Identifying Causal Effects in Information Provision Experiments

Dylan Balla-Elliott

November, 2023

Information provision experiments are a popular way to study causal effects of beliefs on behavior. Researchers estimate these effects using TSLS. I show that existing TSLS specifications do not estimate the average partial effect; they have weights proportional to belief updating in the first-stage. If people whose decisions depend on their beliefs gather information before the experiment, the information treatment may shift beliefs more for people with weak belief effects. This attenuates TSLS estimates. I propose researchers use a local-least-squares (LLS) estimator that I show consistently estimates the average partial effect (APE) under Bayesian updating, and apply it to Settele (2022).

JEL CODES: C26, C9, D83, D9

dbe@uchicago.edu University of Chicago, Kenneth C. Griffin Department of Economics. Thanks to Alex Torgovitsky, Max Tabord-Meehan, Magne Mogstad, Santiago Lacouture, Max Maydanchik, Isaac Norwich, Francesco Ruggieri, Sofia Shchukina, Alex Weinberg, Jun Wong, Itzhak Rasooly, and participants at the University of Chicago Applied Micro Lunch for helpful comments and suggestions. Thanks especially to Zoë Cullen and Ricardo Perez-Truglia. I am also indebted to Sonja Settele for an interesting and well-executed experiment and clear replication package. This material is based on work supported by the National Science Foundation Graduate Research Fellowship under Grant No. DGE 1746045. Any opinion, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

Information provision experiments have become a popular experimental design because the effect of a shift in beliefs on behavior can be informative about the effect of a shift in the underlying economic fundamental. While it may be difficult to vary the returns to education while holding all else fixed, it is easy to generate variation in beliefs by experimentally providing people with new information (Jensen, 2010). While it may not be feasible to change outside wage offers, or someones rank in the income distribution, or the racial composition of welfare recipients, it has proven possible to shift people's perceptions (Jäger et al., 2023; Bottan and Perez-Truglia, 2022b; Akesson et al., 2022).

In these experiments, researchers vary the information (or "signal") shown to participants. Then, they typically estimate the effect of beliefs on behavior using two-stage least square (TSLS) regressions. When the effects of beliefs are heterogeneous, TSLS targets a weighted average of these heterogeneous effects with weights proportional to the heterogeneity in the first stage (Angrist and Imbens, 1995).

I show that in information provision experiments, TSLS puts the most weight on people who update their beliefs the most in response to new information. Coefficients in common specifications will thus differ from the average effect of beliefs to the extent that the heterogeneity in the effect of information provision on beliefs is correlated with the heterogeneous effects of beliefs on behavior. In this sense, TSLS estimates depend on any endogeneity in belief updating in response to new information.

I propose that researchers use an alternative local least squares (LLS) estimator, which I show consistently estimates the average partial effect (APE) under an assumption of Bayesian belief updating. The first stage change in beliefs is used to construct a control for the endogenous component of belief updating. Conditional on this control function, OLS identifies the average effect of beliefs for the subset with a particular value of the control; iterating over all of these subsets identifies the average partial effect (Masten and Torgovitsky, 2016). Since the LLS estimator uses the observed belief update to control for endogeneity in the first stage, it is available in so-called "active" experimental designs, where study participants are randomly assigned to see a "high" or "low" signal (e.g. Akesson et al., 2022; Bottan and Perez-Truglia, 2022b; Roth et al., 2022).[1]

[1]When an active experimental design is not possible, the theoretical characterization of TSLS estimates

The key identification result is robust to deviations from the Bayesian belief updating assumption. The APE will still be identified as long as belief updating satisfies a rank invariance condition (Masten and Torgovitsky, 2014, 2016). Conditional on the prior belief and the sign of the difference between the signal and the prior, the posterior beliefs must be rank invariant across potential signals.

The LLS estimator will be larger in magnitude than TSLS whenever there is negative correlation between the extent of belief updating and the importance of these beliefs for decision-making. I show that this negative correlation will arise in models where beliefs are formed through costly information acquisition. In these models, people with large causal effects of beliefs endogenously form precise beliefs in equilibrium. When they join the sample in an information provision experiment, they are less responsive to new information and have a smaller first stage response. The following example illustrates the key mechanism.

Alice bicycles to work when it is sunny and drives when it rains; Bob always drives. Alice checks the weather since her decision depends on it. Bob does not. Alice and Bob join an experiment and are provided with accurate information about the weather. Bob updates his beliefs since he is uninformed, but Alice is already well informed and so only changes her beliefs slightly. This intervention shifts beliefs but not behaviors, even though Alice bases her actions directly on her beliefs.

This mechanism may explain puzzling results in the literature that find small or insignificant effects of beliefs on behavior despite information treatments that have a large effect on beliefs (Alesina et al., 2023; Kuziemko et al., 2015). In general, the bias of existing estimators can be studied in any model that relates belief updating to the effects of beliefs.

I apply the local least squares estimator to a recent study of the effect of beliefs about the gender wage gap on demand for public policy intervention (Settele, 2022). In this exercise, the average partial effect of beliefs on demand for public policy is roughly 70% larger than the corresponding TSLS estimate. I use the observed change in beliefs caused by the information treatment to order people by their responsiveness to new information,

provides a framework for using theory that relates belief formation to the causal effects of interest to extrapolate from TSLS estimates to average effects in the population.

conditional on their prior beliefs. Average effects are close to zero for the 20% of people who update their beliefs the most, but are nearly three times as large as the original TSLS estimate among the bottom 60%. While these conditional average partial effects are less precisely estimated, they are broadly consistent with the theoretical predictions.

This paper builds on the large literature on instrumental variables with heterogeneous effects (Angrist and Imbens, 1995). This paper also builds on and studies the large and growing applied literature that uses information provision experiments to study the effects of beliefs on behavior and decision-making (Balla-Elliott et al., 2022; Cavallo et al., 2017; Cullen et al., 2022; Cullen and Perez-Truglia, 2022; Fuster et al., 2022; see Haaland et al., 2023, for a recent review). I focus on experimental designs where the information treatment is quantitative, for example, "*12 percent of the US population are immigrants*" (Grigorieff et al., 2020; Hopkins et al., 2019) and not treatments that are qualitative, for example, a video (Alesina et al., 2021; Dechezlepretre et al., 2023; Stantcheva, 2023).[2]

In independent and concurrent work, Vilfort and Zhang (2023) also study the interpretation of common TSLS specifications in information provision experiments. They consider a general non-parametric model and provide conditions for when TSLS will have negative weights. They propose that researchers use information about the priors and signals in passive designs to construct specifications that are guaranteed to have non-negative weights. I show that even when weights are non-negative, they will generally depend on endogeneity in belief updating. For this reason, I propose that researchers use an alternative LLS estimator that consistently estimates an unweighted average effect.

This paper thus makes two main contributions to the existing literature. This first is to show that even when TSLS weights are non-negative, they may still depend on the causal effects of interest in an undesirable way. I show how the workhorse Bayesian updating model can be used to interpret these weights, and to connect the TSLS weights to economic theories of how beliefs are formed. These models of belief formation can then be used to formally study the dependence between the weights and the causal effects of interest.

[2]Examples of the kind of quantitative information presented in these experiments are government statistics (Bottan and Perez-Truglia, 2022a; Kuziemko et al., 2015; Roth et al., 2022; Settele, 2022), previous responses from other participants (Balla-Elliott et al., 2022; Bursztyn et al., 2020; Coibion et al., 2021) or other institutional sources (Cullen et al., 2022; Cullen and Perez-Truglia, 2022). See Haaland et al. (2023) for a systematic treatment.

Certain models can sign the bias of TSLS relative to the unweighted average effect.

The second contribution of this paper is to show that the Bayesian learning assumption on the first stage can be used to construct an alternative estimator that directly targets the average partial effect. This estimator importantly does not depend on the relationship between heterogeneity in belief updating and heterogeneity in the effects of beliefs. I show how this estimator can be applied to active experimental designs and suggest how it can naturally be extended to passive designs, though I leave the formal development of cases other than the active design to future work.

The remainder of this paper has the following structure. Section 1 presents the model framework and shows how the key coefficient in a common linear belief updating model can be micro-founded in a normal-normal Bayesian updating model. Section 2 surveys common TSLS specifications and makes clear the conditions under which they estimate a non-negatively weighted average of heterogeneous causal effects. It then introduces a simple model of belief formation to interpret these weights substantively. Section 3 shows that Bayesian updating is sufficient to identify the average partial effect (APE) in experimental designs with an active control group. Section 4 applies this estimator to Settele (2022) and shows that the empirical results are broadly consistent with the theory. Section 5 concludes.

1. Model Framework

This section defines the notation we will use throughout. We will work with a simple outcome equation that is linear in beliefs and allows for beliefs to have heterogeneous effects on behaviors. Then, I will introduce notation for the experiments we will analyze.

There is now a robust literature that uses information provision experiments, and a range of designs now fall under this broad umbrella (Haaland et al., 2023). For example, some information provision experiments focus on how beliefs are formed in response to new information (Coibion et al., 2021; Fuster et al., 2022). There is also a set of papers with a more qualitative analysis of how beliefs affect behaviors. These papers use information treatments that are more qualitative – for example informational videos (Alesina et al.,

2021; Dechezlepretre et al., 2023; Stantcheva, 2023).

This paper focuses on experiments with two key features. First, they study how beliefs affect behavior and not only how beliefs respond to new information. Additionally, we will focus on experimental designs where the information treatment is quantitative, for example "*12 percent of the US population are immigrants*" (Grigorieff et al., 2020; Hopkins et al., 2019) and not treatments that are qualitative, for example *"[t]he chances of a poor kid staying poor as an adult are extremely large"* (Alesina et al., 2018).

To analyze belief updating in these designs, we will follow the literature in using a model that is linear in the difference between the new signal and the prior. This linear first stage equation can be interpreted in a Bayesian model of belief updating. This will help exposition throughout by allowing us to express TSLS weights in terms of substantive features like the learning rate and the relative precision of prior beliefs.

1.1. Outcomes

The outcome equation is a linear model with heterogeneous coefficients on beliefs:

$$Y_i = \tau_i X_i + U_i \tag{1}$$

This is the canonical random coefficients model where Y_i is the outcome or behavior of interest, X_i is the belief, U_i is the structural error term, and τ_i is the partial effect of X_i on Y_i. We will assume that the beliefs X_i are endogenous $\left(\mathbb{E}\left[X_i U_i\right] \neq 0\right)$; Y_i can be arbitrarily affected by unobservables U_i. The only restriction here is that the causal effect for a given individual is linear.

A natural parameter of interest is the average partial effect (APE) of X_i on Y_i, denoted as $\mathbb{E}\left[\tau_i\right]$. This parameter has a simple interpretation as the average effect of beliefs on behaviors. In a recent example, Jäger et al. (2023) study the effect of workers' beliefs about their outside options on labor market outcomes. They pose the question "how much does a percentage point shift in beliefs about a workers' outside option causally shift workers' intended labor market behavior? " (p. 25). The average partial effect answers this question: on average, a one unit increase in beliefs causally shifts the outcome Y by $\mathbb{E}\left[\tau_i\right]$.

An important special case of the linear in parameters model is the log-log specification, which is often interpreted in applied work as an approximation to an average elasticity. The average coefficient on beliefs in the log-log model – sometimes called a behavioral elasticity (Haaland et al., 2023) – is thus a special case of an APE. Results about the APE therefore apply immediately to settings where the target parameter is a behavioral elasticity.

1.2. Information Treatment and Beliefs

There are two major classes of experiments with quantitative information treatments: "active" designs where all participants see some new information, but the particular signal is randomly assigned, and "passive" designs where only a subset of participants see new information and the remaining control group is not shown any signal (Haaland et al., 2023).

We will denote treatments arms by Z_i. Throughout, we will assume that the researcher randomizes over two arms $Z_i \in \{A, B\}$. In the passive design, arm A will be the treatment arm that receives new information and arm B will be the control arm that does not receive new information. In the active design, arm A will be the treatment arm that receives one signal and arm B will be the treatment arm that receives the alternative signal. The treatment indicator $T_i \equiv \mathbb{1}\left(Z_i = A\right)$ indicates assignment to arm A. This is consistent with the language in active designs that arm B is an "active" control group. In any design, treatment arms will be randomly assigned. Finally, $S_i(z)$ is the signal that is shown to individual i in treatment arm z.

While the treatment Z_i will be randomly assigned, it is important to note that the realized signal $S_i(Z_i)$ can generally vary with individuals in a way that is not assumed to be independent of the structural unobservable U_i. For example, consider when $S_i(A)$ is a high estimate of home value and $S_i(B)$ is a low estimate of the home value as in Bottan and Perez-Truglia (2022a). The researcher will randomly assign an individual to see a high or low signal, but the potential signal values are not random and indeed often depend directly on observable features.[3] Roth et al. (2022) explicitly model the two potential signals as noisy measures of the same object: increases in unemployment during a recession.

[3] Balla-Elliott et al. (2022) use the type of the small-business and region to construct the signals values

This makes clear that the difference between the signals is due to random noise due to difference in the data sources, but the level of the signals is not randomly assigned. The realized signal is only randomly assigned conditional on the potential signal values.

Treatment is assigned randomly in the sense that Z_i is statistically independent of U_i, X_i^0, $S_i(\cdot)$, α_i, τ_i. We will introduce one final notational convention: in passive designs, treatment arm B does not receive any signal. We will define $S_i(B) \equiv X_i^0$ in passive designs. This is a notational convention so that we can write the potential outcomes for beliefs in a unified way. Importantly, this does not imply that the control group in passive designs receives a signal.

With this in mind, we can write the potential outcomes for beliefs as

$$X_i(z) = X_i^0 + \alpha_i\big(S_i(z) - X_i^0\big) \tag{2}$$

In treatment arms that receive information, the posterior is a weighted average of the new signal and the prior. The weight on the signal is given by the heterogeneous learning rate α_i. Notice that in passive designs $X_i(B) = X_i^0$ since we set $S_i(B) \equiv X_i^0$ when treatment arm B receives no information. It is worth emphasizing that this is merely a notational device to ensure that the potential signals $S_i(z)$ are always defined.

This weighted average expression is a workhorse in the applied literature and seems to reflect belief updating well, at least in the context of information provision experiments (Balla-Elliott et al., 2022; Cavallo et al., 2017; Cullen et al., 2022; Cullen and Perez-Truglia, 2022; Fuster et al., 2022; Giaccobasso et al., 2022). We will use the structural equations (1) and (2) to study common empirical specifications.

1.3. Belief Potential Outcomes are Motivated by Bayesian Learning

The literature often motivates this weighted-average expression in (2) in a Bayesian learning model with normally distributed beliefs (Balla-Elliott et al., 2022; Cullen and Perez-Truglia, 2022). Since this Bayesian updating form in the first stage will be helpful to guide exposition throughout, I will briefly show how these potential outcomes for beliefs are generated by a Bayesian learning model and relate the key coefficient α_i to model primitives.

Consider a group of individuals with uncertain prior beliefs. The subjective probability that the variable X_i takes the value x is given by the density of the normal distribution $\mathcal{N}\left(X_i^0, \sigma_{iX}^2\right)$. We thus interpret X_i^0 as the mean of the prior distribution. As shorthand, we will call X_i^0 the prior belief of an individual i.

People then observe a signal S_i, which we model as a draw from a distribution $\mathcal{N}(S_i^*, \sigma_{iS}^2)$. The variances of these distributions to reflect the subjective (inverse) precision of the prior and the signal. These variances are important only in their relative size. People for whom $\sigma_{iS}^2/\sigma_{iX}^2$ is large think their prior is more precise than the signal, whereas those for whom $\sigma_{iS}^2/\sigma_{iX}^2$ is small think that the signal is more precise than their prior.

The posterior is then a distribution

$$\mathcal{N}\left(\left(1-\alpha_i\right) X_i^0 + \alpha_i S_i, \frac{\sigma_{iS}^2\sigma_{iX}^2}{\sigma_{iS}^2+\sigma_{iX}^2}\right) \tag{3}$$

$$\text{where } \alpha_i \equiv \frac{\sigma_{iX}^2}{\sigma_{iS}^2+\sigma_{iX}^2} \tag{4}$$

As with the prior, we will call the mean of this distribution the posterior X. Note that the mean of the posterior distribution is a weighted average of the prior X_i^0 and the signal S_i, where the weights are given by their relative precision.[4] We can also note that $\frac{\sigma_{iS}^2\sigma_{iX}^2}{\sigma_{iS}^2+\sigma_{iX}^2} < \sigma_{iX}^2$; intuitively, the posterior distribution is more precise than the prior distribution.[5] We can then relate the prior X_i^0, the signal S_i and the posterior X_i through the equation

$$X_i = (1-\alpha_i)X_i^0 + \alpha_i S_i \tag{5}$$

which generates the potential outcomes for beliefs in (2). There is some direct empirical support for this Bayesian foundation of the linear updating model. Roth et al. (2022) find that all the belief updating in their study is driven by people who report being "very unsure", "unsure" or "somewhat unsure" and that people who are "sure" or "very sure" do not update their beliefs. Similarly Roth and Wohlfart (2020) find that people who are less confident in their prior beliefs update roughly twice as much as people who are more

[4] A full discussion of this derivation can be found in introductory textbook treatments of Bayesian statistics like Robert (2007) or Hoff (2009).

[5] There is experimental evidence that people randomized to the group receiving a signal report greater confidence in their posterior beliefs (Akesson et al., 2022; Cavallo et al., 2017).

confident. Kerwin and Pandey (2023) also find in a more general model that people with less precise priors update more in response to an information treatment.

2. The TSLS Estimator in Information Experiments

This section applies standard results about the TSLS estimator with heterogeneous effects to information provision experiments. TSLS estimates a weighted combination of individual causal effects with weights proportional to the individual first stage; these weights are non-negative and sum to one when the instrument shifts beliefs monotonically (Angrist and Imbens, 1995). These estimates therefore depend on the covariance between individual causal effects and the strength of the individual first stage.

I then consider when instrument monotonicity is consistent with the workhorse Bayesian updating model. Unconditional instrument monotonicity follows from Bayesian updating in active designs where treatment arms correspond to "high" and "low" signals. In passive designs, it requires the additional strong assumption that the population only contains "overestimators" or "underestimators". However, common specifications that condition on the prior by splitting the sample or by interacting the treatment indicator with "exposure" to treatment have non-negative weights under Bayesian updating.

Even when weights are non-negative, standard TSLS estimates will depend on endogeneity in the first stage effects of information on beliefs. I then turn to the relationship between the causal effects of beliefs and belief formation and updating is thus central to the task of understanding existing TSLS estimates. A model of belief formation through information acquisition at a fixed cost predicts that TSLS estimates will be attenuated relative to the unweighted average partial effect. The key mechanism is that large causal effects of beliefs cause people to form precise beliefs before the experiment and therefore update less in response to the experimental signal.

Any model that makes predictions about the covariance between the learning rate α_i and the causal effect of beliefs on behavior τ_i can be used to interpret how TSLS estimates of belief effects may differ from the APE.

2.1. The Reduced Form Effect of Information Provision

In active designs, the "reduced form" effect of treatment is the effect of being assigned to see the signal in arm A rather than the signal in arm B. In passive designs, this is the effect of being assigned to see new information. Consider the simple OLS regression of the outcome Y_i on the treatment indicator $T_i = \mathbb{1}\left(Z_i = A\right)$.

$$\beta^{RF} \equiv \frac{\text{Cov}\left(T_i, Y_i\right)}{\text{Var}\left(T_i\right)} \tag{6}$$

$$= \mathbb{E}\left[\tau_i\left(X_i(A) - X_i(B)\right)\right] \tag{7}$$

We substitute in the linear outcome equation from (1) to get the second line. The reduced form effect of assignment to arm A on the outcome is the expectation of the individual effect of beliefs on behaviors τ_i scaled by the individual effect of the information treatment on beliefs $X_i(A) - X_i(B)$. If all τ_i have the same sign, the reduced form effect of treatment assignment on the outcome will be informative of the sign of the effect of beliefs on behaviors only if the $X_i(A) - X_i(B)$ are all positive or all negative. If some of the first stage effects of treatment assignment on beliefs $X_i(A) - X_i(B)$ are positive and others are negative, then the reduced form effect of treatment assignment on the outcome can be of either sign, depending on the covariance between the τ_i and the $X_i(A) - X_i(B)$.

This also implies that the reduced form effect of new information can be close to zero if there are shifts in beliefs in opposite directions that partially offset. If the first stage effect on beliefs is positive for some people and negative for others, then the average effect of the information treatment on beliefs can be close to zero, even if the effect of beliefs on behaviors is large and the individual first stage effects of the information treatment on beliefs are large.

From the Effect of Information Provision to the Effect of Beliefs. As has become more widely appreciated recently (for example by Giaccobasso et al., 2022), reduced form estimates can be difficult to interpret since they combine the causal effects of beliefs on behaviors with the first stage effects of the information provision on beliefs. The reduced form can therefore be small if beliefs have only a weak effect on behavior, or if the information

provision has only a weak effect on beliefs.

However, there are important cases when the reduced form effect is the parameter of interest. The reduced form is of direct interest when the counterfactuals of interest are about the effect of *information provision* per se, and not the effects of beliefs more generally. For example, if the information treatment corresponds to a (potential) policy intervention, then the reduced form effect of the information provision on behaviors is directly policy-relevant.

The reduced form is more difficult to interpret when the relationship of interest is the effect of beliefs on behavior. For this reason, researchers will often normalize the reduced form effect of information provision on behaviors by the first stage effect of the information provision on beliefs and report TSLS estimates.

Constructing TSLS Estimates. The TSLS coefficient is the ratio of the reduced form effect in (6) and the "first stage" regression of beliefs X_i on treatment assignment T_i .

$$\beta^{TSLS} \equiv \frac{\beta^{RF}}{\beta^{FS}} = \frac{\text{Cov}\big(T_i, Y_i\big)}{\text{Cov}\big(T_i, X_i\big)} \tag{8}$$

where $\beta^{FS} \equiv \text{Cov}\big(T_i, X_i\big)/\text{Var}\big(T_i\big)$. Since the treatment indicator T_i is a binary variable, this can be rewritten as

$$= \frac{\mathbb{E}\big[Y_i \mid T_i = 1\big] - \mathbb{E}\big[Y_i \mid T_i = 0\big]}{\mathbb{E}\big[X_i \mid T_i = 1\big] - \mathbb{E}\big[X_i \mid T_i = 0\big]} \tag{9}$$

Recall that the treatment indicator T_i is defined such that $T_i = 1 \iff Z_i = A$ and $T_i = 0 \iff Z_i = B$, and substitute in the linear outcome equation in (1). This gives the ratio

$$= \frac{\mathbb{E}\big[\tau_i\big(X_i(A) - X_i(B)\big)\big]}{\mathbb{E}\big[\big(X_i(A) - X_i(B)\big)\big]} \tag{10}$$

To make the role of the heterogeneity clear, suppose that the effects of beliefs are constant such that $\tau_i = \tau$. Then we can simplify (10) further:

$$\frac{\mathbb{E}\big[\tau\big(X_i(A) - X_i(B)\big)\big]}{\mathbb{E}\big[\big(X_i(A) - X_i(B)\big)\big]} = \frac{\tau\,\mathbb{E}\big[\big(X_i(A) - X_i(B)\big)\big]}{\mathbb{E}\big[\big(X_i(A) - X_i(B)\big)\big]} = \tau$$

This is the familiar result that the TSLS identifies the coefficient on the endogenous variable in the outcome equation when this coefficient is a constant. This is an extremely strong restriction in most applied settings, including the ones studied in this paper. For example, in a study of beliefs about home value, constant effects of beliefs would imply that every homeowner would respond the same way to learning that their house is worth $\$100,000$ more than they thought. In reality, people are likely to respond differently to this news: some may choose to sell their house, others may choose to take out a home equity loan, still others may change their consumption or savings behavior, and so on.

In the presence of heterogeneous effects, TSLS is not generally a consistent estimator of the average of the heterogeneous coefficients on the endogenous variable. Instead, the TSLS coefficient (like the reduced form) depends not only on the effects of beliefs on behaviors τ_i but also on the covariance between the individual first stage effects of the information treatment on beliefs $X_i(A) - X_i(B)$ and the individual effect of beliefs τ_i. With the aim of interpreting β^{TSLS} as a weighted average of individual effects of beliefs τ_i, we can rewrite the ratio in (10) as

$$\mathbb{E}\left[\tau_i \cdot \frac{\big(X_i(A) - X_i(B)\big)}{\mathbb{E}\big[\big(X_i(A) - X_i(B)\big)\big]}\right] \tag{11}$$

The weights on the effects of beliefs on behaviors τ_i are proportional to the individual first stage effects of the information treatment on beliefs $X_i(A) - X_i(B)$. The TSLS weights in (11) are the weights in the reduced form (7) after normalizing by their average.

Under instrument monotonicity, the TSLS coefficient is a non-negatively weighted average of individual causal effects τ_i. Formally, the instrument monotonicity assumption is that either $X_i(A) - X_i(B) \geq 0$ or $X_i(A) - X_i(B) \leq 0$ uniformly for all i (Angrist and Imbens, 1995). If some people have higher beliefs in treatment arm A and others have higher beliefs in treatment arm B, then the instrument monotonicity assumption is violated.

Bayesian Updating and Instrument Monotonicity. The TSLS weights in (11) are strictly non-negative (and have the same sign in the reduced form 7) when the instrument monotonicity assumption holds. We can use the Bayesian updating model of the first stage to study the plausibility of this assumption in active and passive designs. Substituting the Bayesian

updating model of the first stage (2) into the expression for the weights in (11) gives $X_i(A) - X_i(B) = \alpha_i\big(S_i(A) - S_i(B)\big)$.

We can use this to write the TSLS coefficient as

$$\beta_{T_i}^{TSLS} = \mathbb{E}\left[\tau_i \cdot \underbrace{\alpha_i\big(S_i(A) - S_i(B)\big)\Omega_T^{-1}}_{\text{weights}}\right] \tag{12}$$

where $\Omega_T \equiv \mathbb{E}\big[\alpha_i(S_i(A) - S_i(B))\big]$ is the average strength of the first stage, which ensures the weights integrate to one.

Bayesian updating implies that $0 < \alpha_i < 1$. The sign of these weights therefore depends on the sign of $S_i(A) - S_i(B)$. In an active design, the monotonicity assumption holds (or is violated) by construction. If the treatment arms are coded such that $S_i(A) > S_i(B)$ or $S_i(A) < S_i(B)$ uniformly for all i, then the monotonicity assumption follows directly from Bayesian updating.

In a passive design, we use $S_i(B)$ to denote the prior X_i^0 for the group that does not receive information. Substituting this in to the expression for the weights, we see that the sign of the weights depends on the sign of $S_i(A) - X_i^0$. These weights are positive when $S_i(A) - X_i^0$ has the same sign for all participants. This is equivalent to the requirement that the signal is either above (or below) the prior for every person, which is often hard to justify empirically. [6] Instead of Bayesian updating, Vilfort and Zhang (2023) directly assume that people update beliefs weakly in the direction of the signal: $S_i(z) \geq X_i^0 \implies X_i \geq X_i^0$ and show that non-negative weights require the same additional assumption about priors. Instrument monotonicity is for this reason difficult to justify in passive designs.

2.2. Constructing TSLS Estimates with Non-Negative Weights

To ensure that the weights are non-negative, researchers will use information on prior beliefs to construct alternative estimators. One strategy uses prior beliefs to split the sample based on the direction that the signal shifts beliefs. Another strategy constructs an exposure-weighted instrument, where the treatment indicator is interacted with the

[6] For example Cullen et al. (2022), Cullen and Perez-Truglia (2022), and Fuster et al. (2022) observe people with prior beliefs above and below the signal.

difference between the difference between the two signals (active designs) or the signal and the prior (passive designs). Vilfort and Zhang (2023) propose a general framework to use priors and signals to construct specifications with non-negative weights. This framework includes the split sample and exposure-weighted designs as special cases.

These strategies have non-negative weights under the same assumption: people update their beliefs towards the signal. This is an immediate implication of the Bayesian updating model, which implies that the learning rate α_i is strictly between zero and one. In Appendix A, I show that the Bayesian updating assumption can be used to construct the split sample estimator even when priors are not elicited. The sign of the shift in beliefs can be inferred from the posterior beliefs and the signal.

Ensuring Non-Negative Weights by Splitting the Sample. Consider estimating the regression in (8) separately for the groups with priors above and below the signal.

$$\beta_+^{\text{split}} \equiv \frac{\text{Cov}\big(T_i, Y_i \mid S_i(A) - X_i^0 > 0\big)}{\text{Cov}\big(T_i, X_i \mid S_i(A) - X_i^0 > 0\big)} \tag{13}$$

$$= \frac{\mathbb{E}\big[\tau_i \alpha_i \big| S_i(A) - X_i^0\big| \;\big|\; S_i(A) - X_i^0 > 0\big]}{\mathbb{E}\big[\alpha_i \big| S_i(A) - X_i^0\big| \mid S_i(A) - X_i^0 > 0\big]} \tag{14}$$

$$= \mathbb{E}\big[\tau_i \cdot \alpha_i \big| S_i(A) - X_i^0\big| \Omega_+^{-1} \;\big|\; S_i(A) - X_i^0 > 0\big] \tag{15}$$

and symmetrically

$$\beta_-^{\text{split}} \equiv \frac{\text{Cov}\big(T_i, Y_i \mid S_i(A) - X_i^0 < 0\big)}{\text{Cov}\big(T_i, X_i \mid S_i(A) - X_i^0 < 0\big)} \tag{16}$$

$$= \frac{\mathbb{E}\big[\tau_i \cdot \alpha_i \big| S_i(A) - X_i^0\big| \;\big|\; S_i(A) - X_i^0 < 0\big]}{\mathbb{E}\big[\alpha_i \big| S_i(A) - X_i^0\big| \mid S_i(A) - X_i^0 < 0\big]} \tag{17}$$

$$= \mathbb{E}\big[\tau_i \cdot \alpha_i \big| S_i(A) - X_i^0\big| \Omega_-^{-1} \;\big|\; S_i(A) - X_i^0 < 0\big] \tag{18}$$

where $\Omega_+ \equiv \mathbb{E}\big[\alpha_i \big| S_i(A) - X_i^0\big| \mid S_i(A) - X_i^0 > 0\big]$ and $\Omega_- \equiv \mathbb{E}\big[\alpha_i \big| S_i(A) - X_i^0\big| \mid S_i(A) - X_i^0 < 0\big]$ are constants that normalize the weights to integrate to one. Under Bayesian updating, β_+^{split} and β_-^{split} can be interpreted as non-negatively weighted averages of individual partial effects τ_i.

Notice that the priors X_i^0 are used only to construct the subsamples, and that the sign of the difference between the signal and prior is sufficient for this conditioning. After conditioning on the sign of $S_i(A) - X_i^0$, the same TSLS regression as in (8) is estimated. Bayesian updating implies that feasible estimators for the split estimators (13), (16) can be constructed even when the prior belief is not directly observed. If the posterior lies between the prior and the signal, the direction that the signal shifted beliefs can be inferred from the posterior and the signal. See Appendix A for further discussion of this strategy to ensure positive weights in a passive design when priors are not elicited.

Ensuring Non-Negative Weights by Constructing an "Exposure" Instrument. Another way that information on the prior can be incorporated is by constructing an exposure-weighted instrument. In an active design, the exposure $(S_i(A) - S_i(B))$ is the difference between the two signals (Bottan and Perez-Truglia, 2022b; Roth et al., 2022). In a passive design, the exposure $(S_i(A) - S_i(B)) = (S_i(A) - X_i^0)$ is the difference between the signal and the prior (Armona et al., 2019; Cullen and Perez-Truglia, 2022). The exposure-weighted instrument interacts the treatment indicator with the exposure to treatment.

$$\Delta_i \equiv T_i(S_i(A) - S_i(B)) \tag{19}$$

In order to be uncorrelated with the structural error term U_i, this instrument needs to be re-centered; this can be done manually or by including $(S_i(A) - S_i(B))$ as a control as in Cullen and Perez-Truglia (2022). After re-centering, the instrument is

$$\widetilde{\Delta}_i \equiv \left(T_i - \mathbb{E}\left[T_i\right]\right)(S_i(A) - S_i(B)) \tag{20}$$

The TSLS target parameter is then

$$\beta_{\Delta}^{TSLS} \equiv \frac{\text{Cov}\left(\widetilde{\Delta}_i, Y_i\right)}{\text{Cov}\left(\widetilde{\Delta}_i, X_i\right)} \tag{21}$$

$$= \frac{\mathbb{E}\left[\tau\alpha_i(S_i(A) - S_i(B))^2\right]}{\mathbb{E}\left[\alpha_i(S_i(A) - S_i(B))^2\right]} \tag{22}$$

$$= \mathbb{E}\left[\tau \cdot \alpha_i(S_i(A) - S_i(B))^2\Omega_{\Delta}^{-1}\right] \tag{23}$$

where $\Omega_\Delta \equiv \mathbb{E}\left[\alpha_i(S_i(A) - S_i(B))^2\right]$ is a normalizing constant. Instead of weighting individuals proportional to the absolute value of the first stage (as in the split sample approach), the exposure-weighted instrument weights them proportional to $\alpha_i(S_i(A) - S_i(B))^2$, which is quadratic in the difference between the signals (active designs) or the difference between the signal and the prior (passive designs). These weights are non-negative, but put even more weight on outliers.

Like in the split sample regressions, these weights are non-negative under the assumption of Bayesian updating. Substantively, this is an assumption that people update beliefs uniformly towards the signal. Bayesian updating can then be relaxed so long as the assumption that $\alpha_i \geq 0$ is maintained.[7]

2.3. Interpreting TSLS Estimates with Non-Negative Weights

$\beta_{\pm}^{\text{split}}$ and β_{Δ}^{TSLS} both aggregate individual partial effects τ_i by weighting them proportionally to α_i. In Bayesian updating model, this puts the most weight on those with the least precise priors. The difference between the parameters targeted by the split regressions and the exposure weighted instrument is whether the term that depends on the difference $S_i(A) - S_i(B)$ is guaranteed to be non-negative by taking the absolute value of the difference ($\beta_{\pm}^{\text{split}}$) or by taking the square of this difference (β_{Δ}^{TSLS}). Relative to the split regressions, the exposure weighted instrument puts more weight on outliers. In either case these estimates depend on endogeneity in both the learning rate α_i and the difference in signals $S_i(A) - S_i(B)$ (active) or the prior $S_i(A) - X_i^0$ (passive).

2.3.1. Endogenous Belief Formation Through Costly Information Acquisition

The parameters identified by TSLS depend on the learning rate α_i. In passive designs, these weights also depend on the content of the prior beliefs. In a broad class of models of endogenous belief formation, people with large causal effects of beliefs $|\tau_i|$ endogenously form precise beliefs. The following example illustrates the key dynamic that causes people

[7] Weights will also be non-negative if every individual updates uniformly *away* from the signal (i.e. $\alpha_i \leq 0$ for all i). Negative weights arise if people have an individual first stage that is of a different sign than the average first stage.

to endogenously form more precise beliefs when these beliefs have a strong effect on their behavior.

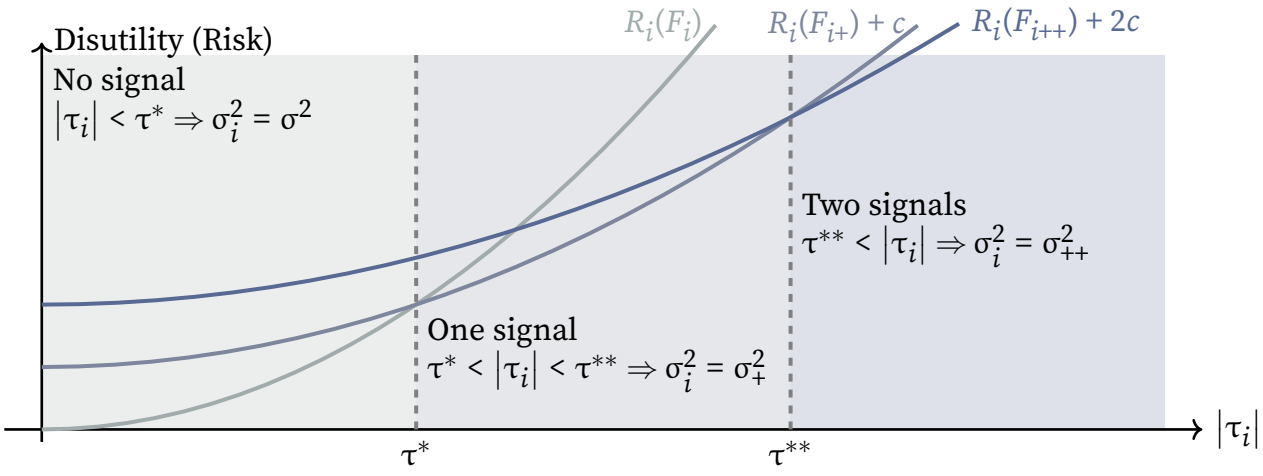
Example: Alice checks the weather because she (sometimes) bikes to work. Bob drives to work every day. However, Alice drives to work only if the weather is bad and bikes when the weather is nice. Alice checks the weather before she leaves the house so that she can decide how she wants to commute to work. Bob gets into his car and starts driving without needing to check the weather. Alice and Bob are sampled into an information provision experiment; the experimenter gives both Alice and Bob an informative signal about the weather for the day. Since Alice had already checked the weather for the day, she changes her beliefs only slightly. When Bob is shown new information, he significantly revises his beliefs since he did not check the weather before the experiment.

On average, this intervention shifts beliefs, but it does not shift behaviors; the experimenter may then erroneously conclude that the mode of transportation does not depend on beliefs about the weather. However, we know that Alice chooses how to commute based on her belief about the weather conditions. People whose actions depend on their beliefs will invest time and energy into forming precise beliefs, before the experiment takes place. Then, at the time of the experiment, the people who are most responsive to new information are the people whose decision-making depends only weakly on their beliefs.

Models of Endogenous Belief Formation via Costly Information Acquisition. People have a subjective belief distribution given by $F_i(\cdot)$. People are uncertain about their beliefs, and this uncertainty about their beliefs generates uncertainty about the action that they would like to take. Let $R_i(F)$ denote the subjective risk or ex-ante regret (for example, the expected loss) that individual i faces when their beliefs are given by F.

We make the following assumptions on $R_i(\cdot)$. First, uncertainty is costly: $\frac{\partial R_i}{\partial \sigma^2} \geq 0$, where $\frac{\partial R_i}{\partial \sigma^2} = 0$ if and only if $\tau_i = 0$. Second, since there is uncertainty in beliefs, it is costly to base behavior on these beliefs: $\frac{\partial R_i}{\partial |\tau_i|} \geq 0$, where $\frac{\partial R_i}{\partial |\tau_i|^2} = 0$ if and only if $\sigma^2 = 0$. Finally, uncertainty is more costly for people whose beliefs affect actions more: $\frac{\partial^2 R_i}{\partial \sigma^2 \partial |\tau_i|} > 0$.

FIGURE 1. People with Large Effects of Beliefs τ_i Form Precise Beliefs



Notes: This figure plots the loss as a function of $|\tau_i|$ after seeing no signals, one signal, and two signals. The assumptions on R_i ensure that each pair of lines crosses exactly once. Since $R_i(F) = R_i(F_+)$ when $\tau_i = 0$, $R_i(F) < R_i(F_+) + c$. If $\sigma_{++}^2 > 0$, these curves are all strictly increasing in $|\tau_i|$ by assumption. Additionally, since $\sigma^2 > \sigma_+^2 > \sigma_{++}^2$, then $R_i(F)$ is steeper than $R_i(F_+)$, which is steeper than $R_i(F_{++})$ by the assumption that $\frac{\partial^2 R_i}{\partial \sigma^2 \partial |\tau_i|} > 0$.

People make a decision to pay a cost $c > 0$ to obtain new information or to do nothing. There is an updating process such that the variance σ_+^2 of beliefs after viewing a signal F_+ is less than the variance σ^2 of the initial beliefs F. People then trade off the reduction in risk from the new information against the cost of the signal. Thus, when person i has beliefs F, her loss can be given recursively by

$$V_i(F) = \min \left\{ R_i(F), V_i(F_+) + c \right\} \tag{24}$$

Given the assumptions we have made on R_i, for any beliefs F with $\sigma^2 > 0$, there is some threshold value τ^* such that people with $|\tau_i| > \tau^*$ prefer to pay c to update their beliefs. That such a threshold exists is guaranteed by the fact that $R_i(F) = R_i(F_+)$ when $\tau_i = 0$, which this implies that $R_i(F) < R_i(F_+) + c$ at $\tau_i = 0$. However, since $\frac{\partial^2 R_i}{\partial \sigma^2 \partial |\tau_i|} > 0$, we also know that $\frac{\partial R_i(F)}{\partial |\tau_i|} > \frac{\partial R_i(F_+)}{\partial |\tau_i|}$ since $\sigma^2 > \sigma_+^2$.

At $\tau_i = 0$, $R_i(F)$ is below $R_i(F_+) + c$. However, $R_i(F)$ is increasing faster than $R_i(F_+)$ in $|\tau_i|$ such that eventually these curves will cross. And since $R_i(F)$ is always increasing faster

than $R_i(F_+)$ in $|\tau_i|$, they will cross exactly once. Figure 1 illustrates this graphically. When beliefs are formed through such a process, people with larger causal effects of beliefs will have (weakly) more precise beliefs in equilibrium.

Using Models of Belief Formation and Updating to Interpret TSLS Estimates. The class of parameters that are targeted by existing TSLS specifications depend not only on the causal effects of beliefs on outcomes τ_i, but also on any endogeneity in the way that beliefs are updated in response to new information. In the toy model discussed in Section 2.3.1, beliefs are formed endogenously through a process of costly information acquisition. In Appendix B, I solve a special case of this model where the subjective risk R_i is given by the expected quadratic loss. Parameterizing the loss function makes it possible to solve analytically for the learning rate α_i and variance of the prior σ_i^2 as a function of the causal effects of beliefs τ_i.

People have inaccurate and imprecise beliefs precisely because they have small individual partial effects (small $|\tau_i|$); when beliefs are an important determinant of the behaviors (large $|\tau_i|$), people exert effort to form accurate and precise beliefs. In this environment, parameters with weights proportional to the strength of the shift in beliefs will be attenuated and underestimate the magnitude of the average effect.

Alternative models of the relationship between belief updating and the effects of beliefs on behaviors can be used to relate causal parameters estimated via TSLS to the APE. For example, Fuster et al. (2022) allow variation in the learning rate to come from a more complicated model that adds dynamics of rational inattention to costly information acquisition. Any model that makes predictions about the covariance between the learning rate α_i and the causal effect of beliefs on behavior τ_i can be used to make predictions about the difference between TSLS estimates and the APE. Only in the special case when the heterogeneity in the first stage is uncorrelated with individual causal effects of interest will TSLS identify the average partial effect.

3. Identifying the Average Partial Effect of Beliefs on Behavior

Since the parameters estimated in common TSLS specifications can understate the causal effects of beliefs on behaviors, researchers hoping to learn about these effects may be interested in targeting alternative parameters that are informative about the general population, and not just the subset whose beliefs are most responsive to new information. One contribution of this paper is establish the conditions under which the average partial effect $\mathbb{E}[\tau_i]$ is point identified and to provide an estimator.

There are two main conditions: first, the identification argument will rely on a conditional rank invariance that is implied by Bayesian updating. Second, since the observed update in beliefs is used to construct a control for the endogeneity in belief updating, the proposed estimator is available in active designs. I suggest how this estimator may be extended to other experimental designs but leave the formal development of these arguments to future work. In the remainder of this section, I develop the identification argument and formally state the necessary assumptions.

3.1. Identifying the Average Partial Effect via Local Least Squares

Recall that people are randomly assigned to a high signal $S_i(A)$ or low signal $S_i(B)$. The potential outcome for the posterior is

$$X_i(z) = X_i^0 + \alpha_i\left(S_i(z) - X_i^0\right)$$

I will use a two-step approach to identify $\mathbb{E}[\tau_i]$ following Masten and Torgovitsky (2016). First, I will construct a control function R_i such that $\frac{\text{Cov}\left(Y_i, X_i | R_i=r\right)}{\text{Var}\left(X_i | R_i=r\right)} = \mathbb{E}[\tau_i \mid R_i = r]$. Then, I will integrate over the support of the control function to identify $\mathbb{E}[\tau_i]$. The key step in constructing this control function is to condition on the prior, the potential signals and the assigned treatment and then use the remaining variation in X to identify the conditional

rank of α_i. [8] We can construct such a control function as follows:

$$R_{i1} = F_{\alpha|W,Z}(\alpha_i) = \begin{cases} F_{X|W,Z}(X_i) & S_i(Z_i) - X_i^0 > 0 \\ 1 - F_{X|W,Z}(X_i) & S_i(Z_i) - X_i^0 < 0 \end{cases} \tag{25}$$

$$W_i \equiv \begin{bmatrix} X_i^0 & S_i(A) & S_i(B) \end{bmatrix} \tag{26}$$

$$R_i \equiv \begin{bmatrix} R_{i1} & W_i \end{bmatrix} = \begin{bmatrix} F_{\alpha|W}(\alpha_i) & X_i^0 & S_i(A) & S_i(B) \end{bmatrix} \tag{27}$$

Conditional on the prior and potential signals W_i, and the assigned treatment Z_i, all the remaining variation in the posterior X_i is generated by the learning rate α_i. Notice that when $S_i(Z_i) = X_i^0$, $X_i = X_i^0$ and thus X_i is not informative about α_i. This means that in passive designs where the control group does not recieve new information, the learning rate is not identified in the control group and so this estimator is not feasible.[9]

If $\text{Var}\big(X_i \mid R_i = r\big) \neq 0$ for all r, then it follows immediately from the law of iterated expectations that $\mathbb{E}\big[\mathbb{E}\big[\tau_i \mid R = r\big]\big] = \mathbb{E}\big[\tau_i\big]$ so $\mathbb{E}\big[\tau_i\big]$ is identified. If $\alpha_i > 0$ for all i, which in the Bayesian model is equivalent to assuming that prior beliefs are not degenerate, then it is sufficient to design the experiment such that $S_i(A) \neq X_i^0$, $S_i(B) \neq X_i^0$, $S_i(A) \neq S_i(B)$, and $0 < \mathbb{P}\big[Z_i = 1\big] < 1$ to ensure that $\text{Var}\big(X_i \mid R_i = r\big) \neq 0$.[10] If any of these do not hold, then the conditional average partial effect can be identified for the subset $\mathcal{R}$ such that $\text{Var}\big(X_i \mid R_i = r\big) \neq 0$ for $r \in \mathcal{R}$. The two-stage control function is an alternative way to use the variation induced by the instrument; instrument relevance is thus still required.

[8] In some experiments, the variance of the prior is elicited. If the variance of the prior is sufficient to explain all of the heterogeneity in the learning rates α_i (i.e. the normal-normal Bayesian micro-foundation of the updating equation is correctly specified), then the variance of the prior can be used directly as a control function instead of the rank of α. I discuss below, using the rank of the learning rate is robust to certain kinds of misspecification in the updating equation. In this sense, the rank of the learning rate can be used to construct a valid control function under weaker assumptions than are needed to directly use the variance of the prior. However, using the variance of the prior may be one way to implement this estimator in passive designs where the learning rate is not identified among the control group that does not receive information.

[9] In passive designs where pre and post-treatment outcomes are available in addition to pre and post-treatment beliefs (Cullen et al., 2022; Wiswall and Zafar, 2015), a closely related strategy that uses the within-person changes in beliefs and outcomes would be feasible . In this case, the pre-treatment observations of beliefs and outcomes are used as the control group. Noting that the extension is immediate, I focus on the more common case where the outcome is only observed post treatment.

[10] Recall that our notation uses $S_i(B) \equiv X_i^0$ in passive designs. This assumption thus rules out passive designs.

ASSUMPTION 1.

(a) ***Instrument Exogeneity*** $Z_i \perp\!\!\!\perp \left(X_i^0, S_i(\cdot), U_i, \alpha_i, \tau_i\right)$

(b) ***Linear Outcome (eq. 1)*** *The outcome is linear in beliefs; partial effects are heterogeneous.*
$Y_i = \tau_i X_i + U_i$

(c) ***Linear Updating (eq. 2)*** *Beliefs are updated following* $X_i(z) = X_i^0 + \alpha_i\left(S_i(z) - X_i^0\right)$

(d) ***Relevance*** *The instrument always induces variation in beliefs.*
$\alpha_i \in (0,1)$; $S_i(A) \neq X_i^0$, $S_i(B) \neq X_i^0$, $S_i(A) \neq S_i(B)$ *and* $\mathbb{P}\left[Z_i = 1\right] \in (0,1)$

(e) ***Existence*** $\mathbb{E}\left[\tau_i\right]$ *and* $\mathrm{Var}\left(X_i\right)$ *exist and are finite*

PROPOSITION 1. *Under Assumption 1,* $\frac{\mathrm{Cov}\left(Y_i, X_i | R_i = r\right)}{\mathrm{Var}\left(X_i | R_i = r\right)} = \mathbb{E}\left[\tau_i \mid R_i = r\right]$.

The proof is an application of the argument in Masten and Torgovitsky (2014) conditional on the sign of the difference between the signal and the prior (Appendix C).

Having identified $\mathbb{E}\left[\tau_i \mid R_i = r\right]$, one can identify $\mathbb{E}\left[\tau_i\right]$ by integrating over the support of R. Since W_i is directly observed, and (25) shows that the remaining entry R_{i1} is also identified, the regression of Y_i on X_i conditional on $R_i = r$ is feasible. Of course, since $\mathbb{E}\left[\tau_i | R_i = r\right]$ is identified, heterogeneity analysis is immediately available by aggregating over only a subset of possible values of the control function. This can be used to empirically study the relationship between the learning rate α_i and the causal effects of beliefs τ_i.

For ease of exposition, I have maintained in Assumption 1.c that belief updating is linear, though it is not necessarily Bayesian since there is no structure on the learning rates α_i. However, it is not necessary for belief updating to be linear.

The APE will still be identified as long as belief updating satisfies a rank invariance condition (Masten and Torgovitsky, 2014, 2016). Conditional on the prior and the sign of the difference between the signal and the prior, the distribution of posterior beliefs must be rank invariant across potential signals. Suppose Chris and Dianne have the same prior and recieve the same signal that is above their prior. This assumption states that if Chris's posterior is higher than Dianne's, then Chris's posterior would also be higher than Dianne's

if they were both to recieve another signal that is also above their prior (and lower than Dianne's if they recieve a signal that is below their prior).

These dynamics follow from Bayesian updating model; the observed ranking would be rationalized by assigning Chris a higher learning rate, which would then generate the needed predictions. This structure is equivalent to the statement in (25) that identifies the conditional rank of α_i through the conditional rank of the posterior X_i. Bayesian updating is a familiar model that implies rank invariance and thus is a sufficient assumption, but rank invariance of this sort can also be assumed directly.

When the outcome equation is nonlinear, the control function will still ensure that there is no bias due to correlation between heterogeneity in the first stage and heterogeneity in the outcome equation. Regression conditional on $R_i = r$ will recover a best linear predictor of the outcome given beliefs; the coefficient of interest is a positively weighted average of the derivative of the conditional expectation function (Yitzhaki, 1996). Integrating again over R_i preserves the interpretation as a positively weighted average.

4. An Application to the Effect of Beliefs on Demand for Public Policy

Settele (2022) studies the effect of beliefs about the gender wage gap on demand for policies aimed at increasing gender equality. Settele randomly assigns some participants to receive a high signal of women's relative wages – that women's wages are 94% of men's wages – and some to a low signal of women's wages – that they are 74% of men's wages. [11] The main estimates use the comparison between the groups that receive high and low signals of relative earnings for women to estimate the causal effects of beliefs.

In my preferred specification, the LLS estimate of the APE of a standard deviation shift in beliefs on a policy demand index is roughly 70% larger in magnitude than the corresponding TSLS estimate. The imprecisely estimated heterogeneous treatment effects are also consistent with the theoretical predictions. The average partial effect for people in the top fifth of the learning rate α_i distribution is close to zero (0.003). The average partial

[11] These estimates come from the most recent CPS and ACS survey at the time of the experiment, and so are generated without deception. Akesson et al. (2022) and Bottan and Perez-Truglia (2022b) also use this source randomization strategy to generate variation in potential signals without deception. See Settele (2022) for more information about the design of the experiment.

effect for the bottom 60% of the learning rate (-0.28) is almost triple the corresponding TSLS estimate (-0.105). However, I estimate that TSLS places slightly more weight on the top 20% of the learning rate distribution than the bottom 60%.

In the rest of this section, I discuss how the LLS estimator can be implemented and applied to this setting. I then present and compare LLS and TSLS estimates.

4.1. Estimating the Control Function

In the first stage, I use a kernel to estimate the CDF of the posterior conditional on a particular value of the prior. I estimate this conditional CDF at every value of the prior in the data and use the these estimates to recover the rank of the unobservable learning rate α_i.

Then, in the second stage, I condition on the first-stage estimates of the conditional rank of the learning rate using a kernel. Then, I estimate the intermediate conditional average partial effect $\mathbb{E}\left[\tau_i \mid F_{\alpha_i|W_i}(\alpha_i) = r_1\right]$ by regressing the policy index Y_i on the posterior belief X_i and the prior X_i^0.[12] Conditional on the learning rate and the prior, the only remaining variation in beliefs comes from the value of the randomly assigned signal. Then, I estimate the average partial effect by numerically integrating over the distribution of ranks. To estimate standard errors, I bootstrap this procedure following Masten and Torgovitsky (2016).

Additional Covariates from Settele (2022) Are Omitted. To avoid complications from including additional covariates, I ignore the additional controls that Settele (2022) includes in the main specification reported in Table 5 of the paper. In Appendix D.1, I compare results in this specification without additional covariates to the full paper specification. Point estimates are somewhat larger in magnitude without these additional controls, but results are qualitatively similar. Thus, comparisons between the APE that I estimate and TSLS

[12] If the linear updating model is correctly specified, then it is sufficient to control for the prior linearly, which eases estimation. It is possible to condition on the prior non-parametrically by using, for example, a two-dimensional kernel to condition on both $F(\alpha_i)$ and the prior X_i^0 and then regressing the outcome directly on the posterior in the intermediate regression. This is a more data intensive approach that will use fewer observations for each intermediate regression.

estimates will use results from this more parsimonious specification, and not the results reported in Settele (2022).

4.2. Empirical Evidence that TSLS Underestimates the APE

The main TSLS specification in Settele (2022) Table 5 uses an indicator for assignment to the "low" signal as an instrument for posterior beliefs. Since the signals are common ($S_i(z) = S(z)$), TSLS targets

$$\mathbb{E}\left[\tau_i \cdot \widetilde{\alpha}_i\right] \tag{28}$$

where $\widetilde{\alpha}_i \equiv \alpha_i / \mathbb{E}\left[\alpha_i\right]$ is the learning rate α_i normalized by the mean $\mathbb{E}\left[\alpha_i\right]$ so that these weights integrate to 1. [13] Under the Bayesian learning assumption maintained throughout (in particular the assumption that $0 \leq \alpha_i$), this is a non-negatively weighted average. Since these weights are given by the learning rate α_i, the toy model in Section 2.3.1 predicts that the TSLS estimates will be smaller in magnitude than the average partial effect. Replicating the main specification from the paper without covariates yields an estimate of –0.105.[14]

In Panel A of Figure 2 I report the local least squares (LLS) estimate of the average partial effect of beliefs on the policy demand index. The control function estimate of the APE is –0.178, which is roughly 70% larger in magnitude than the TSLS estimate of –0.105. I also plot binned estimates of the conditional average partial effect for different ranks of the learning rate α_i. The conditional average partial effect (CAPE) is the average partial effect for people at the r_1th rank of the learning rate $\mathbb{E}\left[\tau_i \mid F_{\alpha_i|W_i}(\alpha_i) = r_1\right]$. For ease of interpretation, I present binned estimates of the CAPE in five equally sized bins.[15] In Appendix Figure D.1 Panel A, I present the full set of estimates of the CAPE at every value of the learning rate in the sample.

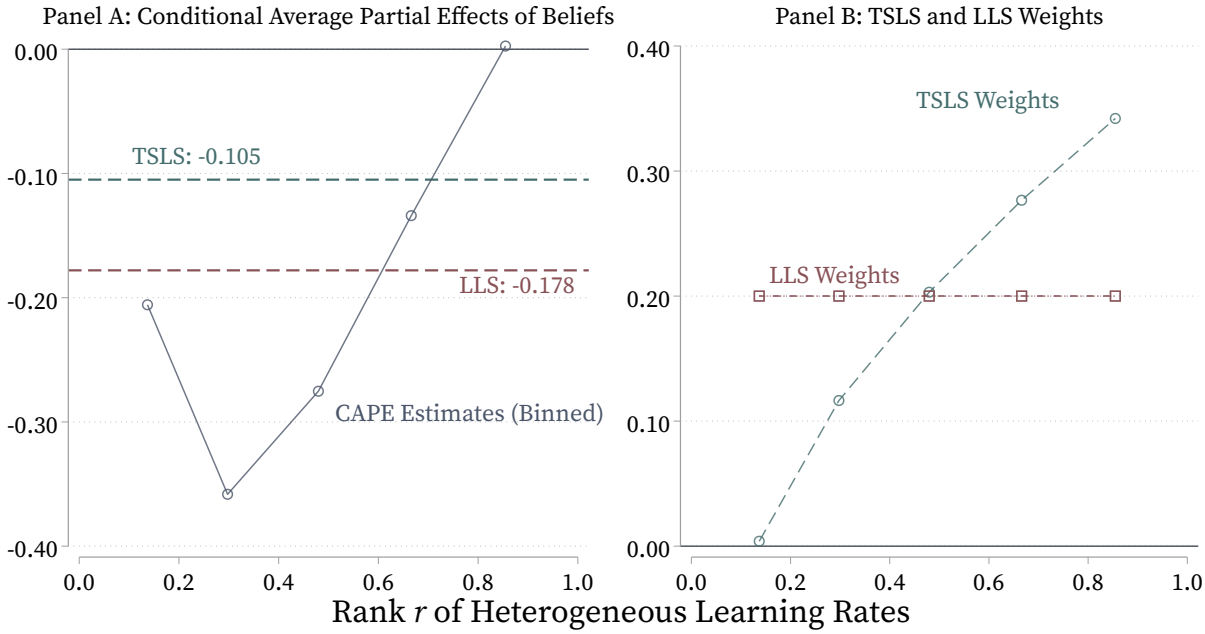
The estimates of the conditional average partial effect are noisy, but the estimates are consistent with the predictions of the model in Section 2.3.1. In particular, notice that

[13] Since the difference in signals is a constant, it cancels from both the individual weights and the normalization term in the general expression in (11).

[14] Following the original paper, estimates throughout are scaled to be relative to a standard deviation change in beliefs.

[15] When constructing these bins, I exclude the top 0.05 and bottom 0.05 ranks of the learning rate, since the bandwidth begins to truncate within this range, making estimates within these ranges particularly noisy.

FIGURE 2. LLS Estimates of the Average Partial Effect are 70% Larger than TSLS



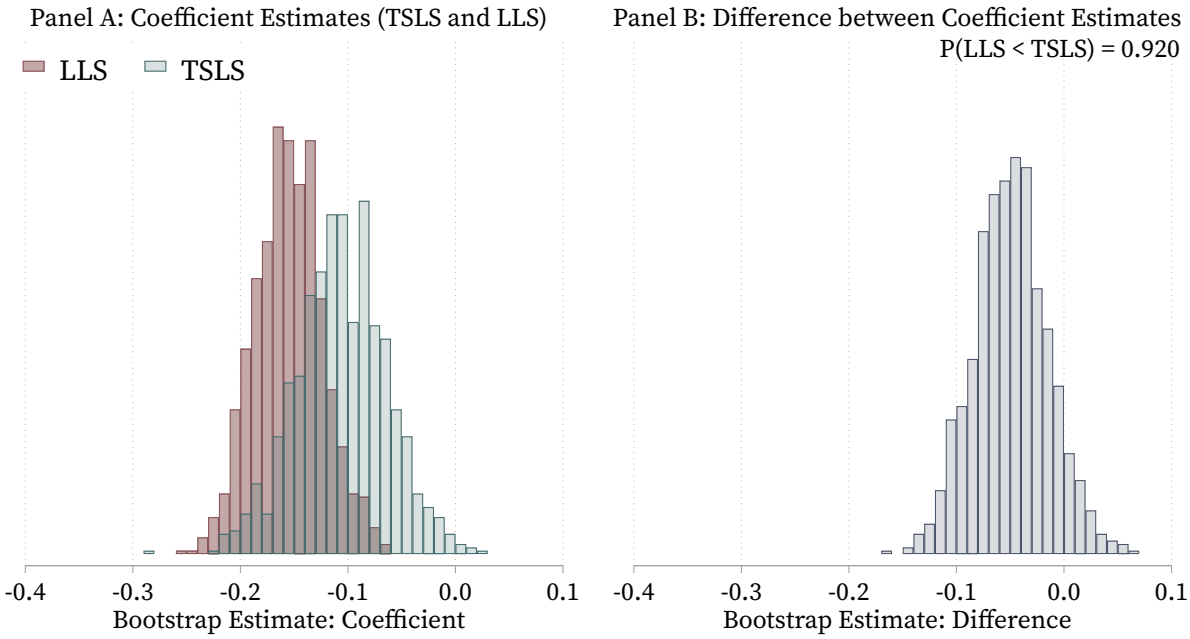
Notes: Panel A presents a binned scatter plot of conditional average partial effect $\mathbb{E}\left[\tau_i \mid F_{\alpha_i|W_i}(\alpha_i) = r_1\right]$ for different ranks r_1 of the learning rate α_i. Panel B presents the weights that the TSLS estimator and the LLS estimators place on each bin. By construction, LLS places equal weight on each bin. Appendix Figure D.1 reports the raw CAPE estimates at every r without binning. Bins are constructed excluding the top and bottom 0.05, since the bandwidth begins to truncate within these ranges as the rank r approaches the boundary.

the conditional average partial effect is close to zero for the top 20% of the learning rate distribution but is quite large for the bottom 40%.

Across 1,000 bootstrap iterations, the best linear predictor of the CAPE curve has a positive slope in 97% of iterations; the fifth percentile is 0.019. Since the LLS estimate of the APE is negative at every iteration, this bootstrap evidence suggests that the magnitude of the causal effects is decreasing as the rank of the learning rate increases.

In Panel B of Figure 2, I estimate the TSLS weights associated with each of the five bins. I also include the LLS weights for each group to highlight the fact that LLS places equal weight on each bin by construction. Comparing Panels A and B shows that the TSLS weights are correlated with the heterogeneity in the estimated casual effects. TSLS places roughly equal weight on the bin with the largest learning rate (and causal effects close

FIGURE 3. TSLS is Attenuated Relative to LLS in 92% of Bootstrap Iterations



Notes: Panel A plots the marginal bootstrap distributions of TSLS and LLS estimates after 1000 iterations. Panel B plots the bootstrap distribution of the difference between the LSS and TSLS estimates. The bootstrap standard deviation is 0.042 for the TSLS estimator and 0.031 for the LLS estimator.

to zero) and the bottom three bins (where the causal effects are large). This is consistent with the theoretical mechanisms discussed throughout and is exactly the dynamic that will attenuate the TSLS estimates.

In Panel A of Figure 3, I plot the marginal distribution of TSLS and LLS estimates across 1,000 bootstrap iterations. With a bandwidth of 0.05, the LLS estimates appear more precisely estimated than the TSLS estimates. The marginal distributions do have some overlap, but are centered in different locations. Since these estimates are correlated, a comparison of the marginal distributions understates the difference. In Panel B of Figure 3, I plot the distribution of the difference between the LLS and TSLS estimates. This reveals that, even though the marginal distributions do overlap, the estimate are correlated such that the difference has the expected sign in 92% of iterations.

Given the power of these estimates, it is difficult to assess the extent to which the CAPE

curve is decreasing at the lowest levels and therefore forming a U shape. Estimates are particularly noisy for the people with the lowest ranks of the learning rate. Intuitively, since the learning rate is low, there is relatively less variation in the posterior belief, which makes estimating the regression of the outcome on beliefs more difficult. However, the true CAPE curve may indeed be U-shaped, so that the largest causal effects of beliefs are found among people around the 20^{th} percentile of the learning rate distribution, and not at the bottom.

A richer model of belief formation and updating may then be necessary to explain the behavior of people with the lowest ranks of the learning rate. Perhaps a model that also includes rational inattention in the spirit of Fuster et al. (2022) can explain the behavior of people with the lowest learning rates. In such a model, some people with weak effects of beliefs on behavior would not update their beliefs at all in response to new information because of costs of processing information.

An interpretation proposed by Settele (2022) is "politically motivated" belief formation; some people form beliefs that allow them to sustain their desired political beliefs. This is an alternative mechanism that would rationalize small causal effects of beliefs among a group that does not update their beliefs very much. Formally modelling these dynamics could rationalize a U-shaped relationship between the learning rate and the effect of beliefs on behavior by introducing additional terms into the expression for the learning rate.

This discussion highlights the power of this approach beyond simply estimating the average partial effect. Regardless of the model that generates heterogeneous learning rates, so long as the unobservable heterogeneity in the first stage can be captured in the learning rate α_i, the conditional average partial effect curve can be estimated. In addition to providing estimates of the APE, these intermediate local least squares estimates allow for a rich empirical investigation of the relationship between belief formation and updating and the causal effect of beliefs on behavior.

5. Conclusion

Two-stage least squares estimates of the average partial effect of beliefs on behavior depend not only on the effects of beliefs, but also on endogeneity in belief updating. The sign of the bias is given by the covariance between the causal effects of beliefs and belief updating in the first stage. A simple model with costly information acquisition predicts that this covariance will be negative. People with strong causal effects endogenously form precise beliefs before the experiment and thus update beliefs less in response to the information treatment.

I confirm these predictions in an application to a recent study of the effect of beliefs about the gender wage gap on demand for public policy (Settele, 2022). Using an alternative local least squares estimator, I estimate that the average partial effect is almost 70% larger than the corresponding TSLS estimate. LLS estimates a larger average effect than TSLS because the people who update their beliefs the most have the smallest causal effects of beliefs and the largest weights in TSLS.

If this mechanism is present more widely, it suggests that common TSLS estimates understate the average strength of the causal effect of beliefs. This may explain puzzling results in the literature that find small or insignificant effects of beliefs on behavior despite having information treatments that have a large effect on beliefs (Alesina et al., 2023; Kuziemko et al., 2015).

References

Akesson, Jesper, Robert Hahn, Robert Metcalfe, and Itzhak Rasooly (2022). "Race and Redistribution in the United States: An Experimental Analysis", w30426. DOI: 10.3386/w30426 (p. 1, 8, 23).

Alesina, Alberto, Matteo F Ferroni, and Stefanie Stantcheva (2021). "Perceptions of Racial Gaps, Their Causes, and Ways to Reduce Them". *Working Paper* (p. 3, 4).

Alesina, Alberto, Armando Miano, and Stefanie Stantcheva (2023). "Immigration and Redistribution". *The Review of Economic Studies* 90.1, pp. 1–39. DOI: 10.1093/restud/rdac011 (p. 2, 29).

Alesina, Alberto, Stefanie Stantcheva, and Edoardo Teso (2018). "Intergenerational Mobility and Preferences for Redistribution". *American Economic Review* 108.2, pp. 521–554. DOI: 10.1257/aer.20162015 (p. 5).

Angrist, Joshua D. and Guido W. Imbens (1995). "Two-Stage Least Squares Estimation of Average Causal Effects in Models with Variable Treatment Intensity". *Journal of the American Statistical Association* 90.430, pp. 431–442. DOI: 10.1080/01621459.1995.10476535 (p. 1, 3, 9, 12).

Armona, Luis, Andreas Fuster, and Basit Zafar (2019). "Home Price Expectations and Behaviour: Evidence from a Randomized Information Experiment". *The Review of Economic Studies* 86.4 (309), pp. 1371–1410 (p. 15).

Balla-Elliott, Dylan, Zoë B. Cullen, Edward L. Glaeser, Michael Luca, and Christopher Stanton (2022). "Determinants of Small Business Reopening Decisions After Covid Restrictions Were Lifted". *Journal of Policy Analysis and Management* 41.1, pp. 278–317. DOI: 10.1002/pam.22355 (p. 3, 6, 7).

Bottan, Nicolas L. and Ricardo Perez-Truglia (2022a). "Betting on the House: Subjective Expectations and Market Choices", p. 53. DOI: 10.3386/w27412 (p. 3, 6).

— (2022b). "Choosing Your Pond: Location Choices and Relative Income". *The Review of Economics and Statistics* 104.5, pp. 1010–1027. DOI: 10.1162/rest_a_00991 (p. 1, 15, 23).

Bursztyn, Leonardo, Alessandra L. González, and David Yanagizawa-Drott (2020). "Misperceived Social Norms: Women Working Outside the Home in Saudi Arabia". *American Economic Review* 110.10, pp. 2997–3029. DOI: 10.1257/aer.20180975 (p. 3).

Cavallo, Alberto, Guillermo Cruces, and Ricardo Perez-Truglia (2017). "Inflation Expectations, Learning, and Supermarket Prices: Evidence from Survey Experiments". *American Economic Journal: Macroeconomics* 9.3, pp. 1–35. DOI: 10.1257/mac.20150147 (p. 3, 7, 8).

Coibion, Olivier, Yuriy Gorodnichenko, Saten Kumar, and Jane Ryngaert (2021). "Do You Know That I Know That You Know…? Higher-Order Beliefs in Survey Data". *The Quarterly Journal of Economics* 136.3, pp. 1387–1446. DOI: 10.1093/qje/qjab005 (p. 3, 4).

Cruces, Guillermo, Ricardo Perez-Truglia, and Martin Tetaz (2013). "Biased Perceptions of Income Distribution and Preferences for Redistribution: Evidence from a Survey Experiment". *Journal of Public Economics* 98, pp. 100–112. DOI: 10.1016/j.jpubeco.2012.10.009 (p. 33).

Cullen, Zoë, Will Dobbie, and Mitchell Hoffman (2022). "Increasing the Demand for Workers with a Criminal Record". *The Quarterly Journal of Economics*. DOI: 10.1093/qje/qjac029 (p. 3, 7, 13, 21).

Cullen, Zoë and Ricardo Perez-Truglia (2022). "How Much Does Your Boss Make? The Effects of Salary Comparisons". *Journal of Political Economy* 130.3, pp. 766–822. DOI: 10.1086/717891 (p. 3, 7, 13, 15, 33).

Dechezlepretre, Antoine, Adrien Fabre, Tobias Kruse, Bluebery Planterose, Ana Sanchez Chico, and Stefanie Stantcheva (2023). "Fighting Climate Change: International Attitudes Toward Climate Policies" (p. 3, 5).

Fuster, Andreas, Ricardo Perez-Truglia, Mirko Wiederholt, and Basit Zafar (2022). "Expectations with Endogenous Information Acquisition: An Experimental Investigation". *The Review of Economics and Statistics* 104.5, pp. 1059–1078. DOI: 10.1162/rest_a_00994 (p. 3, 4, 7, 13, 19, 28).

Giaccobasso, Matias, Brad C. Nathan, Ricardo Perez-Truglia, and Alejandro Zentner (2022). "Where Do My Tax Dollars Go? Tax Morale Effects of Perceived Government Spending". Working Paper Series. DOI: 10.3386/w29789 (p. 7, 10).

Grigorieff, Alexis, Christopher Roth, and Diego Ubfal (2020). "Does Information Change Attitudes Toward Immigrants?" *Demography* 57.3, pp. 1117–1143. DOI: 10.1007/s13524-020-00882-8 (p. 3, 5).

Haaland, Ingar, Christopher Roth, and Johannes Wohlfart (2023). "Designing Information Provision Experiments". *Journal of Economic Literature* 61.1, pp. 3–40. DOI: 10.1257/jel.20211658 (p. 3, 4, 6).

Hoff, Peter D. (2009). *A First Course in Bayesian Statistical Methods*. Springer Texts in Statistics. New York, NY: Springer New York. DOI: 10.1007/978-0-387-92407-6 (p. 8).

Hopkins, Daniel J., John Sides, and Jack Citrin (2019). "The Muted Consequences of Correct Information about Immigration". *The Journal of Politics* 81.1, pp. 315–320. DOI: 10.1086/699914 (p. 3, 5).

Jäger, Simon, Christopher Roth, Nina Roussille, and Benjamin Schoefer (2023). "Worker Beliefs About Outside Options". *Working Paper* (p. 1, 5).

Jensen, Robert (2010). "The (Perceived) Returns to Education and the Demand for Schooling". *Quarterly Journal of Economics* 125.2, pp. 515–548. DOI: 10.1162/qjec.2010.125.2.515 (p. 1).

Kerwin, Jason T and Divya Pandey (2023). "Navigating Ambiguity: Imprecise Probabilities and the Updating of Disease Risk Beliefs". *Working Paper* (p. 9).

Kuziemko, Ilyana, Michael I. Norton, Emmanuel Saez, and Stefanie Stantcheva (2015). "How Elastic Are Preferences for Redistribution? Evidence from Randomized Survey Experiments". *American Economic Review* 105.4, pp. 1478–1508. DOI: 10.1257/aer.20130360 (p. 2, 3, 29).

Masten, Matthew and Alexander Torgovitsky (2014). "Instrumental Variables Estimation of a Generalized Correlated Random Coefficients Model". DOI: 10.1920/wp.cem.2013.0214 (p. 2, 22, 41).

— (2016). "Identification of Instrumental Variable Correlated Random Coefficients Models". *The Review of Economics and Statistics* 98.5, pp. 1001–1005. DOI: 10.1162/REST_a_00603 (p. 1, 2, 20, 22, 24, 38).

Robert, Christian P. (2007). *The Bayesian Choice: From Decision-Theoretic Foundations to Computational Implementation*. 2nd ed. Springer Texts in Statistics. New York: Springer. DOI: 10.1007/0-387-71599-1 (p. 8).

Roth, Christopher, Sonja Settele, and Johannes Wohlfart (2022). "Risk Exposure and Acquisition of Macroeconomic Information". *American Economic Review: Insights* 4.1, pp. 34–53. DOI: 10.1257/aeri.20200662 (p. 1, 3, 6, 8, 15).

Roth, Christopher and Johannes Wohlfart (2020). "How Do Expectations about the Macroeconomy Affect Personal Expectations and Behavior?" *The Review of Economics and Statistics* 102.4, pp. 731–748. DOI: 10.1162/rest_a_00867 (p. 8).

Settele, Sonja (2022). "How Do Beliefs about the Gender Wage Gap Affect the Demand for Public Policy?" *American Economic Journal: Economic Policy* 14.2, pp. 475–508. DOI: 10.1257/pol.20200559 (p. 1–4, 23–25, 28, 29, 40).

Stantcheva, Stefanie (2023). "Understanding of Trade". *Working Paper* (p. 3, 5).

Vilfort, Vod and Whitney Zhang (2023). *Interpreting IV Estimators in Information Provision Experiments*. DOI: 10.48550/arXiv.2309.04793 (p. 3, 13, 14).

Wiswall, Matthew and Basit Zafar (2015). "Determinants of College Major Choice: Identification Using an Information Experiment". *The Review of Economic Studies* 82.2 (291), pp. 791–824 (p. 21).

Yitzhaki, Shlomo (1996). “On Using Linear Regressions in Welfare Economics”. *Journal of Business & Economic Statistics* 12.4, pp. 478–486 (p. 23).

Appendix

A. Ensuring Positive Weights in Passive Designs With Unknown Priors

The dominant approach in the literature to ensure that the weights are non-negative is to incorporate information on the prior belief, either by splitting the sample (Cruces et al., 2013) or constructing an exposure-weighted instrument (Cullen and Perez-Truglia, 2022).

In this appendix, I provide a novel argument to identify a causal parameter that does not require eliciting priors or monotonicity. Instead, we use the assumption that people update their beliefs in the direction of the signal, but not past the signal. This follows from Bayesian updating, but Bayesian updating is not strictly necessary; directly assuming that the posterior belief is a convex combination of the prior and the signal is sufficient.

This assumption identifies the same causal parameters that are targeted by β_{+}^{split} (15) and β_{-}^{split} (18). This is a weaker assumption than instrument monotonicity, which requires that the signal shifts beliefs in the same direction for all participants. This is a slightly stronger assumption than what was needed in Section 2.3 when priors were available. There we assumed that people update (weakly) towards the signal, but did not need to assume that they do not cross it.

Since the control group that is not shown a signal, we directly observe their prior: recall that $X_i(B) = X_i^0$ in passive designs. Since the signal is known, we can directly condition on the sign of $(S_i(A) - X_i^0)$. The prior for the treated group is unknown and we observe only $X_i(A)$. But since we can rewrite the potential outcome equation in 2 as

$$S_i(A) - X_i(A) = (1 - \alpha_i)(S_i(A) - X_i^0)$$

and since $\alpha \in (0, 1)$ then

$$S_i(A) - X_i(A) > 0 \iff (S_i(A) - X_i^0) > 0$$

We used the Bayesian updating structure, but note this could be relaxed to include any model of updating such that the posterior lies between the prior and the signal.

Thus, although the regressions in (13) and (16) are not feasible since they use the prior to split the sample, the following regressions are feasible and equivalent.

$$\beta_{+}^{\text{split}} = \widetilde{\beta}_{+}^{\text{split}} \equiv \frac{\text{Cov}\left(T_i, Y_i \mid S_i(A) - X_i > 0\right)}{\text{Cov}\left(T_i, X_i \mid S_i(A) - X_i > 0\right)} \tag{29}$$

$$\beta_{-}^{\text{split}} = \widetilde{\beta}_{-}^{\text{split}} \equiv \frac{\text{Cov}\left(T_i, Y_i \mid S_i(A) - X_i < 0\right)}{\text{Cov}\left(T_i, X_i \mid S_i(A) - X_i < 0\right)} \tag{30}$$

B. A Toy Model of Belief Formation via Costly Information Acquisition

Let Y be the action (e.g., list price of a house) and X denote beliefs (e.g., about the market value). People start with a prior belief distribution centered around π_i. The initial variance of their beliefs is $\sigma^2_{X_0}$ so that their beliefs are represented by the normal $\mathcal{N}(\pi_i, \sigma^2_{X_0})$. For simplicity, $\sigma^2_{X_0}$ is common. We will consider signals S drawn from a normal distribution $\mathcal{N}(\mu_S, \sigma^2_S)$. This is an assumption that people have the same information environment.

People are uncertain about their beliefs, and this uncertainty about their beliefs generates uncertainty about the action that they would like to take. People act to minimize the loss function $L_i(y, x) = D\big(y, Y_i(x)\big)$, for some distance function D, which is the disutility associated with taking action y when $X = x$. Intuitively, integrating $L_i(y, x)$ over the distribution of beliefs converts uncertainty about beliefs (i.e., what is the probability that $X = x$) into regret about actions (i.e., how far is the choice y from $Y_i(x)$, which is optimal when $X = x$). In this loss function, beliefs affect utility only through their effect on actions. There is no direct "psychic" cost of imprecise beliefs.

People choose $Y_i(x)$ following the rule $Y_i(x) = \tau_i x + U_i$, where τ_i and U_i vary across individuals, and have quadratic loss $D(a, b) = \big(a - b\big)^2$. They act to minimize their expected loss, which is simply the expectation of $L_i(y, x)$ with respect to X $\big($i.e. $\mathbb{E}_X\big[L_i(y, x)\big]\big)$.

When beliefs are given by the normal $\mathcal{N}(\overline{X}, \sigma^2_{X_0})$, the choice of Y that minimizes expected loss is simply $Y^* \equiv Y_i(\overline{X}) = \tau_i \overline{X} + U_i$. We can use this to further simplify the expression for expected loss and write

$$\mathbb{E}_X\big[L_i(Y^*, x)\big] = \mathbb{E}_X\big[D(Y_i(\overline{X}), Y_i(x))\big] \tag{31}$$

$$= \mathbb{E}_X\big[\big(\big(\tau_i \overline{X} + U_i\big) - \big(\tau_i x + U_i\big)\big)^2\big] = \tau_i^2 \sigma^2_X \tag{32}$$

The disutility generated by uncertainty about X is increasing in both the variance of the belief distribution and the magnitude of the causal effect of beliefs on the outcome. This expression allows us to study the information acquisition problem.

I endogenize belief formation by allowing people to pay a fixed cost C to view a signal that is centered around the unknown true value. They then update beliefs following the

normal-normal Bayesian learning formula we have been working with throughout. When a person's beliefs are given by $\mathcal{N}(\overline{X}, \sigma_x^2)$, her loss is given recursively by

$$V_i(\overline{X}, \sigma_x^2) = \min \left\{ \mathbb{E}_X[L_i(Y_i(\overline{X}), x)], \mathbb{E}_S[V_i(X'(s)), \sigma_{X'}^2] + C \right\} \tag{33}$$

Where $\sigma_{X'}^2 = \frac{\sigma_X^2 \sigma_S^2}{\sigma_X^2 + \sigma_S^2}$ and the expectation $\mathbb{E}_S$ is the expectation with respect to the signal. Notice that in this model, the benefit of the signal comes from the fact that the posterior variance is less than the prior variance as long as the prior distribution is not already degenerate.

Solving this recursive problem gives the equilibrium condition

$$\tau_i^2 \sigma_X^2 = \tau_i^2 \sigma_{X'}^2 + C \tag{34}$$

In equilibrium, agents will be indifferent between paying the fixed cost to obtain new information and living with the uncertainty they have.[16] Replacing $\sigma_{X'}^2$ with its definition, and recalling that $1 - \frac{\sigma_S^2}{\sigma_S^2 + \sigma_X^2} = \alpha_i$ we obtain the following equality

$$\alpha_i \tau_i^2 \sigma_X^2 = C \tag{35}$$

Agents for whom the outcome is very sensitive to the beliefs (τ_i^2 is very large) will update their information until $\sigma_X^2 \alpha_i$ is small.[17] On the other hand, agents for whom the outcome is not sensitive to beliefs (τ_i^2 is small) will stop after seeing fewer signals, so that $\sigma_X^2 \alpha_i$ is relatively large.

We can see in this toy model how the causal relationship of interest affects the formation of beliefs before the experiment takes place. People whose actions depend more on their beliefs will be more willing to pay to obtain new information, and will therefore have more

[16] To ease exposition, I have ignored integer constraints that will, in general, prevent this from holding with equality. People will purchase signals until the next signal reduces their expected loss by less than the cost of the signal and will generally be strictly worse off if they buy another signal, not indifferent. This technicality makes exposition more cumbersome without any conceptual payoff.

[17] Notice that since $\alpha_i \equiv \frac{\sigma_X^2}{\sigma_S^2 + \sigma_X^2}$, α_i and σ_X^2 move together. That is, holding fixed σ_S^2, an increase in σ_X^2 implies an increase in α_i and vice-versa.

precise beliefs. In a Bayesian updating model, people with more precise beliefs will be less responsive to new information. In this way, the amount of variation in beliefs that can be induced by experimentally providing new information is directly depends on the causal effects of interest.

C. Proofs

C.1. Identification of the Average Partial Effect

This is a proof of proposition 1 that uses the structure of the specific information provision setting for the sake of exposition. The linear updating and Bayesian learning structure maintained throughout this paper is stronger than is necessary. Interested readers should see Masten and Torgovitsky (2016) for a more general proof and for a formal discussion of the more general conditions under which the APE is identified.

PROOF. First, use assumptions 1.b and 1.c to write

$$\begin{aligned}&\operatorname{Cov}\left(Y_i, X_i \mid R_i = r\right)\\&= \operatorname{Cov}\left(\tau_i\left(X_i^0 + \alpha_i(S_i(Z_i) - X_i^0)\right) + U_i, \left(X_i^0 + \alpha_i(S_i(Z_i) - X_i^0)\right) \mid R_i = r\right)\end{aligned}$$

By definition $r \equiv \begin{bmatrix} r_1 & x^0 & s_1 & s_2 \end{bmatrix}$. Let $\widetilde{\alpha}(r)$ be such that $r_1 = F_{\alpha|w,z}(\widetilde{\alpha})$.[18]Substituting this in gives

$$\begin{aligned}&= \operatorname{Cov}\left(\tau_i\left(x^0 + \widetilde{\alpha}(S_i(Z_i) - x^0)\right) + U_i, \left(x^0 + \widetilde{\alpha}(S_i(Z_i) - x^0)\right) \mid R_i = r\right)\\&= \widetilde{\alpha}(r)x^0 \operatorname{Cov}\left(\tau_i, S_i(Z_i) \mid R_i = r\right) + \widetilde{\alpha}(r)^2 \operatorname{Cov}\left(\tau_i S_i(Z_i), S_i(Z_i) \mid R_i = r\right)\end{aligned}$$

Notice that the remaining random variables are τ_i and the remaining variation in $S_i \mid R_i = r$. Writing S_i in terms of the switching equation makes it clear that all remaining variation in S_i comes from the treatment assignment Z_i. Recall that by definition $S_i = S_i(1)\mathbb{1}\left(Z_i = 1\right) + S_i(2)\mathbb{1}\left(Z_i = 2\right)$ so that $(S_i \mid R_i = r) = s_1\mathbb{1}\left(Z_i = 1\right) + s_2\mathbb{1}\left(Z_i = 2\right)$. From random assignment (1.a), $\operatorname{Cov}\left(\tau_i, S_i(Z_i) \mid R_i = r\right) = \operatorname{Cov}\left(\tau_i, s_1\mathbb{1}\left(Z_i = 1\right) + s_2\mathbb{1}\left(Z_i = 2\right) \mid R = r\right) = 0$. Substituting in the switching equation to the final term gives

$$= \widetilde{\alpha}(r)^2 \operatorname{Cov}\left(\tau_i\left(s_1\mathbb{1}\left(Z_i = 1\right) + s_2\mathbb{1}\left(Z_i = 2\right)\right), \left(s_1\mathbb{1}\left(Z_i = 1\right) + s_2\mathbb{1}\left(Z_i = 2\right)\right) \mid R_i = r\right)$$

so from random assignment

$$= \widetilde{\alpha}^2 \operatorname{Var}\big(\big(s_1 \mathbb{1}(Z_i = 1) + s_2 \mathbb{1}(Z_i = 2)\big) \mid R_i = r\big)\, \mathbb{E}\big[\tau_i \mid R_i = r\big]$$

Which finally gives

$$\operatorname{Cov}\big(Y_i, X_i \mid R_i = r\big) = \widetilde{\alpha}_i^2 \operatorname{Var}\big(S_i(Z) \mid R = r\big)\, \mathbb{E}\big[\tau_i \mid R = r\big]$$
$$\operatorname{Var}\big(X \mid R = r\big) = \operatorname{Var}\big(X_i^0 + \widetilde{\alpha}(S_i(Z) - X_i^0) \mid R = r\big) = \widetilde{\alpha}_i^2 \operatorname{Var}\big(S_i(Z) \mid R = r\big)$$

Combining these results, we have that

$$\frac{\operatorname{Cov}\big(Y_i, X_i \mid R_i = r\big)}{\operatorname{Var}\big(X_i \mid R_i = r\big)} = \mathbb{E}\big[\tau_i \mid R_i = r\big]$$

where assumption 1.d ensures that

$$\operatorname{Var}\big(X_i \mid R_i = r\big) = \operatorname{Var}\big(\big(x^0 + \widetilde{\alpha}(r)(S_i(Z_i) - x^0)\big) \mid R_i = r\big) > 0$$

□

[18] To define $\widetilde{\alpha}(r)$ without needing to restrict $F(\cdot)$, let $\widetilde{\alpha}(r) = \inf\{a \mid F_{\alpha|W}(a) = r\}$.

D. Additional Empirical Results and Alternative Specifications

In this section, I will provide some addition information about estimation of the control function in Section 4 and some results from alternative specifications.

D.1. TSLS Estimates from Settele (2022) Are Similar Without Controls

Since I estimate the control function without additional controls, I also re-estimate the main specification TSLS without these controls to make these estimators more comparable. In Table D.1, I present results comparing the more parsimonious specification used in this paper with the original estimates from Settele (2022). I will then compare these estimates to the estimates of the APE. In Column 1, I replicate the specification in Table 5, Panel C of Settele (2022). In Column 2, I omit the weights used in the original paper. In Column 3, I also omit the covariates used in the original paper. For the sake of simplicity, I will omit the covariates used in the original paper in the control function estimation. Thus, the control function estimates are most comparable to the TSLS estimates in Column 3, and not the TSLS estimates in Column 1 (and the original paper).

There is no conceptual reason that this estimation cannot be done conditional on covariates. However, since the control function estimation is a semi parametric estimator, including these covariates is technically demanding and introduces new researcher degrees of freedom, making this exercise less transparent. For the purposes of the empirical illustration, I use the parsimonious specification without additional covariates.

TABLE D.1. TSLS Estimates are Similar with and without Weights and Covariates

	Paper Specification	No Weights	No Covariates
Policy demand index	-0.087	-0.094	-0.105
	(0.038)	(0.038)	(0.041)

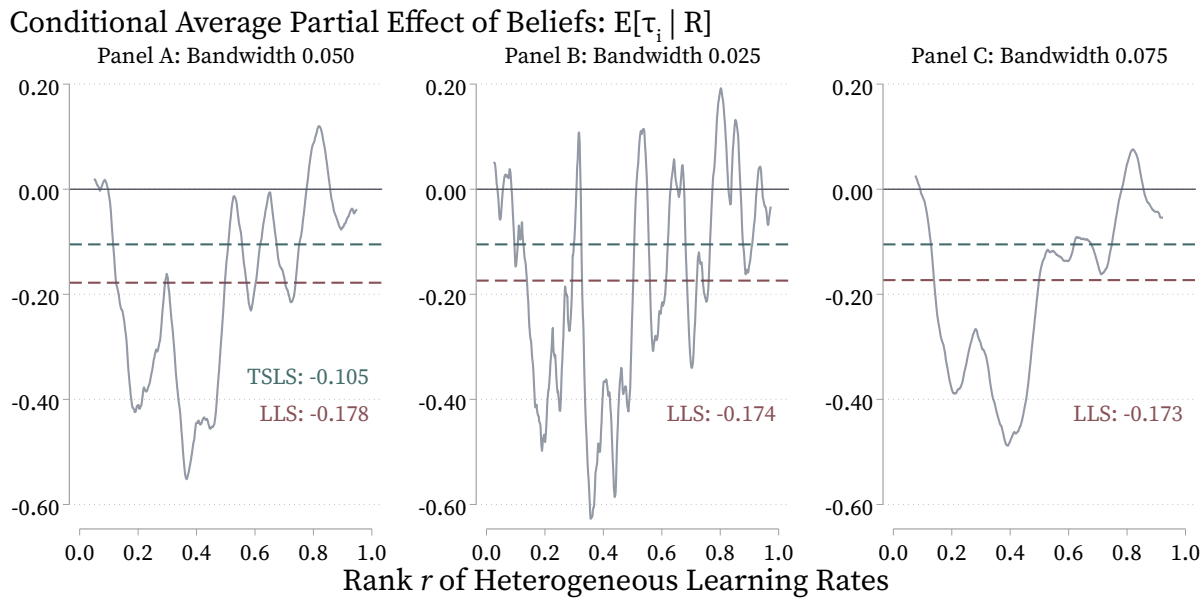
Notes: This table replicates the TSLS estimates from Settele (2022). The first column replicates the specification in Table 5, Panel C of Settele (2022). The second column omits the weights used in the original paper. The third column also omits the covariates used in the original paper. N = 3,022. Standard errors for the TSLS estimates are heteroskedasticity-robust.

D.2. Alternative Choices of the Bandwidth

We need to choose a bandwidth to approximate conditioning on the continuous rank of the learning rate α_i. In the main specification, the bandwidth has has a half-width of 0.050. This is slightly smaller than the rule of thumb bandwidth of 0.0635, which "undersmooths" the intermediate estimates to minimize bias in the final estimate of the APE following Masten and Torgovitsky (2014).

Figure D.1 presents the raw results for a range of bandwidths $h \in \{0.025, 0.05, 0.075\}$. While the intermediate estimates of the conditional average partial effect (CAPE) curve become noisier at the smaller bandwidth and smoother at the larger bandwidth, the final estimate of the APE is remarkably stable and changes only at the 3rd decimal.

FIGURE D.1. LLS Estimates of the APE Are Not Sensitive to Choice of Bandwidth



Notes: This table presents the LLS estimate of the APE, as well as the CAPE estimated estimated at every rank r in the sample at different values of the bandwidth. Panel A reports the underlying CAPE estimates that are used to construct the binned scatter plot in Figure 2 in the main body.