A Framework for Eliciting, Incorporating, and Disciplining Identification Beliefs in Linear Models*

Francis J. DiTraglia †[1] and Camilo García-Jimeno[1,2]

[1]Department of Economics, University of Pennsylvania
[2]NBER

This Version: February 10, 2017, First Version: August 29, 2015

Abstract

To estimate causal effects from observational data, an applied researcher must impose beliefs. The instrumental variables exclusion restriction, for example, represents the belief that the instrument has no direct effect on the outcome of interest. Yet beliefs about instrument validity do not exist in isolation. Applied researchers often discuss the likely direction of selection and the potential for measurement error in their papers but at present lack formal tools for incorporating this information into their analyses. As such they not only leave money on the table, by failing to use all relevant information, but more importantly run the risk of reasoning to a contradiction by expressing mutually incompatible beliefs. In this paper we characterize the sharp identified set relating instrument invalidity, treatment endogeneity, and measurement error in a workhorse linear model, showing how beliefs over these three dimensions are mutually constrained. We consider two cases: in the first the treatment is continuous and subject to classical measurement error; in the second it is binary and subject to non-differential measurement error. In each, we propose a formal Bayesian framework to help researchers elicit their beliefs, incorporate them into estimation, and ensure their mutual coherence. We conclude by illustrating the usefulness of our proposed methods on a variety of examples from the empirical microeconomics literature.

Keywords: Partial identification, Beliefs, Instrumental variables, Measurement error, Bayesian econometrics

JEL Codes: C10, C11, C18, C26

*We thank Daron Acemoglu, Richard Hahn, Hidehiko Ichimura, Laura Liu, Ulrich Müller, Frank Schorfheide, and Ben Ukert, as well as seminar participants at Princeton, Penn State, the 2015 NSF-NBER Seminar on Bayesian Inference, the 2015 Midwest Econometrics Group Meetings, and the 2016 ISBA World Meeting for helpful comments and suggestions. We thank Mallick Hossain and Alejandro Sánchez for excellent research assistance and acknowledge support from a UPenn University Research Foundation award.

†Corresponding Author: fditra@sas.upenn.edu, *3718 Locust Walk, Philadelphia, PA 19104.*

"Belief is so important! A hundred contradictions might be true."

— Blaise Pascal, *Pensées*

1 Introduction

To identify causal effects from observational data, an applied researcher must augment the data with her beliefs. The exclusion restriction in an instrumental variables (IV) regression, for example, represents the belief that the instrument has no direct effect on the outcome of interest. Although this belief can never be tested directly, applied researchers know how to think about it and how to debate it. In practice, however, not all beliefs are treated equally. In addition to "formal beliefs" such as the IV exclusion restriction – beliefs that are directly imposed to obtain identification – researchers often state a number of "informal beliefs." While not directly imposed on the problem, informal beliefs play an important role in interpreting results and reconciling conflicting estimates. Papers that report IV estimates, for example, almost invariably state the authors' belief about the sign of the correlation between the endogenous treatment and the error term but do not exploit this information in estimation.[1] Another common informal belief concerns the extent of measurement error. When researchers observe an ordinary least squares (OLS) estimate that is substantially smaller than, but has the same sign as its IV counterpart, classical measurement error, with its attendant "least squares attenuation bias," often is suggested as the likely cause.

Relegating informal beliefs to second-class status is both wasteful of information and dangerous; beliefs along different dimensions of the problem are mutually constrained by each other, the model, and the data. By failing to explicitly incorporate all relevant information, applied researchers both leave money on the table and, more importantly, risk reasoning to a contradiction by expressing mutually incompatible beliefs. Although this point is general, we illustrate its implications here in the context of a linear model

$$y = \beta T^* + \mathbf{x}'\gamma + u \tag{1}$$

$$T = T^* + w \tag{2}$$

where T^* is a potentially endogenous treatment, y is an outcome of interest, and $\mathbf{x}$ is a vector of exogenous controls. Our goal is to estimate the causal effect of T^* on y, namely β, but

[1] Referring to more than 60 papers published in the top three empirical journals between 2002 and 2005, Moon and Schorfheide (2009) note that "in almost all of the papers the authors explicitly stated their beliefs about the sign of the correlation between the endogenous regressor and the error term; yet none of the authors exploited the resulting inequality moment condition in their estimation."

we observe only T, a noisy measure of T^* polluted by measurement error w. While we are fortunate to have an instrument z at our disposal, it may not satisfy the exclusion restriction: z is potentially correlated with u. This scenario is typical in applied microeconomics: endogeneity is the rule rather than the exception, the treatments of greatest interest are often the hardest to measure, and the validity of a proposed instrument is almost always debatable. We consider two cases. In the first, T^* is continuous and subject to classical measurement error: T^* is independent of w. In the second, z and T^* are binary and T^* is subject to *non-differential* measurement error: the joint distribution of T^* and w is unrestricted, but T is assumed to be conditionally independent of all other variables in the system, given T^*.[2] In each case we derive the identified set relating treatment endogeneity, measurement error, and instrument invalidity in terms of empirically meaningful parameters. We then use this characterization to construct a framework for Bayesian inference that combines the information contained in the data with researcher beliefs in a coherent and transparent way. As we demonstrate through a number of empirical examples, this framework not only allows researchers to incorporate relevant problem-specific beliefs, but, by identifying any inconsistencies that may be present, provides a tool for refining and disciplining these beliefs. Although our method employs Bayesian reasoning, it can be implemented in a number of different ways that should make it appealing both to frequentists and Bayesians.

While measurement error, treatment endogeneity, and invalid instruments have all generated voluminous literatures, to the best of our knowledge this is the first paper to carry out a partial identification exercise in which all three problems can be present simultaneously. Our main point is simple but has important implications for applied work that have been largely overlooked; measurement error, treatment endogeneity, and instrument invalidity are mutually constrained by each other and the data in a manner that can only be made apparent by characterizing the full identified set for the model. Because the dimension of this set is strictly smaller than the number of variables used to describe it, the constraints of the model could easily contradict prior researcher beliefs. Given the shape of the identified set, the belief that z is a valid instrument, for example, could imply an implausible amount of measurement error or a selection effect with the opposite of the expected sign. In this way our framework provides a means of reconciling and refining beliefs that would not be possible based on introspection alone. We are by no means the first to recognize the importance of requiring that beliefs be compatible. Kahneman and Tversky (1974), for example, make a closely related point in their discussion of heuristic decision-making under uncertainty. Even if specific probabilistic assessments appear coherent on their own,

[2]These cases require a separate treatment because, as we discuss below, a binary regressor cannot be subject to classical measurement error.

> an internally consistent set of subjective probabilities can be incompatible with other beliefs held by the individual ... For judged probabilities to be considered adequate, or rational, internal consistency is not enough. The judgements must be compatible with the entire web of beliefs held by the individual. Unfortunately, there can be no simple formal procedure for assessing the compatibility of a set of probability judgements with the judge's total system of beliefs (p. 1130).

Our purpose here is to take up the challenge laid down by Kahneman and Tversky (1974) and provide just such a formal procedure for assessing the compatibility of researcher beliefs over treatment endogeneity, measurement error, and instrument invalidity in linear models. Although the intuition behind our procedure is straightforward, the details are more involved. For this reason we provide free and open-source software in R and Stata to make it easy for applied researchers to implement the methods described in this paper.[3]

Elicitation is a key ingredient of our framework. Before we can impose researcher beliefs we must express them in intuitive, empirically meaningful terms. In the continuous treatment setting, we express instrument invalidity in terms of $\rho_{uz} \equiv \text{Cor}(u, z)$, treatment endogeneity in terms of $\rho_{T^*u} \equiv \text{Cor}(T^*, u)$, and measurement error in terms of $\kappa \equiv \text{Var}(T^*)/\text{Var}(T)$, essentially a signal-to-noise ratio that is conveniently bounded between zero and one. For the binary case, measurement error is parameterized in terms of a pair of misclassification probabilities (α_0, α_1), defined in Section 4, and instrument invalidity and treatment endogeneity are more naturally expressed as differences of conditional means. Specifically, δ_z is the average difference in unobservables u between individuals with high and low value for the instrument while δ_{T^*} is the average difference in u between treated and untreated individuals. In this paper we impose only relatively weak prior beliefs in the form of sign and interval restrictions on the aforementioned parameters.[4] As we discuss further in our empirical examples, these are fairly easy to elicit in practice and can be surprisingly informative about the causal effect of interest.

The addition of researcher beliefs is not only extremely helpful, but unavoidable. As we show below, the data alone provide no restriction on β, although they do bound the maximum possible amount of measurement error. Nevertheless, whenever one imposes information beyond what is contained in the data, it is crucial to make clear how this affects the ultimate result. This motivates our use of a *transparent parameterization*, which decomposes the problem into a vector of partially-identified *structural* parameters θ, and a vector of identified *reduced form* parameters φ in such a way that inference for the identified set Θ for θ depends on the data only through φ. This decomposition has several advantages. First, since the

[3] See https://github.com/fditraglia/ivdoctr and https://github.com/fditraglia/binivdoctr.

[4] Researchers who feel comfortable imposing more finely-grained beliefs can easily do so within our framework, but elicitation of fully-informative priors is more challenging in practice.

reduced form parameters are identified, inference for this part of the problem is completely standard. Second, a transparent parameterization shows us precisely where any identification beliefs we may choose to impose enter the problem: the data rule out certain values of φ, while our beliefs place restrictions on the conditional identified set $\Theta(\varphi)$ which ultimately yields inference for the causal effect β.

This paper contributes to a small but growing literature on the Bayesian analysis of partially-identified models, including Poirier (1998), Gustafson (2005), Richardson et al. (2011), Moon and Schorfheide (2012), Kitagawa (2012), Hahn et al. (2016), and Gustafson (2015). Some recent contributions to the literature on structural vector autoregression models (Amir-Ahmadi and Drautzburg, 2016; Arias et al., 2016; Baumeister and Hamilton, 2015) also explore related ideas. Because we discuss, as part of our exercise, Bayesian inferences for the identified set, our work relates to Kline and Tamer (2016) and Chen et al. (2016) who give sufficient conditions under which such inferences have a valid frequentist interpretation.

Our results also relate to a large literature on estimating the effect of mis-measured binary regressors. An early contribution is Bollinger (1996) who provides partial identification bounds for an exogenous mis-measured regressor. van Hasselt and Bollinger (2012) derive additional bounds for the same model and Bollinger and van Hasselt (2015) propose a Bayesian inference procedure based on these bounds. Because we consider a situation in which an instrumental variable is available, our setting is more closely related to that considered by Kane et al. (1999), Black et al. (2000), Frazis and Lowenstein (2003), Lewbel (2007), Mahajan (2006) and Hu (2008). The key lesson from these papers is that the two-stage least squares (TSLS) estimator is inconsistent even if the instrument is valid. When the treatment is exogenous, however, it is possible to construct a non-linear method of moments estimator that recovers the treatment effect using a discrete instrumental variable.

Unlike these papers, we consider a setting in which the binary treatment of interest may be endogenous. As shown in DiTraglia and García-Jimeno (2016) the usual instrumental variable assumption – conditional mean independence – is insufficient to identify the effect of an endogenous, mis-measured, binary treatment. Although DiTraglia and García-Jimeno (2016) provide a point identification result under a stronger assumption on the instrument – full independence – we do not employ this result here. Instead we allow for an invalid instrument and derive partial identification results.

Two recent papers that similarly consider partial identification under instrument invalidity are Conley et al. (2012) and Nevo and Rosen (2012). Like us, Conley et al. (2012) adopt a Bayesian approach that allows for a violation of the IV exclusion restriction, but they do not explore the relationship between treatment endogeneity and instrument invalidity. In contrast, Nevo and Rosen (2012) derive bounds for a causal effect in the setting

where an endogenous regressor is "more endogenous" than the variable used to instrument it is invalid.[5] Our framework encompasses the settings considered in these two papers, but is strictly more general; we allow for measurement error simultaneously with treatment endogeneity and instrument invalidity. More importantly, the central message of our paper is that it can be misleading to impose beliefs on only one dimension of a partially identified problem unless one has a way of ensuring their mutual consistency with all other relevant researcher beliefs. For example, although a single valid instrument solves both the problem of classical measurement error and treatment endogeneity, we argue that it is insufficient to carry out a partial identification exercise that merely relaxes the exclusion restriction, as in Conley et al. (2012). Values for the correlation between z and u that seem plausible when viewed in isolation could easily imply implausible amounts of measurement error or treatment endogeneity. While our main contribution here is to describe the relationship between measurement error, treatment endogeneity, and instrument invalidity, we also derive sharp bounds on the extent of both classical measurement error when the treatment is continuous, and non-differential measurement error when the treatment is binary. These are, to the best of our knowledge, new to the literature and could be of interest in their own right.

The remainder of this paper is organized as follows. Section 2 derives the identified set for a continuous treatment under classical measurement error and Section 3 describes our approach to inference. Section 4 derives the identified set for a binary instrument and binary treatment subject to non-differential measurement error while Section 4.5 explains the differences between inference for a binary and a continuous treatment. Section 5 presents a number of substantive empirical examples illustrating our procedure in both the binary and continuous–treatment cases, and Section 6 concludes. Proofs and additional empirical examples appear in the appendix.

2 The Identified Set for a Continuous Treatment

2.1 Model and Assumptions

To simplify the notation, suppose either that there are no exogenous control regressors $\mathbf{x}$ (including a constant), or equivalently, that they have been "projected out." In Section 2.4 we explain why this assumption is innocuous and how to accommodate control regressors in practice. With this simplification, Equations 1–2 and the first stage $T^* = \pi z + v$ can be

[5] In our notation, ρ_{T^*u} and ρ_{uz} have the same sign but $|\rho_{uz}| < |\rho_{T^*u}|$.

written in matrix form as

$$\begin{bmatrix} y \\ T \\ T^* \\ z \end{bmatrix} = \Gamma \begin{bmatrix} u \\ w \\ v \\ z \end{bmatrix}, \qquad \Gamma = \begin{bmatrix} 1 & 0 & \beta & \beta\pi \\ 0 & 1 & 0 & \pi \\ 0 & 0 & 1 & \pi \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{3}$$

where we assume, without loss of generality, that all random variables in the system are mean zero or have been de-meaned.[6] Our goal is to learn the parameter β, the causal effect of T^*. In general, T^* is unobserved: we only observe a noisy measure T that has been polluted by classical measurement error w. We call (u, v, w, z) the "primitives" of the system and assume that they satisfy the following assumptions.

Assumption 2.1. *The covariance matrix Ω of (u, v, z, w) is finite and satisfies*

$$\Omega = \begin{bmatrix} \widetilde{\Omega} & \mathbf{0} \\ \mathbf{0}' & \sigma_w^2 \end{bmatrix}, \quad \widetilde{\Omega} = \begin{bmatrix} \sigma_u^2 & \sigma_{uv} & \sigma_{uz} \\ \sigma_{uv} & \sigma_v^2 & 0 \\ \sigma_{uz} & 0 & \sigma_z^2 \end{bmatrix} \tag{4}$$

where $\widetilde{\Omega}$ is positive definite, and σ_w^2 is non-negative.

Because w represents classical measurement error, it is uncorrelated with u, v, and z as well as T^*. The parameter σ_{uz} controls the invalidity of the instrument z: unless $\sigma_{uz} = 0$, z is an invalid instrument. Both σ_{uz} and σ_{uv} control the endogeneity of T^*; σ_{uv} is the component of $Cov(T^*, u)$ that is unrelated to z. The matrix Ω is unobserved. We observe only Σ, the covariance matrix of (T, y, z):

$$\Sigma = \begin{bmatrix} \sigma_T^2 & \sigma_{Ty} & \sigma_{Tz} \\ \sigma_{Ty} & \sigma_y^2 & \sigma_{yz} \\ \sigma_{Tz} & \sigma_{yz} & \sigma_z^2 \end{bmatrix}. \tag{5}$$

To ensure that the IV estimand is well-defined and that the elements of Σ are finite, we impose the following assumption:

Assumption 2.2. *The first-stage coefficient π is non-zero and both β and π are finite.*

Γ is full rank. Moreover, by Assumption 2.1, Σ is positive definite. The system we have just finished describing does not identify the treatment effect β. In particular, neither the

[6] Equivalently, we can treat the constant term in the first-stage and main equation as exogenous regressors that have been projected out.

OLS nor IV estimators converge in probability to β, instead they approach

$$\beta_{OLS} = \frac{\sigma_{Ty}}{\sigma_T^2} = \left(\frac{\sigma_{T^*}^2}{\sigma_{T^*}^2 + \sigma_w^2}\right)\left(\beta + \frac{\sigma_{T^*u}}{\sigma_{T^*}^2}\right) \tag{6}$$

and

$$\beta_{IV} = \frac{\sigma_{zy}}{\sigma_{Tz}} = \beta + \frac{\sigma_{uz}}{\sigma_{Tz}} \tag{7}$$

where $\sigma_{T^*}^2 = \sigma_T^2 - \sigma_w^2$ denotes the variance of the unobserved regressor T^*, and

$$\sigma_{T^*u} = \sigma_{uv} + \pi\sigma_{uz}. \tag{8}$$

Because both σ_{uv} and σ_{uz} are sources of endogeneity for the unobserved regressor T^*, there is an indirect link between instrument invalidity and the OLS estimand. Moreover, while the IV probability limit depends neither on the extent of measurement error, σ_w^2, nor on σ_{uv}, through the model and assumptions it nevertheless contains information about both quantities. As a result, the problems of measurement error, regressor endogeneity, and instrument invalidity are mutually constrained. Our next task is to characterize the relationship between them by exploiting all of the implications of Equation 3 and Assumptions 2.1 and 2.2. To aid in this characterization, we first re-parameterize the problem, expressing it in terms of quantities that are empirically meaningful and thus practical for eliciting researcher beliefs.

2.2 A Convenient Parameterization

In order to elicit and incorporate researcher's beliefs, we work in terms of the following:

$$\rho_{uz} = \text{Cor}(u, z) \tag{9}$$

$$\rho_{T^*u} = \text{Cor}(T^*, u) \tag{10}$$

$$\kappa = \frac{\sigma_{T^*}^2}{\sigma_T^2} = \frac{\sigma_{T^*}^2}{\sigma_{T^*}^2 + \sigma_w^2}. \tag{11}$$

The first quantity, ρ_{uz}, is the correlation between the instrument and the main equation error term u. It measures the endogeneity of the instrument. The exclusion restriction in IV estimation, for example, corresponds to the degenerate belief that $\rho_{uz} = 0$. When critiquing an instrument, researchers often state a belief about the likely sign of this quantity. The second quantity, ρ_{T^*u}, is the correlation between the unobserved regressor T^* and the main equation error term. It measures the overall endogeneity of T^*, taking into account both the effect of σ_{uv} and σ_{uz}. While in practice it would be unusual to be able to articulate a belief about σ_{uv}, researchers almost invariably state their belief about the sign of the quantity ρ_{T^*u}

before undertaking an IV estimation exercise.

The third quantity, κ, may be somewhat less familiar. When there are no covariates and T^* is exogenous, κ measures the degree of attenuation bias present in the OLS estimator: if $\rho_{T^*u} = 0$ then the OLS probability limit is $\kappa\beta$. Equivalently, provided that $\sigma_{yT^*} \neq 0$,

$$\kappa = \left(\frac{\sigma^2_{T^*}}{\sigma^2_T}\right)\left(\frac{\sigma^2_{yT}}{\sigma^2_{yT^*}}\right) = \left(\frac{\sigma^2_{yT}}{\sigma^2_T\sigma^2_y}\right)\left(\frac{\sigma^2_{T^*}\sigma^2_y}{\sigma^2_{yT^*}}\right) = \frac{\rho^2_{yT}}{\rho^2_{yT^*}} \tag{12}$$

so another way to interpret κ is as the ratio of the observed R^2 of the main equation and the unobserved R^2 that we would obtain if our regressor had not been polluted with measurement error.[7] A third way to think about κ is in terms of signal and noise. If $\kappa = 1/2$, for example, this means that half of the variation in the observed regressor T is "signal," T^*, and the remainder is noise, w. While the other two interpretations are specific to the case of no covariates, this third interpretation is general. We consider it much easier to elicit beliefs about κ than about σ^2_w because κ has bounded support: it takes a value in $(0, 1]$. When $\kappa = 1$, $\sigma^2_w = 0$ so there is no measurement error. The limit as κ approaches zero corresponds to taking σ^2_w to infinity.

Expressed in this way, our parameter space is bounded, all of our parameters are scale-free, and most importantly, they are meaningful in real-world applications. Moreover, although the model introduced in the preceding section contains six non-identified parameters – β, σ^2_u, σ_{uv}, σ_{uz}, σ^2_v, and σ^2_w – knowledge of any two of the parameters $(\rho_{uz}, \rho_{T^*u}, \kappa)$ is sufficient to identify the whole system. In the following section we solve for ρ_{uz} in terms of ρ_{T^*u} and κ, and go on to characterize the sharp identified set for $(\rho_{T^*u}, \rho_{uz}, \kappa)$. This fully describes the information contained in the data and our assumptions.

2.3 Deriving the Identified Set for $(\rho_{T^*u}, \rho_{uz}, \kappa)$

We begin by deriving the relationship between ρ_{uz}, ρ_{T^*u}, and κ. The basic idea is to combine the OLS and IV probability limits, Equations 6 and 7, with the variance decomposition for y implied by the linear model. After eliminating β and σ^2_u from the resulting equations, and re-parameterizing as described in the preceding section, we derive a quadratic equation in ρ_{uz} with coefficients that involve κ and ρ_{T^*u}. Solving and simplifying, we show that one of the two roots is extraneous because it implies a negative value for σ_u, leading to the following.[8]

[7] This follows because $\text{Cov}(T, y) = \text{Cov}(T^*, y)$ under classical measurement error.

[8] The convention that standard deviations are positive ensures that correlations have the same sign as covariances.

Proposition 2.1. *Under Equation 3 and Assumptions 2.1 and 2.2,*

$$\rho_{uz} = \left(\frac{\rho_{T^*u}\rho_{Tz}}{\sqrt{\kappa}}\right) - (\rho_{Ty}\rho_{Tz} - \kappa\rho_{zy})\sqrt{\frac{1-\rho^2_{T^*u}}{\kappa\left(\kappa-\rho^2_{Ty}\right)}} \tag{13}$$

Equation 13 allows us to solve for ρ_{uz} in terms of observable correlations – ρ_{Ty}, ρ_{Tz}, and ρ_{zy} – and the unobserved parameters ρ_{T^*u} and κ. Thus, to fully characterize the relationship between measurement error, treatment endogeneity, and instrument invalidity, it suffices to derive the sharp identified set for (ρ_{T^*u}, κ). To this end, we first list a set of simple conditions that are Equivalent to Assumption 2.1.

Lemma 2.1. *The following conditions are equivalent to Assumption 2.1:*

(a) $\sigma^2_u, \sigma^2_v, \sigma^2_z, \sigma^2_w < \infty$

(b) $\sigma^2_u, \sigma^2_v, \sigma^2_z > 0$, $\sigma^2_w \geq 0$

(c) $\rho^2_{uv} + \rho^2_{uz} < 1$

(d) $Cov(w, z) = Cov(w, u) = Cov(w, v) = 0$.

Parts (a) and (b) of Lemma 2.1 are straightforward: all variances must be finite and strictly positive with the exception of σ^2_w, which equals zero in the absence of measurement error. Part (c), however, is somewhat less intuitive. Geometrically, it states that (ρ_{uz}, ρ_{uv}) must lie within the unit circle: if one of the correlations is very large in absolute value, the other cannot be. To understand the intuition behind this constraint, recall that since v is the residual from the projection of T^* onto z, it is uncorrelated with z by construction. If ρ_{uz} and ρ_{uv} were both sufficiently close to one this would require z and v to be correlated, leading to a contradiction. Part (d) of the Lemma is simply the classical measurement error assumption. We now use Lemma 2.1 to derive the sharp identified set for ρ_{T^*u} and κ.

Proposition 2.2 (Sharp Identified Set for ρ_{T^*u} and κ.)**.** *Under Assumptions 2.1–2.2 and Equation 3,* $(\rho_{T^*u}, \kappa) \in (-1, 1) \times (\underline{\kappa}, 1]$ *where*

$$\underline{\kappa} = \frac{\rho^2_{Ty} + \rho^2_{Tz} - 2\rho_{Ty}\rho_{Tz}\rho_{zy}}{1-\rho^2_{zy}} \tag{14}$$

These bounds are sharp.

The proof of Proposition 2.2 proceeds by showing that the two bounds $\underline{\kappa} < \kappa \leq 1$ and $|\rho_{T^*u}| < 1$ are equivalent to Assumption 2.1 given the model from Equation 3 and Assumption

2.2.[9] Thus, the sharp identified set is a rectangular region: for any allowable value of κ, ρ_{T^*u} can take on any value strictly between -1 and 1.

Because Proposition 2.2 provides a lower bound for κ, it places an *upper bound* on the extent of measurement error given the observed covariance matrix Σ. This bound relies on two simpler, but weaker bounds. The first is $\kappa > \rho_{Ty}^2$. In a setting without covariates, this says that the R-squared of a regression of y on T provides an upper bound for the maximum possible amount of measurement error. Although typically stated somewhat differently, this bound is well known: it corresponds to the familiar "reverse regression bound" for β.[10] In our setting this bound is implied by $\sigma_u^2 > 0$ since

$$\sigma_u^2 = \sigma_y^2 \left[\frac{\kappa - \rho_{Ty}^2}{\kappa \left(1 - \rho_{T^*u}^2\right)} \right] \tag{15}$$

by Lemma A.1(c) from the Appendix. Note that $\rho_{Ty}^2 < \kappa$ implies that the solution for ρ_{uz} from Proposition 2.1 is always real-valued. The second of these two weaker bounds is $\kappa > \rho_{Tz}^2$, which says that the R-squared of the first-stage regression of T on z *also* provides an upper bound for the maximum possible amount of measurement error. This follows since, by Lemma A.1(a) of the Appendix,

$$\sigma_v^2 = \sigma_T^2(\kappa - \rho_{Tz}^2) \tag{16}$$

and σ_v^2 must be strictly positive. We doubt that we are the first to notice this bound given its simplicity. Nevertheless, to the best of our knowledge, it has not appeared in the literature. After some algebra, these two weak bounds together with the restriction that $\rho_{uv}^2 + \rho_{uz}^2 < 1$ from Lemma 2.1 and the positive-definiteness of Σ yield the sharp lower bound for κ in Proposition 2.2. This bound is strictly tighter than $\kappa > \max\left\{\rho_{Tz}^2, \rho_{Ty}^2\right\}$ because it incorporates additional information from the reduced form regression of y on z.

The existence of an upper bound on measurement error, one that tightens as the OLS and first-stage R-squared values increase, is important because applied econometricians often explain a substantial discrepancy between OLS and IV estimators by arguing that their data is subject to large measurement errors. We are unaware of any cases in which such a belief has been confronted with these restrictions. In addition to bounding the possible amount of measurement error, our assumptions also bound the instrument invalidity parameter ρ_{uz}, in

[9] Since the coefficient matrix Γ from Equation 3 is full rank, this means that the positive-definiteness of Σ and of the extended covariance matrix $Cov(y, T, z, T^*)$ is implied by Assumption 2.1 and thus provides no additional restrictions.

[10] To see this, suppose that $\rho_{T^*u} = 0$, and without loss of generality that β is positive. Then Equation 6 gives $\beta_{OLS} = \kappa\beta < \beta$. Multiplying both sides of $\kappa > \rho_{Ty}^2$ by β and rearranging gives $\beta < \beta_{OLS}/\rho_{Ty}^2$, and hence $\beta_{OLS} < \beta < \beta_{OLS}/\rho_{Ty}^2$.

spite of the fact that they place no restriction on ρ_{T^*u}.

Corollary 2.1 (Sharp Bounds for ρ_{uz})**.** *Under the conditions of Proposition 2.2, ρ_{uz} has a non-trivial one-sided bound. If $\rho_{Ty}\rho_{Tz} - \underline{\kappa}\rho_{zy} < 0$, then $\rho_{uz} \in (-|\rho_{Tz}|/\sqrt{\underline{\kappa}}, 1)$. Otherwise $\rho_{uz} \in (-1, |\rho_{Tz}|/\sqrt{\underline{\kappa}})$, where $\underline{\kappa}$ is defined in Proposition 2.2. These bounds are sharp.*

Because $\underline{\kappa} > \rho_{Tz}^2$, Corollary 2.1 always rules out a range of values for ρ_{uz}. Notice, however, that it never rules out $\rho_{uz} = 0$. This is unsurprising given that it is known to be impossible to test for instrument validity in the model we consider here.

Together, Propositions 2.1 and 2.2 characterize the sharp identified set for ρ_{uz}, ρ_{T^*u}, and κ showing us exactly how the problems of instrument invalidity, treatment endogeneity, and measurement error are mutually constrained by each other and the data. From the identified set for $(\rho_{uz}, \rho_{T^*u}, \kappa)$ we can easily derive the identified set for any other parameters of the model in Equation 3. In particular we can find the identified set for β, the main object of interest to an applied researcher. Unfortunately, and perhaps unsurprisingly, the model places no restrictions on the causal effect in spite of the bounds it yields for ρ_{uz} and κ.

Corollary 2.2 (No Restriction on β)**.** *Under the conditions of Proposition 2.2, the sharp identified set for β is $(-\infty, \infty)$.*

The only way to learn about β in this model is to impose beliefs. For example, the standard IV identification assumption (belief) imposes $\rho_{uz} = 0$, which point identifies β. But this belief could imply an implausible amount of measurement error or selection effect. This fact highlights the central point of our analysis: our beliefs about ρ_{uz} are constrained by any beliefs we may have about ρ_{T^*u} and κ. This observation has two important consequences. First, it provides us with the opportunity to *incorporate* our beliefs about measurement error and the endogeneity of the treatment to improve our estimates. Failing to use this information is like leaving money on the table. Second, it disciplines our beliefs to prevent us from reasoning to a contradiction. Without knowledge of the form of the identified set, applied researchers could easily state beliefs that are mutually incompatible without realizing it. Our analysis provides a tool for them to realize this and adjust their beliefs accordingly. While we have thus far only discussed beliefs about ρ_{uz}, ρ_{T^*u}, and κ, among other things, one could also work backwards from beliefs about β to see how they constrain the identified set. We explore this possibility in one of our examples below.

2.4 Accommodating Exogenous Controls

At the beginning of Section 2 we assumed either that there were no control regressors or that they had been projected out. Because the control regressors $\mathbf{x}$ are exogenous, this is

innocuous, as we now show. Without loss of generality, suppose that (T^*, z, y) are mean zero or have been demeaned. Let $(\widetilde{T}^*, \widetilde{T}, \widetilde{y}, \widetilde{z})$ denote the residuals from a linear projection of the random variables (T^*, T, y, z) on $\mathbf{x}$, e.g. $\widetilde{T}^* = T^* - \Sigma_{T^*\mathbf{x}}\Sigma_{\mathbf{xx}}^{-1}\mathbf{x}$ and so on, where Σ_{ab} is shorthand for $Cov(a, b)$. Then, provided that $(T^*, T, \mathbf{x}, y, z)$ satisfy the model described above, it follows that

$$\begin{aligned} \widetilde{y} &= \beta\widetilde{T}^* + u \quad (17) \\ \widetilde{T}^* &= \pi\widetilde{z} + v \quad (18) \\ \widetilde{T} &= \widetilde{T}^* + w \quad (19) \end{aligned}$$

since $\mathbf{x}$ is uncorrelated with u and w by assumption and uncorrelated with v by construction. The parameters of this transformed system, β and π, are identical to those of the original system, as are the error terms. And because the transformed system contains no covariates, the analysis presented above applies directly. Projecting out covariates does, however, alter the definition of the structural parameters. The equations for the identified set presented above will involve not $(\rho_{uz}, \rho_{T^*u}, \kappa)$ but their analogues for the transformed system, namely

$$\begin{aligned} \widetilde{\kappa} &= \text{Var}(\widetilde{T}^*)/\text{Var}(\widetilde{T}) \\ \widetilde{\rho}_{\widetilde{T}^*u} &= \text{Cor}(\widetilde{T}^*, u) \\ \widetilde{\rho}_{u\widetilde{z}} &= \text{Cor}(\widetilde{z}, u). \end{aligned}$$

All of the results derived above continue to apply to the transformed system; they simply refer to $(\widetilde{\rho}_{u\widetilde{z}}, \widetilde{\rho}_{\widetilde{T}^*u}, \widetilde{\kappa})$. This is extremely convenient for both of the correlation parameters. In the presence of covariates, $\widetilde{\rho}_{\widetilde{T}^*u}$ is the measure of treatment endogeneity over which researchers are most likely to be able to state a belief because it is *net* of covariates. Similarly, $\widetilde{\rho}_{u\widetilde{z}}$ is the natural measure of instrument invalidity because the usual exclusion restriction is a lack of correlation between the instrument and the error term after accounting for the effect of exogeneous controls. For this reason, we elicit researcher beliefs directly over $\widetilde{\rho}_{u\widetilde{z}}$ and $\widetilde{\rho}_{\widetilde{T}^*u}$.[11]

The situation is different for the measurement error parameter, κ. This quantity is not defined relative to any causal model – it is simply a function of the signal-to-noise ratio for the observed treatment T. Thus, irrespective of whether covariates are available, κ rather than $\widetilde{\kappa}$ is the natural quantity over which to elicit researcher beliefs. Fortunately, there is a

[11] If desired, a researcher could instead state beliefs over $(\rho_{uz}, \rho_{T^*u}, \kappa)$ and then transform them into beliefs over $(\widetilde{\rho}_{u\widetilde{z}}, \widetilde{\rho}_{\widetilde{T}^*u}, \widetilde{\kappa})$ using the results of a regression of T on $\mathbf{x}$. See, e.g. Equation 20.

simple mapping between κ and $\widetilde{\kappa}$, namely

$$\widetilde{\kappa} = \frac{\sigma^2_{T^*} - \Sigma_{T\mathbf{x}}\Sigma^{-1}_{\mathbf{xx}}\Sigma_{\mathbf{x}T}}{\sigma^2_T - \Sigma_{T\mathbf{x}}\Sigma^{-1}_{\mathbf{xx}}\Sigma_{\mathbf{x}T}} = \frac{\sigma^2_{T^*}(1 - \Sigma_{T\mathbf{x}}\Sigma^{-1}_{\mathbf{xx}}\Sigma_{\mathbf{x}T}/\sigma^2_{T^*})}{\sigma^2_T(1 - \Sigma_{T\mathbf{x}}\Sigma^{-1}_{\mathbf{xx}}\Sigma_{\mathbf{x}T}/\sigma^2_T)} = \frac{\kappa - R^2_{T.\mathbf{x}}}{1 - R^2_{T.\mathbf{x}}} \tag{20}$$

where $R^2_{T.\mathbf{x}}$ denotes the population R-squared from a regression of T on $\mathbf{x}$.[12] Equation 20 relates $\kappa \equiv \sigma^2_{T^*}/\sigma^2_T$ for the original system to the analogue $\widetilde{\kappa}$, purely in terms of an identified quantity: $R^2_{T.\mathbf{x}}$. Thus, if a researcher states beliefs over κ, we can easily transform them to the implied beliefs about $\widetilde{\kappa}$ simply by using the R-squared that results from the regression that projects $\mathbf{x}$ out of T.

Since we can always reduce a problem with exogenous covariates to one without, and because we can describe the mapping between the parameters that govern the identified set of the original problem and those of the transformed system, we can easily accommodate control variables in our framework. In practice, one simply projects out $\mathbf{x}$ before proceeding, using the R-squared from a regression of T on $\mathbf{x}$ to transform between κ and $\widetilde{\kappa}$.[13]

3 Inference for a Continuous Treatment

Having characterized the identified set for this problem, we now describe how to use it to carry out statistical inference on quantities of interest. Doing so requires us to tackle two sources of uncertainty. First, while they must satisfy the restrictions given in Propositions 2.1 and 2.2, the true values of ρ_{uz}, ρ_{T^*u}, and κ are unknown. This source of uncertainty is the lack of identification. Second, the covariance matrix Σ of the observables (T, y, z), which pins down the relationship between $(\rho_{uz}, \rho_{T^*u}, \kappa)$, must be estimated from data and is thus subject to sampling uncertainty. We can handle these two sources of uncertainty separately because our parameterization is *transparent*.[14] In a transparent parameterization there are two groups of parameters: "reduced form" and "structural." The reduced form parameters, denoted by φ, are directly identified by the data whereas the structural parameters, denoted by θ, are not: the identified set Θ for θ depends on the data only through φ. Thus we write $\theta \in \Theta(\varphi)$. In the case of a continuous treatment subject to classical measurement error, $\varphi = \Sigma$ while $\theta = (\rho_{uz}, \rho_{T^*u}, \kappa)$.

[12] This expression has appeared elsewhere in the literature, see e.g. Dale and Krueger (2002). It follows from the fact that $\Sigma_{T^*\mathbf{x}} = \Sigma_{T\mathbf{x}}$ since w is classical measurement error.

[13] Throughout our exercise, we hold the set of exogeneous regressors fixed. It could be interesting to investigate the effects of covariate selection on beliefs over endogeneity, instrument validity and measurement error, but this is beyond the scope of our paper. Oster (2017) considers a related question, namely how changes in R^2 and regression coefficients can inform researchers about the extent of omitted variable bias when selecting over control regressors.

[14] See, e.g., Gustafson (2015).

The approach we follow here is Bayesian, which makes the use of a transparent parameterization particularly convenient for inference. We proceed in two steps. First we generate posterior draws $\varphi^{(j)}$ for the reduced form parameters. Each of these draws determines a *conditional* identified set $\Theta(\varphi^{(j)})$ for the structural parameters θ. Because this identified set does not restrict β, producing meaningful inferences for the causal effect of interest requires a second step in which we impose researcher beliefs on $\Theta(\varphi^{(j)})$. The usual large-sample equivalence between Bayesian posterior credible intervals and frequentist confidence intervals holds for φ, because the reduced form parameters are identified (Moon and Schorfheide, 2012; Poirier, 1998). This makes the first step uncontroversial. The second step, in contrast, imposes researcher beliefs that can never be directly falsified by data. Nevertheless, our use of a transparent parameterization makes clear precisely where any identification beliefs we may choose to impose enter the problem: the data rule out certain values of φ, while our beliefs amount to placing restrictions on the conditional identified set $\Theta(\varphi)$. Whenever one imposes information beyond what is contained in the data, it is crucial to make clear how this affects the ultimate result.

3.1 Inference for the Reduced Form Parameters

For the model described in Section 2, the first step of our inference procedure requires producing posterior draws for the covariance matrix Σ of (T, y, z).[15] Because inference for this part of the problem is standard, the researcher can effectively "drop in" any procedure that generates posterior draws for Σ. Here we propose two simple possibilities. The first is based on a large-sample approximation that works well in sufficiently large samples. This method conditions on T and z and incorporates sampling uncertainty in σ_{Ty} and σ_{zy} only, by applying the Central Limit Theorem exactly as one does when deriving the frequentist large-sample distribution of IV and OLS estimators. Specifically, we draw $(\sigma_{Ty}^{(j)}, \sigma_{zy}^{(j)})$ from a normal distribution centered at the corresponding maximum likelihood estimates $(\widehat{\sigma}_{Ty}, \widehat{\sigma}_{zy})$ with a variance matrix estimated from the residuals we obtain by running two auxiliary regressions: y on T and y on z. Details appear in Appendix A.2.1. An advantage of this approach is that it is robust to heteroskedasticity without requiring us to model the conditional variance of the errors. Although it does not involve an explicit prior and likelihood, one can view this method as an approximation to a non-informative Bayesian analysis.

Unfortunately the large-sample approximation we have just outlined is not guaranteed to produce positive definite draws for Σ. When the sample size is large this is extremely unlikely to occur, but in small samples, such as our example from Section 5.1, this can be

[15] For simplicity we suppress exogenous covariates throughout this section. If they are present we simply project them out, as described in Section 2.4, and apply the methods described here to the resulting residuals.

problematic. A solution to this problem is to proceed in a fully Bayesian fashion rather than using an approximation based on the Central Limit Theorem. There are many possible ways to accomplish this. One simple method is to posit a joint normal likelihood for (T, y, z) and place a Jeffrey's prior on Σ, a benchmark noninformative prior that is often used in practice. Under this model the marginal posterior for Σ is inverse Wishart.[16] Draws $\Sigma^{(j)}$ produced in this way are guaranteed to be positive definite. Notice that this approach models the *joint* distribution of (T, z, y), which may seem odd given that the typical regression problem, Bayesian or frequentist, models only the conditional distribution of y given T and z. This is less of a concern in examples featuring a large number of exogenous control regressors. These are projected out before proceeding, so we are in effect positing a normal distribution only for the residuals of the regressions of (T, y, z) on $\mathbf{x}$.

3.2 Inference for the Structural Parameters

Every draw $\Sigma^{(j)}$ from the first step of our inference procedure determines a conditional identified set $\Theta(\Sigma^{(j)})$ for the structural parameters $(\rho_{uz}, \rho_{T^*u}, \kappa)$. We now discuss several ways to summarize the information contained in $\Theta(\Sigma^{(j)})$, proceeding from most conservative to least. Whichever summary one chooses, the resulting inference is obtained by averaging over the reduced form draws $\Sigma^{(j)}$.

Recall from our discussion in Section 2 above that Σ restricts neither ρ_{T^*u} nor β. It does however provide a lower bound for κ via Proposition 2.2. Computing this bound at each draw $\Sigma^{(j)}$ provides posterior inference for the maximum amount of measurement error compatible with our assumptions, given the data. One could proceed similarly for the one-sided bound for ρ_{uz} using Corollary 2.1. Going beyond this, however, requires imposing beliefs.

Sign and interval restrictions on the degree of measurement error, treatment endogeneity, and instrument invalidity are often straightforward to elicit in practice. By imposing such restrictions we can add relatively weak prior information to the problem and restrict Θ accordingly. In the discussion that follows we denote by $\mathcal{R} \subset (-1, 1) \times (-1, 1) \times (0, 1]$ a user-imposed restriction on the domain of $(\rho_{uz}, \rho_{T^*u}, \kappa)$. Incorporating beliefs in this way has the potential to bound the treatment effect. Calculating the bounds implied by each $\Theta(\Sigma^{(j)}) \cap \mathcal{R}$ provides posterior inference for the identified set for β under our beliefs over $(\rho_{uz}, \rho_{T^*u}, \kappa)$. Yet, even when $\Theta(\Sigma^{(j)}) \cap \mathcal{R}$ is not particularly informative about β, it can easily rule out a wide range of values for ρ_{uz}, ρ_{T^*u} and κ. For example, suppose a researcher

[16] Specifically we suppose that $(T_i, y_i, z_i) \overset{iid}{\sim} N(\boldsymbol{\mu}, \Sigma)$ and place the prior $\pi(\boldsymbol{\mu}, \Sigma) \propto |\Sigma|^{-2}$ on the mean vector and variance matrix. A standard calculation shows that the marginal posterior for Σ is $\Sigma|\mathbf{T}, \mathbf{y}, \mathbf{z} \sim$ Inverse-Wishart$(n-1, S)$ where the scale matrix S equals $n-1$ times the sample covariance matrix of (T, y, z).

strongly believes that $\rho_{T^*u} < 0$. At a given draw $\Theta(\Sigma^{(j)})$ this restriction could very easily rule out $\rho_{uz} = 0$, as we see from Equation 13. Calculating the proportion of draws $\Sigma^{(j)}$ that are compatible with $\rho_{uz} = 0$ gives the posterior probability of a valid instrument under the belief that $\rho_{T^*u} < 0$. If one imposes beliefs over two or more of $(\rho_{uz}, \rho_{T^*u}, \kappa)$, $\Theta(\Sigma^{(j)}) \cap \mathcal{R}$ could even be empty for certain draws $\Sigma^{(j)}$. Calculating the proportion of such empty identified sets gives the posterior probability that our beliefs are mutually incompatible, given the data. This illustrates an important general point of our approach. By making explicit the relationship between measurement error, treatment endogeneity, and instrument invalidity, our method allows researchers to learn whether their beliefs over these different dimensions of the problem cohere.

An important advantage of the inferences we have described thus far is that, while Bayesian, they can be given a valid frequentist interpretation under mild regularity conditions.[17] This is because they do not impose a prior on the conditional identified set; they merely intersect it with researcher beliefs, stated as interval restrictions. A more thoroughly Bayesian treatment, on the other hand, will impose a fully-fledged prior on $\Theta(\Sigma^{(j)})$, putting the two sources of uncertainty – lack of identification, and sampling uncertainty in the reduced form parameters – on equal footing. Although more controversial because it can no longer be given a frequentist interpretation, this approach has a key advantage: rather than summarizing only the most extreme points of $\Theta(\Sigma^{(j)}) \cap \mathcal{R}$, it provides a more complete picture by averaging over this set. Inferences that rely only on interval restrictions are necessarily very sensitive to small changes in $\mathcal{R}$ and inherently pessimistic. In contrast, because any reasonable prior will place only a small amount of probability density near the boundaries, averaging over $\Theta(\Sigma^{(j)}) \cap \mathcal{R}$ can produce more robust inferences.

In most cases it will not be feasible to elicit a fully informative prior over the conditional identified set. Its support, for example, changes with each draw $\Sigma^{(j)}$. For this reason we suggest an approach based on a conditionally uniform reference prior that gives equal weight to regions of the support with equal area.[18] Specifically, we draw uniformly over the intersection of $\mathcal{R}$ with the manifold $(\rho_{uz}, \rho_{T^*u}, \kappa)$ that describes the identified set for instrument invalidity, treatment endogeneity, and measurement error, a two-dimensional manifold embedded in three-dimensional space.[19]

While a uniform distribution seems like the natural choice for representing prior ignorance some caution is warranted: uniformity in one parameterization could imply a highly

[17] These conditions concern the first step of the procedure: inference for the reduced form parameters. See Moon and Schorfheide (2012) and Kline and Tamer (2016).

[18] Moon and Schorfheide (2012) likewise employ a conditionally uniform reference prior in their example of a two-player entry game.

[19] For details see Appendix A.2.2.

informative prior in some different parameterization. This is unavoidable. We emphasize, however, that the uniform serves here as a reference prior only. As such, one need not take it completely literally but could instead consider, for example, what kind of deviation from uniformity would be necessary to support a particular belief about β. We explore this possibility in our examples below.

4 The Case of a Binary Treatment

4.1 Model and Assumptions

Although the logic of our approach from above is general, our characterization of the identified set does not apply when the treatment of interest is binary. This is because, as we mentioned in the introduction, a binary instrument cannot be subject to classical measurement error. Accordingly, our characterization of the identified set for this common setting will require a different approach. Let T, T^*, and z be binary variables.[20] We continue to allow for treatment endogeneity and instrument invalidity: both T^* and z are potentially correlated with u. For convenience we absorb the intercept into the error term u as follows

$$y = \beta T^* + u \tag{21}$$

$$u = c + \varepsilon \tag{22}$$

where ε is mean zero but u may not be. For simplicity, we begin by assuming that there are no covariates. In Section 4.4 we show how to account for the effect of covariates by transforming the geometry of the problem.[21] This is important for elicitation because researcher beliefs over treatment endogeneity and instrument invalidity are typically conditional on covariates.

4.1.1 Non-differential Measurement Error

Since T and T^* are both binary, measurement error is governed by two probabilities:

$$\alpha_0 = \mathbb{P}(T = 1|T^* = 0) \tag{23}$$

$$\alpha_1 = \mathbb{P}(T = 0|T^* = 1) \tag{24}$$

These mis-classification probabilities replace κ from the case of a continuous treatment. As mentioned above, it is impossible for a binary regressor to be subject to classical measurement

[20] If a continuous instrument is available, it can always be binarized.

[21] Observe that this is different from our treatment of the continuous treatment case. There we could simply project out any exogenous covariates from all other observables.

error: the true value T^* must be *negatively* correlated with the measurement error w. To see why, first note that since T and T^* are both binary, $w = T - T^*$ can only take on values in the set $\{-1, 0, 1\}$. The conditional distribution of w given T^* is as follows:

$$T^* = 0 \implies \begin{cases} T = 0 & \text{with prob. } 1 - \alpha_0 \iff w = 0 \\ T = 1 & \text{with prob. } \alpha_0 \iff w = 1 \end{cases}$$

$$T^* = 1 \implies \begin{cases} T = 0 & \text{with prob. } \alpha_1 \iff w = -1 \\ T = 1 & \text{with prob. } 1 - \alpha_1 \iff w = 0 \end{cases}$$

Hence $\mathbb{E}[w|T^* = 0] = \alpha_0$, while $\mathbb{E}[w|T^* = 1] = -\alpha_1$. Since classical measurement error is impossible, we assume instead that the measurement error is *non-differential*, the closest assumption to classical measurement error in the context of a binary treatment. Non-differential measurement error requires that w be conditionally independent of all other random variables in the system given knowledge of true treatment status T^*. Consider, for example, self-reports of smoking behavior. The non-differential measurement error assumption allows for the possibility that smokers are more likely to mis-represent their true smoking status than nonsmokers. After controlling for true smoking status, however, it rules out any relationship between measurement error and the instrument, as well as any other unobserved characteristics that determine the outcome y. The precise assumption we use below takes the following form:

Assumption 4.1 (Non-differential Measurement Error)**.**

(i) $\mathbb{E}[\varepsilon|T, T^*, z] = \mathbb{E}[\varepsilon|T^*, z]$, $\mathbb{E}[\varepsilon^2|T, T^*, z] = \mathbb{E}[\varepsilon^2|T^*, z]$

(ii) $\mathbb{P}(T = 1|T^*, z) = \mathbb{P}(T = 1|T^*)$

Because we only work with first and second moments of the observables, Assumption 4.1, rather than full conditional independence, suffices. We also impose an assumption about the extent of measurement error, which is standard in the literature on mis-classified binary regressors.

Assumption 4.2 (Extent of Measurement Error)**.** *Assume that* $\alpha_0 + \alpha_1 < 1$.

As shown in Lemma B.3 , $Cov(T, T^*) = (1 - \alpha_0 - \alpha_1)\text{Var}(T^*)$ so Assumption 4.2 amounts to asserting that T and T^* are positively correlated, or equivalently that the misclassification is "not so bad ... that the effective definition of the classification has been reversed" (Bollinger, 1996, p. 389).

Assumptions 4.1 and 4.2 have two implications that contrast sharply with those of the classical measurement error case from above. While these have been known in the literature for some time, they do not appear to be very widely appreciated. First, while the IV estimator is unaffected by classical measurement error (Equation 7), it is affected by non-differential measurement error. In particular,

$$\beta_{IV} = \frac{\beta}{1-\alpha_0-\alpha_1} + \frac{\sigma_{zu}}{\sigma_{zT}}. \tag{25}$$

as explained in Lemma B.7. Indeed, under a valid instrument β_{IV} is necessarily an *overestimate* of β: the opposite of the familiar OLS attenuation bias logic for the case of classical measurement error.[22] Second, under non-differential measurement error it is no longer true that $\text{Var}(T^*) \leq \text{Var}(T)$. Let $p = \mathbb{P}(T = 1)$ and $p^* = \mathbb{P}(T^* = 1)$. By the law of total probability,

$$\sigma^2_{T^*} = \text{Var}(T^*) = p^*(1-p^*) = \frac{(p-\alpha_0)(1-p-\alpha_1)}{(1-\alpha_0-\alpha_1)^2} \tag{26}$$

whereas $\sigma^2_T = \text{Var}(T) = p(1-p)$. The probability limit of the OLS estimator in this case is accordingly more complicated. In particular,

$$\beta_{OLS} = \frac{\sigma^2_{T^*}}{\sigma^2_T}\left[\beta\left(1-\alpha_0-\alpha_1\right) + \frac{\sigma_{T^*u}}{\sigma^2_{T^*}}\right] \tag{27}$$

as explained in Lemma B.9. Again, contrast this with the case of classical measurement error from Equation 6 from above. Although this is not immediately apparent from the form of Equation 27, if $\sigma_{T^*u} = 0$ then OLS is attenuated towards zero whenever $\alpha_0 + \alpha_1 < 1$.[23]

4.1.2 Notation: Observables and Unobservables

Unlike their counterparts for the continuous case, Equations 25 and 27 do not allow us to recover β even if both the instrument and regressor are exogenous. This is because they do not incorporate all information contained in the data for the binary case. Kane et al. (1999), Black et al. (2000) and Frazis and Lowenstein (2003) show, however, that if the instrument and regressor are jointly exogenous then β can be consistently estimated via a method of moments approach that uses strictly more information than is contained in the OLS and IV

[22] Without Assumption 4.2, β_{IV} could have the wrong sign even if the instrument is valid.

[23] To see why note that, by Lemma B.1, $p^* = (p-\alpha_0)/(1-\alpha_0-\alpha_1)$ and $1-p^* = (1-p-\alpha_1)/(1-\alpha_0-\alpha_1)$. Since both p^* and $1-p^*$ must be positive, $\alpha_0+\alpha_1 < 1$ implies $\alpha_0 < p$ and $\alpha_1 < 1-p$. After expanding, the term that multiples β in Equation 27 equals $[p(1-p)-p\alpha_1-\alpha_0(1-p)]/[p(1-p)-p(1-p)\alpha_1-\alpha_0 p(1-p)]$. The result follows since the second term in the numerator is greater than the second term in the denominator and the same holds for the third terms.

estimators.[24] Although we do not assume that the treatment and instrument are exogenous, a full characterization of the identified set relies on the additional information exploited by these method of moments estimators.

The simplest way to incorporate this additional information is to work with the joint probability distribution of (z, T) and the conditional means $\bar{y}_{tk} \equiv \mathbb{E}[y|T = t, z = k]$ for $t, k \in \{0, 1\}$ as depicted in Table 1a. First, define $p_k^* = \mathbb{P}(T^* = 1|z = k)$ and $p_k = \mathbb{P}(T = 1|z = k)$. Although p_k^* is unobserved, it is related to the observed probability p_k by $p_k^* = (p_k - \alpha_0)/(1 - \alpha_0 - \alpha_1)$ as shown in Lemma B.1. This means that p_k^* is observed up to knowledge of α_0, α_1 so we need not consider it a separate unknown. Now, as shown in Lemma B.11, the observed conditional means can be related to the unobservable ones by

$$\widetilde{y}_{0k} \equiv (1 - p_k)\bar{y}_{0k} = (\beta + m_{1k}^*)\alpha_1 p_k^* + (1 - \alpha_0)(1 - p_k^*)m_{0k}^* \tag{28}$$

$$\widetilde{y}_{1k} \equiv p_k \bar{y}_{1k} = (\beta + m_{1k}^*)(1 - \alpha_1)p_k^* + \alpha_0(1 - p_k^*)m_{0k}^* \tag{29}$$

where $m_{tk}^* \equiv \mathbb{E}[u|T = t, z = k]$ for $t, k \in \{0, 1\}$, as depicted in Table 1b. Because of the mis-classification, each of the means in Table 1a contains a mixture of treated and untreated individuals, depending on the values of α_0 and α_1. The expression for $\bar{y}_{00}$, for example, involves not only m_{00}^* but also β and m_{10}^*.

In addition to conditional means, $\bar{y}_{tk}$, we assume that conditional variances of the outcome $\sigma_{tk}^2 = \text{Var}(y|T = t, z = k)$ are likewise observed. This information has not been used in the existing literature because, under joint exogeneity of the instrument and treatment, one obtains point identification from conditional means alone. Without this assumption the model is unidentified, and conditional variance information becomes useful. Let $s_{tk}^{*2} = \text{Var}(u|T^* = t, z = k)$. Equations B.13 and B.15 in the Appendix are the counterparts of Equations 28 and 29 for second moments: they relate the observable variances σ_{tk}^2, depicted in Table 1a to the unobservable variances s_{tk}^{*2} depicted in Table 1b. As we show below, the restriction $s_{tk}^{*2} > 0$ will allow us to tighten our bounds for the mis-classification probabilities.

4.2 A Convenient Parameterization

While the m_{tk}^* provide a very convenient way of expressing how misclassification pollutes the conditional means of y, they depend simultaneously on both the extent of treatment endogeneity and instrument invalidity. The assumption of an exogenous treatment, $\sigma_{T^*u} = 0$, is equivalent to

$$\frac{1}{\mathbb{P}(T^* = t)} \sum_k p_{tk}^* m_{tk}^* = c$$

[24] The required condition is slightly stronger than $\sigma_{T^*u} = \sigma_{zu} = 0$. It is in fact $\mathbb{E}[\varepsilon|T, z] = 0$.

(a) Observables

	$z=0$	$z=1$
$T=0$	$\bar{y}_{00}$, σ^2_{00}, p_{00}	$\bar{y}_{01}$, σ^2_{01}, p_{01}
$T=1$	$\bar{y}_{10}$, σ^2_{10}, p_{10}	$\bar{y}_{11}$, σ^2_{11}, p_{11}

(b) Unobservables

	$z=0$	$z=1$
$T^*=0$	m^*_{00}, s^{*2}_{00}, p^*_{00}	m^*_{01}, s^{*2}_{01}, p^*_{01}
$T^*=1$	m^*_{10}, s^{*2}_{10}, p^*_{10}	m^*_{11}, s^{*2}_{11}, p^*_{11}

Table 1: We observe $p_{tk} = \mathbb{P}(T = t, z = k)$, $\bar{y}_{tk} = \mathbb{E}[y|T = t, z = k]$, and $\sigma^2_{tk} = \text{Var}(y|T = t, z = k)$. In contrast, $p^*_{tk} = \mathbb{P}(T^* = t, z = k)$, $m^*_{tk} = \mathbb{E}[u|T = t, z = k]$, and $s^{*2}_{tk} = \text{Var}(u|T^* = t, z = k)$ are unobservables.

for $t = 0, 1$ while that of an exogenous instrument, $\sigma_{zu} = 0$, is equivalent to

$$(1 - p^*_k)m^*_{0k} + p^*_k m^*_{1k} = c$$

for $k = 0, 1$ where c is the constant term from Equation 22.[25] This shows that the objects over which researchers often hold and express beliefs – treatment exogeneity and instrument invalidity – are not the m^*_{tk} themselves, but rather certain functions of them. For this reason, in the case of a binary treatment and instrument it is more natural to elicit researcher beliefs in terms of the following quantities

$$\delta_{T^*} \equiv \mathbb{E}[u|T^* = 1] - \mathbb{E}[u|T^* = 0] \tag{30}$$

$$\delta_z \equiv \mathbb{E}[u|z = 1] - \mathbb{E}[u|z = 0] \tag{31}$$

The first, δ_{T^*}, measures the average difference in unobservables between the treated and untreated; the second, δ_z, measures the average difference in unobservables between those with the high value of the instrument and those with the low. Both quantities are linear functions of m^*_{tk} with coefficients that depend on α_0, α_1, and observables, as shown in Lemma B.12. Although δ_{T^*} and δ_z are not scale-free, both are empirically meaningful and conveniently measured in units of y. For example, consider the smoking cessation randomized controlled trial studied in Courtemanche et al. (2016) where z is the randomized offer to participate in a smoking cessation program, T^* is an indicator of true smoking cessation, T is self-reported smoking cessation, and y is body-mass index (BMI). Here δ_{T^*} is the selection effect. For example, those who succeed in quitting smoking are likely more health-conscious overall. We would expect them to have a lower BMI on average even if they had not quit smoking,

[25] Without covariates, the exogeneity assumption used by Kane et al. (1999), Black et al. (2000), Frazis and Lowenstein (2003) is equivalent to $m^*_{tk} = c$, for $t, k \in \{0, 1\}$.

leading to a negative value of δ_{T^*}. We would also expect a knowledgeable obesity researcher to be able to put a lower bound on δ_{T^*}. But what about δ_z? Courtemanche et al. (2016) point out that, in spite of being randomized, the offer to participate in a smoking cessation program may have a direct effect on BMI, making it an invalid instrument. For example, those who participate in the smoking cessation program may be led to smoke less even if they fail to quit entirely. Because nicotine is an appetite suppressant, this would likely lead to a *positive* value of δ_z. We would also expect our researcher to be able to provide at least a reasonable order of magnitude for δ_z, although in most applications of our framework, researchers will likely prefer to compute the value of δ_z consistent with their other beliefs rather than the reverse.

In addition to δ_{T^*} and δ_z, the identified set will also depend on α_0 and α_1. Fortunately, both of these quantities are probabilities so they are already directly intelligible, unitless, and bounded.

4.3 Deriving the Identified Set for $(\delta_{T^*}, \delta_z, \alpha_0, \alpha_1)$

We begin by deriving the relationship between $\delta_{T^*}, \delta_z, \alpha_0$ and α_1, the objects over which we can elicit researcher beliefs, by eliminating m^*_{tk} and β. The derivation proceeds in two steps. We first manipulate Equations 28 and 29 to yield an expression for $m^*_{10} - m^*_{11}$ that depends only on observables and α_0. Combining this with the definitions of δ_{T^*} and δ_z in terms of m^*_{tk} from Lemma B.12 gives an overdetermined linear system of three equations in (m^*_{10}, m^*_{11}) given $(\alpha_0, \alpha_1, \delta_{T^*}, \delta_z)$ and observables. Solving to eliminate m^*_{10} and m^*_{11} we derive a linear relationship between δ_z and δ_{T^*} given α_0, α_1 and observables.

Proposition 4.1. *Under Assumption 4.1*

$$\delta_z = B(\alpha_0, \alpha_1) + S(\alpha_0, \alpha_1)\delta_{T^*}. \tag{32}$$

where

$$B(\alpha_0, \alpha_1) = \frac{g(\alpha_1) - (p_0 - p_1)h(\alpha_1)}{1 - \alpha_0 - \alpha_1} - \frac{(p_0 - \alpha_0)(p_1 - \alpha_0)\Delta(\alpha_0)}{(p - \alpha_0)(1 - \alpha_0 - \alpha_1)}$$
$$S(\alpha_0, \alpha_1) = \frac{p_1 - p_0}{1 - \alpha_0 - \alpha_1},$$

and g, h, and Δ are simple functions of α_0, α_1 and observables defined in the proof.

Note that the slope of the relationship between δ_z and δ_{T^*} is directly proportional to the strength of the instrument: $p_1 - p_0$. The mis-classification probabilities, on the other hand, enter in nonlinear fashion in both the slope and intercept.

Proposition 4.1 allows us to express δ_z in terms of δ_{T^*}, α_0, α_1 and the observable probabilities and conditional means from Table 1a. Thus, to fully characterize the relationship between instrument invalidity, treatment endogeneity, and measurement error it suffices to derive the sharp identified set for δ_{T^*}, α_0 and α_1. Just as we were able to construct bounds for κ in the case of classical measurement error, we can bound the mis-classification error rates α_0 and α_1 in the binary regressor case.

Proposition 4.2 (Sharp Identified Set for δ_{T^*}, α_0, and α_1)**.** *Suppose that $s^{*2}_{tk} > 0$ for all t, k. Then, under Assumptions 4.1 and 4.2, $(\delta_{T^*}, \alpha_0, \alpha_1) \in (-\infty, \infty) \times [0, \bar{\alpha}_0) \times [0, \bar{\alpha}_1)$ where*

$$\bar{\alpha}_0 = \min_k \left\{ \bar{\alpha}_0^k \right\}, \quad \bar{\alpha}_1 = \min_k \left\{ f_{0k}(\bar{\alpha}_0^k) \right\},$$

$\bar{\alpha}_0^k$ is the smallest solution to $f_{0k}(\alpha_0) = f_{1k}(\alpha_0)$, and

$$f_{0k}(\alpha_0) = \frac{(p_k - \alpha_0)\sigma^2_{0k}}{(p_k - \alpha_0)\sigma^2_{0k} + (\bar{y}_{1k} - \bar{y}_{0k})^2 p_k^2 (1 - \alpha_0)} \tag{33}$$

$$f_{1k}(\alpha_0) = \frac{(1 - p_k)(p_k - \alpha_0)\sigma^2_{1k} - (\bar{y}_{1k} - \bar{y}_{0k})^2 (1 - p_k)^2 \alpha_0}{(p_k - \alpha_0)\sigma^2_{1k} - (\bar{y}_{1k} - \bar{y}_{0k})^2 (1 - p_k)^2 \alpha_0} \tag{34}$$

for $k = 0, 1$. These bounds are sharp.

The result of Proposition 4.2 is analogous to that of Proposition 2.2 for the continuous treatment case; in each case we obtain a non-trivial upper bound on the extent of measurement error, but no restriction on the extent of treatment endogeneity. The proof in the binary case proceeds by showing that $\alpha_0 < \bar{\alpha}_0$ and $\alpha_1 < \bar{\alpha}_1$ is equivalent to $s^{*2}_{tk} > 0$ for all $t, k = 0, 1$. As a result, the variance information places no restrictions on m^*_{tk}. Since the conditional mean information from Equations 28 and 29 also places no restrictions on m^*_{tk}, it follows that δ_{T^*} is unbounded.

Figure 1 illustrates how to construct the bounds for α_0 and α_1, using data from one of our empirical examples below. The region in which Assumption 4.2 holds ($\alpha_0 + \alpha_1 < 1$) is shaded in light gray. Under this assumption, the equality $p^*_k = (p_k - \alpha_0)/(1 - \alpha_0 - \alpha_1)$ implies that $\alpha_0 \leq \min_k\{p_k\}$ and $\alpha_1 \leq \min_k\{1 - p_k\}$, as shown in Lemma B.14. These bounds, which do not incorporate the information contained in the conditional variances of y, are depicted in light blue. The sharp bounds, depicted in dark blue, are determined by the intersection of f_{11} with f_{01} and f_{10} with f_{00}. Each point of intersection provides a bound for both α_0 and α_1. Since all of these bounds must hold simultaneously, however, only the smallest of each binds. In Figure 1 the intersection of f_{10} with f_{00} determines the binding constraint for α_1 while the intersection of f_{11} with f_{01} determines the binding constraint for α_0.

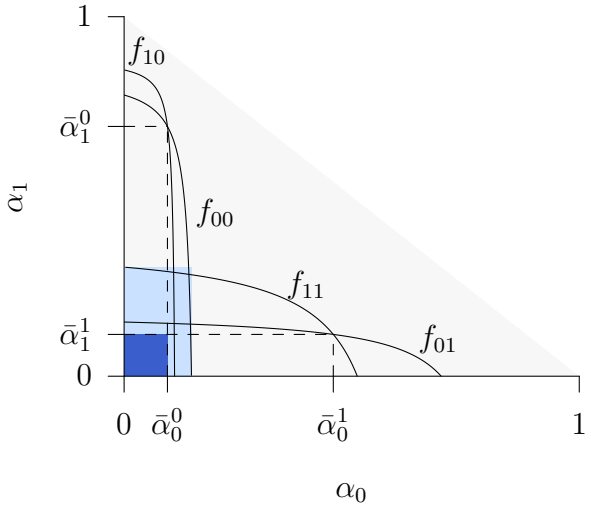


Figure 1: The identified set for (α_0, α_1) from Proposition 4.2. The region where Assumption 4.2 is satisfied ($\alpha_0 + \alpha_1 < 1$) is depicted in light gray and the weak bounds $\alpha_0 < \min_k\{p_k\}$, $\alpha_1 < \min_k\{1 - p_k\}$ are shown in light blue. The region in dark blue gives the sharp bounds $\alpha_0 < \min_k\{\bar{\alpha}_0^k\}$ and $\alpha_1 < \min\{\bar{\alpha}_1^k\}$, where $(\bar{\alpha}_0^k, \bar{\alpha}_1^k)$ is the intersection of f_{0k} with f_{1k}, as defined in Proposition 4.2. This figure uses estimates of the observables from the example in Section 5.3.

Finally, as in the continuous treatment case, the assumptions of our model place no restrictions on β.

Corollary 4.1. *Under the Assumptions of Proposition 4.2, the sharp identified set for β and δ_z are both $(-\infty, \infty)$.*

The careful reader may wonder whether an analogue of the condition that Ω is positive semi-definite, from Assumption 2.1 in the continuous treatment case, provides any additional restrictions when the treatment is binary. The answer is no. Because T^*, z, w and the analogue of v are all discrete, the analogue of Ω in the binary case is guaranteed to be positive semi-definite provided that all probabilities are between zero and one, and all conditional variances of y are positive.[26]

Now that we have the identified set for the case of a binary treatment, we can proceed in the same way as we did for the continuous treatment case described above in Section 2. In particular, we can intersect researcher beliefs over measurement error, treatment endogeneity and instrument invalidity with the identified set itself to check whether these

[26]Let $\boldsymbol{\eta}$ be a random vector, ξ be a binary random variable, and define $\Omega = \text{Var}(\boldsymbol{\eta})$, $\Omega_0 = \text{Var}(\boldsymbol{\eta}|\xi = 0)$, and $\Omega_1 = \text{Var}(\boldsymbol{\eta}|\xi = 1)$. A straightforward calculation shows that if Ω_0 and Ω_1 are positive semi-definite, so is Ω. In our case, we simply apply this fact recursively to condition on both T^* and z.

beliefs are mutually consistent given the data and, if so, harness them to learn about the treatment effect. As in the continuous treatment case, each point on the identified set implies a corresponding value for the treatment effect β. A convenient way to compute this value is to use the IV probability limit from Equation 25. Since $\sigma_{zu}/\sigma_{zT} = \delta_z/(p_1 - p_0)$, $\beta = (1 - \alpha_0 - \alpha_1)[\beta_{IV} - \delta_z/(p_1 - p_0)]$ as explained in Lemma B.10.

4.4 Accommodating Exogenous Covariates

In the presence of covariates we redefine u from Equation 22 as

$$u = c + \mathbf{x}'\boldsymbol{\gamma} + \varepsilon \tag{35}$$

and the following replaces Assumption 4.1:

Assumption 4.3 (Non-differential Measurement Error with Exogenous Covariates)**.**

(i) $\mathbb{E}[\mathbf{x}|T, T^*, z] = \mathbb{E}[\mathbf{x}|T^*, z]$

(ii) $\mathbb{E}[\varepsilon|T, T^*, z] = \mathbb{E}[\varepsilon|T^*, z]$, $\mathbb{E}[\varepsilon^2|T, T^*, z] = \mathbb{E}[\varepsilon^2|T^*, z]$

(iii) $\mathbb{P}(T = 1|T^*, z) = \mathbb{P}(T = 1|T^*)$.

Assumption 4.3 allows us to employ the results from above in the presence of exogenous covariates: the error term u is merely re-defined to make explicit the role of $\mathbf{x}$. In principle we could continue to express the identified set in terms of δ_{T^*} and δ_z with the understanding that they refer to a u with a slightly different meaning. In practice, however, researchers' beliefs about regressor endogeneity and instrument validity are likely to be *conditional* on covariates. Because the point is to study the effect of T^* net of $\mathbf{x}$, the more natural objects over which to elicit researcher beliefs regarding treatment endogeneity and instrument invalidity are

$$\widetilde{\delta}_z \equiv \mathbb{E}[\varepsilon|z = 1] - \mathbb{E}[\varepsilon|z = 0] \tag{36}$$

$$\widetilde{\delta}_{T^*} \equiv \mathbb{E}[\varepsilon|T^* = 1] - \mathbb{E}[\varepsilon|T^* = 0]. \tag{37}$$

Unlike the continuous treatment case, here we cannot simply project $\mathbf{x}$ out of the system if we wish to work with the information contained in the four cells from Table 1a. Instead, we now show how to re-express the identified set from the preceding section in terms of $\widetilde{\delta}_{T^*}$ and $\widetilde{\delta}_z$ rather than δ_{T^*} and δ_z.

As shown in Lemma B.22, we can relate $\widetilde{\delta}_{T^*}$ to δ_{T^*} and $\widetilde{\delta}_z$ to δ_z using the following

expressions:

$$\delta_z = \left[\mathbb{E}\left(\mathbf{x}|z=1\right) - \mathbb{E}\left(\mathbf{x}|z=0\right)\right]'\boldsymbol{\gamma} + \widetilde{\delta}_z \tag{38}$$

$$\delta_{T^*} = \frac{p(1-p)(1-\alpha_0-\alpha_1)}{(p-\alpha_0)(1-p-\alpha_1)}\left[E\left(\mathbf{x}|T=1\right) - E\left(\mathbf{x}|T=0\right)\right]'\boldsymbol{\gamma} + \widetilde{\delta}_{T^*} \tag{39}$$

If $\boldsymbol{\gamma}$ were known, these expressions would immediately allow us to re-write the identified set as desired. The problem, of course, is that $\boldsymbol{\gamma}$ is unknown. By an argument related to that of Frazis and Lowenstein (2003, p. 158), we show in Lemma B.1 that the probability limits of the IV estimators for β and $\boldsymbol{\gamma}$ are

$$\begin{bmatrix} \beta_{IV} \\ \boldsymbol{\gamma}_{IV} \end{bmatrix} = \begin{bmatrix} \beta/(1-\alpha_0-\alpha_1) \\ \boldsymbol{\gamma} \end{bmatrix} + \widetilde{\delta}_z q(1-q) \begin{bmatrix} \sigma^{zT} \\ \boldsymbol{\sigma}^{xT} \end{bmatrix} \tag{40}$$

where $q = \mathbb{P}(z=1)$ and σ^{zT} and $\boldsymbol{\sigma}^{xT}$, defined in the statement of Lemma B.1, depend only on covariances of the observables $(z, T, \mathbf{x})$. Rearranging Equation 40, we can write $\boldsymbol{\gamma}$ solely in terms of observable quantities and $\widetilde{\delta}_z$, namely $\boldsymbol{\gamma} = \boldsymbol{\gamma}_{IV} - \widetilde{\delta}_z q(1-q)\boldsymbol{\sigma}^{xT}$. Using this fact, we can eliminate $\boldsymbol{\gamma}$ from Equations 38 and 39. After doing so, both equations involve $\widetilde{\delta}_z$ but the relationship is linear. Accordingly, using Lemma B.12, we obtain a linear relationship between $\widetilde{\delta}_z$ and $\widetilde{\delta}_{T^*}$. As shown in Lemma B.22,

$$\widetilde{\delta}_z = \widetilde{B}(\alpha_0, \alpha_1) + \widetilde{S}(\alpha_0, \alpha_1)\widetilde{\delta}_{T^*} \tag{41}$$

where $\widetilde{B}$ and $\widetilde{S}$ are functions of α_0, α_1 and reduced form parameters defined in the Lemma. Thus, in the presence of exogenous covariates Equation 41 replaces Equation 32 and we calculate the same bounds for α_0 and α_1 as given in Proposition 4.2.[27]

4.5 Inference for a Binary Treatment

Our inference procedure for a binary treatment closely parallels the continuous treatment case described in Section 3, so we describe here only the differences. As above, we rely upon a transparent parameterization. In the binary case, the structural parameters are $\theta = (\delta_{T^*}, \delta_z, \alpha_0, \alpha_1)$, while the reduced form parameters, φ, are the joint probability distribution of (z, T) along with the conditional means and variances of y given z and T, as shown in

[27] In the presence of covariates, the bounds for α_0 and α_1 from Proposition 4.2 are technically no longer sharp, as one could in principle exploit the additional information contained in $\mathbf{x}$ to tighten them. If $\mathbf{x}$ is discrete and sufficient data are available, the sharp bounds can be obtained as follows: simply apply our bounds separately at every value in the support of the covariates and report the tightest. When one or more covariates are continuous, as is the case in each of our examples, one would need to model the first-stage relationship between $\mathbf{x}$ and T^*, which we prefer to avoid here.

Table 1a. We again propose a simple method for generating posterior draws for the reduced from parameters that matches the usual large-sample frequentist treatment of estimation error in IV and OLS regression. Accordingly, we condition on z and T and incorporate sampling uncertainty in the conditional means of y only, applying the central limit theorem. In the presence of covariates, we also require posterior draws for $\widehat{\gamma}_{IV}$. These must be made jointly with those for the conditional means of y as they are necessarily correlated. Appendix B.2 describes in detail how to make draws in this fashion, again appealing to a large-sample approximation based on the central limit theorem. Using these reduced form draws, we can conduct the same inference exercises for the structural parameters described in Section 3. The only difference is that there are now four rather than three elements in θ. As in the case of a continuous treatment we consider fully Bayesian inference under a uniform prior on the intersection of the conditional identified set and any user restrictions $\mathcal{R}$.[28] We elaborate further in our discussion of the empirical examples presented below.

5 Empirical Examples

We now present a number of empirical examples illustrating how the framework described above can be applied in practice. The examples in Sections 5.1 and 5.2 involve a continuous treatment while those in Sections 5.3 and 5.4 involve a binary treatment. In the interest of space, we only sketch the examples in Sections 5.2 and 5.4. Full details and results for this pair of examples appears in Online Appendix C.

5.1 The Colonial Origins of Comparative Development

Acemoglu et al. (2001) study the effect of institutions on GDP per capita using a cross-section of 64 countries.[29] Because institutional quality is endogenous, they use differences in the mortality rates of early western settlers across colonies as an instrumental variable. We consider their benchmark specification

$$\begin{aligned} \log \text{ GDP/capita} &= \text{constant} + \beta\,(\text{Institutions}) + u \\ \text{Institutions} &= \text{constant} + \pi\,(\log \text{ Settler Mortality}) + v \end{aligned}$$

[28] Because the geometry of the problem is slightly more complex in the binary case, however, we employ a slightly different method of making the uniform draws. Although this makes little difference in practice it is more convenient computationally. For details, see https://github.com/binivdoctr.

[29] Because the sample size is so small in this example, we generate posterior draws for Σ using the Jeffreys Prior approach described in Section 3 to avoid non-positive definite draws.

which does not include covariates.[30] This yields an IV estimate of 0.94 with a standard error of 0.16 – nearly twice as large as the corresponding OLS estimate of 0.52 with a standard error of 0.06. The authors attribute this disparity to classical measurement error:

> This estimate is highly significant ... and in fact larger than the OLS estimates ... This suggests that measurement error in the institutions variables that creates attenuation bias is likely to be more important that reverse causality and omitted variables biases. (Acemoglu et al., 2001, p. 1385)

Acemoglu et al. (2001) state two beliefs that are relevant for our partial identification exercise. First, their discussion implies there is likely a positive correlation between "true" institutions and the main equation error term u. This could arise from reverse causality – wealthier societies can afford better institutions – or omitted variables, such as legal origin or British culture, which are likely to be positively correlated with present-day institutional quality. We encode this belief using the prior restriction $0 < \rho_{T^*u} < 0.9$ below, ruling out only unreasonably large values of treatment endogeneity.[31] Second, in a footnote that uses an alternative measure of institutions as an instrument for the first, the authors argue that measurement error could be substantial.[32] Taken at face value, the calculations from this footnote imply a point estimate of $\kappa = 0.6$ which would mean that 40 percent of the variation in measured institutions is noise.[33] Below we consider two alternative ways of encoding this auxiliary information about κ.

Results for the Colonial Origins example appear in rows 1–3 of Table 2. Estimates and bounds for β in these rows indicate the percentage increase in GDP per capita that would result from a one point increase in the quality of institutions, as measured by average protection against expropriation risk.[34] All other values in the table are unitless: they are either probabilities, correlations, or variance ratios. OLS and IV estimates and standard errors, along with estimates of the lower bounds for κ and ρ_{uz}, appear in the first row of

[30] Additional results, available upon request, consider alternative specifications that include covariates. The results are essentially unchanged.

[31] Note that, from Corollary 2.2, the identified set for β is $(-\infty, \infty)$ unless ρ_{T^*u} is restricted. In this example, we impose the researchers' stated belief that $\rho_{T^*u} > 0$ along with an extremely conservative upper bound for ρ_{T^*u} of 0.9. Even these relatively weak restrictions are quite informative about β.

[32] Footnote #19 of Acemoglu et al. (2001) states "We can ascertain, to some degree, whether the difference between OLS and 2SLS estimates could be due to measurement error by making use of an alternative measure of institutions ... This suggests that 'measurement error' in the institutions variables ... is of the right order of magnitude to explain the difference between the OLS and 2SLS estimates."

[33] Suppose T_1 and T_2 are two measures of institutions that are subject to classical measurement error: $T_1 = T^* + w_1$ and $T_2 = T^* + w_2$. Both T_1 and T_2 suffer from precisely the same degree of endogeneity, because they inherit this problem from T^* alone under the assumption of classical measurement error. Thus, the OLS estimator based on T_1 converges to $\kappa(\beta + \sigma_{T^*u}/\sigma^2_{T^*})$ while the IV estimator that uses T_2 to instrument for T_1 converges to $\beta + \sigma_{T^*u}/\sigma^2_{T^*}$. The ratio identifies κ: $0.52/0.87 \approx 0.6$.

[34] See Acemoglu et al. (2001) for a detailed explanation of this measure of institutions.

Panel (I). The first column of Panel (II) gives the fraction of posterior draws for the reduced form parameters that yield an empty identified set, while the second column gives the fraction that are compatible with a valid instrument: $\rho_{uz} = 0$. Panel (III), along with the third and fourth columns of Panel (II), present posterior medians and accompanying 90 percent highest posterior density intervals. The results in Panel (II) are marked "Frequentist-Friendly" because they do not involve placing a prior on the conditional identified set: they only average over reduced form parameter draws under the restriction listed in the corresponding row label.[35] In contrast, those in Panel (III) are "Fully Bayesian"; they place a uniform prior on the conditional identified set (see Section 3.2).

We first consider a prior under which 0.6 is an upper bound for κ and thus a *lower* bound on the extent of measurement error.[36] Under this restriction, approximately 29 percent of the draws for the reduced form parameters Σ yield an *empty* identified set, as shown in the first column of Panel (II). Intuitively, this means that there are covariance matrices Σ that are close to the sample estimate $\widehat{\Sigma}$ but which rule out the region $(\kappa, \rho_{T^*u}) \in (0, 0.6] \times [0, 0.9]$. The problem is not the restriction on ρ_{T^*u} but on κ: the data place no restrictions on the extent of treatment endogeneity although they do provide an upper bound on the extent of measurement error, as shown in Proposition 2.2. Indeed, the proposed *a priori* upper bound of 0.6 for κ is only slightly larger than our point estimate of 0.54 for $\underline{\kappa}$. After accounting for uncertainty in Σ, we find that 29 percent of the posterior density for $\underline{\kappa}$ lies above 0.6. As such, our framework strongly suggests that the belief $\kappa < 0.6$ is incompatible with the data, and we cannot proceed further under this prior.

Instead we consider a second candidate prior that takes 0.6 as a lower bound on κ and thus an *upper* bound on the extent of measurement error. We continue to impose $\rho_{T^*u} \in [0, 0.9]$. Results for this prior specification appear in the third row of Table 2. Unlike the specification considered above, this prior does not yield empty identified sets, as we see from the first column of Panel (II). It does however, strongly suggest that settler mortality is an invalid instrument: 70 percent of the posterior draws for the reduced form parameters Σ exclude $\rho_{uz} = 0$ under the restriction $(\kappa, \rho_{T^*u}) \in (0.6, 1] \times [0, 0.9]$. Figure 2a makes this point in a slightly different way, by depicting the identified set for $(\kappa, \rho_{T^*u}, \rho_{uz})$, evaluated at the maximum likelihood $\widehat{\Sigma}$ of the reduced form parameters, in the region where ρ_{T^*u} is positive. The gray region corresponds to $\underline{\kappa} < \kappa < 0.6$, the largest amount of measurement error consistent with $\widehat{\Sigma}$. We see from the figure that the plane $\rho_{uz} = 0$ only intersects the identified set in the region where measurement error is extremely severe. Moreover, unless

[35] See Section 3 for details.

[36] This interpretation comes from personal communication with one of the authors of Acemoglu et al. (2001). Based on footnote 19 of the paper, he expressed the belief that at least 40 percent of the measured variation in quality of institutions was likely to be noise.

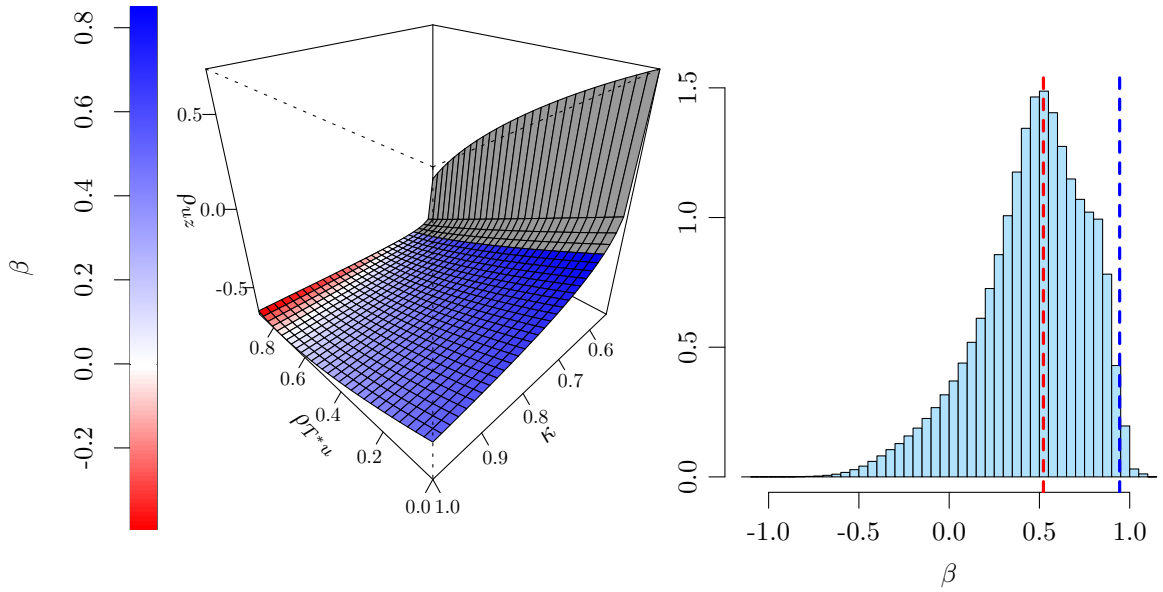
	(I) Summary Statistics				(II) Frequentist-Friendly				(III) Full Bayesian	
	OLS	IV	$\underline{\kappa}$	$\underline{\tilde{\rho}}_{uz}$	$\mathbb{P}(\varnothing)$	$\mathbb{P}$(Valid)	$\underline{\beta}$	$\bar{\beta}$	ρ_{uz}	β
Colonial Origins ($n = 64$)	0.52	0.94	0.54	−0.71						
	(0.06)	(0.16)								
$(\kappa, \rho_{T^*u}) \in (0, 0.6] \times [0, 0.9]$					0.29	–	–	–	–	–
							[–, –]	[–, –]	[–, –]	[–, –]
$(\kappa, \rho_{T^*u}) \in (0.6, 1] \times [0, 0.9]$					0.00	0.30	−0.46	0.85	−0.57	0.49
							[−0.68, −0.23]	[0.71, 1.00]	[−0.82, −0.16]	[0.00, 0.94]

Table 2: Results for Colonial Origins Example. Panel (I) contains OLS and IV estimates and standard errors, and estimates of the bounds for κ and ρ_{uz} from Proposition 2.2 and Corollary 2.1. Panels (II) and (III) present posterior inferences under interval restrictions on (κ, ρ_{T^*u}). The column $\mathbb{P}(\varnothing)$ gives the fraction of reduced form parameter draws that yield an empty identified set, while $\mathbb{P}$(Valid) gives the fraction of reduced form parameter draws compatible with a valid instrument.($\rho_{uz} = 0$). The remaining columns give posterior medians with 90 percent highest posterior density intervals in square brackets. In Panel (II), $\underline{\beta}$ and $\bar{\beta}$ report inferences for the lower and upper boundaries of the identified set for β. In contrast, Panel (III) reports fully Bayesian inference for β and ρ_{uz} under a uniform prior on the intersection between the restrictions and the conditional identified set. See Section 3.2 for details.

	(I) Summary Statistics				(II) Frequentist-Friendly				(III) Fully Bayesian	
	OLS	IV	$\bar{\alpha}_0$	$\bar{\alpha}_1$	$\underline{\delta}_{T^*/z}$	$\bar{\delta}_{T^*/z}$	$\underline{\beta}$	$\bar{\beta}$	$\delta_{T^*/z}$	β
Afgan Girls RCT ($n = 687$)	0.86	1.30	0.10	0.12						
	(0.06)	(0.12)								
$\delta_z = 0$					−0.38	0.11	1.03	1.29	−0.14	1.16
					[−0.54, −0.23]	[0.03, 0.22]	[0.89, 1.19]	[1.11, 1.48]	[−0.36, 0.08]	[0.97, 1.36]
$\delta_{T^*} \in [0, 1]$					−0.10	0.74	−0.21	1.19	0.39	0.42
					[−0.16, −0.04]	[0.67, 0.85]	[−0.37, −0.05]	[1.03, 1.32]	[0.09, 0.69]	[−0.12, 0.98]
$\delta_{T^*} \in [0, 1]$, $p = p^*$					−0.06	0.75	−0.25	1.18	0.41	0.40
					[−0.12, 0.01]	[0.67, 0.84]	[−0.40, −0.09]	[1.02, 1.32]	[0.11, 0.71]	[−0.16, 0.95]

Table 3: Results for the Afghan Girls RCT. Panel (I) contains OLS and IV estimates and standard errors along with estimates of the upper bounds α_0 and α_1 from Proposition 4.2. Panels (II) and (III) present posterior medians and 90 percent highest posterior density intervals under the restrictions on $\alpha_0, \alpha_1, \delta_{T^*}$ and δ_z indicated in the row labels: e.g. a row marked $\delta_z = 0$ assumes that the instrument is valid but does not restrict α_0, α_1 or δ_{T^*u}. The restriction $p = p^*$ in row four imposes $\alpha_1 = \alpha_0(1 - p)/p$. In columns 1–2 of (II) and 1 of (III), the subscript T^*/z indicates that we report inference for δ_{T^*} when δ_z is restricted *a priori* and vice-versa. In a row marked $\delta_z = 0$, these columns report inference for δ_{T^*}; in a row marked $\delta_{T^*} \in [a, b]$ they report inference for δ_z. The inferences in Panels (II) and (III) correspond to those of Table 2.

$\kappa = \underline{\kappa}$, $\rho_{uz} = 0$ implies that ρ_{T^*u} must be close to zero, which would require that institutions are approximately exogenous.



(a) Identified Set at MLE for Σ **(b)** Posterior for Treatment Effect

Figure 2: Results for the Colonial Origins example from Section 5.1. Panel (a) plots the identified set for $(\rho_{uz}, \rho_{T^*u}, \kappa)$ evaluated at the maximum likelihood estimate for Σ in the region corresponding to a positive selection effect: $\rho_{T^*u} \in [0, 0.9]$. The region in which $0.6 > \kappa$ is shaded in gray while the colors on the remainder of the surface correspond to the implied value of the treatment effect β. Panel (b) gives the posterior for β under a uniform prior on the intersection of the restriction $(\kappa, \rho_{T^*u}) \in [0.6, 1] \times [0, 0.9]$ with the conditional identified set (see Section 3.2 for details). The dashed red line gives the OLS estimate and the blue line the IV estimate.

Indeed, under the prior $(\kappa, \rho_{T^*u}) \in (0.6, 1] \times [0, 0.9]$ depicted in shades of red and blue in Figure 2a, the identified set resides exclusively below the plane $\rho_{uz} = 0$, suggesting that log settler mortality is *negatively* correlated with the unobservables in u. The Bayesian posterior inference for ρ_{uz} in column one of Panel (III) shows that, even after accounting for uncertainty in the reduced form parameters Σ, the sign of ρ_{uz} is still almost certainly negative. The primary question of interest, of course, is not the validity of settler mortality as an instrumental variable, but the causal effect of institutions on development. The colored region in Figure 2a shows how κ, ρ_{T^*u} and ρ_{uz} map into corresponding values for β. Blue indicates a positive treatment effect, red a negative treatment effect, and white a zero treatment effect. In both directions, darker colors indicate larger magnitudes. As seen from the figure, we cannot rule out negative values for β. The posterior inference for the boundaries of the identified set for β from columns 3–4 of Panel (II) tell the same story, while accounting for sampling uncertainty in Σ: the highest posterior density interval for $\underline{\beta}$ is comfortably to

the left of zero, while that for $\bar{\beta}$ is comfortably to the right of zero. Notice from Figure 2a, however, that at least when evaluated at $\widehat{\Sigma}$, the identified set only implies negative values for β when ρ_{T^*u} is extremely large and there is very little measurement error (κ is close to one). Because the posterior for $\underline{\beta}$ is determined *entirely* from these extreme points, the resulting inference is very conservative, a concern that we raised above in Section 3.2. This observation motivates the idea of averaging not only over reduced form draws Σ but also over the conditional identified set itself, as we do in Figure 2b and the second column of Panel (III), under a uniform reference prior. These results indicate that the conditional identified sets for $(\kappa, \rho_{T^*u}, \rho_{uz})$ do not contain more than a very small region in which β is negative.[37] Indeed, the posterior median for β is 0.49, very close to the OLS estimate from Acemoglu et al. (2001), while the corresponding 90 percent highest posterior density interval includes only positive values. In spite of the likely negative correlation between settler mortality and u under reasonable prior beliefs that accord with the data, the main result of Acemoglu et al. (2001) continues to hold: it appears that the effect of institutions on income per capita is almost certainly positive.

5.2 Was Weber Wrong?

In the preceding example, our framework revealed an inconsistency among researcher beliefs. Online Appendix C.1 presents an additional continuous–treatment example with very different results. The example is based on Becker and Woessmann (2009) who study the long run effect of the Protestant share in 1870 on literacy rates in Prussia. Their instrument is the distance to Wittenberg, the city where Martin Luther introduced his ideas. The authors explicitly express beliefs about the nature of selection, namely that the adoption of Protestantism is likely *negatively* correlated with unobservables that cause higher literacy rates. This is because Protestantism began as a protest movement in opposition to richer Catholic regions. And because the data for this study comes from a highly reliable Prussian census, they suggest that measurement error in Protestant share is likely to be modest. In this example, we find that the authors beliefs are consistent with distance to Wittenberg being a valid instrument. Moreover, even if the instrument is invalid, under the authors' stated beliefs the effect of Protestant share on literacy remains positive. Additional details appear in Online Appendix C.1.

[37] Because the prior is uniform, "small" refers to the relative area of a region on the identified set: in Figure 2a, for example, the red region is small compared to the blue and white regions.

5.3 Afghan Girls RCT

Burde and Linden (2013) study the effect of village schools on the academic performance of children in rural northwestern Afghanistan, using data from a randomized controlled trial. Both test scores and reported enrollment rates increased significantly in villages that were randomly allocated to receive a school compared to those that were not. The effects were particularly striking for girls, whose enrollment increased by 52 percentage points and test scores by 0.65 standard deviations. Both effects are statistically significant at the 1 percent level and remain essentially unchanged after controlling for a host of demographic covariates.

These results quantify the causal effect of establishing a school in a rural village. But the data from Burde and Linden (2013) are rich enough for us to pose a more specific question that the authors do not directly address in their paper: what is the causal effect of school attendance on the test scores of Afghan girls? With school enrollment as our treatment of interest, the 0.65 standard deviation increase in test scores becomes an intent to treat (ITT) effect, while the 52 percent increase in reported enrollment becomes an IV first stage. In this example we consider the specification

$$\text{Test score} = \text{constant} + \beta\,(\text{Enrollment}) + \mathbf{x}'\gamma + \varepsilon$$

and instrument enrollment using the experimental randomization: Girls in a village where a school was established have $z = 1$ and girls in a village where none was have $z = 0$. The vector $\mathbf{x}$ contains the same covariates used by Burde and Linden (2013).[38] Because this example includes exogenous controls, we define treatment endogeneity and instrument invalidity *net* of these covariates, as detailed in Section 4.4. To simplify the notation we write δ_z and δ_{T^*} rather than $\widetilde{\delta}_z$ and $\widetilde{\delta}_{T^*}$ below but both of these should be understood as being net of $\mathbf{x}$. In contrast, α_0 and α_1 are not defined net of covariates, again as detailed in Section 4.4. As such, they continue to refer to the probabilities $\mathbb{P}(T = 1|T^* = 0)$ and $\mathbb{P}(T = 0|T^* = 1)$.

This dataset has three features that make it an ideal candidate for the methods we have developed above. First, the enrollment variable measures not whether a girl attended the newly-established village school, but whether she attended a school of any kind. This means that our treatment of interest, enrollment, is endogenous: the sample contains 248 girls who did not enroll despite a school being established in their village, and 49 who attended school despite the lack of one in their village. In this example a prior that imposes positive

[38] These are: an indicator for whether the girl is a child of the household head, the girl's age, the number of years the household has lived in the village, a Farsi dummy, a Tajik dummy, a farmers dummy, the age of the household head, years of education of the household head, the number of people in the household, Jeribs of land, number of sheep, distance to the nearest formal school, and a dummy for Chagcharan province.

selection, $\delta_{T^*} > 0$, seems uncontroversial: parents who enroll their daughter in school are likely to have other unobserved characteristics favorable for their academic performance. Second, although the allocation of village schools was randomized, this does not necessarily make it a valid instrument. Indeed, the authors argue that establishing a village school may affect performance through channels other than increased enrollment alone if, for example,

> the village-based schools were of lower quality than the traditional public schools, and some treatment students who would have otherwise attended traditional public schools attended village-based schools instead, or if children who were not enrolled in the treatment group experienced positive spillovers from enrolled siblings or other peers. (Burde and Linden (2013), p. 36.)

Third, school enrollment status is determined from a household survey and, as such, could be subject to substantial mis-reporting. Note that non-differential measurement error in enrollment would not affect the ITT estimate but would bias the estimated causal effect of establishing a school on enrollment. Although Burde and Linden (2013) concede that misreporting is a possibility, they point out that the observed enrollment rates in their sample are comparable to official Afghan government estimates for the region. Even if *aggregate* enrollment is correctly measured, as the authors suggest, individual mis-reporting can still bias the IV estimate. Nevertheless, the prior belief that $p = p^*$ does impose an informative restriction on α_0 and α_1 which we explore below.

Results for the Afghan Girls RCT example appear in the first four rows of Table 3. All values other than $\bar{\alpha}_0$ and $\bar{\alpha}_1$ in columns 3–4 of panel (I) are measured in standard deviations of test scores. IV and OLS estimates, along with lower and upper bounds $\bar{\alpha}_0$ and $\bar{\alpha}_1$ for the mis-classification probabilities appear in panel (I). Posterior medians and 90 percent highest posterior density intervals appear in Panels (II) and (III). The results in Panel (II) are labeled "Frequentist-Friendly" because they do not involve placing a prior on the conditional identified set: they average only over reduced form parameter draws under the restriction listed in the corresponding row label.[39] In contrast, those in Panel (III) are "Fully Bayesian" in that they place a uniform prior on the conditional identified set.

The OLS estimate in this example is quite large, 0.86 standard deviations, but the IV estimate is even larger: 1.3 standard deviations. Notice that our bounds on the mis-classification error rates in this example are very tight: our point estimate of $\bar{\alpha}_0$ is 0.10 while that of $\bar{\alpha}_1$ is 0.12 as shown in Table 3 and depicted in Figure 1. To implement our framework, we consider three prior restrictions. The first assumes that z is a valid instrument, $\delta_z = 0$, but places no *a priori* restrictions on the extent of misclassification, α_0 and α_1, or the sign or extent

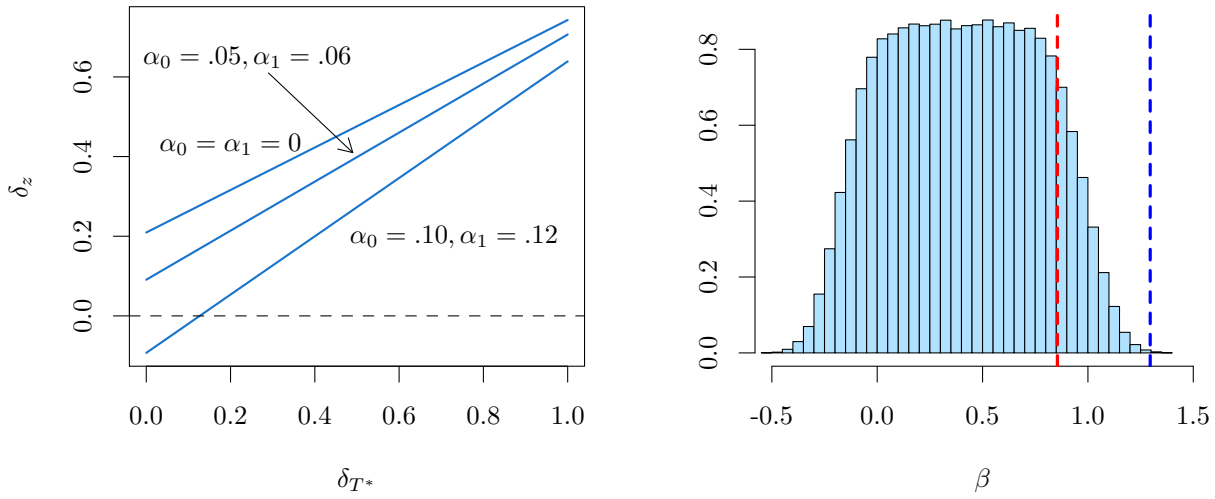
[39] See Sections 3 and 4.5 for details.

of treatment endogeneity, δ_{T^*}. Results for this prior appear in the second row of Table 3. Even under this fairly strong prior restriction, the IV estimate could still show substantial bias because the measurement error is non-differential rather than classical. The first two columns of Panel (II) present posterior inference for the boundaries of the identified set for δ_{T^*} under the assumption that z is a valid instrument while the first column for Panel (III) presents analogous fully Bayesian inference for the *parameter* δ_{T^*}. In this example both tell a similar story; even allowing for measurement error, the assumption that z is a valid instrument requires us to accept the possibility of substantial *negative* selection into treatment, and rules out anything beyond a very modest degree of positive selection. This is precisely the opposite of what most researchers would consider reasonable in this setting.

Figure 3a illustrates this point in a slightly different way, by plotting selected contours of the identified set for $(\delta_{T^*}, \delta_z, \alpha_0, \alpha_1)$ in the region where δ_{T^*} is positive, by evaluating Equation 41 at the maximum likelihood estimates for the reduced form parameters. If δ_{T^*} is assumed to be positive, the only way to sustain a valid instrument is by assuming both that there is essentially zero selection into treatment, and that mis-classification is extremely severe. These results suggest that Burde and Linden (2013) were right to be suspicious of the IV exclusion restriction. Note, moreover, that both the inference for the upper and lower bounds of the identified set for β in columns 3–4 of Panel (II) and the corresponding fully Bayesian inference for the parameter β in the second column of Panel (III) point to a large causal effect of enrollment on test scores. This indicates that, provided one is willing to assume that the instrument is valid, the effect of measurement error on the IV estimate is modest in this example.

The third and fourth rows of Table 3 relax the assumption that $\delta_z = 0$ and instead impose $\delta_{T^*} \in [0, 1]$. This amounts to assuming that the selection effect is positive and no greater than one standard deviation of test scores, after controlling for covariates. The third row imposes no *a priori* restrictions on α_0 and α_1 while the fourth assumes that $p = p^*$ so that $\alpha_1 = \alpha_0(1-p)/p$.[40] The results for these two specifications are practically identical, indicating that $p^* = p$ is not a particularly informative restriction given the tight bounds the data already place on α_0 and α_1. When $\delta_{T^*} \in [0, 1]$, the identified set for δ_z includes a very small range of negative values only, and a much wider range of comparatively large positive values, as we see from the second and third columns of Panel (II). The fully Bayesian inference for δ_z from the first column on Panel (III) is more conclusive, assigning 90 percent probability *a posteriori* to the event that δ_z is between 0.11 and 0.71 standard deviations. The difference between "Frequentist-Friendly" and fully Bayesian inferences in this case indicates that, while all draws for the reduced form parameters are compatible with $\delta_z < 0$,

[40] This follows from $p^* = (p - \alpha_0)/(1 - \alpha_0 - \alpha_1)$.



(a) Contours of Identified Set at MLE

(b) Posterior for Treatment Effect

Figure 3: Results for Afghan Girls RCT example from Section 5.3. Panel (a) illustrates Equation 41 for $\delta_{T^*} \in [0, 1]$ at three pairs of values for (α_0, α_1). Both δ_{T^*} and δ_z are expressed in standard deviations of the test score distribution, and the reduced form parameters are set equal to their maximum likelihood estimates. Panel (b) gives the posterior distribution for β under a uniform prior on the intersection between the restrictions $\delta_{T^*} \in [0, 1], p = p^*$ and the conditional identified set (see Section 4.5). The dashed red line gives the OLS estimate and the blue line the IV estimate.

there is only a very limited combination of values for δ_{T^*}, α_0, and α_1 at which this can occur. This fact is also apparent from Figure 3a: it is only at extremely small values for δ_{T^*} and extremely large values of α_0 and α_1 that δ_z can be negative. One need not take our uniform reference prior from Panel (III) literally: the point is that one would need to place large prior probability on a very small and implausible region of the identified set in order to obtain a substantial posterior probability on the proposition that $\delta_z < 0$. In light of our discussion from Burde and Linden (2013) from above, this suggests that positive peer effects are more plausible than negative village-school quality effects.

Inferences for the causal effect of enrollment are less conclusive. Under the restriction $\delta_{T^*} \in [0, 1]$, the identified set for β comfortably contains zero, although it does extend farther in the positive than the negative direction, as shown in the last two columns of panel (II) in Table 3. The fully Bayesian inferences from the second column of Panel (III) are somewhat more suggestive. Although the 90 percent highest posterior density interval for β does include zero, it is fairly close to the lower limit of the interval. As seen from the

posterior distribution in Figure 3b, the causal effect of enrollment is very likely positive, although substantially smaller than either the OLS or the IV estimate. Again, the difference between the inferences for the identified set and the Bayesian posterior for β under a uniform reference prior indicate that there is only a small region of values for α_0 and α_1 that are compatible with a negative value for β, given that $\delta_{T^*} \in [0, 1]$

5.4 Smoking and BMI

In an additional application, discussed in detail in Online Appendix C.2, we use data from the Lung Health Study (LHS), a well-known randomized clinical trial, to examine the effect of quitting smoking on body mass index (BMI). Following Courtemanche et al. (2016), we use the randomized offer to participate in a smoking cessation program as part of the LHS for self-reported quitting. In this setting, participants who failed to quit smoking may still have claimed that they succeeded, leading to potential one-sided misclassification. Naturally, the decision to quit is endogenous. In particular, selection may be negative if those who are more likely to quit are also more health-conscious. Moreover, the randomized offer of a smoking cessation program may be an invalid instrument if some individuals offered the smoking cessation program cut back on their smoking without quitting entirely. This would suggest a *positive* correlation between the instrument and unobservables. In this example our framework reveals that negative selection into quitting and positive correlation of the instrument with unobservables together imply implausibly large values for the treatment effect. This suggests that selection may in fact be *positive*. Full details appear in Online Appendix C.2.

6 Conclusion and Extensions

Causal inference relies on researcher beliefs. The main message of this paper is that imposing them requires a formal framework, both to guard against contradiction and to ensure that we learn everything that the data have to teach us. While this point is general, we have focused here on a simple but common setting, that of a linear model with a mis-measured, endogenous treatment and a potentially invalid instrument, presenting both results for the case of a continuous treatment subject to classical measurement error and that of a binary treatment subject to non-differential measurement error. By characterizing the relationship between measurement error, treatment endogeneity, and instrument invalidity in terms of intuitive and empirically meaningful parameters, we have developed a Bayesian tool for eliciting, disciplining, and incorporating credible researcher beliefs in the form of sign and

interval restrictions. As we have demonstrated through a wide range of illustrative empirical examples, even relatively weak researcher beliefs can be surprisingly informative in practice.

The methods we describe above could be extended in a number of directions. One possibility is to allow for multiple instrumental variables, expanding the range of examples to which our framework could be applied. There is no serious theoretical obstacle to this extension, although it would likely make prior elicitation more challenging. Another possibility is to consider a wider range of prior specifications on the conditional identified set. One could, for example, explore more informative priors than a uniform distribution, or undertake a formal prior robustness exercise, perhaps along the lines of the ε-contaminated class of priors described by Berger and Berliner (1986) or "posterior lower probability" as in Kitagawa (2012). A limitation of the results presented here is that they assume the treatment effect is homogeneous. While it would likely be difficult to accommodate heterogeneous treatment effects when the treatment is continuous, the binary treatment case shows more promise. Under appropriate modifications it may be possible to extend our framework to the estimation of a local average treatment effect (LATE), possibly by leveraging the testable implications of the LATE model under a binary treatment described by Kitagawa (2015) and Huber and Mellace (2015) among others. We leave this possibility for future research.

References

Acemoglu, D., Johnson, S., Robinson, J. A., 2001. The colonial origins of comparative development: An empirical investigation. The American Economic Review 91 (5), 1369–1401.

Amir-Ahmadi, P., Drautzburg, T., 2016. Identification through heterogeneity, Working Paper.

Apostol, T. M., 1969. Calculus, 2nd Edition. Vol. II. John Wiley and Sons, New York.

Arias, J. E., Rubio-Ramírez, J. F., Waggoner, D. F., 2016. Inference based on SVARs identified with sign and zero restrictions: Theory and applications, Working Paper.

Baumeister, C., Hamilton, J. D., September 2015. Sign restrictions, structural vector autoregressions, and useful prior information. Econometrica 83 (5), 1963–1999.

Becker, S. O., Woessmann, L., 2009. Was Weber wrong? A human capital theory of Protestant economic history. Quarterly Journal of Economics 124 (2), 531–596.

Berger, J., Berliner, L. M., 1986. Robust bayes and empirical bayes analysis with ε-contaminated priors. The Annals of Statistics, 461–486.

Black, D. A., Berger, M. C., Scott, F. A., 2000. Bounding parameter estimates with nonclassical measurement error. Journal of the American Statistical Association 95 (451), 739–748.

Bollinger, C. R., 1996. Bounding mean regressions when a binary regressor is mismeasured. Journal of Econometrics 73, 387–399.

Bollinger, C. R., van Hasselt, M., 2015. Bayesian moment-based inference in a regression models with misclassification error, Working Paper.

Burde, D., Linden, L., 2013. Bringing education to Afghan girls: A randomized controlled trial of village-based schools. American Economic Journal: Applied Economics 5 (3), 27–40.

Chen, X., Christensen, T., O'Hara, K., Tamer, E., 2016. MCMC confidence sets for identified sets, arXiv:1605.00499.

Chou, S.-Y., Grossman, M., Saffer, H., 2004. An economic analysis of adult obesity: results from the behavioral risk factor surveillance system. Journal of health economics 23 (3), 565–587.

Chou, S.-Y., Grossman, M., Saffer, H., 2006. Reply to Jonathan Gruber and Michael Frakes. Journal of Health Economics 25 (2), 389–393.

Conley, T. G., Hansen, C. B., Rossi, P. E., 2012. Plausibly exogenous. The Review of Economics and Statistics 94 (1), 260–272.

Courtemanche, C., Tchernis, R., Ukert, B., 2016. The effect of smoking on obesity: Evidence from a randomized trial, Working Paper.

Dale, S. B., Krueger, A. B., 2002. Estimating the payoff to attending a more selective college: An application of selection on observables and unobservables. Quarterly Journal of Economics, 1491–1527.

DiTraglia, F., García-Jimeno, C., 2016. On mis-measured binary regressors: New results and some comments on the literature, Working Paper.

Frazis, H., Lowenstein, M. A., 2003. Estimating linear regressions with mismeasured, possibly endogenous, binary explanatory variables. Journal of Econometrics 117 (1), 151–178.

Gruber, J., Frakes, M., 2006. Does falling smoking lead to rising obesity? Journal of health economics 25 (2), 183–197.

Gustafson, P., May 2005. On model expansion, model contraction, identifiability and prior information: Two illustrative examples involving mismeasured variables. Statistical Science 20 (2), 111–140.

Gustafson, P., 2015. Bayesian Inference for Partially Identified Models: Exploring the Limits of Limited Data. No. 141 in Monographs on Statistics and Applied Probability. CRC Press, Boca Raton.

Hahn, P. R., Murray, J. S., Manolopoulou, I., 2016. A Bayesian partial identification approach to inferring the prevalence of accounting misconduct. Journal of the American Statistical Association 111 (513).

Hu, Y., 2008. Identification and estimation of nonlinear models with misclassification error using instrumental variables: A general solution. Journal of Econometrics 144 (1), 27–61.

Huber, M., Mellace, G., 2015. Testing instrument validity for late identification based on inequality moment constraints. Review of Economics and Statistics 97 (2), 398–411.

Kahneman, D., Tversky, A., 1974. Judgement under uncertainty: Heuristics and biases. Science 185 (4157), 1124–1131.

Kane, T., Rouse, C. E., Staiger, D., July 1999. Estimating the returns to schooling when schooling is misreported, NBER Working Paper # 7235.

Kitagawa, T., July 2012. Estimation and inference for set-identified parameters using posterior lower probability, Working Paper.
URL http://www.homepages.ucl.ac.uk/~uctptk0/Research/LowerUpper.pdf

Kitagawa, T., 2015. A test for instrument validity. Econometrica 83 (5), 2043–2063.

Kline, B., Tamer, E., July 2016. Bayesian inference in a class of partially identified models. Quantitative Economics 7 (2).

Lewbel, A., March 2007. Estimation of average treatment effects with misclassification. Econometrica 75 (2), 537–551.

Mahajan, A., 2006. Identification and estimation of regression models with misclassification. Econometrica 74 (3), 631–665.

Melfi, G., Schoier, G., 2004. Simulation of random distributions on surfaces. Societa Italiana di Statistica, 173–176.

Moon, H. R., Schorfheide, F., 2009. Estimation with overidentifying inequality moment conditions. Journal of Econometrics 153, 136–154.

Moon, H. R., Schorfheide, F., 2012. Bayesian and frequentist inference in partially identified models. Econometrica 80 (2), 755–782.

Nevo, A., Rosen, A. M., 2012. Identification with imperfect instruments. The Review of Economics and Statistics 94 (3), 659–671.

Ohara, P., Grill, J., Rigdon, M., Connett, J., Lauger, G., Johnston, J., 1993. Design and results of the initial intervention program for the lung health study. Preventive medicine 22 (3), 304–315.

Oster, E., January 2017. Unobservable selection and coefficient stability: Theory and evidence. Forthcoming: Journal of Business and Economic Statistics.

Poirier, D. J., 1998. Revising beliefs in nonidentified models. Econometric Theory 14, 483–509.

Richardson, T. S., Evans, R. J., Robins, J. A., 2011. Transparent parameterizations of models for potential outcomes. In: Bayesian Statistics. Vol. 9. pp. 569–610.

van Hasselt, M., Bollinger, C. R., 2012. Binary misclassification and identification in regression models. Economics Letters 115, 81–84.

A Appendices for Continuous Treatment Case

A.1 Proofs

Lemma A.1. *Under Equation 3 and Assumptions 2.1–2.2,*

(a) $\sigma_v^2 = \sigma_T^2(\kappa - \rho_{Tz}^2)$

(b) $\sigma_w^2 = \left(\frac{1-\kappa}{\kappa}\right)\left(\sigma_v^2 + \pi^2\sigma_z^2\right)$

(c) $\sigma_u^2 = \sigma_y^2\left[\frac{\kappa - \rho_{Ty}^2}{\kappa(1-\rho_{T^*u}^2)}\right]$

(d) $\rho_{uv} = \frac{\rho_{T^*u}\sqrt{\kappa} - \rho_{uz}\rho_{Tz}}{\sqrt{\kappa - \rho_{Tz}^2}}$.

Proof of Lemma A.1(a). First, $\sigma_T^2 = \sigma_w^2 + \pi^2\sigma_z^2 + \sigma_v^2$ and $\rho_{Tz}^2 = \pi^2\sigma_z^2/\sigma_T^2$. The result follows by combining these and rearranging, using the definition of κ and the fact that $\sigma_{T^*}^2 = \sigma_T^2 - \sigma_w^2 = \kappa\sigma_T^2$. □

Proof of Lemma A.1(b). By definition $\kappa = \sigma_{T^*}^2/\sigma_T^2$. Since $\sigma_T^2 = \sigma_{T^*}^2 + \sigma_w^2$, we have $\sigma_w^2 = \sigma_{T^*}^2(1-\kappa)/\kappa$. The result follows since $\sigma_{T^*}^2 = \sigma_v^2 + \pi^2\sigma_z^2$. □

Proof of Lemma A.1(c). The result follows by squaring Equation A.11 in the proof of Proposition 2.1. □

Proof of Lemma A.1(d). From Equation 8, $\rho_{T^*u} = (\sigma_v\rho_{uv} + \pi\sigma_z\rho_{uz})/\sigma_{T^*}$. and by Lemma A.1(a) and the definition of κ, $\sigma_v/\sigma_{T^*} = \sqrt{1 - \rho_{Tz}^2/\kappa}$ and $\pi\sigma_z/\sigma_T = \rho_{Tz}$. Thus,

$$\rho_{T^*u} = \rho_{uv}\sqrt{1 - \rho_{Tz}^2/\kappa} + \rho_{uz}\rho_{Tz}/\sqrt{\kappa}$$

The result follows by solving for ρ_{uv}. □

Corollary A.1. *Under Equation 3 and Assumptions 2.1–2.2,* $\kappa > \max\{\rho_{Ty}^2, \rho_{Tz}^2\}$.

Proof of Corollary A.1. Since $\sigma_v^2 > 0$, $\kappa > \rho_{Tz}^2$ by Lemma A.1(a). Similarly, since $\sigma_u^2 > 0$, $\kappa > \rho_{Ty}^2$ by Lemma A.1(c). □

Proof of Proposition 2.1. Rewriting Equation 6 and using $\sigma_{T^*}^2 = \sigma_T^2 - \sigma_w^2$

$$\beta = (\sigma_{Ty} - \sigma_{T^*u})/\kappa\sigma_T^2 \tag{A.1}$$

and proceeding similarly for Equation 7, we have

$$\beta = (\sigma_{zy} - \sigma_{uz})/\sigma_{zT} \tag{A.2}$$

Combining these,

$$(\sigma_{zy} - \sigma_{uz})/\sigma_{zT} = (\sigma_{Ty} - \sigma_{T^*u})/\kappa\sigma_T^2 \tag{A.3}$$

Now, using Equation 3 and Assumption 2.1, $\sigma_y^2 = \sigma_u^2 + \beta\left(2\sigma_{T^*u} + \beta\kappa\sigma_T^2\right)$. Substituting Equation A.2 for β, Equation A.1 for $\beta\kappa\sigma_T^2$, and rearranging,

$$\left(\sigma_u^2 - \sigma_y^2\right) + \left(\frac{\sigma_{zy} - \sigma_{uz}}{\sigma_{zT}}\right)(\sigma_{T^*u} + \sigma_{Ty}) = 0 \tag{A.4}$$

The next step is to eliminate σ_u from our system of equations. First we substitute $\sigma_{T^*u} = \sigma_u\sqrt{\kappa}\sigma_T\rho_{T^*u}$ and $\sigma_{uz} = \sigma_u\sigma_z\rho_{uz}$ into Equations A.3 and A.4, yielding

$$(\sigma_{zy} - \sigma_u\sigma_z\rho_{uz})/\sigma_{zT} = (\sigma_{Ty} - \sigma_T\sigma_u\rho_{T^*u})/(\kappa\sigma_T^2) \tag{A.5}$$

and

$$\left(\sigma_u^2 - \sigma_y^2\right) + \left(\frac{\sigma_{zy} - \sigma_u\sigma_z\rho_{uz}}{\sigma_{zT}}\right)\left(\sigma_u\sigma_T\sqrt{\kappa}\rho_{T^*u} + \sigma_{Ty}\right) = 0. \tag{A.6}$$

Rearranging Equation A.5 and solving for σ_u, we find that

$$\sigma_u = \frac{\sigma_{zT}\sigma_{Ty} - \kappa\sigma_T^2\sigma_{zy}}{\sigma_T\sqrt{\kappa}\sigma_{Tz}\rho_{T^*u} - \sigma_z\kappa\sigma_T^2\rho_{uz}} \tag{A.7}$$

Since we have stated the problem in terms of *scale-free* structural parameters, namely $(\rho_{uz}, \rho_{T^*u}, \kappa)$, we may assume without loss of generality that $\sigma_T = \sigma_y = \sigma_z = 1$. Even if the raw data do not satisfy this assumption, the relationship between the structural parameters ρ_{uz}, ρ_{T^*u} and κ is unchanged. Imposing this normalization, Equation A.6 becomes

$$\left(\widetilde{\sigma}_u^2 - 1\right) + \left(\frac{\rho_{zy} - \widetilde{\sigma}_u\rho_{uz}}{\rho_{zT}}\right)\left(\widetilde{\sigma}_u\sqrt{\kappa}\rho_{T^*u} + \rho_{Ty}\right) = 0 \tag{A.8}$$

where

$$\widetilde{\sigma}_u = \frac{\rho_{Tz}\rho_{Ty} - \kappa\rho_{zy}}{\sqrt{\kappa}\rho_{Tz}\rho_{T^*u} - \kappa\rho_{uz}}. \tag{A.9}$$

We use the notation $\widetilde{\sigma}_u$ to indicate that normalizing y to have unit variance *does* change the scale of σ_u. Specifically, $\widetilde{\sigma}_u = \sigma_u/\sigma_y$. This does not introduce any complications because we eliminate $\widetilde{\sigma}_u$ from the system by substituting Equation A.9 into Equation A.8. After eliminating $\widetilde{\sigma}_u$, Equation A.8 becomes a quadratic in ρ_{uz} that depends on the structural parameters (ρ_{T^*u}, κ) and the reduced form correlations $(\rho_{Ty}, \rho_{Tz}, \rho_{zy})$. Solving and simplifying the result, we find that

$$(\rho_{uz}^+, \rho_{uz}^-) = \left(\frac{\rho_{T^*u}\rho_{Tz}}{\sqrt{\kappa}}\right) \pm (\rho_{Ty}\rho_{Tz} - \kappa\rho_{zy})\sqrt{\frac{1 - \rho_{T^*u}^2}{\kappa\left(\kappa - \rho_{Ty}^2\right)}} \tag{A.10}$$

Although the preceding expression yields two solutions, one of these is extraneous as it implies a *negative* value for $\widetilde{\sigma}_u$ and hence σ_u. To see why, substitute each solution into the reciprocal of Equation A.9 to yield

$$\widetilde{\sigma}_u^{-1} = \frac{\sqrt{\kappa}\rho_{Tz}\rho_{T^*u}}{\rho_{Ty}\rho_{Tz} - \kappa\rho_{zy}} - \left[\left(\frac{\sqrt{\kappa}\rho_{Tz}\rho_{T^*u}}{\rho_{Ty}\rho_{Tz} - \kappa\rho_{zy}}\right) \pm \sqrt{\frac{\kappa(1 - \rho_{T^*u}^2)}{\kappa - \rho_{Ty}^2}}\right] = \mp\sqrt{\frac{\kappa(1 - \rho_{T^*u}^2)}{\kappa - \rho_{Ty}^2}}$$

and hence

$$\sigma_u = \mp\sigma_y\sqrt{\frac{\kappa - \rho_{Ty}^2}{\kappa(1 - \rho_{T^*u}^2)}}. \tag{A.11}$$

This implies that ρ_{uz}^+ is always extraneous. □

Proof of Lemma 2.1. Assumption 2.1 requires that all the elements of Ω are finite, which implies that $\sigma_w^2, \sigma_v^2, \sigma_z^2, \sigma_u^2 < \infty$. The reverse implication follows since $Cov(u, v) \leq \sqrt{Var(u)Var(v)}$ and $Cov(u, z) \leq \sqrt{Var(u)Var(z)}$ by the Cauchy-Schwarz inequality. This establishes that condition (a) of Lemma 2.1 is equivalent to all of the entries of Ω being finite.

Now, $\widetilde{\Omega}$ is positive definite if and only if each of its three leading principle minors are positive:

$$\sigma_u^2 > 0 \tag{A.12}$$

$$\sigma_u^2\sigma_v^2 - \sigma_{uv}^2 > 0 \tag{A.13}$$

$$\sigma_z^2(\sigma_u^2\sigma_v^2 - \sigma_{uv}^2) - \sigma_v^2\sigma_{uz}^2 > 0. \tag{A.14}$$

Thus it is sufficient to show that Equations A.12–A.14 are equivalent to $\sigma_u^2, \sigma_v^2, \sigma_z^2 > 0$ and $\rho_{uv}^2 + \rho_{uz}^2 < 1$. The equivalence is obvious for A.12. By A.12 we can rearrange A.13 to yield $\sigma_v^2 > \sigma_{uv}^2/\sigma_u^2 \geq 0$ implying that σ_v^2 is strictly positive. Dividing through by σ_v^2, this implies that $|\rho_{uv}| < 1$. Now, since both σ_u^2 and σ_v^2 are strictly positive, we can divide both sides of A.14 through by $\sigma_v^2\sigma_u^2$ to obtain $\sigma_z^2(1 - \rho_{uv}^2) > \sigma_{uz}^2/\sigma_u^2 \geq 0$. Since

$\rho_{uv}^2 < 1$, this implies $\sigma_z^2 > 0$. Thus, dividing Equation A.14 through by $\sigma_v^2\sigma_u^2\sigma_z^2$ and rearranging we find that $\rho_{uv}^2 + \rho_{uz}^2 < 1$, establishing the "if" direction of the equivalence. For the "only if" direction, $\rho_{uv}^2 + \rho_{uz}^2 < 1$ implies $\rho_{uv}^2 < 1$. Mutiplying both sides by $\sigma_u^2\sigma_v^2$ gives $\sigma_u^2\sigma_v^2\rho_{uv}^2 < \sigma_u^2\sigma_v^2$ since $\sigma_u^2, \sigma_v^2 > 0$. Substituting $\rho_{uv}^2 = \sigma_{uv}^2/(\sigma_u^2\sigma_v^2)$ and rearranging implies A.13. Equation A.14 follows similarly, by multiplying both sides of $\rho_{uv}^2 + \rho_{uz}^2 < 1$ by $\sigma_u^2\sigma_v^2\sigma_z^2$ and rearranging. □

Proof of Proposition 2.2. We first derive the bounds; at the end of the proof we show that they are sharp. To show that $|\rho_{T^*u}| < 1$ we combine the assumption that $\sigma_u^2 < \infty$ with the expression from Lemma A.1(c). The fact that $\kappa \leq 1$ follows from $\sigma_w^2 \geq 0$ and the definition of κ.

The lower bound for κ is more involved. We first restate the inequality from Lemma 2.1(c) so that it no longer involves ρ_{uv} by substituting Lemma A.1(d), yielding

$$\left(\frac{\rho_{T^*u}\sqrt{\kappa} - \rho_{uz}\rho_{Tz}}{\sqrt{\kappa - \rho_{Tz}^2}}\right)^2 + \rho_{uz}^2 < 1 \tag{A.15}$$

Using the fact that $\kappa > \rho_{Tz}^2$ from Corollary A.1, putting the terms of Equation A.15 over a common denominator and rearranging,

$$\rho_{T^*u}^2 + \rho_{zu}^2 - \frac{2\rho_{T^*u}\rho_{zu}\rho_{Tz}}{\sqrt{\kappa}} < \frac{\kappa - \rho_{Tz}^2}{\kappa}.$$

Completing the square, we find

$$\left(\rho_{zu} - \frac{\rho_{T^*u}\rho_{Tz}}{\sqrt{\kappa}}\right)^2 < \left(1 - \rho_{T^*u}^2\right)\left(\frac{\kappa - \rho_{Tz}^2}{\kappa}\right)$$

Now, using Equation 13 to substitute for $(\rho_{zu} - \rho_{T^*u}\rho_{Tz}/\sqrt{\kappa})$, we find that

$$(\rho_{Ty}\rho_{Tz} - \kappa\rho_{zy})^2\left[\frac{1 - \rho_{T^*u}^2}{\kappa(\kappa - \rho_{Ty}^2)}\right] < \left(1 - \rho_{T^*u}^2\right)\left(\frac{\kappa - \rho_{Tz}^2}{\kappa}\right)$$

Cancelling a factor of $(1 - \rho_{T^*u}^2)/\kappa$ from each side and rearranging yields an expression that does not involve ρ_{T^*u}, namely

$$(\rho_{Ty}\rho_{Tz} - \kappa\rho_{zy})^2 - (\kappa - \rho_{Ty}^2)(\kappa - \rho_{Tz}^2) < 0 \tag{A.16}$$

again using the fact that $\kappa > \rho_{Ty}^2$. Expanding and simplifying,

$$(\rho_{zy}^2 - 1)\kappa^2 + (\rho_{Ty}^2 + \rho_{Tz}^2 - 2\rho_{Ty}\rho_{Tz}\rho_{zy})\kappa < 0 \tag{A.17}$$

Since $\rho_{zy}^2 < 1$ by positive definiteness, the preceding inequality defines an interval of values that κ *cannot* take on, an interval bounded by the roots of a quadratic function that opens downwards. Factoring the quadratic to solve for these roots gives

$$\kappa\left[(\rho_{zy} - 1)\kappa + \left(\rho_{Ty}^2 + \rho_{Tz}^2 - 2\rho_{Ty}\rho_{Tz}\rho_{zy}\right)\right] = 0$$

Thus one root is zero and the other is $\underline{\kappa}$, as defined in the statement of the proposition. We have shown that κ cannot take on a value between the two roots. We do not yet know, however, which is larger. The result of the proposition follows after we establish two more claims. First $\underline{\kappa} < 1$, and second $\underline{\kappa} > \max\{\rho_{Ty}^2, \rho_{Tz}^2\}$. For the first claim, note that the correlation matrix of (y, T, z) must be positive-definite, implying that

$$1 - \rho_{Ty}^2 - \rho_{Tz}^2 - \rho_{zy}^2 + 2\rho_{Ty}\rho_{Tz}\rho_{zy} > 0$$

Rearranging this inequality using the fact that $\rho_{zy}^2 < 1$ establishes that $\underline{\kappa} < 1$. For the second claim notice that by evaluating the quadratic on the left hand side of A.16 at $\max\{\rho_{Ty}^2, \rho_{zT}^2\}$, the second term vanishes leaving us with a squared term. Since this cannot be less than zero, the inequality is violated at

$\kappa = \max\{\rho_{Ty}^2, \rho_{zT}^2\} > 0$. This combined with the fact that the parabola opens downwards establishes that $\underline{\kappa}$ is greater than both zero and $\max\{\rho_{Ty}^2, \rho_{zT}^2\}$. This establishes the lower bound for κ.

Now that we have derived the bounds, we explain why they are sharp. We need to show that for any values of ρ_{T^*u} and κ within our bounds, there exist values for all the parameters of Ω, defined in Assumption 2.1, that satisfy the first three equivalent conditions from Lemma 2.1 and generate the observed covariance matrix Σ. First, from A.1(a), $\rho_{Tz}^2 < \kappa \leq 1$ implies that σ_v^2 is strictly positive and finite. The bound $\rho_{Tz}^2 < \kappa$ is implied by $\kappa > \underline{\kappa}$, as explained in our derivation above. Second, by A.1(b), the fact that σ_v^2 is positive and finite, along with $0 < \underline{\kappa} < \kappa < 1$, implies that σ_w^2 is non-negative and finite. Third, by A.1(c), $\rho_{Ty}^2 < \underline{\kappa} < \kappa < 1$ along with $|\rho_{T^*u}| \neq 1$ implies that σ_u^2 is strictly positive and finite. It remains only to verify that $\rho_{uv}^2 + \rho_{uz}^2 < 1$, but this is immediate from our derivation of $\underline{\kappa}$ from above, since all of our steps were reversible. Now, to construct Ω, set σ_v^2, σ_w^2, etc. according to Lemma A.1 with ρ_{uz} as in Proposition 2.1. So long as κ and ρ_{T^*u} satisfy the stated bounds, this choice of Ω satistfies Assumption 2.1 and generates the observed covariance matrix Σ of (y, T, z) via Equation 3. □

Proof of Corollary 2.1. Let $f(\rho_{T^*u}, \kappa)$ denote the right hand side of Equation 13. We begin by finding the optima of f as a function of ρ_{T^*u}, holding κ fixed. The first derivative of f with respect to ρ_{T^*u} is

$$\frac{\partial f}{\partial \rho_{T^*u}} = \frac{\rho_{Tz}}{\sqrt{\kappa}} + \frac{\rho_{T^*u}(\rho_{Ty}\rho_{Tz} - \kappa\rho_{zy})}{\sqrt{\kappa(\kappa - \rho_{Ty}^2)(1 - \rho_{T^*u}^2)}} \tag{A.18}$$

so the first-order condition for ρ_{T^*u} is

$$\rho_{T^*u} = -\frac{a}{\sqrt{1+a^2}}, \qquad a = \frac{\rho_{Tz}\sqrt{\kappa - \rho_{Ty}^2}}{(\rho_{Ty}\rho_{Tz} - \kappa\rho_{zy})} \tag{A.19}$$

The second derivative of f with respect to ρ_{T^*u} is

$$\frac{\partial^2 f}{\partial \rho_{T^*u}^2} = \frac{(\rho_{Ty}\rho_{Tz} - \kappa\rho_{zy})}{\sqrt{\kappa(\kappa - \rho_{Ty}^2)}} \left[\frac{1}{(1 - \rho_{T^*u}^2)^{3/2}}\right]. \tag{A.20}$$

Note that the sign of Equation A.20 depends only on the sign of $(\rho_{Ty}\rho_{Tz} - \kappa\rho_{zy})$. Define $\widehat{\kappa} = \rho_{Ty}\rho_{Tz}/\rho_{zy}$ and suppose first that $\kappa \neq \widehat{\kappa}$. If $\kappa < \widehat{\kappa}$ then $f(\rho_{T^*u}, \kappa)$ is a strictly convex function of ρ_{T^*u} and thus, holding κ fixed, has a unique global minimum at the solution to A.19. In contrast, the global maximum is a corner solution: it occurs either at $\rho_{T^*u} = -1$ or 1. Similarly, if $\kappa > \widehat{\kappa}$ then $f(\rho_{T^*u}, \kappa)$ is a strictly concave function of ρ_{T^*u} and thus, holding κ fixed, has a unique global maximum at the solution to A.19. In this case the global minimum is a corner solution: it occurs either at $\rho_{T^*u} = -1$ or 1. In either case, the interior solution is strictly less than one in absolute value, as we see from Equation A.19 using the fact that $\kappa > \rho_{Ty}^2$ by Corollary A.1. If $\kappa = \widehat{\kappa}$, then f reduces to $\rho_{T^*u}\rho_{Tz}/\sqrt{\widehat{\kappa}}$. In this case both extrema are corner solutions: they occur at $\rho_{T^*u} = -1$ and 1.

We have now fully characterized the values of ρ_{T^*u} that optimize f for any *fixed* value of κ. It remains to find the optimal values of κ within the feasible set $(\underline{\kappa}, 1]$. Using Equation A.19 we can concentrate ρ_{T^*u} out of f to yield a new function g that imposes the first-order condition for ρ_{T^*u}, namely

$$g(\kappa) = -\text{sign}\{\rho_{Ty}\rho_{Tz} - \kappa\rho_{zy}\}\sqrt{\frac{(\rho_{Ty}\rho_{Tz} - \kappa\rho_{zy})^2 + \rho_{Tz}^2(\kappa - \rho_{Ty}^2)}{\kappa(\kappa - \rho_{Ty}^2)}} \tag{A.21}$$

and calculate its derivative as follows

$$g'(\kappa) = -\frac{(\underline{\kappa} - \rho_{Ty}^2)(1 - \rho_{zy}^2)}{2g(\kappa)(\kappa - \rho_{Ty}^2)^2}. \tag{A.22}$$

Note that g' is positive whenever $\kappa < \widehat{\kappa}$ and negative whenever $\kappa > \widehat{\kappa}$. Moreover, recall from the proof of Proposition 2.2 (see Equation A.17) that $\underline{\kappa}$ satisfies $(\rho_{Ty}\rho_{Tz} - \underline{\kappa}\rho_{zy})^2 = (\underline{\kappa} - \rho_{Ty}^2)(\underline{\kappa} - \rho_{Tz}^2)$. Substituting

this into Equation A.21 we find that

$$g(\underline{\kappa}) = -\text{sign}\{\rho_{Ty}\rho_{Tz} - \underline{\kappa}\rho_{zy}\} \tag{A.23}$$

Equation A.22 applies at any interior optimum for ρ_{T^*u}. At a corner solution, $\rho_{T^*u} = -1$ or 1 and the objective function reduces to

$$h(\kappa) = \begin{cases} -\rho_{Tz}/\sqrt{\kappa}, & \text{if } \rho_{T^*u} = -1 \\ \rho_{Tz}/\sqrt{\kappa}, & \text{if } \rho_{T^*u} = 1 \end{cases} \tag{A.24}$$

Hence the extrema of h always occur at $\kappa = \underline{\kappa}$. We now have all the ingredients needed to find the bounds for ρ_{zu}. The rest of the proof proceeds in cases, depending on the values of $(\rho_{Ty}, \rho_{Tz}, \rho_{zy})$.

Case I: $\widehat{\kappa} \notin (\underline{\kappa}, 1]$. In this case, the second derivative of f with respect to ρ_{T^*u} has the same sign for all $\kappa \in (\underline{\kappa}, 1]$. There are two sub-cases.

(a) Suppose first that the second derivative of f is positive. Since this occurs when $\kappa < \widehat{\kappa}$, it is equivalent to $\rho_{Ty}\rho_{Tz} - \kappa\rho_{zy} > 0$ for all $\kappa \in (\underline{\kappa}, 1]$ given that $\widehat{\kappa} \notin (\underline{\kappa}, 1]$. In this case the function g from Equation A.21 describes the global *minimum* for ρ_{uz} and is a strictly increasing function of κ. Thus, the minimum value for ρ_{uz} equals $g(\underline{\kappa}) = -1$. The global maximum thus occurs at either $\rho_{T^*u} = -1$ or 1 and requires us to make $h(\kappa)$ *positive*. Thus, the upper bound for ρ_{uz} equals $|\rho_{Tz}|/\sqrt{\underline{\kappa}}$.

(b) Suppose next that the second derivative of f is negative. Since this occurs when $\kappa > \widehat{\kappa}$, it is equivalent to $\rho_{Ty}\rho_{Tz} - \kappa\rho_{zy} < 0$ for all $\kappa \in (\underline{\kappa}, 1]$ given that $\widehat{\kappa} \notin (\underline{\kappa}, 1]$. In this case the function g from Equation A.21 describes the global *maximum* for ρ_{uz} and is a strictly decreasing function of κ. Thus, the maximum value for $\rho_{uz} = g(\underline{\kappa}) = 1$. The global minimum thus occurs at either $\rho_{T^*u} = -1$ or 1 and requires us to make $h(\kappa)$ *negative*. Thus, the lower bound for ρ_{uz} equals $-|\rho_{Tz}|/\sqrt{\underline{\kappa}}$.

Case II: $\widehat{\kappa} \in (\underline{\kappa}, 1]$. This case is more complicated because the sign of the second derivative of f with respect to ρ_{T^*u} now depends on κ. Again there are two sub-cases.

(a) Suppose first that $\rho_{Ty}\rho_{Tz} - \kappa\rho_{zy} > 0$ for $\kappa < \widehat{\kappa}$ and $\rho_{Ty}\rho_{Tz} - \kappa\rho_{zy} < 0$ for $\kappa > \widehat{\kappa}$. In this case, f is strictly convex in ρ_{T^*u} for $\kappa < \widehat{\kappa}$. Accordingly, for a fixed $\kappa < \widehat{\kappa}$, g gives the *minimum* value of ρ_{uz} over all ρ_{T^*u}. Moreover, g is strictly *increasing* for $\kappa < \widehat{\kappa}$ thus giving us a candidate minimum at $\underline{\kappa}$. Since $\rho_{Ty}\rho_{Tz} - \kappa\rho_{zu} > 0$ for $\kappa < \widehat{\kappa}$, we see from Equation A.23 that $g(\underline{\kappa}) = -1$ hence $\rho_{uz} = -1$ is indeed the minimum. Now, when $\kappa > \widehat{\kappa}$, f is a strictly concave function of ρ_{T^*u} so for any fixed $\kappa > \widehat{\kappa}$, g gives the *maximum* value of ρ_{uz} over all ρ_{T^*u}. Moreover, g is strictly *decreasing* for $\kappa > \widehat{\kappa}$. Thus there cannot be an interior maximum in this region. As mentioned above, when $\kappa = \widehat{\kappa}$ the extrema occur either at $\rho_{T^*u} = -1$ or 1, so applying h we find that the maximum value of ρ_{uz} at $\kappa = \widehat{\kappa}$ is $|\rho_{Tz}|/\widehat{\kappa}$. Notice that this is equal to the limit of $g(\kappa)$ as κ approaches $\widehat{\kappa}$ from the right. We have thus identified a candidate maximum. It is not the global maximum, however, since $h(\underline{\kappa}) > h(\widehat{\kappa})$. Thus, ρ_{uz} is maximized at $|\rho_{Tz}|/\sqrt{\underline{\kappa}}$.

(b) Suppose next that $\rho_{Ty}\rho_{Tz} - \kappa\rho_{zy} < 0$ for $\kappa < \widehat{\kappa}$ and $\rho_{Ty}\rho_{Tz} - \kappa\rho_{zy} > 0$ for $\kappa > \widehat{\kappa}$. In this case, f is strictly concave in ρ_{T^*u} for $\kappa < \widehat{\kappa}$. Accordingly, for a fixed $\kappa < \widehat{\kappa}$, g gives the *maximum* value of ρ_{uz} over all ρ_{T^*u}. Moreover, g is strictly *decreasing* for $\kappa < \widehat{\kappa}$ thus giving us a candidate maximum at $\underline{\kappa}$. Since $\rho_{Ty}\rho_{Tz} - \kappa\rho_{zu} < 0$ for $\kappa < \widehat{\kappa}$, we see from Equation A.23 that $g(\underline{\kappa}) = 1$ hence $\rho_{uz} = 1$ is indeed the maximum. Now, when $\kappa > \widehat{\kappa}$, f is a strictly convex function of ρ_{T^*u} so for any fixed $\kappa > \widehat{\kappa}$, g gives the *minimum* value of ρ_{uz} over all ρ_{T^*u}. Moreover, g is strictly *increasing* for $\kappa > \widehat{\kappa}$. Thus there cannot be an interior minimum in this region. As mentioned above, when $\kappa = \widehat{\kappa}$ the extrema occur either at $\rho_{T^*u} = -1$ or 1, so applying h we find that the minimum value of ρ_{uz} at $\kappa = \widehat{\kappa}$ is $-|\rho_{Tz}|/\widehat{\kappa}$. Notice that this is equal to the limit of $g(\kappa)$ as κ approaches $\widehat{\kappa}$ from the right. We have thus identified a candidate minimum. It is not the global minimum, however, since $h(\underline{\kappa}) < h(\widehat{\kappa})$. Thus, ρ_{uz} is minimized at $-|\rho_{Tz}|/\sqrt{\underline{\kappa}}$.

From our proof of Proposition 2.2, we know that $\rho_{Tz}^2 < \underline{\kappa}$ which implies $|\rho_{Tz}|/\sqrt{\underline{\kappa}} < 1$, thus the one-sided bounds are non-trivial. The result follows since $\rho_{Ty}\rho_{Tz} - \underline{\kappa}\rho_{zy} > 0$ implies that we are either in case I(a) or II(a), while $\rho_{Ty}\rho_{Tz} - \underline{\kappa}\rho_{zy} > 0$ implies that we are either in case I(b) or II(b).

□

Proof of Corollary 2.2. First, by Equation 7 $\beta = \beta_{IV} - \sigma_{uz}/\sigma_{Tz} = (\rho_{zy}\sigma_y - \rho_{uz}\sigma_u)/(\rho_{Tz}\sigma_T)$. Now, combining Proposition 2.1 and Equation 15,

$$\rho_{uz}\sigma_u = \frac{\sigma_y}{\kappa}\left[\rho_{Tz}\sqrt{\kappa - \rho_{Ty}^2}\left(\frac{\rho_{T^*u}}{\sqrt{1-\rho_{T^*u}^2}}\right) - (\rho_{Ty}\rho_{Tz} - \kappa\rho_{zy})\right]$$

The result follows since ρ_{T^*u} can take on any value in $(-1, 1)$ and $\rho_{T^*u}/(1-\rho_{T^*u}^2)^{1/2}$ tends to $+\infty$ when ρ_{T^*u} approaches 1 and $-\infty$ when it approaches -1. □

A.2 Inference

A.2.1 Draws for the Reduced Form Parameters

This appendix provides details of our first proposal for drawing the reduced form parameters Σ from Section 3.1 using on a large-sample approximation. Let

$$\begin{aligned} \varepsilon_T &= (y - \mathbb{E}[y]) - \beta_T(T - \mathbb{E}[T]) \\ \varepsilon_z &= (y - \mathbb{E}[y]) - \beta_z(z - \mathbb{E}[z]) \end{aligned}$$

where $\beta_T = \sigma_{Ty}/\sigma_T^2$, and $\beta_z = \sigma_{zy}/\sigma_z^2$. While neither β_T nor β_z equals the true treatment effect β, the parameters of both of these regressions are identified. Under the standard regularity conditions for linear regression, we have

$$\begin{bmatrix} \sqrt{n}\left(\widehat{\beta}_T - \beta_T\right) \\ \sqrt{n}\left(\widehat{\beta}_z - \beta_z\right) \end{bmatrix} \to_d B \begin{bmatrix} M_T \\ M_z \end{bmatrix} \tag{A.25}$$

where $\widehat{\beta}_T = \widehat{\sigma}_{Ty}/\widehat{\sigma}_T^2$ and $\widehat{\beta}_z = \widehat{\sigma}_{zy}/\widehat{\sigma}_z^2$ are the OLS estimators of β_T and β_z, $(M_T, M_z)' \sim N(0, V)$, and

$$B = \begin{bmatrix} 1/\sigma_T^2 & 0 \\ 0 & 1/\sigma_z^2 \end{bmatrix},\ V = \mathbb{E}\begin{bmatrix} T^2\varepsilon_T^2 & zT\varepsilon_z\varepsilon_T \\ zT\varepsilon_z\varepsilon_T & z^2\varepsilon_z^2 \end{bmatrix}. \tag{A.26}$$

Note that V depends not on the *structural* error u but on the *reduced form* errors $\varepsilon_T, \varepsilon_z$. By construction ε_T is uncorrelated with T and ε_z is uncorrelated with z but the reduced form errors are *necessarily* correlated with each other. Now, using Equations A.25 and A.26 we see that

$$\begin{bmatrix} \sqrt{n}\left(\widehat{\sigma}_{Ty} - \sigma_{Ty}\right) \\ \sqrt{n}\left(\widehat{\sigma}_{zy} - \sigma_{zy}\right) \end{bmatrix} \to_d B^{-1}B \begin{bmatrix} M_T \\ M_z \end{bmatrix} = \begin{bmatrix} M_T \\ M_z \end{bmatrix} \tag{A.27}$$

and thus, in large samples

$$\begin{bmatrix} \widehat{\sigma}_{Ty} \\ \widehat{\sigma}_{zy} \end{bmatrix} \approx N\left(\begin{bmatrix} \sigma_{Ty} \\ \sigma_{zy} \end{bmatrix}, \widehat{V}/n\right) \tag{A.28}$$

where $\widehat{V}$ is the textbook robust variance matrix estimator, namely

$$\widehat{V} = \frac{1}{n}\sum_{i=1}^{n}\begin{bmatrix} T_i^2\hat{\varepsilon}_{Ti}^2 & z_iT_i\hat{\varepsilon}_{zi}\hat{\varepsilon}_{Ti} \\ z_iT_i\hat{\varepsilon}_{zi}\hat{\varepsilon}_{Ti} & z_i^2\hat{\varepsilon}_{zi}^2 \end{bmatrix}$$

where $\widehat{\varepsilon}_{Ti}$ denotes the i^{th} residual from the β_T regression and $\widehat{\varepsilon}_{zi}$ the i^{th} residual from the β_z regression. Since we are working solely with identified parameters, the usual large-sample equivalence between a Bayesian posterior and frequentist sampling distribution holds. Accordingly, we propose to generate draws for σ_{Ty} and σ_{zy} according to

$$\begin{bmatrix} \sigma_{Ty}^{(j)} \\ \sigma_{zy}^{(j)} \end{bmatrix} \sim N\left(\begin{bmatrix} \widehat{\sigma}_{Ty} \\ \widehat{\sigma}_{zy} \end{bmatrix}, \widehat{V}/n\right) \tag{A.29}$$

Combining these draws with the *fixed* values $\widehat{\sigma}_T^2, \widehat{\sigma}_z^2$ and $\widehat{\sigma}_{zT}$, since we are conditioning on z and T, yields posterior draws for Σ based on a large-sample normal approximation, namely

$$\Sigma^{(j)} = \begin{bmatrix} \widehat{\sigma}_T^2 & \sigma_{Ty}^{(j)} & \widehat{\sigma}_{Tz} \\ \sigma_{Ty}^{(j)} & \widehat{\sigma}_y^2 & \sigma_{zy}^{(j)} \\ \widehat{\sigma}_{Tz} & \sigma_{zy}^{(j)} & \widehat{\sigma}_z^2 \end{bmatrix} \tag{A.30}$$

A.2.2 Uniform Draws on the Conditional Identified Set

There are at least two different methods of placing a conditionally uniform prior on Θ. The first, and simplest, draws ρ_{T^*u} and κ uniformly and independently on $(-1,1) \times (\underline{\kappa}, 1] \cap \mathcal{R}$ and then solves for ρ_{uz} at each draw using Equation 13. The second method, which we use in this paper, draws uniformly on intersection of the *manifold* $(\rho_{uz}, \rho_{T^*u}, \kappa)$ – defined by 13 and the identified set for (ρ_{T^*u}, κ) described in Proposition 2.2 – with any user restrictions $\mathcal{R}$. This method begins by making the same draws as described in the first method, but proceeds to re-weight them based on the local surface area of the identified set at that point (Melfi and Schoier, 2004). By local surface area we refer to the quantity

$$M\left(\rho_{T^*u}, \kappa\right) = \sqrt{1 + \left(\frac{\partial \rho_{uz}}{\partial \rho_{T^*u}}\right)^2 + \left(\frac{\partial \rho_{uz}}{\partial \kappa}\right)^2} \tag{A.31}$$

which Apostol (1969) calls the "local magnification factor" of a parametric surface. The derivatives required to evaluate the function M are

$$\frac{\partial \rho_{uz}}{\partial \rho_{T^*u}} = \frac{\rho_{Tz}}{\sqrt{\kappa}} + \frac{\rho_{T^*u}\left(\rho_{Ty}\rho_{Tz} - \kappa\rho_{zy}\right)}{\sqrt{\kappa\left(\kappa - \rho_{Ty}^2\right)\left(1 - \rho_{T^*u}^2\right)}}$$

$$\frac{\partial \rho_{uz}}{\partial \kappa} = -\frac{\rho_{T^*u}\rho_{Tz}}{2\kappa^{3/2}} + \sqrt{\frac{1 - \rho_{T^*u}^2}{\kappa\left(\kappa - \rho_{Ty}^2\right)}}\left\{\rho_{zy} + \frac{1}{2}\left(\rho_{Ty}\rho_{Tz} - \kappa\rho_{zy}\right)\left[\frac{1}{\kappa} + \frac{1}{\kappa - \rho_{Ty}^2}\right]\right\}.$$

To accomplish the re-weighting, we first evaluate $M^{(\ell)} = M(\rho_{T^*u}^{(\ell)}, \kappa^{(\ell)})$ at each draw ℓ that was accepted in the first step. We then calculate $M_{max} = \max_{\ell=1,\ldots,L} M^{(\ell)}$ and *resample* the draws $\left(\rho_{uz}^{(\ell)}, \rho_{T^*u}^{(\ell)}, \kappa^{(\ell)}\right)$ with probability $p^{(\ell)} = M^{(\ell)}/M_{max}$.

B Appendices for Binary Treatment Case

B.1 Proofs

Lemma B.1. *For $k = 0, 1$,*

$$p^* = \frac{p - \alpha_0}{1 - \alpha_0 - \alpha_1}, \quad p_k^* = \frac{p_k - \alpha_0}{1 - \alpha_0 - \alpha_1}, \quad 1 - p^* = \frac{1 - p - \alpha_1}{1 - \alpha_0 - \alpha_1} \quad 1 - p_k^* = \frac{1 - p_k - \alpha_1}{1 - \alpha_0 - \alpha_1}$$

Proof. By the Law of Total Probability,

$$p = (1 - \alpha_1)p^* + \alpha_0(1 - p^*) = (1 - \alpha_0 - \alpha_1)p^* + \alpha_0$$

The result for p^* and $1 - p^*$ follows after rearranging. The argument for p_k^* and $1 - p_k^*$ is identical. □

Lemma B.2.

$$\begin{aligned}
\mathbb{P}(T^* = 0|T = 0, z_k) &= (1 - \alpha_0)(1 - p_k^*)/(1 - p_k) \\
\mathbb{P}(T^* = 1|T = 0, z_k) &= \alpha_1 p_k^*/(1 - p_k) \\
\mathbb{P}(T^* = 0|T = 1, z_k) &= \alpha_0(1 - p_k^*)/p_k \\
\mathbb{P}(T^* = 1|T = 1, z_k) &= (1 - \alpha_1)p_k^*/p_k.
\end{aligned}$$

Proof. The result follows from Bayes' rule and Assumption 4.1. □

Lemma B.3. $Cov(T^*, T) = Var(T^*)(1 - \alpha_0 - \alpha_1)$

Proof. By the Law of Total Probability $p = (1 - \alpha_1)\, p^* + \alpha_0\, (1 - p^*)$. Hence,

$$\mathrm{Cov}(T, T^*) = \mathbb{P}(T = 1, T^* = 1) - pp^* = (1 - \alpha_1)\, p^* - [(1 - \alpha_1)\, p^* + \alpha_0\, (1 - p^*)]\, p^* = \mathrm{Var}(T^*)(1 - \alpha_0 - \alpha_1).$$

□

Lemma B.4. $Cov(T^*, w) = Cov(T, T^*) - Var(T^*) = -Var(T^*)(\alpha_0 + \alpha_1)$

Proof. Since $w = T - T^*$, by Lemma B.3 $\mathrm{Cov}(T^*, w) = \mathrm{Cov}(T, T^*) - \mathrm{Var}(T^*) = -\mathrm{Var}(T^*)(\alpha_0 + \alpha_1)$. □

Lemma B.5. $Cov(z, w) = -(\alpha_0 + \alpha_1)Cov(z, T^*)$

Proof. By iterated expectations and Assumption 4.1 $\mathbb{E}(zw|T^* = 0) = \alpha_0\mathbb{E}(z|T^* = 0)$ and similarly $\mathbb{E}(zw|T^* = 1) = -\alpha_1\mathbb{E}(z|T^* = 1)$. Hence, $\mathbb{E}(zw) = \alpha_0(1 - p^*)\mathbb{E}(z|T^* = 0) - \alpha_1 p^* E(z|T^* = 1)$. Now, since $E(zT^*) = p^*\mathbb{E}(z|T^* = 1)$, we have $\mathbb{E}(z)\mathbb{E}(T^*) = p^*\mathbb{E}(z|T^* = 1) - \mathrm{Cov}(z, T^*)$ Expanding $\mathbb{E}(z)\mathbb{E}(T)$ using Law of Total Probability,

$$\begin{aligned}
\mathbb{E}(z)\mathbb{E}(T) &= [p^*\mathbb{E}(z|T^* = 1) + (1 - p^*)\mathbb{E}(z|T^* = 0)]\,[p^*(1 - \alpha_1) + (1 - p^*)\alpha_0] \\
&= (1 - \alpha_1)p^*(1 - p^*)\mathbb{E}[z|T^* = 0] + \alpha_0 p^*(1 - p^*)\mathbb{E}[z|T^* = 1] \\
&\quad + (1 - \alpha_1)(p^*)^2\mathbb{E}[z|T^* = 1] + \alpha_0(1 - p^*)^2\mathbb{E}[z|T^* = 0]
\end{aligned}$$

Finally, substituting the expressions for $\mathbb{E}(zw), \mathbb{E}(z)\mathbb{E}(T^*)$ and $\mathbb{E}(z)\mathbb{E}(T)$ and simplifying,

$$\begin{aligned}
\mathrm{Cov}(z, w) &= \mathbb{E}(z, w) - \mathbb{E}(z)\mathbb{E}(w) = \mathbb{E}(z, w) - [\mathbb{E}(z)\mathbb{E}(T) - \mathbb{E}(z)\mathbb{E}(T^*)] \\
&= \mathbb{E}(z, w) - \mathbb{E}(z)\mathbb{E}(T) + [p^*\mathbb{E}(z|T^* = 1) - \mathrm{Cov}(z, T^*)] \\
&= (1 - \alpha_0 - \alpha_1)p^*(1 - p^*)\,[\mathbb{E}(z|T^* = 1) - \mathbb{E}(z|T^* = 0)] - \mathrm{Cov}(z, T^*) \\
&= (1 - \alpha_0 - \alpha_1)\mathrm{Var}(T^*)\mathrm{Cov}(z, T^*)/\mathrm{Var}(T^*) - \mathrm{Cov}(z, T^*) \\
&= -(\alpha_0 + \alpha_1)\mathrm{Cov}(z, T^*)
\end{aligned}$$

□

Lemma B.6. $(1-\alpha_0-\alpha_1)\,Cov(z,T^*) = Cov(z,T)$

Proof. This follows from $\text{Cov}(z,T) = \text{Cov}(z,T^*) + \text{Cov}(z,w)$ by substituting Lemma B.5. □

Lemma B.7. *In the absence of covariates* $\beta_{IV} = \beta/(1-\alpha_0+\alpha_1) + \sigma_{zu}/\sigma_{zT}$.

Proof. Using Lemma B.6, the result follows from $\beta_{IV} = [\beta\text{Cov}(z,T^*) + \text{Cov}(z,u)]/\text{Cov}(z,T)$. □

Lemma B.8. $Cov(w,u) = -(\alpha_0+\alpha_1)\,Cov(T^*,u)$

Proof. By iterated expectations $\mathbb{E}[wu] = -\alpha_1 p^* \mathbb{E}[u|T^*=1] + \alpha_1(1-p^*)\mathbb{E}[u|T^*=0]$, using Assumption 4.1. Applying iterated expectations two more times, we find that

$$\begin{aligned}\text{Cov}(T^*,u) = \mathbb{E}[T^*u] - \mathbb{E}[u]\mathbb{E}[T^*] = p^*\mathbb{E}[u|T^*=1] - cp^* \\ \mathbb{E}[u] = c = p^*\mathbb{E}[u|T^*=1] + (1-p^*)\mathbb{E}[u|T^*=0].\end{aligned}$$

Solving the first of the two preceding equalities for $p^*\mathbb{E}[u|T^*=1]$, the second for $(1-p^*)\mathbb{E}[u|T^*=0]$, and substituting into the expression for $\mathbb{E}[wu]$ gives

$$\begin{aligned}\mathbb{E}[wu] &= -\alpha_1\left[\text{Cov}(T^*,u) + cp^*\right] + \alpha_0\left(c - p^*\mathbb{E}\left[u|T^*=1\right]\right) \\ &= -\alpha_1\left[\text{Cov}(T^*,u) + cp^*\right] + \alpha_0\left\{c - \left[\text{Cov}(T^*,u) + cp^*\right]\right\} \\ &= -(\alpha_0+\alpha_1)\text{Cov}(T^*,u) + \alpha_0 c - cp^*(\alpha_0+\alpha_1)\end{aligned}$$

Now, since $\mathbb{E}[w|T^*=0] = \alpha_0$ and $\mathbb{E}[w|T^*=1] = -\alpha_1$, we have $\mathbb{E}[w] = (1-p^*)\alpha_0 - p^*\alpha_1$. Therefore,

$$\begin{aligned}\text{Cov}(w,u) = \mathbb{E}[wu] - \mathbb{E}[w]\mathbb{E}[u] &= -(\alpha_0+\alpha_1)\text{Cov}(T^*,u) + \alpha_0 c - cp^*(\alpha_0+\alpha_1) - \left[(1-p^*)\alpha_0 - p^*\alpha_1\right]c \\ &= -(\alpha_0+\alpha_1)\text{Cov}(T^*,u).\end{aligned}$$

□

Lemma B.9. *In the absence of covariates,*

$$\beta_{OLS} = \frac{1}{p(1-p)}\left\{\left[\frac{(p-\alpha_0)(1-p-\alpha_1)}{1-\alpha_0-\alpha_1}\right]\beta + (1-\alpha_0-\alpha_1)\sigma_{T^*u}\right\}.$$

Proof. The result follows from $\beta_{OLS} = [\beta\text{Cov}(T,T^*) + \text{Cov}(T^*,u) + \text{Cov}(w,u)]/\text{Var}(T)$ by substituting Lemmas B.3 and B.8. □

Lemma B.10. $Cov(z,u)/Cov(z,T) = \delta_z/(p_1-p_0)$

Proof. By iterated expectations, $\mathbb{E}(zu) = q\mathbb{E}(u|z=1)$ and $\mathbb{E}(u) = q\delta_z + \mathbb{E}(u|z=0)$. Thus,

$$\text{Cov}(z,u) = q\mathbb{E}(u|z=1) - q\left[q\delta_z + \mathbb{E}(u|z=0)\right] = q(1-q)\delta_z.$$

Similarly, $\mathbb{E}(zT) = qp_1$ and $\mathbb{E}(T) = p_1 q + p_0(1-q)$. Thus,

$$\text{Cov}(z,T) = qp_1 - q\left[p_1 q + p_0(1-q)\right] = q(1-q)(p_1-p_0).$$

□

Lemma B.11 (Equations 28 and 29)**.** *Under Assumption 4.1,*

$$\begin{aligned}\widetilde{y}_{0k} &= (\beta + m^*_{1k})\alpha_1 p^*_k + (1-\alpha_0)(1-p^*_k)m^*_{0k} \\ \widetilde{y}_{1k} &= (\beta + m^*_{1k})(1-\alpha_1)p^*_k + \alpha_0(1-p^*_k)m^*_{0k}.\end{aligned}$$

Proof. By the iterated expectations and Lemma B.2,

$$\begin{aligned}\mathbb{E}\left[u|T=0,z_k\right] &= \mathbb{E}_{T^*|T=0,z_k}\left[\mathbb{E}\left[u|T^*,T=0,z_k\right]\right] = \mathbb{E}_{T^*|T=0,z_k}\left[\mathbb{E}\left[u|T^*,z_k\right]\right] \\ &= \mathbb{P}(T^*=1|T=0,z_k)m^*_{1k} + \mathbb{P}(T^*=0|T=0,z_k)m^*_{0k} \\ &= \frac{\alpha_1 p^*_k}{1-p_k}m^*_{1k} + \frac{(1-\alpha_0)(1-p^*_k)}{1-p_k}m^*_{0k}.\end{aligned}$$

Analogously,

$$\mathbb{E}\left[u|T=1,z_k\right] = \frac{(1-\alpha_1)p^*_k}{p_k}m^*_{1k} + \frac{\alpha_0(1-p^*_k)}{p_k}m^*_{0k}.$$

The result follows by combining these expressions with $\bar{y}_{tk} = \beta\mathbb{E}[T^*|T=t,z_k] + \mathbb{E}\left[u|T=t,z_k\right]$. □

Lemma B.12 (Relating δ_{T^*} and δ_z to m^*_{tk}.)**.**

$$\delta_{T^*} = (1-q)\left[\frac{p_0-\alpha_0}{p-\alpha_0}m^*_{10} - \frac{1-p_0-\alpha_1}{1-p-\alpha_1}m^*_{00}\right] + q\left[\frac{p_1-\alpha_0}{p-\alpha_0}m^*_{11} - \frac{1-p_1-\alpha_1}{1-p-\alpha_1}m^*_{01}\right] \quad \text{(B.1)}$$

$$\delta_z = \frac{1}{1-\alpha_0-\alpha_1}\left[(p_1-\alpha_0)m^*_{11} - (p_0-\alpha_0)m^*_{10} + (1-p_1-\alpha_1)m^*_{01} - (1-p_0-\alpha_1)m^*_{00}\right] \quad \text{(B.2)}$$

Proof. By iterated expectations and Bayes' Rule,

$$\begin{aligned}\mathbb{E}[u|T^*=1] &= \mathbb{E}_{z|T^*=1}\left[\mathbb{E}\left[u|T^*=1,z\right]\right] = \mathbb{P}(z=0|T^*=1)m^*_{10} + \mathbb{P}(z=1|T^*=1)m^*_{11} \\ &= \frac{p^*_0(1-q)}{p^*}m^*_{10} + \frac{p^*_1 q}{p^*}m^*_{11} = \frac{1}{p-\alpha_0}\left[(p_0-\alpha_0)(1-q)m^*_{10} + (p_1-\alpha_0)qm^*_{11}\right]\end{aligned}$$

where the final equality follows from Lemma B.1. Similarly

$$\mathbb{E}[u|T^*=0] = \frac{1}{1-p-\alpha_1}\left[(1-p_0-\alpha_1)(1-q)m^*_{00} + (1-p_1-\alpha_1)qm^*_{01}\right]$$

The result for δ_{T^*} now follows from $\delta_{T^*} = \mathbb{E}[u|T^*=1] - \mathbb{E}[u|T^*=0]$. Proceeding in the same way

$$\mathbb{E}[u|z=k] = \mathbb{E}_{T^*|z=k}\left[\mathbb{E}\left(u|z=k,T^*\right)\right] = \frac{1}{1-\alpha_0-\alpha_1}\left[(p_k-\alpha_0)m^*_{1k} + (1-p_k-\alpha_1)m^*_{0k}\right]$$

and the result for δ_z follows from $\delta_z = \mathbb{E}[u|z=1] - \mathbb{E}[u|z=0]$. □

Proof of Proposition 4.1. To begin, define

$$\begin{aligned}g(\alpha_1) &= (\widetilde{y}_{01}-\widetilde{y}_{00}) - \alpha_1[(\widetilde{y}_{01}-\widetilde{y}_{00}) + (\widetilde{y}_{11}-\widetilde{y}_{10})] \\ h(\alpha_1) &= \frac{[(1-q)\widetilde{y}_{00} + q\widetilde{y}_{01}] - \alpha_1\left\{[(1-q)\widetilde{y}_{00} + q\widetilde{y}_{01}] + [(1-q)\widetilde{y}_{10} + q\widetilde{y}_{11}]\right\}}{1-p-\alpha_1} \\ \Delta(\alpha_0) &= \frac{(1-\alpha_0)\widetilde{y}_{10} - \alpha_0\widetilde{y}_{00}}{p_0-\alpha_0} - \frac{(1-\alpha_0)\widetilde{y}_{11} - \alpha_0\widetilde{y}_{01}}{p_1-\alpha_0}\end{aligned}$$

The proof proceeds as follows. First substitute for p^*_k using Lemma B.1 in the expression for $\widetilde{y}_{0k}$ from Equation 28 and then solve for $(\beta + m_{1k^*})$ to yield

$$\beta + m^*_{1k} = \frac{(1-\alpha_0-\alpha_1)\widetilde{y}_{0k} - (1-\alpha_0)(1-p_k-\alpha_1)m^*_{0k}}{\alpha_1(p_k-\alpha_0)}$$

Now, substituting the preceding equality into the expression for $\widetilde{y}_{1k}$ from Equation 29, again replacing for

p_k^* using Lemma B.1 and rearranging, we find that

$$m_{0k}^* = \frac{(1-\alpha_1)\widetilde{y}_{0k} - \alpha_1 \widetilde{y}_{1k}}{1 - p_k - \alpha_1} \tag{B.3}$$

Next, summing Equations 28 and 29, and solving for $(\beta + m_{1k}^*)$ we obtain

$$(\beta + m_{1k}^*) = \frac{(\widetilde{y}_{0k} + \widetilde{y}_{1k}) - (1-p_k^*)m_{0k}^*}{p_k^*}. \tag{B.4}$$

Now we subtract the preceding equation evaluated at $k=1$ from the same evaluated at $k=0$, yielding

$$m_{10}^* - m_{11}^* = \left[\frac{\widetilde{y}_{00} + \widetilde{y}_{10}}{p_0^*} - \frac{\widetilde{y}_{01} + \widetilde{y}_{11}}{p_1^*}\right] + \left[\frac{(1-p_\ell^*)m_{01}^*}{p_1^*} - \frac{(1-p_0^*)m_{00}^*}{p_0^*}\right].$$

Now we eliminate m_{00}^* and m_{01}^* from the preceding using Equation B.3 to obtain

$$m_{10}^* - m_{11}^* = \frac{(1-\alpha_0)\widetilde{y}_{10} - \alpha_0\widetilde{y}_{00}}{p_0 - \alpha_0} - \frac{(1-\alpha_0)\widetilde{y}_{11} - \alpha_0\widetilde{y}_{01}}{p_1 - \alpha_0}. \tag{B.5}$$

Next we eliminate m_{0k}^* from Equations B.1 and B.2 again using Equation B.3. We obtain,

$$\delta_{T^*} = \left\{ m_{10}^* \left[\frac{(1-q)(p_0-\alpha_0)}{p-\alpha_0}\right] + m_{11}^* \left[\frac{q(p_1-\alpha_0)}{p-\alpha_0}\right]\right\} - h(\alpha_1) \tag{B.6}$$

and

$$\delta_z = \frac{(p_1-\alpha_0)m_{11}^* - (p_0-\alpha_0)m_{10}^*}{1-\alpha_0-\alpha_1} + \frac{g(\alpha_1)}{1-\alpha_0-\alpha_1} \tag{B.7}$$

Equations B.5, B.6 and B.7 constitute an over-determined system of linear equations in (m_{10}^*, m_{11}^*), namely

$$m_{10}^* - m_{11}^* = \Delta(\alpha_0) \tag{B.8}$$

$$(1-q)(p_0-\alpha_0)m_{10}^* + q(p_1-\alpha_0)m_{11}^* = (p-\alpha_0)\left[\delta_{T^*} + h(\alpha_1)\right] \tag{B.9}$$

$$(p_0-\alpha_0)m_{10}^* - (p_1-\alpha_0)m_{11}^* = g(\alpha_1) - (1-\alpha_0-\alpha_1)\delta_z \tag{B.10}$$

Substituting Equation B.8 into B.10 to eliminate m_{10}^* and rearranging yields

$$(p_0-\alpha_0)\left[m_{11}^* + \Delta(\alpha_0)\right] - (p_1-\alpha_0)m_{11}^* = g(\alpha_1) - (1-\alpha_0-\alpha_1)\delta_z$$

and thus

$$m_{11}^* = \left[\frac{g(\alpha_1) - (1-\alpha_0-\alpha_1)\delta_z - (p_0-\alpha_0)\Delta(\alpha_0)}{p_0-p_1}\right] \tag{B.11}$$

while making the same substitution into Equation B.9 yields

$$m_{11}^* = \left[\frac{(p-\alpha_0)\left[\delta_{T^*} + h(\alpha_1)\right] - (1-q)(p_0-\alpha_0)\Delta(\alpha_0)}{(1-q)(p_0-\alpha_0) + q(p_1-\alpha_0)}\right]. \tag{B.12}$$

Finally, equating the two preceding expressions we see that

$$\delta_{T^*} + h(\alpha_1) - \left[\frac{(1-q)(p_0-\alpha_0)\Delta(\alpha_0)}{p-\alpha_0}\right] = \left[\frac{g(\alpha_1) - (1-\alpha_0-\alpha_1)\delta_z - (p_0-\alpha_0)\Delta(\alpha_0)}{p_0-p_1}\right]$$

using the fact that $(1-q)p_0 + qp_1 = p$. □

Lemma B.13. *Under Assumption 4.1,*

$$s_{0k}^{*2} = \frac{(1-\alpha_1)(1-p_k)\sigma_{0k}^2 - \alpha_1 p_k \sigma_{1k}^2}{1-p_k-\alpha_1} - \frac{\alpha_1(1-\alpha_1)p_k(1-p_k)(\bar{y}_{1k}-\bar{y}_{0k})^2}{(1-p_k-\alpha_1)^2}$$
$$s_{1k}^{*2} = \frac{(1-\alpha_0)p_k\sigma_{1k}^2 - \alpha_0(1-p_k)\sigma_{0k}^2}{p_k-\alpha_0} - \frac{\alpha_0(1-\alpha_0)p_k(1-p_k)(\bar{y}_{1k}-\bar{y}_{0k})^2}{(p_k-\alpha_0)^2}.$$

Proof of Lemma B.13. First $\mathbb{E}(y^2|T,z) = \mathbb{E}_{T^*|T,z}\left[\mathbb{E}(y^2|T^*,z)\right]$ and $\mathbb{E}(y|T,z) = \mathbb{E}_{T^*|T,z}\left[\mathbb{E}(y|T^*,z)\right]$ by iterated expectations. Next, by Lemma B.2,

$$\begin{aligned}
\mathbb{E}(y^2|T=0,z_k) &= \frac{\alpha_1 p_k^*}{1-p_k}\mathbb{E}\left[(\beta+u)^2|T^*=1,z_k\right] + \frac{(1-\alpha_0)(1-p_k^*)}{1-p_k}\mathbb{E}\left[u^2|T^*=0,z_k\right] \\
\mathbb{E}(y|T=0,z_k) &= \frac{\alpha_1 p_k^*}{1-p_k}\mathbb{E}\left[\beta+u|T^*=1,z_k\right] + \frac{(1-\alpha_0)(1-p_k^*)}{1-p_k}\mathbb{E}\left[u|T^*=0,z_k\right]
\end{aligned}$$

using the fact that $y = \beta T^* + u$. Combining these and simplifying yields,

$$\sigma_{0k}^2 = \frac{\alpha_1 p_k^*}{1-p_k}s_{1k}^{*2} + \frac{(1-\alpha_0)(1-p_k^*)}{1-p_k}s_{0k}^{*2} + V_{0k}(\alpha_0,\alpha_1)(\beta + m_{1k}^* - m_{0k}^*)^2 \tag{B.13}$$

where

$$V_{0k}(\alpha_0,\alpha_1) = \frac{\alpha_1(1-\alpha_0)(p_k-\alpha_0)(1-p_k-\alpha_1)}{(1-p_k)^2(1-\alpha_0-\alpha_1)^2} \tag{B.14}$$

Similarly,

$$\begin{aligned}
\mathbb{E}(y^2|T=1,z_k) &= \frac{(1-\alpha_1)p_k^*}{p_k}\mathbb{E}\left[(\beta+u)^2|T^*=1,z_k\right] + \frac{\alpha_0(1-p_k^*)}{p_k}\mathbb{E}\left[u^2|T^*=0,z_k\right] \\
\mathbb{E}(y|T=1,z_k) &= \frac{(1-\alpha_1)p_k^*}{p_k}\mathbb{E}\left[\beta+u|T^*=1,z_k\right] + \frac{\alpha_0(1-p_k^*)}{p_k}\mathbb{E}\left[u|T^*=0,z_k\right]
\end{aligned}$$

and thus

$$\sigma_{1k}^2 = \frac{(1-\alpha_1)p_k^*}{p_k}s_{1k}^{*2} + \frac{\alpha_0(1-p_k^*)}{p_k}s_{0k}^{*2} + V_{1k}(\alpha_0,\alpha_1)(\beta + m_{1k}^* - m_{0k}^*)^2 \tag{B.15}$$

where

$$V_{1k}(\alpha_0,\alpha_1) = \frac{\alpha_0(1-\alpha_1)(p_k-\alpha_0)(1-p_k-\alpha_1)}{p_k^2(1-\alpha_0-\alpha_1)^2} \tag{B.16}$$

Now, combining Equations B.3 and B.4 from the proof of Proposition 4.1,

$$\beta + m_{1k}^* - m_{0k}^* = (1-\alpha_0-\alpha_1)\frac{(1-p_k)\widetilde{y}_{1k} - p_k\widetilde{y}_{0k}}{(p_k-\alpha_0)(1-p_k-\alpha_1)}. \tag{B.17}$$

Substituting Equations B.14 and B.17 into Equation B.13, and Equations B.16 and B.17 into Equation B.15,

$$\begin{aligned}
\sigma_{0k}^2 &= \frac{\alpha_1 p_k^*}{1-p_k}s_{1k}^{*2} + \frac{(1-\alpha_0)(1-p_k^*)}{1-p_k}s_{0k}^{*2} + \frac{\alpha_1(1-\alpha_0)p_k^2\left(\bar{y}_{1k}-\bar{y}_{0k}\right)^2}{(p_k-\alpha_0)(1-p_k-\alpha_1)} \\
\sigma_{1k}^2 &= \frac{(1-\alpha_1)p_k^*}{p_k}s_{1k}^{*2} + \frac{\alpha_0(1-p_k^*)}{p_k}s_{0k}^{*2} + \frac{\alpha_0(1-\alpha_1)(1-p_k)^2\left(\bar{y}_{1k}-\bar{y}_{0k}\right)^2}{(p_k-\alpha_0)(1-p_k-\alpha_1)}
\end{aligned}$$

The result follows by solving these equations for s_{0k}^{*2} and s_{1k}^{*2}. □

Lemma B.14. *Under Assumptions 4.1 and 4.2,*

$$\alpha_0 \leq \min_k\{p_k\}, \quad \alpha_1 \leq \min_k\{1-p_k\}. \tag{B.18}$$

These inequalities are strict unless p_k^* is zero or one.

Proof of Lemma B.14. Rearranging the result of Lemma B.1,

$$p_k - \alpha_0 = (1 - \alpha_0 - \alpha_1)p_k^*$$
$$(1 - p_k) - \alpha_1 = (1 - \alpha_0 - \alpha_1)(1 - p_k^*).$$

Now, since p_k^* and $(1 - p_k^*)$ are probabilities they are between zero and one which means that the sign of $p_k - \alpha_0$ as well as that of $(1 - p_k) - \alpha_1$ are both determined by that of $1 - \alpha_0 - \alpha_1$. Thus, under Assumption 4.2, $\alpha_0 \leq p_k$ and $\alpha_1 \leq (1 - p_k)$ for all k. □

Proof of Proposition 4.2. We first derive the bounds for α_0 and α_1. We then show that these bounds are sharp and that our assumptions imply no restrictions on δ_{T^*}. By assumption, $s_{0k}^{*2}, s_{1k}^{*2} > 0$. Thus, by Lemma B.13,

$$(p_k - \alpha_0)\left[(1 - \alpha_0)p_k\sigma_{1k}^2 - \alpha_0(1 - p_k)\sigma_{0k}^2\right] > \alpha_0(1 - \alpha_0)p_k(1 - p_k)(\bar{y}_{1k} - \bar{y}_{0k})^2 \tag{B.19}$$
$$(1 - p_k - \alpha_1)\left[(1 - \alpha_1)(1 - p_k)\sigma_{0k}^2 - \alpha_1 p_k\sigma_{1k}^2\right] > \alpha_1(1 - \alpha_1)p_k(1 - p_k)(\bar{y}_{1k} - \bar{y}_{0k})^2 \tag{B.20}$$

Thus, for each k we obtain a pair of quadratic inequalities that bound α_0 and α_1.

Consider first Inequality B.19. Notice that both $\alpha_0 = 0$ and $\alpha_0 = 1$ satisfy the inequality. In contrast $\alpha_0 = p_k$ does not: the left hand side becomes zero while the right hand side is *strictly* positive. This implies that the quadratic equation defining the boundary of the inequality crosses the α_0-axis and thus has two real roots. Moreover, one of these is strictly less than p_k. Rearranging Inequality B.19, we have $A_k\alpha_0^2 + B_k^0\alpha_0 + C_k^0 > 0$ where

$$A_k = p_k(1 - p_k)(\bar{y}_{1k} - \bar{y}_{0k})^2 + (1 - p_k)\sigma_{0k}^2 + p_k\sigma_{1k}^2$$
$$B_k^0 = -\left[\sigma_{1k}^2 p_k(1 + p_k) + p_k(1 - p_k)\sigma_{0k}^2 + p_k(1 - p_k)(\bar{y}_{1k} - \bar{y}_{0k})^2\right]$$
$$C_k^0 = p_k^2\sigma_{1k}^2.$$

Since $A_k > 0$, the quadratic equation defined by $A_k\alpha_1^2 + B_k^0\alpha_1 + C_k^0 = 0$ opens upwards. By Lemma B.14, $\alpha_0 < p_k$ so we need only consider the smaller of the two roots. Our bound imposes that α_0 be strictly less than this quantity. Analogous reasoning for Inequality B.20, using $1 - p_k$ rather than p_k, shows that the smaller of the two roots of $A_k\alpha_1^2 + B_k^1\alpha_1 + C_k^1 = 0$ bounds α_1 from above, where

$$A_k = p_k(1 - p_k)(\bar{y}_{1k} - \bar{y}_{0k})^2 + (1 - p_k)\sigma_{0k}^2 + p_k\sigma_{1k}^2$$
$$B_k^1 = -\left[p_k(1 - p_k)\sigma_{1k}^2 + (1 - p_k)(2 - p_k)\sigma_{0k}^2 + p_k(1 - p_k)(\bar{y}_{1k} - \bar{y}_{0k})^2\right]$$
$$C_k^1 = (1 - p_k)^2\sigma_{0k}^2.$$

Notice that the coefficient of the squared term, A_k, is common to both quadratics. Because the bounds for α_0 and α_1 hold for each k, we can take the tighter of each. Now, equating $f_{1k}(\alpha_0)$ and $f_{0k}(\alpha_0)$ and rearranging gives precisely $A_k\alpha_0^2 + B_k^0\alpha_0 + C_k^0 = 0$. Equating the inverse functions

$$f_{0k}^{-1}(\alpha_1) = \frac{p_k(1 - p_k - \alpha_1)\sigma_{0k}^2 - p_k^2(\bar{y}_{1k} - \bar{y}_{0k})^2\alpha_1}{(1 - p_k - \alpha_1)\sigma_{0k}^2 - p_k^2(\bar{y}_{1k} - \bar{y}_{0k})^2\alpha_1}$$
$$f_{1k}^{-1}(\alpha_1) = \frac{p_k(1 - p_k - \alpha_1)\sigma_{1k}^2}{(1 - p_k - \alpha_1)\sigma_{1k}^2 + (1 - p_k)^2(\bar{y}_{1k} - \bar{y}_{0k})^2(1 - \alpha_1)}$$

and rearranging gives precisely $A_k\alpha_1^2 + B_k^1\alpha_1 + C_k^1 = 0$.

We now show that $(-\infty, \infty) \times [0, \bar{\alpha}_0) \times [0, \bar{\alpha}_1)$ is the sharp identified set for $(\delta_{T^*}, \alpha_0, \alpha_1)$. Because all of our steps from above are reversible, so long as $\bar{y}_{1k} \neq \bar{y}_{0k}$, any $(\alpha_0, \alpha_1) \in [0, \bar{\alpha}_0) \times [0, \bar{\alpha}_1)$ implies $s_{0k}^{*2}, s_{1k}^{*2} > 0$. Moreover, at any pair (α_0, α_1) within these bounds, p^* and p_k^* all lie in the interval $[0, 1]$. It remains only to show that δ_{T^*} is unrestricted. Notice that the expressions for s_{0k}^{*2} and s_{1k}^{*2} from Lemma B.13 involve only α_0, α_1 and observables: they do not involve m_{tk}^*. The m_{tk}^* are only constrained by observables through Equations 28 and 29. For any fixed values of (α_0, α_1) these constitute a linear system of four equations in five unknowns: β, $m_{00}^*, m_{01}^*, m_{10}^*$ and m_{11}^*. This means that we can solve for each of the other unobservables

for *any* value of the free parameter m_{11}^* and still satisfy the assumptions of the model. The result follows since δ_{T^*} can be written as a linear function of m_{11}^* and observables only, for any fixed values of (α_0, α_1), using Equation B.12. □

Lemma B.15. *Under Assumption 4.3, (i)* $\mathbb{P}(z=1|T^*,T) = \mathbb{P}(z=1|T^*)$*, and (ii)* $\mathbb{E}[\mathbf{x}|T^*,T] = \mathbb{E}[\mathbf{x}|T^*]$.

Proof. By Bayes' Rule and Assumption 4.3 (iii),

$$\mathbb{P}(z=1|T^*=t^*,T=t) = \frac{\mathbb{P}(T=t|T^*=t^*)P(z=1|T^*=t^*)}{\mathbb{P}(T=t|T^*=t^*)} = \mathbb{P}(z=1|T^*=1)$$

which proves part (i). Now, by iterated expectations and Assumption 4.3 (i)

$$\mathbb{E}[\mathbf{x}|T^*,T] = \mathbb{E}_{z|T^*,T}[\mathbb{E}(\mathbf{x}|T^*,T,z)] = \mathbb{E}_{z|T^*,T}[\mathbb{E}(\mathbf{x}|T^*,z)]$$

and

$$\mathbb{E}[\mathbf{x}|T^*] = \mathbb{E}_{z|T^*}[\mathbb{E}(\mathbf{x}|T^*,z)]$$

Part (ii) now follows by part (i), which shows that that the conditional distribution of z given T^*, T is the same as the conditional distribution of z given T^*. □

Lemma B.16.

$$\mathbb{E}(\mathbf{x}|T^*=1) - \mathbb{E}(\mathbf{x}|T^*=0) = \frac{p(1-p)(1-\alpha_0-\alpha_1)}{(p-\alpha_0)(1-p-\alpha_1)}[\mathbb{E}(\mathbf{x}|T=1) - \mathbb{E}(\mathbf{x}|T=0)]$$

Proof. By Lemma B.15 (ii), $\mathbb{E}[\mathbf{x}|T^*,T] = \mathbb{E}[\mathbf{x}|T^*]$ and hence

$$\begin{aligned}
\mathbb{E}[\mathbf{x}|T=1] &= \mathbb{P}(T^*=1|T=1)\mathbb{E}[\mathbf{x}|T^*=1] + \mathbb{P}(T^*=0|T=1)\mathbb{E}[\mathbf{x}|T^*=0] \\
\mathbb{E}[\mathbf{x}|T=0] &= \mathbb{P}(T^*=1|T=0)\mathbb{E}[\mathbf{x}|T^*=1] + \mathbb{P}(T^*=0|T=0)\mathbb{E}[\mathbf{x}|T^*=0]
\end{aligned}$$

by the law of iterated expectations. Now, by Bayes' rule,

$$\begin{aligned}
&\mathbb{P}(T^*=1|T=1) = (1-\alpha_1)p^*/p, \quad \mathbb{P}(T^*=0|T=1) = \alpha_0(1-p^*)/p \\
&\mathbb{P}(T^*=1|T=0) = \alpha_1 p^*/(1-p), \quad \mathbb{P}(T^*=0|T=0) = (1-\alpha)(1-p^*)/(1-p)
\end{aligned}$$

and substituting these four expressions into the preceding two along with Lemma B.1 we find that

$$\begin{aligned}
p\mathbb{E}[\mathbf{x}|T=1] &= \frac{(1-\alpha_1)(p-\alpha_0)}{1-\alpha_0-\alpha_1}\mathbb{E}[\mathbf{x}|T^*=1] + \frac{\alpha_0(1-p-\alpha_1)}{1-\alpha_0-\alpha_1}\mathbb{E}[\mathbf{x}|T^*=0] \\
(1-p)\mathbb{E}[\mathbf{x}|T=0] &= \frac{\alpha_1(p-\alpha_0)}{1-\alpha_0-\alpha_1}\mathbb{E}[\mathbf{x}|T^*=1] + \frac{(1-\alpha_0)(1-p-\alpha_1)}{1-\alpha_0-\alpha_1}\mathbb{E}[\mathbf{x}|T^*=0]
\end{aligned}$$

This is a system of two equations in two unknowns. The result follows by solving and rearranging. □

Lemma B.17.

$$\delta_z = [\mathbb{E}(\mathbf{x}|z=1) - \mathbb{E}(\mathbf{x}|z=0)]'\gamma + \widetilde{\delta}_z \tag{B.21}$$

$$\delta_{T^*} = \frac{p(1-p)(1-\alpha_0-\alpha_1)}{(p-\alpha_0)(1-p-\alpha_1)}[\mathbb{E}(\mathbf{x}|T=1) - \mathbb{E}(\mathbf{x}|T=0)]'\gamma + \widetilde{\delta}_{T^*} \tag{B.22}$$

Proof. Since $u = c + \mathbf{x}'\boldsymbol{\gamma} + \varepsilon$ where c is a constant, $\delta_z = [\mathbb{E}(\mathbf{x}|z=1) - \mathbb{E}(\mathbf{x}|z=0)]'\boldsymbol{\gamma} + \widetilde{\delta}_z$ and similarly

$$\begin{aligned}
\delta_{T^*} &= \mathbb{E}[u|T^*=1] - \mathbb{E}[u|T^*=0] = [\mathbb{E}(\mathbf{x}|T^*=1) - \mathbb{E}(\mathbf{x}|T^*=0)]'\boldsymbol{\gamma} + \widetilde{\delta}_{T^*} \\
&= \frac{p(1-p)(1-\alpha_0-\alpha_1)}{(p-\alpha_0)(1-p-\alpha_1)}[\mathbb{E}(\mathbf{x}|T=1) - \mathbb{E}(\mathbf{x}|T=0)]'\gamma + \widetilde{\delta}_{T^*}
\end{aligned}$$

where the final equality follows from Lemma B.16. □

Lemma B.18. *The probability limit of the IV estimators of β and $\boldsymbol{\gamma}$ is given by*

$$\begin{bmatrix} \beta_{IV} \\ \boldsymbol{\gamma}_{IV} \end{bmatrix} = \begin{bmatrix} \left(\sigma^{zT}\sigma_{zT^*} + \boldsymbol{\sigma}^{zx'}\boldsymbol{\sigma}_{xT^*}\right)\beta \\ \gamma + \left(\boldsymbol{\sigma}^{xT}\sigma_{zT^*} + \Sigma^{xx}\boldsymbol{\sigma}_{xT^*}\right)\beta \end{bmatrix} + \sigma_{z\varepsilon}\begin{bmatrix} \sigma^{zT} \\ \boldsymbol{\sigma}^{xT} \end{bmatrix}$$

where

$$\begin{bmatrix} \sigma^{zT} & \boldsymbol{\sigma}^{zx'} \\ \boldsymbol{\sigma}^{xT} & \Sigma^{xx} \end{bmatrix} \equiv \begin{bmatrix} \sigma_{zT} & \boldsymbol{\sigma}'_{zx} \\ \boldsymbol{\sigma}_{xT} & \Sigma_{xx} \end{bmatrix}^{-1}$$

Proof. Let $\widetilde{\mathbf{y}}$ be the de-meaned version of y and so on. Then we have $\widetilde{y} = \beta\widetilde{T}^* + \widetilde{\mathbf{x}}'\gamma + \varepsilon$ using the fact that, since ε is mean zero $\varepsilon = \widetilde{\varepsilon}$. Stacking observations in the usual way,

$$\begin{aligned} \begin{bmatrix} \widehat{\beta}_{IV} \\ \widehat{\gamma}_{IV} \end{bmatrix} &= \begin{bmatrix} \widetilde{\mathbf{z}}'\widetilde{\mathbf{T}}/n & \widetilde{\mathbf{z}}'\widetilde{X}/n \\ \widetilde{X}'\widetilde{\mathbf{T}}/n & \widetilde{X}'\widetilde{X}/n \end{bmatrix}^{-1} \left\{ \begin{bmatrix} \widetilde{\mathbf{z}}'\widetilde{\mathbf{T}}^*/n & \widetilde{\mathbf{z}}\widetilde{X}/n \\ \widetilde{X}'\widetilde{\mathbf{T}}^*/n & \widetilde{X}\widetilde{X}/n \end{bmatrix} \begin{bmatrix} \beta \\ \gamma \end{bmatrix} + \begin{bmatrix} \widetilde{\mathbf{z}}'\boldsymbol{\varepsilon}/n \\ \widetilde{X}'\boldsymbol{\varepsilon}/n \end{bmatrix} \right\} \\ &\to_p \begin{bmatrix} \sigma^{zT} & \boldsymbol{\sigma}^{zx'} \\ \boldsymbol{\sigma}^{xT} & \Sigma^{xx} \end{bmatrix} \left\{ \begin{bmatrix} \sigma_{zT^*} & \boldsymbol{\sigma}'_{zx} \\ \boldsymbol{\sigma}_{xT^*} & \Sigma_{xx} \end{bmatrix} \begin{bmatrix} \beta \\ \boldsymbol{\gamma} \end{bmatrix} + \begin{bmatrix} \sigma_{z\varepsilon} \\ \mathbf{0} \end{bmatrix} \right\} \end{aligned}$$

under standard regularity conditions. The result follows since

$$\begin{bmatrix} \sigma^{zT} & \boldsymbol{\sigma}^{zx'} \\ \boldsymbol{\sigma}^{xT} & \Sigma^{xx} \end{bmatrix} \begin{bmatrix} \boldsymbol{\sigma}'_{zx} \\ \Sigma_{xx} \end{bmatrix} = \begin{bmatrix} \mathbf{0}' \\ \mathbf{I} \end{bmatrix}$$

by the definition of an inverse matrix. □

Lemma B.19. $\boldsymbol{\sigma}_{xT^*} = \boldsymbol{\sigma}_{xT}(\sigma_{zT^*}/\sigma_{zT})$

Proof. Since the result is an element-wise equality between two vectors we need only show that

$$\mathrm{Cov}(x, T^*)/\mathrm{Cov}(x, T) = \mathrm{Cov}(z, T^*)/\mathrm{Cov}(z, T)$$

for an arbitrary element x of $\mathbf{x}$. By the Law of Iterated Expectations,

$$\begin{aligned} \mathrm{Cov}(z, T^*) &= p^* \left[\mathbb{E}\left(z|T^* = 1\right) - \mathbb{E}(z)\right], \quad \mathrm{Cov}(z, T) = p\left[\mathbb{E}\left(z|T = 1\right) - \mathbb{E}(z)\right] \\ \mathrm{Cov}(x, T^*) &= p^* \left[\mathbb{E}\left(x|T^* = 1\right) - \mathbb{E}(x)\right], \quad \mathrm{Cov}(x, T) = p\left[\mathbb{E}\left(x|T = 1\right) - \mathbb{E}(x)\right] \end{aligned}$$

and hence we have

$$\frac{\mathrm{Cov}(z, T^*)}{\mathrm{Cov}(x, T^*)} = \frac{\mathbb{E}(z|T^* = 1) - \mathbb{E}(z)}{\mathbb{E}(x|T^* = 1) - \mathbb{E}(x)}, \quad \frac{\mathrm{Cov}(z, T)}{\mathrm{Cov}(x, T)} = \frac{\mathbb{E}(z|T = 1) - \mathbb{E}(z)}{\mathbb{E}(x|T = 1) - \mathbb{E}(x)}$$

By iterated expectations and Lemma B.15 (i),

$$\mathbb{E}(z|T = 1) = (1 - \alpha_1)\frac{p^*}{p}\mathbb{E}(z|T^* = 1) + \frac{\alpha_0(1 - p^*)}{p}\mathbb{E}(z|T = 0)$$

but by the Law of total probability $(1-p^*)\mathbb{E}(z|T^*=0) = \mathbb{E}(z) - p^*\mathbb{E}(z|T^*=1)$ and hence

$$\begin{aligned}
\mathbb{E}(z|T=1) &= \frac{1}{p}\left[(1-\alpha_1)p^*\mathbb{E}(z|T^*=1) + \alpha_0(1-p^*)\mathbb{E}(z|T=0)\right] \\
&= \frac{1}{p}\left[(1-\alpha_1)p^*\mathbb{E}(z|T^*=1) + \alpha_0\left\{\mathbb{E}(z) - p^*\mathbb{E}(z|T^*=1)\right\}\right] \\
&= \frac{1}{p}\left[(1-\alpha_0-\alpha_1)p^*\mathbb{E}(z|T^*=1) + \alpha_0\mathbb{E}(z)\right] \\
&= \frac{1}{p}\left[(1-\alpha_0-\alpha_1)\frac{p-\alpha_0}{1-\alpha_0-\alpha_1}\mathbb{E}(z|T^*=1) + \alpha_0\mathbb{E}(z)\right] = \frac{1}{p}\left[(p-\alpha_0)\mathbb{E}(z|T^*=1) + \alpha_0\mathbb{E}(z)\right]
\end{aligned}$$

which implies

$$\begin{aligned}
\mathbb{E}(z|T=1) - \mathbb{E}(z) &= \frac{1}{p}\left[(p-\alpha_0)\mathbb{E}(z|T^*=1) + \alpha_0\mathbb{E}(z)\right] - \frac{p}{p}\mathbb{E}(z) \\
&= \frac{1}{p}\left[(p-\alpha_0)\mathbb{E}(z|T^*=1) - (p-\alpha_0)\mathbb{E}(z)\right] = \frac{p-\alpha_0}{p}\left[\mathbb{E}(z|T^*=1) - \mathbb{E}(z)\right]
\end{aligned}$$

An identical argument with x in place of z gives

$$\mathbb{E}(x|T=1) - \mathbb{E}(x) = \frac{p-\alpha_0}{p}\left[\mathbb{E}(z|T^*=1) - \mathbb{E}(x)\right]$$

Therefore

$$\begin{aligned}
\frac{\mathrm{Cov}(z,T)}{\mathrm{Cov}(x,T)} &= \frac{\mathbb{E}(z|T=1) - \mathbb{E}(z)}{\mathbb{E}(x|T=1) - \mathbb{E}(x)} = \frac{(p-\alpha_0)\left[\mathbb{E}(z|T^*=1) - \mathbb{E}(z)\right]/p}{(p-\alpha_0)\left[\mathbb{E}(x|T^*=1) - \mathbb{E}(x)\right]/p} \\
&= \frac{\mathbb{E}(z|T^*=1) - \mathbb{E}(z)}{\mathbb{E}(x|T^*=1)} = \frac{\mathrm{Cov}(z,T^*)}{\mathrm{Cov}(x,T^*)}
\end{aligned}$$

□

Lemma B.20. $\left(\sigma^{zT}\sigma_{zT^*} + \boldsymbol{\sigma}^{zx'}\boldsymbol{\sigma}_{xT^*}\right) = 1/(1-\alpha_0-\alpha_1)$.

Proof. By the partitioned matrix inverse formula,

$$\begin{aligned}
\sigma^{zT} &= \left(\sigma_{zT} - \boldsymbol{\sigma}'_{zx}\Sigma_{xx}\boldsymbol{\sigma}_{xT}\right)^{-1} \\
\boldsymbol{\sigma}^{zx'} &= -\sigma_{zT}^{-1}\boldsymbol{\sigma}'_{zx}\left(\Sigma_{xx} - \boldsymbol{\sigma}_{xT}\sigma_{zT}^{-1}\boldsymbol{\sigma}'_{zx}\right)^{-1}
\end{aligned}$$

Now, by the Woodbury matrix identity $(A-BCD)^{-1} = A^{-1} + A^{-1}B\left(C^{-1} - DA^{-1}B\right)^{-1}DA^{-1}$ and hence

$$\left(\Sigma_{xx} - \boldsymbol{\sigma}_{xT}\sigma_{zT}^{-1}\boldsymbol{\sigma}'_{zx}\right)^{-1} = \Sigma_{xx}^{-1} + \Sigma_{xx}^{-1}\boldsymbol{\sigma}_{xT}\left(\sigma_{zT} - \boldsymbol{\sigma}'_{zx}\Sigma_{xx}^{-1}\boldsymbol{\sigma}_{xT}\right)^{-1}\boldsymbol{\sigma}'_{zx}\Sigma_{xx}^{-1} = \Sigma_{xx}^{-1}\left(\mathbf{I} + \frac{\boldsymbol{\sigma}_{xT}\boldsymbol{\sigma}'_{zx}\Sigma_{xx}^{-1}}{\sigma_{zT} - \boldsymbol{\sigma}'_{zx}\Sigma_{xx}^{-1}\boldsymbol{\sigma}_{xT}}\right)$$

Substituting this into our expression for $\boldsymbol{\sigma}^{zx'}$ gives

$$\sigma^{zT}\sigma_{zT^*} + \boldsymbol{\sigma}^{zx'}\boldsymbol{\sigma}_{xT^*} = \frac{\sigma_{zT^*}}{\sigma_{zT} - \boldsymbol{\sigma}'_{zx}\Sigma_{xx}\boldsymbol{\sigma}_{xT}} - \sigma_{zT}^{-1}\boldsymbol{\sigma}'_{zx}\Sigma_{xx}^{-1}\left(\mathbf{I} + \frac{\boldsymbol{\sigma}_{xT}\boldsymbol{\sigma}'_{zx}\Sigma_{xx}^{-1}}{\sigma_{zT} - \boldsymbol{\sigma}'_{zx}\Sigma_{xx}^{-1}\boldsymbol{\sigma}_{xT}}\right)\boldsymbol{\sigma}_{xT^*}$$

Expressing the right-hand side over a common denominator and simplifying gives

$$\sigma^{zT}\sigma_{zT^*} + \boldsymbol{\sigma}^{zx'}\boldsymbol{\sigma}_{xT^*} = \frac{\sigma_{zT^*} - \boldsymbol{\sigma}'_{zx}\Sigma_{xx}\boldsymbol{\sigma}_{xT^*}}{\sigma_{zT} - \boldsymbol{\sigma}'_{zx}\Sigma_{xx}\boldsymbol{\sigma}_{xT}} \tag{B.23}$$

using the fact that σ_{zT}^{-1} is a scalar and hence commutes. By Lemma B.6 we have $\sigma_{zT}/(1-\alpha_0-\alpha_1) = \sigma_{zT^*}$ and by Lemma B.19 we have $\mathrm{Cov}(z,T)/\mathrm{Cov}(z,T^*) = \mathrm{Cov}(x,T)/\mathrm{Cov}(x,T^*)$ for each element x of $\mathbf{x}$ which

implies $\boldsymbol{\sigma}_{xT}/(1-\alpha_0-\alpha_1) = \boldsymbol{\sigma}_{xT^*}$. The result follows by substituting into Equation B.23. □

Lemma B.21. $(\boldsymbol{\sigma}^{xT}\sigma_{zT^*} + \Sigma^{xx}\boldsymbol{\sigma}_{xT^*}) = \mathbf{0}$

Proof. By the partitioned matrix inverse formula

$$\boldsymbol{\sigma}^{xT}\sigma_{zT^*} + \Sigma^{xx}\boldsymbol{\sigma}_{xT^*} \quad = \quad -\Sigma^{xx}\frac{\boldsymbol{\sigma}_{xT}}{\sigma_{zT}}\sigma_{zT^*} + \Sigma^{xx}\boldsymbol{\sigma}_{xT^*} = \Sigma^{xx}\left(\boldsymbol{\sigma}_{xT^*} - \frac{\boldsymbol{\sigma}_{xT}}{\sigma_{zT}}\sigma_{zT^*}\right)$$

The result follows since $\boldsymbol{\sigma}_{xT^*} = \boldsymbol{\sigma}_{xT}(\sigma_{zT^*}/\sigma_{zT})$ by Lemma B.19. □

Proposition B.1. *The probability limit of the IV estimators of β and $\boldsymbol{\gamma}$ is given by*

$$\begin{bmatrix} \beta_{IV} \\ \boldsymbol{\gamma}_{IV} \end{bmatrix} = \begin{bmatrix} \beta/(1-\alpha_0-\alpha_1) \\ \boldsymbol{\gamma} \end{bmatrix} + \widetilde{\delta}_z q(1-q)\begin{bmatrix} \sigma^{zT} \\ \boldsymbol{\sigma}^{xT} \end{bmatrix}$$

where $\sigma^{zT} = \left(\sigma_{zT} - \boldsymbol{\sigma}'_{zx}\Sigma_{xx}\boldsymbol{\sigma}_{xT}\right)^{-1}$ *and* $\boldsymbol{\sigma}^{xT} = -\left(\Sigma_{xx} - \boldsymbol{\sigma}_{xT}\sigma_{zT}^{-1}\boldsymbol{\sigma}'_{zx}\right)^{-1}\boldsymbol{\sigma}_{xT}\sigma_{zT}^{-1}$.

Proof. We show that $\widetilde{\delta}_z q(1-q) = \sigma_{z\varepsilon}$, after which the result follows by combining Lemmas B.18, B.20 and B.21. By iterated expectations, $0 = \mathbb{E}[\varepsilon] = q\mathbb{E}[\varepsilon|z=1] + (1-q)\mathbb{E}[\varepsilon|z=0]$. Thus, it follows that $\mathbb{E}[\varepsilon|z=0] = -q\mathbb{E}[\varepsilon|z=1]/(1-q)$ and substituting the definition of $\widetilde{\delta}_z$,

$$\widetilde{\delta}_z = \mathbb{E}[\varepsilon|z=1] - \mathbb{E}[\varepsilon|z=0] = \left(\frac{1}{1-q}\right)\mathbb{E}[\varepsilon|z=1]$$

Now, since ε is mean zero $\sigma_{z\varepsilon} = \mathbb{E}[z\varepsilon] = \mathbb{E}[z\mathbb{E}[\varepsilon|z]] = q\mathbb{E}[\varepsilon|z=1]$ so that $\widetilde{\delta}_z q(1-q) = q\mathbb{E}[\varepsilon|z=1] = \sigma_{z\varepsilon}$. □

Lemma B.22.

$$\widetilde{\delta}_z = \widetilde{B}(\alpha_0,\alpha_1) + \widetilde{S}(\alpha_0,\alpha_1)\widetilde{\delta}_{T^*}$$

where

$$\widetilde{B}(\alpha_0,\alpha_1) = \frac{S(\alpha_0,\alpha_1)F(\alpha_0,\alpha_1)C_3 + B(\alpha_0,\alpha_1) - C_1}{S(\alpha_0,\alpha_1)F(\alpha_0,\alpha_1)C_4 - C_2 + 1}, \quad \widetilde{S}(\alpha_0,\alpha_1) = \frac{S(\alpha_0,\alpha_1)}{S(\alpha_0,\alpha_1)F(\alpha_0,\alpha_1)C_4 - C_2 + 1},$$

$$C_1 = \frac{\boldsymbol{\sigma}'_{zx}\boldsymbol{\gamma}_{IV}}{q(1-q)}, \quad C_2 = \boldsymbol{\sigma}'_{zx}\boldsymbol{\sigma}^{xT}, \quad C_3 = \boldsymbol{\sigma}'_{xT}\boldsymbol{\gamma}_{IV}, \quad C_4 = q(1-q)\boldsymbol{\sigma}'_{xT}\boldsymbol{\sigma}^{xT},$$

$F(\alpha_0,\alpha_1) = (1-\alpha_0-\alpha_1)/[(p-\alpha_0)(1-p-\alpha_1)]$, *and* $B(\alpha_0,\alpha_1)$, $S(\alpha_0,\alpha_1)$ *are as defined in Proposition 4.1.*

Proof. Rearranging the result of Lemma B.1, $\boldsymbol{\gamma} = \widehat{\gamma}_{IV} - \widetilde{\delta}_z q(1-q)\boldsymbol{\sigma}^{xT}$. Substituting this into Equation B.21 from Lemma B.22,

$$\begin{aligned}\delta_z &= [\mathbb{E}(\mathbf{x}|z=1) - \mathbb{E}(\mathbf{x}|z=0)]'\left(\widehat{\gamma}_{IV} - \widetilde{\delta}_z q(1-q)\boldsymbol{\sigma}^{xT}\right) + \widetilde{\delta}_z \\ &= \frac{\boldsymbol{\sigma}'_{zx}}{q(1-q)}\left(\widehat{\gamma}_{IV} - \widetilde{\delta}_z q(1-q)\boldsymbol{\sigma}^{xT}\right) + \widetilde{\delta}_z = \frac{\boldsymbol{\sigma}'_{zx}\widehat{\gamma}_{IV}}{q(1-q)} - \left(\boldsymbol{\sigma}'_{zx}\boldsymbol{\sigma}^{xT} - 1\right)\widetilde{\delta}_z\end{aligned}$$

since $\mathbb{E}(\mathbf{x}|z=1) - \mathbb{E}(\mathbf{x}|z=0)$, the coefficient from a regression of $\mathbf{x}$ on z, equals $\boldsymbol{\sigma}_{zx}/[q(1-q)]$. Thus,

$$\delta_z = C_1 - (C_2 - 1)\widetilde{\delta}_z. \tag{B.24}$$

Similarly, substituting into Equation B.22 from Lemma B.22,

$$\begin{aligned}
\delta_{T^*} &= \left[\mathbb{E}\left(\mathbf{x}|T^*=1\right)-\mathbb{E}\left(\mathbf{x}|T^*=0\right)\right]'\left(\widehat{\gamma}_{IV}-\widetilde{\delta}_z q(1-q)\boldsymbol{\sigma}^{xT}\right)+\widetilde{\delta}_{T^*} \\
&= \frac{p(1-p)(1-\alpha_0-\alpha_1)}{(p-\alpha_0)(1-p-\alpha_1)}\left[\mathbb{E}\left(\mathbf{x}|T=1\right)-\mathbb{E}\left(\mathbf{x}|T=0\right)\right]'\left(\widehat{\gamma}_{IV}-\widetilde{\delta}_z q(1-q)\boldsymbol{\sigma}^{xT}\right)+\widetilde{\delta}_{T^*} \\
&= \frac{p(1-p)(1-\alpha_0-\alpha_1)}{(p-\alpha_0)(1-p-\alpha_1)}\left[\frac{\boldsymbol{\sigma}_{xT}}{p(1-p)}\right]'\left(\widehat{\gamma}_{IV}-\widetilde{\delta}_z q(1-q)\boldsymbol{\sigma}^{xT}\right)+\widetilde{\delta}_{T^*} \\
&= \frac{(1-\alpha_0-\alpha_1)\boldsymbol{\sigma}'_{xT}\widehat{\gamma}_{IV}}{(p-\alpha_0)(1-p-\alpha_1)}-\left[\frac{q(1-q)(1-\alpha_0-\alpha_1)\boldsymbol{\sigma}'_{xT}\boldsymbol{\sigma}^{xT}}{(p-\alpha_0)(1-p-\alpha_1)}\right]\widetilde{\delta}_z+\widetilde{\delta}_{T^*} \\
&= \left[\frac{(1-\alpha_0-\alpha_1)}{(p-\alpha_0)(1-p-\alpha_1)}\right]C_3-\left[\frac{(1-\alpha_0-\alpha_1)}{(p-\alpha_0)(1-p-\alpha_1)}\right]C_4\widetilde{\delta}_z+\widetilde{\delta}_{T^*}
\end{aligned}$$

since $\mathbb{E}(\mathbf{x}|T=1)-\mathbb{E}(\mathbf{x}|T=0)$, the coefficient from a regression of $\mathbf{x}$ on T, equals $\boldsymbol{\sigma}_{xT}/[p(1-p)]$. Thus,

$$\delta_{T^*} = F(\alpha_0,\alpha_1)C_3 - F(\alpha_0,\alpha_1)C_4\widetilde{\delta}_z + \widetilde{\delta}_{T^*}. \tag{B.25}$$

Now, substituting Equations B.24 and B.25 into Equation 32 and re-arranging,

$$\widetilde{\delta}_z = \left[\frac{S(\alpha_0,\alpha_1)F(\alpha_0,\alpha_1)C_3+B(\alpha_0,\alpha_1)-C_1}{S(\alpha_0,\alpha_1)F(\alpha_0,\alpha_1)C_4-C_2+1}\right]+\left[\frac{S(\alpha_0,\alpha_1)}{S(\alpha_0,\alpha_1)F(\alpha_0,\alpha_1)C_4-C_2+1}\right]\widetilde{\delta}_{T^*}.$$

□

B.2 Inference: Draws for the Reduced Form Parameters

In this appendix we provide details of our algorithm for producing posterior samples for the reduced from parameters described in Section 4.5. Define $W = \left[\begin{array}{ccc}\mathbf{1} & T & X\end{array}\right]$, $R = \left[\begin{array}{ccc}\mathbf{1} & \mathbf{z} & X\end{array}\right]$, and $\widehat{\mathbf{b}} = (R'W)^{-1}R'\mathbf{y}$ where the estimator $\widehat{\mathbf{b}}$ converges in probability to the parameter $\mathbf{b}$ from the reduced form model $\mathbf{y} = W\mathbf{b}+\boldsymbol{\rho}$ and $\boldsymbol{\rho}$ is a reduced form error. Note that the first element of $\mathbf{b}$ is the constant term and the second element is the reduced-form coefficient for T. The remaining elements are the reduced-form coefficients for $\mathbf{x}$, namely γ_{IV}. Define the residuals $\widehat{\rho}_i = y_i - \mathbf{w}_i'\widehat{\mathbf{b}}$. Using the definition of $\widehat{\mathbf{b}}$ and the reduced form errors $\boldsymbol{\rho}$,

$$\widehat{\mathbf{b}} = (R'W)^{-1}R'\mathbf{y} = (R'W)^{-1}R'(W\mathbf{b}+\boldsymbol{\rho}) = \mathbf{b}+(R'W)^{-1}R'\boldsymbol{\rho}$$

and thus $\sqrt{n}(\mathbf{b}-\widehat{\mathbf{b}}) = (R'W/n)^{-1}(R'\boldsymbol{\rho}/\sqrt{n})$. Now, the conditional means of y given z and T can be constructed from the parameters of the following regression model:

$$y_i = \xi_0 + \xi_T T_i + \xi_z z_i + \xi_{Tz} T_i \times z_i + \omega_i$$

where ω_i is a reduced-form error that is correlated with $\mathbf{x}_i$. Defining $A = \left[\begin{array}{cccc}\mathbf{1} & T & \mathbf{z} & T\mathbf{z}\end{array}\right]$ we can write this as $\mathbf{y} = A\boldsymbol{\xi}+\boldsymbol{\omega}$. We estimate $\boldsymbol{\xi}$ by OLS leading to residuals $\widehat{\omega}_i = y_i - \mathbf{a}_i'\widehat{\xi}$. The cell means $\bar{y}_{tk}$ and the estimated parameters of this regression $\widehat{\boldsymbol{\xi}}$ are related as follows:

$$\underbrace{\begin{bmatrix}\bar{y}_{00}\\ \bar{y}_{01}\\ \bar{y}_{10}\\ \bar{y}_{11}\end{bmatrix}}_{\bar{\mathbf{y}}} = \underbrace{\begin{bmatrix}1&0&0&0\\1&0&1&0\\1&1&0&0\\1&1&1&1\end{bmatrix}}_{Q}\underbrace{\begin{bmatrix}\widehat{\xi}_0\\ \widehat{\xi}_T\\ \widehat{\xi}_z\\ \widehat{\xi}_{Tz}\end{bmatrix}}_{\widehat{\boldsymbol{\xi}}}$$

in other words $\bar{\mathbf{y}} = Q\widehat{\boldsymbol{\xi}}$ and similarly for the population parameters: $\boldsymbol{\mu}_y = Q\boldsymbol{\xi}$. Hence,

$$\sqrt{n}\left(\bar{\mathbf{y}}-\boldsymbol{\mu}_y\right) = Q\left(A'A/n\right)^{-1}\left(A'\boldsymbol{\omega}/\sqrt{n}\right).$$

To determine the joint distribution of $\widehat{\mathbf{b}}$ and $\bar{\mathbf{y}}$ we need to study the joint limiting behavior of $A'\omega/\sqrt{n}$ and $R'\boldsymbol{\rho}/\sqrt{n}$. By the Central Limit Theorem for iid observations

$$\left[\begin{array}{c} R'\boldsymbol{\rho}/\sqrt{n} \\ A'\boldsymbol{\omega}/\sqrt{n} \end{array}\right] = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left[\begin{array}{c} \mathbf{r}_i\rho_i \\ \mathbf{a}_i\omega_i \end{array}\right] \to^d \left[\begin{array}{c} M_\rho \\ M_\omega \end{array}\right] \sim N(0,\Xi).$$

We estimate Ξ as follows

$$\widehat{\Xi} = \frac{1}{n-1}\sum_{i=1}^{n}\left[\begin{array}{cc} \mathbf{r}_i\mathbf{r}_i'\widehat{\rho}_i^2 & \mathbf{r}_i\mathbf{a}_i'\widehat{\rho}_i\widehat{\omega}_i \\ \mathbf{a}_i\mathbf{r}_i'\widehat{\rho}_i\widehat{\omega}_i & \mathbf{a}_i\mathbf{a}_i'\widehat{\omega}_i^2 \end{array}\right] = \left[\begin{array}{cc} \widehat{\Xi}_{\rho\rho} & \widehat{\Xi}_{\rho\omega} \\ (\widehat{\Xi}_{\rho\omega})' & \widehat{\Xi}_{\omega\omega} \end{array}\right]$$

where

$$\widehat{\Xi}_{\rho\rho} = \frac{R'\mathrm{diag}\left\{\widehat{\rho}^2\right\}R}{n-1}, \quad \widehat{\Xi}_{\omega\omega} = \frac{A'\mathrm{diag}\left\{\widehat{\omega}^2\right\}A}{n-1}, \quad \widehat{\Xi}_{\rho\omega} = \frac{R'\mathrm{diag}\left\{\widehat{\rho}\right\}\mathrm{diag}\left\{\widehat{\omega}\right\}A}{n-1}.$$

Stacking the two estimators on top of one another,

$$\left[\begin{array}{c} \sqrt{n}\left(\widehat{\mathbf{b}}-\mathbf{b}\right) \\ \sqrt{n}\left(\bar{\mathbf{y}}-\boldsymbol{\mu}_y\right) \end{array}\right] = \left[\begin{array}{cc} (R'W/n)^{-1} & 0 \\ 0 & Q(A'A/n)^{-1} \end{array}\right]\left[\begin{array}{c} R'\boldsymbol{\rho}/\sqrt{n} \\ A'\boldsymbol{\omega}/\sqrt{n} \end{array}\right] \quad \stackrel{d}{\to} \left[\begin{array}{cc} \Sigma_{RW}^{-1} & 0 \\ 0 & Q\Sigma_{AA}^{-1} \end{array}\right]\left[\begin{array}{c} M_\rho \\ M_\omega \end{array}\right]$$

so we see that the joint limit distribution is $N(0,H)$ with

$$H = \left[\begin{array}{cc} \Sigma_{RW}^{-1} & 0 \\ 0 & Q\Sigma_{AA}^{-1} \end{array}\right]\underbrace{\left[\begin{array}{cc} \Xi_{\rho\rho} & \Xi_{\rho\omega} \\ \Xi_{\omega\rho} & \Xi_{\omega\omega} \end{array}\right]}_{\Xi}\left[\begin{array}{cc} (\Sigma_{RW}^{-1})' & 0 \\ 0 & (Q\Sigma_{AA}^{-1})' \end{array}\right].$$

We only require the joint sampling distribution for $\bar{\mathbf{y}}$ and $\widehat{\boldsymbol{\gamma}}_{IV}$, but the first element of $\widehat{\mathbf{b}}$ is the constant term while the second corresponds to $\widehat{\beta}_{IV}$. We do not need to work explicitly with the constant since in the equations we use covariances hence in effect de-mean everything. Accordingly, define S to be the submatrix of H that contains everything *except* the first and second rows and columns. We have

$$\left[\begin{array}{c} \sqrt{n}\left(\widehat{\beta}_{IV}-\beta_{IV}\right) \\ \sqrt{n}\left(\widehat{\boldsymbol{\gamma}}_{IV}-\boldsymbol{\gamma}_{IV}\right) \\ \sqrt{n}\left(\bar{\mathbf{y}}-\boldsymbol{\mu}_y\right) \end{array}\right] \to^d N(0,S)$$

Treating this as a Bayesian posterior, we draw according to

$$\left[\begin{array}{c} \beta_{IV} \\ \boldsymbol{\gamma}_{IV} \\ \boldsymbol{\mu}_y \end{array}\right] \sim \left[\begin{array}{c} \widehat{\beta}_{IV} \\ \widehat{\boldsymbol{\gamma}}_{IV} \\ \bar{\mathbf{y}} \end{array}\right] + N(0,\widehat{S}/n)$$

Note the block of $\widehat{S}$ corresponding to $\bar{\boldsymbol{y}}$ has a very simple structure: since we assume that our data are iid and the cell means are calculated for non-overlapping groups of individuals, the $\bar{y}_{tk}$ are all independent and

$$n\widehat{H}_{yy} = n\widehat{S}_{yy} = n(Q\widehat{\Sigma}_{AA}^{-1})\widehat{\Xi}_{\omega\omega}(Q\widehat{\Sigma}_{AA}^{-1})' = n\ \mathrm{diag}\left\{\widehat{\sigma}_{00}^2/n_{00}, \widehat{\sigma}_{01}^2/n_{01}, \widehat{\sigma}_{10}^2/n_{10}, \widehat{\sigma}_{11}^2/n_{11}\right\}$$

where $\widehat{\sigma}_{tk}^2$ is the sample variance of those y observations for which $T=t$ and $z=k$ and n_{tk} is the corresponding sample size. Recall that we make our draws not from $\widehat{S}$ but rather from $\widehat{S}/n$. This final division by n yields the familiar variance of the sampling distribution of the sample mean. The other blocks of $\widehat{H}$, from which we must remove some elements as described above to yield the corresponding blocks for $\widehat{S}$, are as follows:

$$\widehat{H}_{bb} = \widehat{\Sigma}_{RW}^{-1}\widehat{\Xi}_{\rho\rho}\left(\widehat{\Sigma}_{RW}^{-1}\right)', \quad \widehat{H}_{by} = \widehat{\Sigma}_{RW}^{-1}\widehat{\Xi}_{\rho\omega}\left(Q\Sigma_{AA}^{-1}\right)', \quad \widehat{H}_{yb} = \widehat{H}_{by}'.$$

C Additional Empirical Examples (Online Only)

C.1 Was Weber Wrong?

Becker and Woessmann (2009) study the long-run effect of the adoption of Protestantism in sixteenth-century Prussia on a number of economic and educational outcomes, using variation across counties in their distance to Wittenberg – the city where Martin Luther introduced his ideas and preached – as an instrument for the Protestant share of the population in the 1870s. Here we consider their estimates of the effect of Protestantism on literacy, based on the specification

$$\text{Literacy rate} = \text{constant} + \beta\,(\text{Protestant share}) + \mathbf{x}'\gamma + u$$
$$\text{Protestant Share} = \text{constant} + \pi\,(\text{Distance to Wittenberg}) + \mathbf{x}'\delta + v$$

where $\mathbf{x}$ is a vector of demographic and regional controls.[41] Because this example includes exogenous controls, we define treatment endogeneity and instrument invalidity *net* of these covariates, as detailed in Section 2.4. To simplify the notation we write ρ_{uz} and ρ_{T^*u} rather than $\widetilde{\rho}_{uz}$ and $\widetilde{\rho}_{T^*u}$ below but both of these should be understood as being net of $\mathbf{x}$. In contrast, κ is not defined net of covariates, again as detailed in Section, 2.4 so it continues to refer to the ratio $\sigma^2_{T^*}/\sigma^2_T$ below.

Becker and Woessmann (2009) express beliefs about the three key parameters in our framework. First, their IV strategy relies on the assumption that $\rho_{uz} = 0$, an assumption that we will relax below. Second, the authors argue that the 1870 Prussian Census is regarded by historians to be highly accurate. As such, measurement error in the Protestant share should be fairly small. Finally, Becker and Woessmann (2009) go through a lengthy discussion of the nature of the endogeneity of the Protestant share, suggesting that it is most likely that Protestantism is *negatively* correlated with the unobservables:

> wealthy regions may have been less likely to select into Protestantism at the time of the Reformation because they benefited more from the hierarchical Catholic structure, because the opportunities provided by indulgences allured to them, and because the indulgence costs weighted less heavily on them ... The fact that "Protestantism" was initially a "protest" movement involving peasant uprisings that reflected social discontent is suggestive of such a negative selection bias (pp. 556-557).

[41] In this exercise we include the controls listed in Section III of Becker and Woessmann (2009), specifically: the fraction of the population younger than age 10, of Jews, of females, of individuals born in the municipality, of individuals of Prussian origin, the average household size, log population, population growth in the preceding decade, the fraction of the population with unreported education information, and fraction of the population that was blind, deaf-mute, and insane.

Results for the "Was Weber wrong?" example appear in Table C.1. Estimates and bounds for β in these rows indicate the percentage point change in literacy that a county would experience if its share of Protestants were to increase by one percentage point. All other values in the table are unitless: they are either probabilities, correlations, or variance ratios. OLS and IV estimates and standard errors, along with the estimates of the lower bounds for κ and ρ_{uz}, appear in row four of Panel (I). The first column of Panel (II) gives the fraction of posterior draws for the reduced form parameters that yield an empty identified set, while the second column gives the fraction that are compatible with a valid instrument: $\rho_{uz} = 0$. Panel (III), along with the third and fourth columns of Panel (II), present posterior medians and accompanying 90 percent highest posterior density intervals. The results in Panel (II) are marked "Frequentist-Friendly" because they do not involve placing a prior on the conditional identified set: they average only over reduced form parameter draws under the restriction listed in the corresponding row label.[42] In contrast, those in Panel (III) are "Fully Bayesian" in that they place a uniform prior on the conditional identified set.

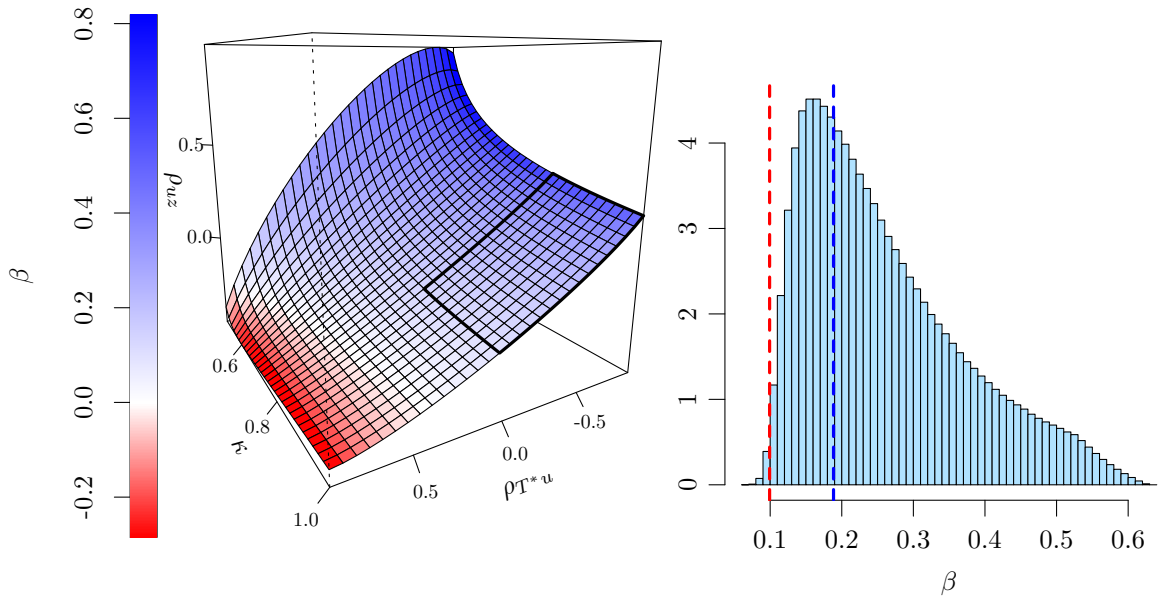
As we see from Table C.1, Becker and Woessmann (2009) obtain an OLS estimate of 0.10 and an IV estimate that is nearly twice as large: 0.19 with a standard error of 0.03. If the instrument is valid, this corresponds to just under a 0.2 percentage point increase in literacy from each percentage point increase in the prevalence of Protestantism in a given county. The estimated lower bound for κ in this example is just under a half, which means that at most 50 percent of the measured variation in the Protestant share can be attributed to measurement error. Notice that this bound is somewhat weak: it allows for far more measurement error than one might consider reasonable given the author's arguments concerning the accuracy of the Prussian census data.

Figure C.1a depicts the identified set for $(\kappa, \rho_{T^*u}, \rho_{uz})$ evaluated at the maximum likelihood estimate of Σ. As above, the surface is colored to indicate the corresponding value of β: blue indicates a positive treatment effect, red a negative effect, and zero no effect. In both directions, darker colors indicate larger magnitudes. We see immediately from the figure, that unless ρ_{T^*u} is large and *positive*, the treatment effect will be positive, irrespective of the amount of measurement error. The rectangular region surrounded by thick black boundaries indicates our approximation to the prior beliefs of Becker and Woessmann (2009): negative selection, and measurement error that is not too severe. This area is well within the blue region, corresponding to a positive treatment effect. Although it is somewhat harder to see from the figure, the region enclosed in the black boundary also contains $\rho_{uz} = 0$. The belief that $\rho_{T^*u} < 0$ and measurement error is modest indeed appears to be compatible with a valid instrument in this example.

[42] See Section 3 for details.

	(I) Summary Statistics				(II) Frequentist-Friendly				(III) Full Bayesian	
	OLS	IV	$\underline{\kappa}$	$\underline{\tilde{\rho}}_{uz}$	$\mathbb{P}(\varnothing)$	$\mathbb{P}(\text{Valid})$	$\underline{\beta}$	$\bar{\beta}$	ρ_{uz}	β
Was Weber Wrong? $(n = 452)$	0.10	0.19	0.49	-0.76						
	(0.01)	(0.03)								
$(\kappa, \rho_{T^*u}) \in (0, 1] \times [-0.9, 0]$					0.00	1.00	0.10	0.88	0.31	0.37
							$[0.08, 0.12]$	$[0.74, 1.03]$	$[-0.10, 0.82]$	$[0.12, 0.60]$
$(\kappa, \rho_{T^*u}) \in (0.8, 1] \times [-0.9, 0]$					0.00	1.00	0.10	0.63	0.06	0.22
							$[0.08, 0.12]$	$[0.58, 0.67]$	$[-0.15, 0.28]$	$[0.10, 0.42]$

Table C.1: Results for "Was Weber Wrong?" (Section C.1). Panel (I) contains OLS and IV estimates and standard errors, and estimates of the bounds for κ and ρ_{uz} from Proposition 2.2 and Corollary 2.1. Panels (II) and (III) present posterior inferences under interval restrictions on (κ, ρ_{T^*u}). The column $\mathbb{P}(\varnothing)$ gives the fraction of reduced form parameter draws that yield an empty identified set, while $\mathbb{P}(\text{Valid})$ gives the fraction of reduced form parameter draws compatible with a valid instrument.$(\rho_{uz} = 0)$. The remaining columns give posterior medians with 90 percent highest posterior density intervals in square brackets. In Panel (II), $\underline{\beta}$ and $\bar{\beta}$ report inferences for the lower and upper boundaries of the identified set for β. In contrast, Panel (III) reports fully Bayesian inference for β and ρ_{uz} under a uniform prior on the intersection between the restrictions and the conditional identified set. See Section 3.2 for details.



(a) Identified Set at MLE for Σ **(b)** Posterior for Treatment Effect

Figure C.1: Results for the "Was Weber Wrong?" example from Section C.1. Panel (a) plots the identified set for $(\rho_{uz}, \rho_{T^*u}, \kappa)$ evaluated at the maximum likelihood estimate for Σ. The color of the surface corresponds to the implied value of the treatment effect β. Panel (b) gives the posterior for β under a uniform prior on the intersection of the restriction $(\kappa, \rho_{T^*u}) \in [0.8, 1] \times [-0.9, 0]$ with the conditional identified set (see Section 3.2 for details). The dashed red line gives the OLS estimate and the blue line the IV estimate.

Although the substance of this example is apparent from Figure C.1a, merely examining the identified set evaluated at the MLE is insufficient, as it fails to account for uncertainty in the reduced form parameters Σ. Row 3 of Table C.1 completes our analysis by providing Bayesian inference for the Weber example under the prior indicated by the black boundary in Figure C.1a: $\kappa > 0.8$ and $-0.9 < \rho_{T^*u} < 0$. In this example one need not even consult the fully Bayesian results from Panel (III): the identified set for β comfortably excludes zero, as we see from columns 3–4 of Panel (II). Indeed, the posterior median for the *lower bound* for β equals the OLS estimate which already implies a substantial causal effect of Protestantism on literacy. This is related to the fact that, as we see from columns 1–2 of the same panel, 100 percent of the reduced form draws for this prior yield an identified set that contains $\rho_{uz} = 0$. Similarly, the fully Bayesian inference for ρ_{uz} in Panel (III) yields a point estimate of 0.06 and a fairly tight highest posterior density interval to accompany it. If we wish to report a point estimate for β, the posterior median from our uniform reference prior in the second column of Panel (III) suggests that the IV estimate is approximately correct, although the highest posterior density interval is skewed somewhat towards even *larger* causal effects. Moreover, none of these results is sensitive to the restriction $\kappa > 0.8$,

as we see from row 2 of Table C.1 which imposes only $-0.9 < \rho_{T^*u} < 0$. In this example, the authors beliefs are mutually consistent and their result is extremely robust.

C.2 Smoking and BMI

This example is based on data from the Lung Health Study (LHS), a well-known randomized clinical trial carried out between 1986 and 1994.[43] The LHS recruited a sample of smokers between the ages of 35 and 59, and offered a smoking cessation program to a random subset. The cessation program consisted of free nicotine gum, an intensive quit week, and access to support personnel, along with invitations to bring a family member to the meetings. Some of the individuals offered treatment also were given an inhaled bronchodilator; the control group received no such offer. The LHS then tracked these subjects over time, recording information on a variety of clinical outcomes. Our outcome of interest here is body mass index (BMI), a measure of obesity defined as weight (in kilograms) divided by squared height (in meters).[44] Following Courtemanche et al. (2016), our objective is to determine the causal effect of quitting on BMI. Our specification is

$$\text{BMI} = \text{constant} + \beta(\text{Quit Smoking}) + \varepsilon$$

where we instrument for the self-reported treatment variable "Quit Smoking" with the randomized offer of participation in the smoking cessation program: $z = 1$ for those in the treatment arm of the LHS while $z = 0$ for those in the control arm. The effect of quitting smoking on BMI is a question of some interest to health researchers, because those who quit smoking are known to experience increases in anxiety and appetite along with physiological changes that may lead to weight gain. Indeed, some suggest that the marked decrease in smoking that has occurred in the U.S. over the past 30 years may be partly to blame for the contemporaneous increase in obesity.[45] Access to the raw data for the LHS is strictly controlled, so we work here with summary statistics provided to us by the authors of Courtemanche et al. (2016). Specifically, we observe conditional means and variances of BMI at the five-year horizon for all combinations of T (Quit Smoking) and z (Offered Smoking Cessation), as well as the empirical joint distribution of T and z. Because our framework for inference relies only on the moments that these quantities estimate, as illustrated in Table

[43] See Ohara et al. (1993) and https://www.clinicaltrials.gov for more information on the LHS.

[44] According to the World Health Organization, individuals whose BMI falls below 18.5 are considered underweight, those whose BMI lies between 18.5 and 25 are fall in the normal range, those whose BMI lies between 25 and 30 are classified as overweight, and those whose BMI exceeds 30 are considered obese. See http://apps.who.int/bmi/index.jsp?introPage=intro_3.html for further details.

[45] For evidence in favor of this claim, see Chou et al. (2004) and Chou et al. (2006); for a contrary view, see Gruber and Frakes (2006).

1a, we can proceed just as if we observed the micro-data.[46]

Measurement error, treatment endogeneity, and instrument invalidity are all serious concerns in this example. First, our measure of whether an individual quit smoking is self-reported. It seems quite likely that some people who have failed to quit will nevertheless claim they have succeeded.[47] Fortunately, the mis-classification is almost certainly one-sided in this example: it is difficult to imagine that someone who successfully quit smoking would report that she did not. Second, while the offer of smoking cessation is randomized, the decision to quit smoking is clearly endogenous. Out of the 5446 subjects in the LHS, 451 report quitting smoking despite not being offered the cessation program, while 2018 report not quitting even though they were offered the program. We might expect subjects who successfully quit smoking to be more health-conscious overall, and thus, thinner than those who do not quit. Courtemanche et al. (2016) also suggest that the offer of a smoking cessation program may not constitute a valid instrument:

> The validity of the [IV] estimator therefore hinges on the assumption that the randomized intervention only affected the BMIs of people who fully quit smoking. To the extent that the intervention also affected the BMIs of those who cut back on smoking but did not quit entirely, the difference in BMI will be scaled by too small a number... (Courtemanche et al. (2016), p. 9.)

As we mentioned in Section 4.2, this logic would imply $\delta_z > 0$.

Results for the Smoking and BMI example appear in rows 2–3 of Table C.2. All values other than those in columns 3–4 of Panel (I) are measured in units of BMI. IV and OLS estimates, along with lower and upper bounds $\bar{\alpha}_0$ and $\bar{\alpha}_1$ for the mis-classification probabilities appear in Panel (I), while posterior medians and 90 percent highest posterior density intervals appear in Panels (II) and (III). Recall the results in Panel (II) are labeled "Frequentist-Friendly" because they do not involve placing a prior on the conditional identified set: they average only over reduced form parameter draws under the restriction listed in the corresponding row label.[48] In contrast, those in Panel (III) are "Fully Bayesian;" they place a uniform prior on the conditional identified set.

Both the OLS and IV estimates in this example are large, positive, and precisely estimated. While one of the mis-classification error bounds for this example is very tight,

[46] One issue with using aggregated moments is the inability to include covariates directly. Nevertheless, if key covariates have discrete support and the sample size is large enough, the analysis can be performed for each cell in the distribution of the support of the covariates.

[47] The LHS data contain three measures of whether an individual has quit smoking, all of which are subject to measurement error: self-reported quit status, self-reported number of cigarettes smoked per day, and the results of salivary cotinine tests administered as part of the study. While Courtemanche et al. (2016) consider all three measures in detail, we focus here on the first for simplicity.

[48] See Sections 3 and 4.5 for details.

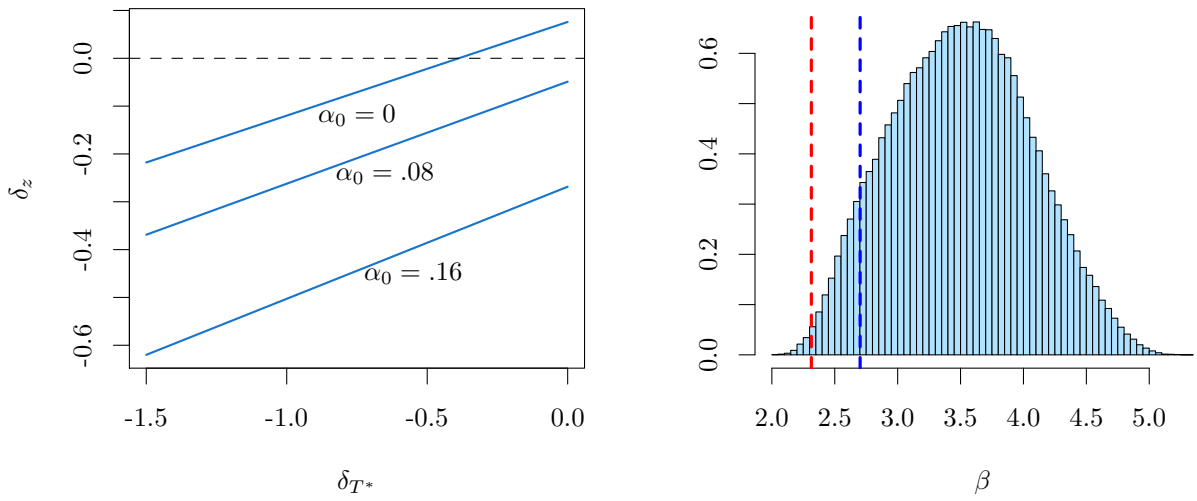
	(I) Summary Statistics				(II) Frequentist-Friendly				(III) Fully Bayesian	
	OLS	IV	$\bar{\alpha}_0$	$\bar{\alpha}_1$	$\underline{\delta}_{T^*/z}$	$\bar{\delta}_{T^*/z}$	$\underline{\beta}$	$\bar{\beta}$	$\delta_{T^*/z}$	β
Smoking & BMI ($n = 5446$)	2.31	2.70	0.16	0.43						
	(0.12)	(0.65)								
$\delta_z = 0, \alpha_1 = 0$					-0.37	1.17	2.24	2.68	0.32	2.44
					$[-1.37, 0.76]$	$[0.35, 2.15]$	$[1.27, 3.11]$	$[1.53, 3.73]$	$[-0.93, 1.53]$	$[1.42, 3.49]$
$\delta_{T^*} \in [-1.5, 0], \alpha_1 = 0$					-0.62	0.07	2.37	5.19	-0.22	3.50
					$[-0.83, -0.43]$	$[-0.13, 0.26]$	$[2.16, 2.57]$	$[4.87, 5.42]$	$[-0.53, 0.07]$	$[2.56, 4.37]$

Table C.2: Results the Smoking and Birthweight example (Section C.2). Panel (I) contains OLS and IV estimates and standard errors along with estimates of the upper bounds α_0 and α_1 from Proposition 4.2. Panels (II) and (III) present posterior medians and 90 percent highest posterior density intervals under the restrictions on $\alpha_0, \alpha_1, \delta_{T^*}$ and δ_z indicated in the row labels: e.g. a row marked $\delta_z = 0$ assumes that the instrument is valid but does not restrict α_0, α_1 or δ_{T^*u}. In columns 1–2 of (II) and 1 of (III), the subscript T^*/z indicates that we report inference for δ_{T^*} when δ_z is restricted *a priori* and vice-versa. In a row marked $\delta_z = 0$, these columns report inference for δ_{T^*}; in a row marked $\delta_{T^*} \in [a, b]$ they report inference for δ_z. As in Table 2, Panel (II) reports inferences for the boundaries of the identified sets for β, δ_{T^*} and/or δ_z while Panel (III) reports fully Bayesian inference under a uniform prior on the intersection between the restrictions and the conditional identified set. See Sections 3.2 and 4.5 for details.

$\alpha_0 < 0.16$, the other is not: $\alpha_1 < 0.43$. Fortunately, α_1 denotes the fraction of true quitters who mis-report and claim they did not quit smoking. As discussed above, the true value of this probability is almost certainly zero so we fix $\alpha_1 = 0$ throughout the remainder of our analysis. Because the partial identification bound for α_1 is so wide, this prior restriction adds a considerable amount of identifying information to the problem.

As a benchmark, row 2 of Table C.2 explores the implications of assuming that $\delta_z = 0$, so that the randomized offer of a smoking cessation program constitutes a valid instrument. Under these beliefs, the inference for the identified set for β from columns 3 and 4 of Panel (II) and the posterior for the parameter β from the second column of Panel (III) indicate a large, positive causal effect of smoking on BMI, but an effect that is somewhat smaller than the IV estimate due to the effect of misclassification error. As discussed above, there are good reasons to doubt the validity of the instrument in this example. Accordingly, row 3 of Table C.2 relaxes the assumption that $\delta_z = 0$ and imposes instead that $\delta_{T^*} \in [-1.5, 0]$. Figure C.2a likewise depicts selected contours of the identified set in this range, evaluated at the maximum likelihood estimates of the reduced form parameters, while Figure C.2b plots the posterior distribution for β corresponding to the inferences from column two of Panel (III) in the table. The sign of δ_{T^*} under this prior represents the belief that those who quit successfully are likely to be more health conscious. The lower bound of 1.5 for the magnitude of the BMI difference corresponds to a third of the distance between the upper end of the "healthy" weight range and the lower end of the "obese" range. At the average U.S. height, $\delta_{T^*} = -1.5$ would say that successful quitters are around 10 pounds lighter on average, a moderate amount of negative selection.

Both the results in the table and the figure imply that δ_z is very likely *negative*, the exact opposite of what we would have expected from above. From Figure C.2a, we see that one would require almost no mis-reporting of true smoking status along with hardly any selection to obtain a postitive value of δ_z. The corresponding inferences for β point to a very large treatment effect: the 90 percent highest posterior density interval for the lower bound for β, for example, ranges from about 2.2 to 2.6, while the median of the posterior for β in Figure C.2b is 3.5. The message of this example is somewhat nuanced compared to our previous ones. If one is certain that δ_{T^*} is negative, our framework implies that δ_z too is almost certainly negative, and the causal effect of quitting smoking on BMI is very large relative to estimates from the existing literature. If on the other hand one feels confident that δ_z should be positive, as the discussion from Courtemanche et al. (2016) suggests, our framework implies that δ_{T^*} must be *positive*: successful quitters are heavier on average.



(a) Contours of Identified Set at MLE

(b) Posterior for Treatment Effect

Figure C.2: Results for Smoking and BMI example from Section C.2. Panel (a) illustrates Equation 32 for $\delta_{T^*} \in [-1, 0], \alpha_1 = 0$ at three of values for α_0. Both δ_{T^*} and δ_z are given in units of BMI, and the reduced form parameters are set equal to their maximum likelihood estimates. Panel (b) gives the posterior distribution for β under a uniform prior on the intersection between the restriction $\delta_{T^*} \in [-1, 0], \alpha_1 = 0$ and the conditional identified set (see Section 4.5 for details). The red dashed line gives the OLS estimate and the blue line the IV estimate.