews to create a "gold standard" measure of female empowerment.

The literature on qualitative methodology is voluminous but as of now exists outside economics (Small, 2011; Gerson and Damaske, 2020; Small and Calarco, 2022; Small and Cook, 2021; Knott et al., 2022). Apart from the historic distrust for "cheap talk", that might be driven by the costs of personnel and lack of replicability and transparency in the procedures. We hypothesize that the use of generative large language models might alleviate these issues. For example, automated questions-answering agent might offer a scalable and relatively more transparent alternative to a human interviewer.[3] It is also possible that an LLM can productively augment a human interviewer. Argyle et al. (2023) show that GPT-3 can help subjects to rephrase their replies in conversations on controversial political topics to be more polite or less defensive. This in turn decreases toxicity in the conversation. To generate the suggestions, they simply use an elaborate prompt. It is reasonable to expect that prompting LLMs to ask questions is a comparable task, hence some relatively simple prompt engineering might produce good results.

It would be particularly interesting to run a proof-of-concept experiment explaining some actions in the lab. Asking experiment subjects about their strategies in an open-ended way is frequent in experimental economics (Danz et al., 2022, is an example), but the effectiveness of such elicitations hasn't been formally studied. One could set up a lab experiment with subjects, for example, choosing among lotteries and use a decoder LLM to ask the subjects questions about their decision-making process and an encoder to predict the next actions with the textual information. However, this design might have multiple issues. First, similarly to the paragraph on generating answers, it is not obvious which fine-tuning or prompt-tuning strategy is optimal. Second, the conversation might end up with the text corpus that is longer than the maximum capacity of BERT model (512 tokens, or approximately 250-300 words for English language). The task of analyzing longer texts presents a significant challenge.

[3]We discuss the relative transparency advantages of the two strategies in the next section.

On the one hand, the computational complexity of the usual self-attention mechanism grows quadratically with the number of tokens. On the other hand, alternative models with less computationally demanding attention mechanisms are currently less effective (Beltagy et al., 2020).

3.4.3 Hypothesis generation

Ludwig and Mullainathan (2023) propose using machine learning algorithms for sourcing hypotheses. They use an example of the effects of appearance on parole decisions. They predict the decisions using mugshots available in the administrative data and then use a generative model to create the pairs of faces that differ in the directions of the highest increase in the probability of denying parole. They later show the pairs of images to a set of human subjects tasked to describe the differences. They find that the most important features are "grooming quality" and "round face shape". These descriptions allow for creation of hypotheses that can later be formally tested. This procedure is based on saliency analysis, which is a technique allowing to understand which parts of the text (or in their example, image) are the most influential for the prediction.

A simple way of interpreting the way treatments change beliefs would be just looking at the most likely text by treatment. This is what Bursztyn et al. (2022) present to illustrate the shift in perceptions. If we try to compare the texts, we might come up with hypotheses on the ways the treatment affects the beliefs. For instance, the examples they give of the answers that are most likely to be in the group evaluating the "after crisis" vignette: *I'm sure Mike saw or read unsavory coverage that said there was an influx of immigrants coming to America and we didn't have the infrastructure to support it. He needed a person to blame for his job loss.* The example in evaluating the "before crisis" vignette is: *He probably felt that immigrants would take his job. Either that or he's most likely racist or prejudiced.* This shows a finer picture than what would have been obtained simply by word counts. Based on the single example it looks like the "after crisis" treatment still allows for some cover through ex-ante anxiety and ex-post rationalization: realized employment shock might appear as a more valid rationale than an unrealized one. This might give us an additional hypothesis that adverse economic events not only give excuses for xenophobic behavior but also make past xenophobic behavior more acceptable. Although a more thorough review of the answer features is warranted, this discussion shows how one might use the large language model analysis to create hypotheses about the effects of the treatment on beliefs. However, Ding and Koehn (2021) provide a discussion of saliency analysis with BERT that might allow for a finer detection of the changing features.

3.4.4 Representing subjects

Some of the recent work has attempted using large language models to represent research subjects. For example, Bybee (2023) shows that the market and macroeconomic expectations generated by the GPT-3.5 model are consistent with a broad range of expert and consumer surveys. Horton (2023) shows that the GPT-3 behaves similarly to humans in a set of common lab experiments. Brand et al. (2023) show that GPT-3 also behaves similarly to humans when deciding on purchases of common consumer goods, including downward-sloping demand and demand for attributes. Argyle et al. (2022) show that GPT-3 can behave as a reasonable approximation for opinion surveys. They construct prompts describing characters they want to survey and ask the model to complete the sequence with a response to the question of interest. The model performs well when predicting vote shares and across group stereotypes.

The LLM modeling offers an intriguing possibility for hypothesis generation. It would be useful to catalog the precision of the LLM answers depending on the task, domain, and social group. Still, using the models for anything beyond hypothesis generation as of now seems unreliable.

3.5 ChatGPT example

In this section, we discuss two simple tests of ChatGPT performance as an interviewer and as a representation of respondents. Appendix 3.A provides the transcript of a conversation between two ChatGPT agents. One agent is pretending to be a qualitative interviewer. Another agent is pretending to be an MBA student replying to a qualitative interview. As we can see, the Interviewer agent is performing surprisingly well: it starts with a reasonable question about career goals, then prompts to select priorities when the respondent gives an overwhelming list of goals, then continues to practical ways of searching for the job.

The interviewee agent, however, does not seem natural: it generates an overwhelming amount of information and rationales, which is unlikely to match real behavior. To remove this, one can either try tweaking the prompts, or using straight GPT-3: it is possible that the overzealous behavior is an artifact of the ChatGPT fine-tuning.

Of course, this exercise is better to be reviewed by an expert. One can also imagine a pipeline where the model learns from the feedback of an expert interviewer.

3.6 Conclusion

We have reviewed the current state of the belief elicitation procedures in economics. Then, we presented the main architectures and fine-tuning strategies for large language models. After that, we discussed the potential applications of LLM for belief elicitation. This includes using LLM to represent elicited beliefs in a way that can be used in estimation; interaction with research subjects to extract information; recognize patterns that could be used in formulating new hypotheses. As a more risky application, we discussed using the decoder model itself to represent research subjects.

This review has focused only on belief elicitation. However, it is reasonable to expect LLMs to be useful in the analysis of communication more broadly. Argyle et al. (2023) is a notable example. We hope that more such examples will arise.

We conclude with some general remarks. Machine learning has been viewed by some as a substitute for economic theory. However, this impact of ML on economics is likely to be similar to the way field experiments affected it earlier. While at first experiments were viewed as substitutes for economic theory, Card et al. (2011) argue that they are in fact complements and allow to test and even structurally estimate theories that used to be out of the scope of rigorous empirical work. Our review presents evidence supporting the analogous impact of ML, as it expands the empirical toolbox to the analysis of beliefs and communication in natural language.

3.A ChatGPT conversation

The conversation starts by initial prompts asking ChatGPT to play either the role of the Interviewer, or an MBA student.

Agent 1, Interviewer prompt

Author (MG): Pretend to be a qualitative researcher trying to learn how MBA students decide on their first job after the MBA program. I will be pretending to be a current MBA student. Ask questions to understand my rationales and motives. Please ask 1 question at a time.

Interviewer: Thank you for participating in this study. To begin, could you please describe the key factors you consider when thinking about your first job post-MBA?

Agent 2, MBA prompt

Author (MG): Pretend to be an MBA student. I will ask you a sequence of questions about the way you choose a job after you complete the MBA program. You will try to answer as if you are a real MBA student with some motivations and interests.

MBA: Sure, I'm ready to step into the role of an MBA student. Feel free to begin your questions whenever you're ready.

After that I copy and paste the responses to make the agents engage in a conversation:

Interviewer: Thank you for participating in this study. To begin, could you please describe the key factors you consider when thinking about your first job post-MBA?

MBA: Certainly, I appreciate your interest in my career planning process. As an MBA student, my career decision-making process is significantly influenced by the following key factors:

Alignment with Career Goals: One of the primary factors is how well the job aligns with my long-term career goals. Having pursued an MBA, my objectives likely extend beyond the immediate role, so I'm interested in positions that offer the potential for professional growth and advancement.

Industry: The industry in which the job is located is also significant. Based on my MBA specialization and personal interests, I might be drawn more towards certain sectors, for instance, technology, consulting, finance, healthcare, or startups.

Role and Responsibilities: The specifics of the job role are also crucial. I would look for a job that leverages the skills I've gained during my MBA program, such as strategic planning, leadership, and financial analysis.

Culture and Values: The organizational culture and values are important to me. I want to be part of an organization that promotes diversity, fosters innovation, values its employees, and is committed to ethical practices.

Compensation and Benefits: While not the sole factor, the salary, bonus potential, and benefits package is certainly a consideration. I've made a substantial investment in my education, and I'd like a return on that investment.

Geographical Location: The location of the job could also play a part in my decision, as it may affect my lifestyle and work-life balance.

Impact: I am particularly interested in roles that allow me to make a significant impact, either in the organization or in society at large.

Work-life Balance: Finally, I also value work-life balance, so I would look at the expected working hours, vacation policies, and the possibility of remote work.

Remember that these factors can vary in importance depending on the individual. For me, alignment with career goals and the ability to make a meaningful impact are particularly important.

Interviewer: Thank you for your detailed response; it gives a good overview of the factors influencing your decision. Now, I'd like to understand more about the weightage you assign to these factors. Could you please tell me which among these factors is most crucial to you, and why?

MBA: Certainly. While all of these factors play an important role in my decision-making process, if I were to choose one as most crucial, it would be the 'Alignment with Career Goals'.

Having invested substantial time, effort, and financial resources in an MBA program, it's important to me that my first job post-MBA sets the foundation for my long-term career trajectory. The job I take needs to offer the right experiences and opportunities for professional growth that align with my goals.

A role that aligns with my long-term career goals is more likely to keep me engaged and motivated, as it would be directly contributing towards my future career aspirations. This does not necessarily mean being on a fast track to a top executive position, but it could mean gaining experience in strategic decision-making, leading a team, managing complex projects, or anything else that is in line with what I envision for my future.

Also, a job that matches my career goals is likely to satisfy other important factors too. For instance, a job that aligns with my long-term goals would ideally be in an industry I'm passionate about, provide a role and responsibilities that I find challenging and exciting, and offer a competitive compensation package.

Of course, this is not to undermine the importance of other factors like work-life balance, company culture, or location. It's about finding the right balance, but for me, career alignment is a starting point from which I can consider the other factors.

Interviewer: Thank you for clarifying the significance of alignment with your career goals. Now, let's delve deeper into your decision-making process. Could you describe how you typically go about researching potential roles and companies to find ones that align with your long-term career goals?

MBA: Of course, I'd be happy to describe my process. It's multi-faceted and involves various resources:

Career Services at my University: Most MBA programs offer robust career services that include job listings, career counseling, resume reviews, and workshops. These resources are usually tailored to the kinds of roles MBA graduates typically pursue, making them a valuable starting point.

Online Research: I spend a significant amount of time researching on the internet. This includes visiting company websites to understand more about their values, culture, and current initiatives. It also means utilizing professional social networking sites like LinkedIn to see where alumni from my program have ended up, job descriptions they've posted, and the types of roles available in the companies I'm interested in.

Job Portals: I use job portals like Indeed, Glassdoor, and LinkedIn Jobs to find available positions. These sites also often provide insights into company culture, employee satisfaction, and compensation ranges.

Networking: Building relationships with professionals in my desired industry is also a crucial part of my research. Networking can provide insider perspectives about the day-to-day realities of the roles and companies I'm considering. This can be through university events, professional networking events, or informal settings.

Informational Interviews: When possible, I conduct informational interviews with professionals in roles or companies that I'm interested in. These conversations can offer valuable insights into the realities of the job and the company culture.

Internships: If it's a possibility, internships during the MBA program are a great way to get firsthand experience in a specific industry or role. They can also provide a foot in the door at a potential employer.

News and Industry Reports: Staying informed about trends, challenges, and opportunities in my desired industry helps me understand potential career paths. Websites, podcasts, trade journals, and business publications are all useful for this kind of insight.

By using a combination of these methods, I can gather a comprehensive understanding of potential roles and companies, helping me identify those that align best with my career goals.

Bibliography

Aaronson, D., S. Agarwal, and E. French (2012). The spending and debt response to minimum wage hikes. American Economic Review 102(7), 3111–3139.

Abadie, A., M. M. Chingos, and M. R. West (2018, October). Endogenous Stratification in Randomized Experiments. The Review of Economics and Statistics 100(4), 567–580.

Abeler, J., D. Nosenzo, and C. Raymond (2019). Preferences for truthâĂŘtelling. Econometrica : journal of the Econometric Society 87(4), 1115–1153.

Adams-Prassl, A. (2021). The Gender Wage Gap in an Online Labour Market: The Cost of Interruptions.

Adams-Prassl, A. and J. Berg (2017). When home affects pay: An analysis of the gender pay gap among crowdworkers. Available at SSRN 3048711.

Agarwal, S., C. Liu, and N. S. Souleles (2007, December). The reaction of consumer spending and debt to tax rebates - evidence from consumer credit data. Journal of Political Economy 115(6), 986–1019.

Agarwal, S. and W. Qian (2014). Consumption and debt response to unanticipated income shocks: Evidence from a natural experiment in Singapore. American Economic Review 104(12), 4205–30. Publisher: American Economic Association.

Aguirregabiria, V. (2012). Empirical Industrial Organization: Models, Methods, and Applications.

Ahler, D. J. and D. E. Broockman (2018). The delegate paradox: Why polarized politicians can represent citizens best. The Journal of Politics 80(4), 1117–1133.

Anderson, S. P., A. De Palma, and J.-F. Thisse (1992). Discrete choice theory of product differentiation. MIT press.

Andre, P., I. Haaland, C. Roth, and J. Wohlfart (2021, November). Narratives about the Macroeconomy. Number: 18/21 Place: Rochester, NY Series: CEBI Working Paper.

Andre, P., I. Haaland, C. Roth, and J. Wohlfart (2022). Inflation narratives. SSRN Electronic Journal.

Argyle, L. P., E. Busby, J. Gubler, C. Bail, T. Howe, C. Rytting, and D. Wingate (2023). AI Chat Assistants can Improve Conversations about Divisive Topics. arXiv preprint arXiv:2302.07268.

Argyle, L. P., E. C. Busby, N. Fulda, J. Gubler, C. Rytting, and D. Wingate (2022). Out of One, Many: Using Language Models to Simulate Human Samples. arXiv preprint arXiv:2209.06899.

Armantier, O., W. Bruine de Bruin, G. Topa, W. van der Klaauw, and B. Zafar (2015). Inflation expectations and behavior: Do survey respondents act on their beliefs?'. International Economic Review 56(2), 505–536.

Armantier, O., S. Nelson, G. Topa, W. van der Klaauw, and B. Zafar (2016). The price is right: Updating inflation expectations in a randomized price information experiment. Review of Economics and Statistics 98(3), 503–523.

Ash, E., G. Gauthier, and P. Widmer (2021). Relatio: Text semantics capture political and economic narratives. arXiv preprint arXiv:2108.01720.

Ash, E. and S. Hansen (2023). Text algorithms in economics. Unpublished manuscript.

Attenberg, J., P. Ipeirotis, and F. Provost (2015). Beat the machine: Challenging humans to find a predictive model's âĂIJunknown unknownsâĂİ. Journal of Data and Information Quality (JDIQ) 6(1), 1–17.

Azrieli, Y., C. P. Chambers, and P. J. Healy (2018, August). Incentives in Experiments: A Theoretical Analysis. Journal of Political Economy 126(4), 1472–1503.

Bachmann, R., T. O. Berg, and E. R. Sims (2015). Inflation expectations and readiness to spend: Cross-sectional evidence. American Economic Journal: Economic Policy 7(1), 1–35.

Beltagy, I., M. E. Peters, and A. Cohan (2020). Longformer: The long-document transformer. arXiv preprint arXiv:2004.05150.

Bergman, P., R. Chetty, S. DeLuca, N. Hendren, L. F. Katz, and C. Palmer (2019). Creating moves to opportunity: Experimental evidence on barriers to neighborhood choice. Technical report, National Bureau of Economic Research.

Bernanke, B. S. (2007). Inflation expectations and inflation forecasting. Speech at the Monetary Economics Workshop of the NBER Summer Institute. Place: Cambridge, MA Publisher: Speech at the Monetary Economics Workshop of the NBER Summer Institute.

Beshears, J., J. J. Choi, D. Laibson, B. C. Madrian, and K. L. Milkman (2015). The effect of providing peer information on retirement savings decisions. Journal of Finance 70(3), 1161–1201.

Beshears, J., J. J. C. Choi, D. Laibson, and B. C. Madrian (2018). Behavioral household finance. In D. Bernheim, S. DellaVigna, and D. Laibson (Eds.), Handbook of behavioral economics, pp. 177–276. Amsterdam: Elsevier.

Bottan, N. and R. Perez-Truglia (2020a). Betting on the house: Subjective expectations and market choices. NBER Working Paper No. 27412.

Bottan, N. L. and R. Perez-Truglia (2020b). Choosing your pond: Location choices and relative income. The Review of Economics and Statistics, 1–46. tex.eprint: https://direct.mit.edu/rest/article-pdf/doi/10.1162/rest_a_00991/1891330/rest_a_00991.pdf.

Brand, J., A. Israeli, and D. Ngwe (2023). Using gpt for market research. Available at SSRN 4395751.

Brown, T., B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, and A. Askell (2020). Language models are few-shot learners. Advances in neural information processing systems 33, 1877–1901.

Brunnermeier, M., E. Farhi, R. S. Koijen, A. Krishnamurthy, S. C. Ludvigson, H. Lustig, S. Nagel, and M. Piazzesi (2021). Perspectives on the future of asset pricing. The Review of Financial Studies 34(4), 2126–2160. Publisher: Oxford University Press.

Buinskaya, A. and M. Galashin (2023). Survey crowdsourcing: integrating community information in research design.

Bursztyn, L., F. Ederer, B. Ferman, and N. Yuchtman (2014). Understanding mechanisms underlying peer effects: Evidence from a field experiment on financial decisions. Econometrica : journal of the Econometric Society 82(4), 1273–1301.

Bursztyn, L., G. Egorov, I. Haaland, A. Rao, and C. Roth (2022, May). Scapegoating during Crises. In AEA Papers and Proceedings, Volume 112, virtual meeting, pp. 151–155.

Bursztyn, L., S. Fiorin, D. Gottlieb, and M. Kanz (2019). Moral incentives in credit card debt repayment: Evidence from a field experiment. Journal of Political Economy 127(4), 1641–1683.

Bursztyn, L. and D. Y. Yang (2022). Misperceptions about others. Annual Review of Economics 14, 425–452.

Bybee, L. (2023). Surveying Generative AI's Economic Expectations. arXiv preprint arXiv:2305.02823.

Camerer, C. F., A. Dreber, F. Holzmeister, T.-H. Ho, J. Huber, M. Johannesson, M. Kirchler, G. Nave, B. A. Nosek, T. Pfeiffer, A. Altmejd, N. Buttrick, T. Chan, Y. Chen, E. Forsell,

A. Gampa, E. Heikensten, L. Hummer, T. Imai, S. Isaksson, D. Manfredi, J. Rose, E.-J. Wagenmakers, and H. Wu (2018, September). Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. Nature Human Behaviour 2(9), 637–644. tex.copyright: 2018 The Author(s).

Campbell, J. Y., H. E. Jackson, B. C. Madrian, and P. Tufano (2011). Consumer financial protection. Journal of Economic Perspectives 25(1), 91–114.

Card, D., S. DellaVigna, and U. Malmendier (2011). The role of theory in field experiments. Journal of Economic Perspectives 25(3), 39–62. Publisher: American Economic Association.

Cavallo, A., G. Cruces, and R. Perez-Truglia (2016). Learning from potentially biased statistics. Brookings Papers on Economic Activity 47(1), 59–108.

Cavallo, A., G. Cruces, and R. Perez-Truglia (2017). Inflation expectations, learning, and supermarket prices: Evidence from survey experiments. American Economic Journal: Macroeconomics 9(3), 1–35.

Charness, G., U. Gneezy, and V. Rasocha (2021, September). Experimental methods: Eliciting beliefs. Journal of Economic Behavior & Organization 189, 234–256.

Chernozhukov, V., M. Demirer, E. Duflo, and I. FernÃąndez-Val (2018, June). Generic Machine Learning Inference on Heterogeneous Treatment Effects in Randomized Experiments, with an Application to Immunization in India. Series: Working Paper Series Type: Working Paper.

Chetty, R., J. N. Friedman, N. Hendren, M. Stepner, and The Opportunity Insights Team (2020). How did COVID-19 and stabilization policies affect spending and employment? A new real-time economic tracker based on private sector data. NBER Working Paper No. 27431.

Chiricos, T., M. Hogan, and M. Gertz (1997). Racial Composition of Neighborhood and Fear of Crime. Criminology : an interdisciplinary journal 35(1), 107–132.

Christiano, L. J., M. Eichenbaum, and C. L. Evans (1999). Chapter 2 Monetary policy shocks: What have we learned and to what end? Volume 1 of Handbook of macroeconomics, pp. 65–148. Elsevier. ISSN: 1574-0048.

Coibion, O., D. Georgarakos, Y. Gorodnichenko, and M. Van Rooij (2019). How does consumption respond to news about inflation? Field evidence from a randomized control trial. NBER Working Paper No. 26106.

Coibion, O., D. Georgarakos, Y. Gorodnichenko, and M. Weber (2023, January). Forward guidance and household expectations. Journal of the European Economic Association.

Coibion, O., Y. Gorodnichenko, E. S. Knotek, and R. Schoenle (2021). Average inflation targeting and household expectations. NBER Working Paper No. 27836. Publisher: National Bureau of Economic Research.

Coibion, O., Y. Gorodnichenko, and M. Weber (2022). Monetary policy communications and their effects on household inflation expectations. Journal of Political Economy 130(6).

Converse, J. M. (1984). Strong Arguments and Weak Evidence: The Open/Closed Questioning Controversy of the 1940s. The Public Opinion Quarterly 48(1), 267–282.

Cravino, J. and A. A. Levchenko (2017). The distributional consequences of large devaluations. American Economic Review 107(11), 3477–3509.

Cullen, Z. and R. Perez-Truglia (2022). How much does your boss make? The effects of salary comparisons. Journal of Political Economy 130(3), 766–822.

D'Acunto, F., D. Hoang, M. Paloviita, and M. Weber (2023). IQ, expectations, and choice. Review of Economic Studies, forthcoming.

Danz, D., L. Vesterlund, and A. J. Wilson (2022, September). Belief Elicitation and Behavioral Incentive Compatibility. American Economic Review 112(9), 2851–2883.

DellaVigna, S. and M. Gentzkow (2010). Persuasion: empirical evidence. Annual Review of Economics, Vol 3 2(1), 643–669.

DellaVigna, S., J. A. List, U. Malmendier, and G. Rao (2022, March). Estimating Social Preferences and Gift Exchange at Work. American Economic Review 112(3), 1038–1074.

DellaVigna, S. and D. Pope (2018a, December). Predicting Experimental Results: Who Knows What? Journal of Political Economy 126(6), 2410–2456.

DellaVigna, S. and D. Pope (2018b, April). What Motivates Effort? Evidence and Expert Forecasts. The Review of Economic Studies 85(2), 1029–1069.

DellaVigna, S., D. Pope, and E. Vivalt (2019). Predict science to improve science. Science 366(6464), 428–429. Publisher: American Association for the Advancement of Science.

Devlin, J., M.-W. Chang, K. Lee, and K. Toutanova (2019, May). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.

Ding, S. and P. Koehn (2021, June). Evaluating Saliency Methods for Neural Language Models. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Online, pp. 5034–5052. Association for Computational Linguistics.

Dornbusch, R. (1987). Exchange rates and prices. The American Economic Review 77(1), 93–106. Publisher: American Economic Association.

Engelmann, J., M. Lebreton, P. Schwardmann, J. J. van der Weele, and L.-A. Chang (2019). Anticipatory anxiety and wishful thinking.

Fehr, D., J. Mollerstrom, and R. Perez-Truglia (2019). Your place in the world: The demand for national and global redistribution. NBER Working Paper No. 26555.

Ferrario, B. and S. Stantcheva (2022, May). Eliciting People's First-Order Concerns: Text Analysis of Open-Ended Survey Questions. In AEA Papers and Proceedings, Volume 112, pp. 163–169.

Franco, A., N. Malhotra, and G. Simonovits (2014, September). Publication bias in the social sciences: Unlocking the file drawer. Science (New York, N.Y.) 345(6203), 1502–1505.

Fuster, A., R. Perez-Truglia, M. Wiederholt, and B. Zafar (2020). Expectations with endogenous information acquisition: An experimental investigation. The Review of Economics and Statistics, 1–54. tex.eprint: https://direct.mit.edu/rest/article-pdf/doi/10.1162/rest_a_00994/1891360/rest_a_00994.pdf.

Gabaix, X. and D. Laibson (2006). Shrouded attributes, consumer myopia, and information suppression in competitive markets. Quarterly Journal of Economics 121(2), 505–540.

Ganong, P. and P. Noel (2019). Consumer spending during unemployment: Positive and normative implications. American Economic Review 109(7), 2383–2424.

Gathergood, J., N. Mahoney, N. Stewart, and J. Weber (2019). How do individuals repay their debt? The balance-matching heuristic. American Economic Review 109(3), 844–875.

Gentzkow, M., B. Kelly, and M. Taddy (2019). Text as data. Journal of Economic Literature 57(3), 535–74.

Gerson, K. and S. Damaske (2020). The science and art of interviewing. Oxford University Press.

Giglio, S., M. Maggiori, J. Stroebel, and S. Utkus (2020). Inside the mind of a stock market crash. NBER Working Paper No. 27272.

Giglio, S., M. Maggiori, J. Stroebel, and S. Utkus (2021). Five facts about beliefs and portfolios. American Economic Review 111(5), 1481–1522.

Goeree, M. S. (2008). Limited information and advertising in the US personal computer industry. Econometrica : journal of the Econometric Society 76(5), 1017–1074.

Gouvea, R. R. (2020). Large devaluations and inflation inequality: Evidence from brazil. UMass Economics Working Papers.

Guest, G., E. Namey, and M. Chen (2020). A simple method to assess and report thematic saturation in qualitative research. PloS one 15(5), e0232076.

Haaland, I., C. Roth, and J. Wohlfart (2023). Designing information provision experiments. Journal of economic literature 61(1), 3–40.

Hajdini, I., E. Knotek, J. Leer, M. Pedemonte, R. Rich, and R. Schoenle (2022). Low passthrough from inflation expectations to income growth expectations: Why people dislike inflation. Working Paper.

Hampole, M., F. Truffa, and A. Wong (2021). Peer effects and the gender gap in corporate leadership: Evidence from MBA students. Technical report, Working Paper.

Hennink, M. and B. N. Kaiser (2022). Sample sizes for saturation in qualitative research: A systematic review of empirical tests. Social Science & Medicine 292, 114523.

Hoff, P. D. (2009). A first course in Bayesian statistical methods, Volume 580. Springer.

Horton, J. J. (2023). Large Language Models as Simulated Economic Agents: What Can We Learn from Homo Silicus? arXiv preprint arXiv:2301.07543.

Hossain, T. and R. Okui (2013). The binarized scoring rule. Review of Economic Studies 80(3), 984–1001.

Hussam, R., N. Rigol, and B. N. Roth (2022, March). Targeting High Ability Entrepreneurs Using Community Information: Mechanism Design in the Field. American Economic Review 112(3), 861–898.

Jayachandran, S., M. Biradavolu, and J. Cooper (2021). Using machine learning and qualitative interviews to design a five-question women's agency index. Technical report, National Bureau of Economic Research.

Jun, S. J. and S. Lee (2018). Identifying the effect of persuasion. arXiv preprint arXiv:1812.02276.

Katz, L. F. and A. B. Krueger (2019). Understanding trends in alternative work arrangements in the United States. RSF: The Russell Sage Foundation Journal of the Social Sciences 5(5), 132–146.

Keeter, S., C. Kennedy, M. Dimock, J. Best, and P. Craighill (2006). Gauging the impact of growing nonresponse on estimates from a national RDD telephone survey. International Journal of Public Opinion Quarterly 70(5), 759–779. Publisher: Oxford University Press England.

Knott, E., A. H. Rao, K. Summers, and C. Teeger (2022). Interviews in the social sciences. Nature Reviews Methods Primers 2(1), 73. Publisher: Nature Publishing Group UK London.

Krosnick, J. A. (1999). Survey Research. Annual Review of Psychology 50(1), 537–567.

Lazarsfeld, P. F. (1944). The Controversy Over Detailed Interviews-An Offer for Negotiation. The Public Opinion Quarterly 8(1), 38–60.

Lintner, J. (1956). Distribution of incomes of corporations among dividends, retained earnings, and taxes. The American economic review 46(2), 97–113. Publisher: JSTOR.

Lu, J. and K. Saito (2022). Mixed Logit and Pure Characteristics Models.

Ludwig, J. and S. Mullainathan (2023). Machine Learning as a Tool for Hypothesis Generation. Technical report, National Bureau of Economic Research.

Mankiw, N. G. and R. Reis (2002). Sticky information versus sticky prices: A proposal to replace the new keynesian phillips curve. The Quarterly Journal of Economics 117(4), 1295–1328.

Mankiw, N. G., R. Reis, and J. Wolfers (2003). Disagreement about inflation expectations. In M. Gertler and K. Rogoff (Eds.), NBER macroeconomics annual, pp. 209–248.

Manski, C. F. (2004). Measuring Expectations. Econometrica : journal of the Econometric Society 72(5), 1329–1376.

McKenzie, D. (2012). Beyond baseline and follow-up: The case for more T in experiments. Journal of Development Economics 99(2), 210–221.

Milkman, K. L., L. Gandhi, M. S. Patel, H. N. Graci, D. M. Gromet, H. Ho, J. S. Kay, T. W. Lee, J. Rothschild, J. E. Bogard, I. Brody, C. F. Chabris, E. Chang, G. B. Chapman, J. E. Dannals, N. J. Goldstein, A. Goren, H. Hershfield, A. Hirsch, J. Hmurovic, S. Horn, D. S. Karlan, A. S. Kristal, C. Lamberton, M. N. Meyer, A. H. Oakes, M. E. Schweitzer, M. Shermohammed, J. Talloen, C. Warren, A. Whillans, K. N. Yadav, J. J. Zlatev, R. Berman, C. N. Evans, R. Ladhania, J. Ludwig, N. Mazar, S. Mullainathan, C. K. Snider, J. Spiess, E. Tsukayama, L. Ungar, C. Van den Bulte, K. G. Volpp, and A. L. Duckworth (2022, February). A 680,000-person megastudy of nudges to encourage vaccination in pharmacies. Proceedings of the National Academy of Sciences 119(6), e2115126119.

Nadeau, R., R. G. Niemi, and J. Levine (1993). Innumeracy About Minority Populations. The Public Opinion Quarterly 57(3), 332–347.

Otis, N. G. (2021). Forecasting in the Field. Type: working paper.

Otis, N. G. (2022). The Efficacy of Crowdsourced Nudges: Experimental Evidence. Type: Working Paper.

Ouyang, L., J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, and A. Ray (2022). Training language models to follow instructions with human feedback. Advances in Neural Information Processing Systems 35, 27730–27744.

Pearl, J. (2009). Causality. Cambridge university press.

Perez-Truglia, R. (2020). The effects of income transparency on well-being: Evidence from a natural experiment. American Economic Review 110, 1019–54.

Ponce, A., E. Seira, and G. Zamarripa (2017). Borrowing on the wrong credit card? Evidence from Mexico. American Economic Review 107(4), 1335–1361. Publisher: American Economic Association.

Prelec, D. (2004, October). A Bayesian Truth Serum for Subjective Data. Science (New York, N.Y.) 306(5695), 462–466.

Radanovic, G. and B. Faltings (2013). A robust bayesian truth serum for non-binary signals. Volume 27, pp. 833–839.

Reimers, N. and I. Gurevych (2019, August). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks.

Roth, C. and J. Wohlfart (2020). How do expectations about the macroeconomy affect personal expectations and behavior? Review of Economics and Statistics 102(4), 731–748.

Salganik, M. J. and K. E. C. Levy (2015, May). Wiki Surveys: Open and Quantifiable Social Data Collection. PLOS ONE 10(5), e0123483.

Saunders, B., J. Sim, T. Kingstone, S. Baker, J. Waterfield, B. Bartlam, H. Burroughs, and C. Jinks (2018). Saturation in qualitative research: exploring its conceptualization and operationalization. Quality & quantity 52, 1893–1907.

Schotter, A. and I. Trevino (2014). Belief elicitation in the laboratory. Annual Review of Economics, Vol 3 6(1), 103–128.

Schuman, H. and S. Presser (1979). The Open and Closed Question. American Sociological Review 44(5), 692–712.

Selten, R., A. Sadrieh, and K. Abbink (1999). Money does not induce risk neutral behavior, but binary lotteries do even worse. Theory and Decision. An International Journal for Multidisciplinary Advances in Decision Science 46(3), 213–252.

Sigelman, L. and R. G. Niemi (2001). Innumeracy about Minority Populations: African Americans and Whites Compared. The Public Opinion Quarterly 65(1), 86–94.

Small, M. L. (2011). How to conduct a mixed methods study: Recent trends in a rapidly growing literature. Annual review of sociology 37, 57–86. Publisher: Annual Reviews.

Small, M. L. and J. M. Calarco (2022). Qualitative literacy: A guide to evaluating ethnographic and interview research. Univ of California Press.

Small, M. L. and J. M. Cook (2021). Using interviews to understand why: Challenges and strategies in the study of motivated action. Sociological Methods & Research, 0049124121995552. Publisher: SAGE Publications Sage CA: Los Angeles, CA.

Stantcheva, S. (2021, November). Understanding Tax Policy: How do People Reason? The Quarterly Journal of Economics 136(4), 2309–2369.

Stantcheva, S. (2022, September). How to Run Surveys: A Guide to Creating Your Own Identifying Variation and Revealing the Invisible. Series: Working Paper Series Type: Working Paper.

Touvron, H., T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. RoziÃÍre, N. Goyal, E. Hambro, and F. Azhar (2023). Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971.

Weber, M., F. D'Acunto, Y. Gorodnichenko, and O. Coibion (2022). The subjective inflation expectations of households and firms: Measurement, determinants, and implications. Journal of Economic Perspectives 36(3), 157–184. Publisher: American Economic Association 2014 Broadway, Suite 305, Nashville, TN 37203-2418.

Witkowski, J. and D. Parkes (2012). A robust bayesian truth serum for small populations. In Proceedings of the AAAI Conference on Artificial Intelligence, Volume 26, pp. 1492–1498.

Wolfers, J. and D. Rothchild (2011). Forecasting elections: Voter intentions versus expectations.

Wong, C. J. (2007, January). "Little" and "Big" Pictures in Our Heads: Race, Local Context, and Innumeracy About Racial Groups in the United States. Public Opinion Quarterly 71(3), 392–412.

Zhong, Q., L. Ding, J. Liu, B. Du, and D. Tao (2023). Can chatgpt understand too? a comparative study on chatgpt and fine-tuned bert. arXiv preprint arXiv:2302.10198.